Multi-modal pain intensity recognition based on the SenseEmotion database

Patrick Thiam, Viktor Kessler, Mohammadreza Amirian, Peter Bellmann, Georg Layher, Yan Zhang, Maria Velana, Sascha Gruss, Steffen Walter, Harald C. Traue, Daniel Schork, Jonghwa Kim, Elisabeth André, Heiko Neumann, Friedhelm Schwenker

Angaben zur Veröffentlichung / Publication details:

Thiam, Patrick, Viktor Kessler, Mohammadreza Amirian, Peter Bellmann, Georg Layher, Yan Zhang, Maria Velana, et al. 2021. “Multi-modal pain intensity recognition based on the SenseEmotion database.” *IEEE Transactions on Affective Computing* 12 (3): 743–60. https://doi.org/10.1109/taffc.2019.2892090.

Multi-Modal Pain Intensity Recognition Based on the *SenseEmotion* Database

Patrick Thiam, Viktor Kessler, Mohammadreza Amirian, Peter Bellmann, Georg Layher, Yan Zhang, Maria Velana, Sascha Gruss, Steffen Walter, Harald C. Traue, Daniel Schork, Jonghwa Kim, Elisabeth André, Heiko Neumann, and Friedhelm Schwenker

Abstract—The subjective nature of pain makes it a very challenging phenomenon to assess. Most of the current pain assessment approaches rely on an individual's ability to recognise and report an observed pain episode. However, pain perception and expression are affected by numerous factors ranging from personality traits to physical and psychological health state. Hence, several approaches have been proposed for the automatic recognition of pain intensity, based on measurable physiological and audiovisual parameters. In the current paper, an assessment of several fusion architectures for the development of a multi-modal pain intensity classification system is performed. The contribution of the presented work is two-fold: (1) 3 distinctive modalities consisting of audio, video and physiological channels are assessed and combined for the classification of several levels of pain elicitation. (2) An extensive assessment of several fusion strategies is carried out in order to design a classification architecture that improves the performance of the pain recognition system. The assessment is based on the *SenseEmotion Database* and experimental validation demonstrates the relevance of the multi-modal classification approach, which achieves classification rates of respectively 83.39%, 59.53% and 43.89% in a 2-class, 3-class and 4-class pain intensity classification task.

Index Terms—Pain intensity recognition, multiple classifier systems, multi-modal information fusion, signal processing

1 Introduction

Effective pain management implies reliable and valid assessment of pain. However, pain is a complex and highly subjective phenomenon [1], [2] which is commonly associated with unpleasant psycho-physiological and physical experiences. Furthermore, pain is an individually unique experience which varies from one individual to the next [3]. This particular aspect further increases the complexity of pain assessment. Hence, self-report is considered to be the gold standard in pain assessment and has been successful in providing valuable insights for effective pain management [4], [5]. However, self-reporting tools such as the Visual Analogue Scale (VAS) or the Numerical Rating Scale (NRS) for pain [6], [7] strongly rely on an individual's ability to recognise, assess and communicate an observed pain episode. Thus, self-report would provide inconsistent and unreliable information in cases where an individual is suffering from a form of cognitive impairment which impedes the individual's ability to reliably and systemically perceive, assess and share informative insights about the experienced pain episode. Hence, relying uniquely on self-report could lead to unsuitable and inadequate pain management.

Various studies have investigated the feasibility and relevancy of automatic pain assessment systems based on measurable audiovisual and physiological parameters (see Section 2). These studies show that such systems are able to provide valuable insights for the assessment of pain intensities by automatically analysing non-verbal pain indicators including pain related facial expressions, paralinguistic vocalisations, body postures and changes in physiological parameters. Therefore, the combination of self-reporting tools with a reliable and automatic pain assessment system could potentially improve the robustness as well as the effectiveness of pain management.

Moreover, the huge diversity of pain related expressions within each specific modality (e.g., frowning (facial expressions), moaning (paralinguistic vocalisations), changes in body posture (behavioural pain responses), changes in physiological parameters (autonomic pain responses)) suggests that pain intensity classification should be approached as a multi-modal pattern recognition problem. Instead of

- *P. Thiam, V. Kessler, P. Bellmann, G. Layher, Y. Zhang, H. Neumann, and F. Schwenker are with the Institut of Neural Information Processing, Ulm University, James-Franck-Ring, Ulm 89081, Germany. E-mail: {patrick.thiam, viktor.kessler, peter.bellmann, georg.layher, yan.zhang, heiko.neumann, friedhelm.schwenker}@uni-ulm.de.*
- *M. Amirian is with the Zurich University of Applied Sciences, Winterthur 8400, Switzerland. E-mail: amir@zhaw.ch.*
- *M. Velana, S. Gruss, S. Walter, and H. C. Traue are with the University Clinic for Psychosomatic Medicine and Psychotherapy, Medical Psychology, Ulm University, Frauensteige 6, Ulm 89075, Germany. E-mail: {maria.velana, sascha.gruss, steffen.walter, harald.traue}@uni-ulm.de.*
- *D. Schork and E. André are with the Department of Computer Science, Human-Centered Multimedia, University of Augsburg, Universitätstr. 6a, Augsburg 86159, Germany. E-mail: {schork, andre}@informatik.uni-augsburg.de.*
- *J. Kim is with the Department of Information and Communication Technology, Cheju Halla University, Jeju 63092, Korea. E-mail: kim@ieee.org.*

relying on the information provided by a single modality, a well designed fusion approach should be able to appropriately combine complementary information from multiple sources in order to improve both the robustness of a classification system as well as its performance.

In the following work, several fusion approaches are proposed and assessed within the scope of the development of an automatic pain intensity recognition system. The assessment is performed on the recently recorded *SenseEmotion Database* [8], which consists of 45 individuals subjected to a series of artificially induced pain stimuli, elicited through temperature elevation. Several modalities were synchronously acquired during the experiments including audio streams, video streams, respiration (RSP), electrocardiography (ECG), electromyography (EMG) and electrodermal activity (EDA) signals. A broad spectrum of descriptors is extracted from each involved modality followed by an evaluation of an uni-modal pain intensity classification system based on the set of features extracted from each single modality. Subsequently, several fusion architectures performing the combination of the extracted descriptors at different levels of abstraction based on various aggregation rules are evaluated. The goal here is to design an effective fusion architecture that is able to significantly outperform the best performing single modality, through an adequate combination of information extracted from each specific modality.

The remainder of this work is organised as follows. In Section 2, an overview of the related research on automatic pain recognition is provided. In Section 3, the recently recorded *SenseEmotion Database* is described. A description of the sensor system used for the data acquisition, followed by a description of the recorded data and the features extracted from each involved modality is provided respectively in Section 4 for the audio modality, Section 5 for the video modality and Section 6 for each physiological modality. The proposed fusion architectures are described in Section 7 and a thorough description of the performed experiments as well as the yielded results is provided in Section 8. Finally, the current work is concluded in Section 9 with a discussion about the findings as well as an overview about potential future works.

2 Related Work

The following section provides an overview of related research and proposed approaches for the development of automatic pain assessment and pain intensity recognition systems.

The recent advancements in the domain of automatic pain assessment have been possible thanks to the availability of a few databases containing specific and representative pain related data. One of the first and very prominent databases specific to pain made available to the research community is the *UNBC-McMaster Shoulder Pain Expression Archive Database* [9]. It consists of 129 participants suffering from shoulder pain and performing specific motion exercises with both affected and unaffected limbs. During the exercises, video sequences of the spontaneous facial expressions of the participants were recorded. Each frame of the recorded video sequences was subsequently annotated using Ekman's Facial Action Unit System (FACS) [10] and the Prkachin and Solomon Pain Intensity (PSPI) [11] metric. The recordings were also annotated at the sequence level based on each participant's self-report and observer measures. This database focuses specifically on the analysis of facial expressions and does not involve any other modality. No external stimulus was used to trigger the pain episode, but rather the exercises conducted with the affected limb triggered genuine pain related facial expressions.

Lately, Walter et al. proposed the *BioVid Heat Pain Database* [12], which is a multi-modal database consisting of 87 healthy participants submitted to four gradually increasing levels of artificially induced pain through temperature elevation. During the experiments, several modalities were synchronously recorded including video streams, EMG, ECG and EDA data. The labels of the acquired data consist of the four different levels of pain elicitation. In contrast to the *UNBC-McMaster Shoulder Pain Expression Archive Database*, the *BioVid Heat Pain Database* is multi-modal since the data acquired stems from at least two different modalities (video and physiology). Furthermore, pain was elicited artificially even though the recorded pain related expressions were genuine.

Most recently, Aung et al. introduced the *Multimodal EmoPain Dataset* [13], which is a collection of data specific to chronic pain. The database consists of 22 individuals suffering from chronic lower back pain and 28 healthy individuals, each performing various physical exercises in a realistic physical rehabilitation setting. High resolution multi-view video streams were recorded during the experiments, as well as multi-directional audio streams, full body three dimensional motion capturing data and EMG signals of back muscles. The recorded data was annotated using two different sets of labels. The first set of labels consists of a continuous rating of the level of pain perceived by an annotator while observing the participants' facial expressions. The assigned rating values ranged between 0 (lowest level of pain) and 1 (highest level of pain). This specific annotation was conducted by eight different annotators. The second set of labels is based on the occurrence of six pain-related body behaviours (*guarding or stiffness*, *hesitation*, *bracing or support*, *abrupt action*, *limping*, *rubbing or stimulating*) that was previously defined by six experts in the field of physical rehabilitation.

Concordantly to the released databases, several approaches for the automatic recognition of pain related expressions have been developed, based either on single modalities or on a combination of several modalities. Many of the proposed approaches focus uniquely on the facial area [14], [15], [16], [17], since a huge amount of information related to an individual's affective state is conveyed throughout facial expressions. These approaches consist of manually or automatically defining and extracting several descriptors from the recorded facial area and performing the classification of the processed data by using common classifiers (e.g., Support Vector Machine (SVM), Random Forests (RF)) or deep learning architectures (e.g., Deep Belief Networks (DBN) [17]).

Moreover, several approaches based on the analysis of physiological modalities as EMG, ECG, RSP and EDA have been proposed [18], [19], [20], [21]. These approaches have

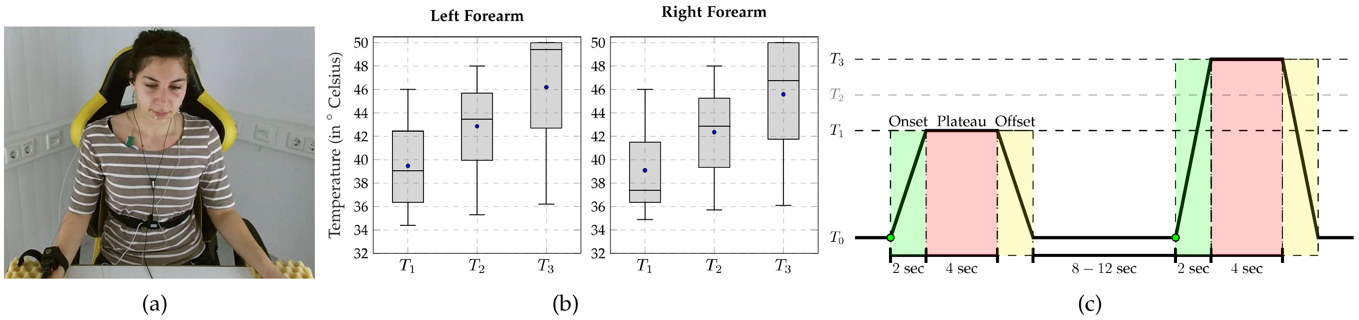


Fig. 1. *(a) Experimental settings.* The participants remained seated during the whole experiment with both forearms resting on a desk in front of them. The picture depicts a session of the experiment during which the thermal simulator is attached to the right forearm. *(b) Temperatures (heat stimulation).* For each level of pain elicitation, the subjective nature of pain is reflected into the large variance of elicitation temperatures across a set of 40 participants selected for the evaluation of the designed classification approaches (see Section 3). *(c) Artificially induced pain stimulation through temperature elevation.* T_0: baseline temperature ($32\ ^\circ\mathrm{C}$); T_1: pain threshold temperature; T_2: intermediate elicitation temperature; T_3: pain tolerance temperature. The green dot symbolises the onset starting point in time which is later used in Section 8.1 as a reference point to define the windows from which features are extracted from each modality.

shown that each modality provides specific insights that can be used in order to adequately assess pain intensity in a realistic setting. However, single modality recognition approaches are known to be inflexible and need extra adjustments in order to deal with missing or erroneous data [22]. Approaches based on the analysis of facial expressions rely strongly on an accurate localisation of the facial area in each frame of a video sequence. This task is known to be very difficult in a natural setting due to unconstrained movements of a monitored participant. Sensors used to record physiological modalities are quite sensitive and might sometimes record unreliable signals due to unconstrained body motion during the acquisition of the data, with the eventuality that the sensors get completely disconnected from the subject's skin, resulting in missing data. This issue can be alleviated by using several modalities and performing the assessment based on an appropriate combination of the data provided by the most reliable ones [23], [24].

Several studies [25], [26], [27], [28] have shown that an adequate combination of information extracted from several modalities might improve the robustness (against noisy inputs) as well as the overall performance of a pain classification system. The most prominent combination approaches consist of the early fusion strategy [29], [30] and several late fusion strategies, which consist of combining the decision of individual models trained on different sets of features by using fixed combination rules (e.g., product rule) or trainable combination rules (e.g., pseudo-inverse) [31], [32]. Furthermore, the combination can occur at different levels of abstraction [33] and also in a hierarchical manner by using a cascade of different aggregation strategies [34]. Multiple Kernel Learning (MKL) [35] and multi-modal deep autoencoders [36] have also been employed as fusion strategies for emotion recognition. In [37], the authors combine both audio and video modalities in order to proceed with pain recognition in real clinical settings, using early and late fusion strategies. The labels used for the assessment of the proposed pain recognition system consist of the recorded subjective pain intensities (defined on the NRS scale), grouped in three pain severity categories (mild, moderate and severe). The proposed late fusion strategy consists in fusing the decision scores from each individual channel using logistic regression.

Analogously to [37], the audio modality is assessed in the following work, in addition to both video and physiological modalities. However, the data assessed in the current work is recorded in an experimental setting and the labels consist of three levels of artificially induced pain elicitation. Moreover, we investigate multiple classifier architectures for the combination of paralinguistic descriptors with bio-visual modalities at different levels of abstraction, and in both user dependent and independent settings.

3 Dataset Description

The following section provides a short description of the *SenseEmotion Database* (the reader is referred to [8] for more details).

The database consists of 45 healthy participants, each subjected to a series of artificially induced pain stimuli. The pain stimuli were elicited through moderate temperature elevation using a Medoc pathway thermal simulator.[1] The experiments were conducted in accordance with the ethical guidelines defined in the Declaration of Helsinki, developed by the World Medical Association (WMA).[2] During the experiments, several modalities were synchronously recorded using several sensors integrated within the Social Signal Integration (SSI) framework [38] including audio streams, high resolution video streams, trapezius EMG, RSP, ECG and EDA. The experiments were conducted in two sessions, each of them lasting approximately 40 minutes, with the pain elicitation sensor attached throughout each session to a different forearm (left and right). The participants remained seated during each experiment with the arms resting on a desk in front of them (see Fig. 1a).

Before the data was recorded, each participant's specific pain threshold temperature (T_1) and pain tolerance temperature (T_3) were calibrated based on the individual's self-reports. The range of calibration of the temperatures was set to a minimum of $32\ ^\circ\mathrm{C}$ and a maximum of $50.5\ ^\circ\mathrm{C}$. An intermediate elicitation temperature (T_2) was computed by taking the average of both temperatures T_1 and T_3. These temperatures formed the three gradually increasing levels of artificial pain elicitation used throughout the experiments (see Fig. 1b). The baseline temperature (T_0) corresponding to no pain stimulation was set to $32\ ^\circ\mathrm{C}$ for all participants. Each temperature

[1] http://medoc-weg.com/products/pathway-model-ats/
[2] Ethics Committee Approval: 196/10-UBB/bal

was applied randomly 30 times with a pause of 8 to 12 seconds (sec) between consecutive stimuli. Each stimulation consisted of a 2 sec onset during which the temperature was gradually elevated starting from the baseline until the target temperature was attained. Subsequently, the target temperature was maintained for 4 sec before being gradually dropped to the baseline (see Fig. 1c for more details).

In the current work, the proposed classification approaches are evaluated on a subset of the dataset consisting of 40 participants (20 male and 20 female). Five of the 45 participants were not included in the assessment because of missing or erroneous data due to technical issues during the recordings. The data specific to each of the remaining 40 participants is complete for each modality and for each experimental session. Moreover, each participant is represented by two sets of data, each one specific to one experimental session (left forearm and right forearm) and consisting of 120 instances of artificial pain stimuli (30 elicitations for each T_0, T_1, T_2, and T_3 temperature).

4 Audio Channel Assessment

The following section provides a description of the experimental settings specific to the audio channel. A description of each single step involved in the assessment of the data is also provided.

Throughout the conducted experiments, three audio streams were synchronously recorded using a digital wireless headset microphone (Line6 XD-V75HS), a directional microphone (Rode M3) and the integrated microphone of the Microsoft Kinect v2. The wireless headset microphone allowed unconstrained head movements and recorded any sound emitted by the participants. The directional microphone as well as the integrated Kinect microphone recorded ambient acoustic sounds. All recordings were performed at a fixed sample rate of 48 kHz. Since the experiments did not involve any type of verbal interaction, the recorded audio data consists mostly of breathing, moaning and sighing sounds, as well as ambient noises.

Since the headset microphone was located in the vicinity of the facial nasolabial area, it was capable to appropriately capture the breathing and moaning sounds emitted by the participants, thus, its recordings were more suitable for the task at hand. Therefore, the current assessment of the audio channel is based uniquely on the recordings from the headset microphone. Those from both directional and Kinect microphones are not further analysed since they were unable to capture the breathing and moaning noises satisfactorily (both sensors were placed at a distance of approximately 1 meter from the participants).

The first step in the processing pipeline of the audio recordings consists of the extraction of several low-level descriptors from the raw audio signal. The resulting signals are further preprocessed using bandpass-filtering, signal smoothing and detrending. Subsequently, several high-level descriptors are extracted from the preprocessed signals. In the following sections, each single step of the pipeline is described.

4.1 Low-Level Descriptors

The first step of the audio data processing pipeline consists of the extraction of Low-Level Descriptors (LLDs) from the raw audio signal. LLDs are parameters computed from short time frames of a whole signal. Such parameters describe temporal and spectral properties of the signals, while significantly reducing the amount of data to be processed. In the current work, all LLDs are extracted from 25 milliseconds (ms) frames with a 10 ms shift between consecutive frames. The extraction is performed by using the openSMILE feature extraction toolkit [39].

Commonly used LLDs in speech processing are the *Mel Frequency Cepstral Coefficients (MFCCs)* [40]. MFCCs have proven to be very effective in tasks such as automatic speech recognition, emotion recognition or speaker identification [41], [42], [43]. For the present work, 13 MFCCs were extracted, each combined with its first and second order temporal derivatives, resulting in a total of 39 MFCC-based LLDs. Another set of commonly used LLDs is computed by using the *Relative Spectral Perceptual Linear Predictive Coding (RASTA-PLP)* [44]. RASTA-PLP is an extension of Perceptual Linear Predictive (PLP) [45] analysis which improves the robustness of the computed coefficients against linear spectral distortions. For the present work, 6 RASTA-PLP coefficients were extracted, each in combination with its first and second order temporal derivatives, resulting in a total of 18 RASTA-PLP-based LLDs.

Finally, a third set of LLDs from the time domain was extracted, consisting of the *root mean square signal energy* and the *logarithmic signal energy*, in combination with their first and second order temporal derivatives. Additionally, the following descriptors were extracted: *loudness contour*, *zero-crossing rate*, *mean-crossing rate*, *maximum absolute sample value*, *minimum* and *maximum sample value* and *arithmetic mean of the sample values*. This last set represents a total of 13 LLDs.

4.2 Signal Processing

Following the extraction of LLDs, an additional signal processing step is undertaken in order to substantially reduce the amount of noise within the signal spawned by each single LLD. Much of this noise is related to the recorded ambient sounds in the room where the experiments were undertaken, since no precaution was taken to avoid them, resulting in a more realistic experimental setting. Therefore, in order to attenuate these noises, a third order Butterworth bandpass filter with a frequency range of $[5, 500]$ Hz is applied on each individual low-level descriptor signal. Next, each filtered signal is smoothed using a Gaussian filter with a 30-point window, and subsequently mean centered.

4.3 High-Level Descriptors and Feature Vectors

Once the LLDs have been extracted and preprocessed, a set of high-level descriptors (HLDs) is extracted from each signal within a predefined and specific temporal window. The preprocessed LLD signals are segmented based on a fixed window and HLDs are extracted from these specific segments before being used as feature vectors for the classification tasks. In the current work, the following set of 14 statistical functions is applied on the segmented LLD signals for the extraction of HLDs: *mean*, *median*, *standard deviation*, *maximum*, *minimum*, *range*, *skewness*, *kurtosis*, *first* and *second quartiles*, *interquartile*, *1%-percentile*, *99%-percentile*, *range from 1%- to 99%-percentile*.

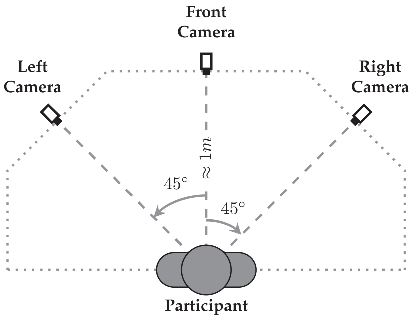


Fig. 2. *Multiple view camera set-up.* The multiple view camera set-up consists of one front camera placed at approximately 1 meter from the participant and two additional cameras placed each in a 45° angle at the left and right hand-side of the participant. Hence, the facial area can still be recorded in a frontal view for a maximal angle of head rotation of 45° to the left or to the right.

Feature Vectors. The MFCC-based feature vectors have a total dimensionality of $14 \times 39 = 546$. The RASTA-PLP-based feature vectors have a total dimensionality of $14 \times 18 = 252$, and the last set of feature vectors from the temporal domain has a total dimensionality of $14 \times 13 = 182$. Subsequently, the HLDs are standardised individually and per participant using the z-score.

5 Video Channel Assessment

This section provides a description of each single step involved in the assessment of the recorded video data. First, a short description of the camera set-up used to perform the recordings is provided. Next, the recorded data is described. Last, a description of the processes undertaken to extract several descriptors from the recorded data is provided.

5.1 Camera Set-Up Description

A multiple view camera set-up was constructed in order to capture the facial expressions of the participants throughout the experiments. It consisted of three identical high resolution cameras (iDS UI-3060CP-C-HQ) equipped with identical lenses (Tevidon $1.8/16$). Each camera recorded a video stream at a resolution of 1600×1200 pixels. The first camera was positioned directly in front of the participant at a distance of approximately 1 meter. The two other cameras were placed respectively at the right and the left hand-side of the participant, each in a 45 Degree angle (see Fig. 2 for an overview of the set-up). In this way, the facial area could still be captured frontally in case it went beyond the scope of the frontal camera, due to relative large head rotations in both left and right directions. Sufficient illumination was provided throughout the experiments by three LED panels mounted respectively at the front, left and right side of the participant. The three cameras synchronously recorded facial expressions displayed by the participants from three different perspectives and additionally allowed unconstrained natural head movements. The recordings of the first 24 participants were performed with a fixed frame rate of 60 frames per second (fps) and involved all three cameras, while the recordings of the next 21 participants were performed at a fixed frame rate of 30 fps, and involved uniquely the frontal camera.

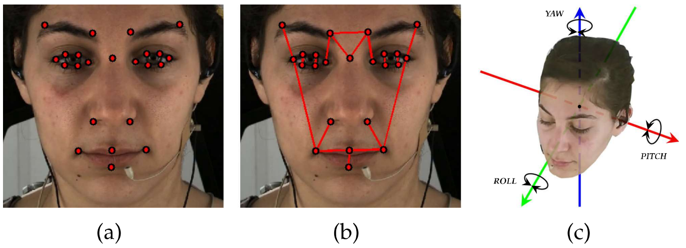


Fig. 3. *Facial area and head pose data. (a)* Using the toolkit OpenFace [47] a set of 23 two-dimensional facial landmarks, which characterises eyebrows, eyes, nose and mouth, is tracked from one video frame to the next. *(b)* The frame level descriptors consist of 17 euclidean distances computed between specific facial landmarks. These distances capture the dynamic of the facial expressions at the frame level. *(c)* The orientation of the head (head pose) can be described by three angles of rotation around three orthogonal axis: roll, pitch and yaw.

5.2 Signal Processing

Prior to the assessment of the recorded data, all recordings were first converted into full color videos using demosaicking [46], since the recordings were performed using a Bayer pattern color filter array (CFA). Then, the full color videos were compressed using the codec H.264. Missing frames were reconstructed using temporal interpolation according to the cameras' time stamps. For the current work, the processed recordings were subsequently converted into a unique frame rate of 30 fps, in order to involve all recorded participants in the current assessment. Moreover, the current work focuses uniquely on the recordings performed with the frontal camera.

Based on the processed video recordings, several descriptors of the facial area are extracted from fixed temporal windows in order to discriminate between the different levels of pain elicitation. Before these descriptors can be computed, the facial area in each video frame has to be localised, aligned and normalised. For this work, the facial behaviour analysis toolkit OpenFace [47] (which uses Constrained Local Neural Fields (CLNF) [48] for facial landmarks detection and tracking) is used for the automatic detection, alignment and normalisation of the facial area. Based on the extracted and preprocessed facial area, the same tool is used for the extraction of a set of two-dimensional facial landmarks and for the estimation of the head pose.

5.3 Feature Extraction

Several descriptors are computed from the two-dimensional location estimations of the facial landmarks, as well as from the head pose estimation data and the preprocessed facial area.

Geometric and Head Pose Descriptors. According to Prkachin et al. [11], [49], four specific facial movements are consistently associated with pain and carry most of the pain related information: *brow lowering*, *tightening of the eye lids in combination with raising cheeks*, *closing of the eyes* and *nose wrinkling in combination with upper lip raising*. Each of these movements involves one or several of the following regions of interest: mouth, nose, eyes and eyebrows. Therefore, a set of 23 two-dimensional facial landmarks (see Fig. 3a), characterising each of the defined regions of interest, are detected and tracked from one video frame to the next. Based on these landmarks, a set of 17 euclidean distances are computed at the frame level

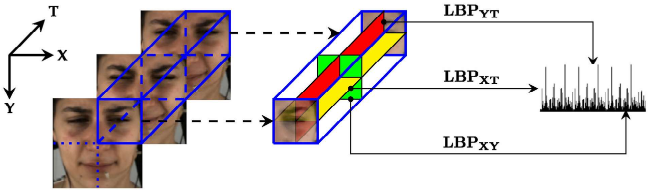


Fig. 4. *Local binary patterns from three orthogonal planes (*LBP-TOP*).* Given a fixed size video sequence, a cuboid consisting of a specific region of interest is extracted. LBP-TOP are subsequently computed based on the cuboid by combining local binary patterns (*LBP*) extracted from the spatial plane *XY*, with those extracted from both spatio-temporal planes *XT* and *YT*. In this way, motion and appearance are both combined and used for the description of facial expressions.

(see Fig. 3b). Each distance characterises a facial dynamic specific to each of the four pain related facial movements described earlier. Hence, each video frame is represented by a 17 dimensional feature vector. Moreover, pain is not only associated with specific facial movements. Intense pain causes sporadic changes of the head orientation and position [50]. Therefore, a three dimensional estimation of the head position as well as an estimation of the head orientation described by three angles of orientation (pitch, raw, roll) (see Fig. 3c), is computed at the frame level. The resulting 6 dimensional frame level vector is used to assess the relevance of head motion for the classification of the different levels of pain elicitation.

Each of these features in the span of a fixed temporal window yields a specific time series, generated by considering the corresponding feature values for all frames within the window. These time series are smoothed by applying a third order low-pass Butterworth filter with a cut-off frequency of $3\,\text{Hz}$. The first and second order derivatives of the filtered time series are also computed.

Feature Vectors. By applying the same set of statistical functions defined in Section 4.3 on each signal, a total of $14 \times 17 \times 3 = 714$ features are extracted from the set of landmark distances and $14 \times 6 \times 3 = 252$ features are extracted from the head pose estimations. The extracted features are subsequently standardised per participant using the z-score.

Appearance-Based Descriptors. Spatio-temporal texture properties of the aligned and normalised facial areas are also assessed and dynamic texture descriptors are extracted using *local binary patterns from three orthogonal planes (LBP-TOP)* [51]. LBP-TOP extend the ordinary *local binary patterns (LBP)* [52] for static images to the spatio-temporal domain (see Fig. 4). They incorporate the temporal component into the description of dynamic textures and therefore combine motion and appearance to describe facial expressions in video sequences. This is done by concatenating local binary patterns extracted from the spatial plane *XY* and from both spatio-temporal planes *XT* and *YT*. The LBP operator can be further extended by using *uniform patterns*. A binary pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa. Subsequently, all non-uniform patterns are assigned the same and unique label, while each uniform pattern is assigned a single and specific label. Hence, the dimensionality of LBP can be substantially reduced by using uniform patterns without any significant loss of information.

In this work, each detected facial region within a fixed temporal window is divided into a 4×4 grid of cells with a 25 percent overlap from one cell to the next. Furthermore, the temporal window is divided in 3 temporal blocks with a 20 percent overlap from one block to the next. This segmentation results in the generation of a total of $4 \times 4 \times 3 = 48$ spatio-temporal cuboids. From each cuboid, uniform LBP-TOP descriptors are extracted. The number of neighbourhood points in each of the three planes (XY, XT, YT) is set to $n = 4$. The radius in both spatial directions r_x and r_y is set to 1, while the radius in the temporal direction r_t is set to 2. This setting results in normalized histograms on each plane with a fixed dimensionality of 15. After concatenating the extracted patterns from each plane, the LBP-TOP descriptor extracted from each cuboid has a dimensionality of $3 \times 15 = 45$.

Feature Vector. To form the dynamic texture descriptor of the whole temporal window, the descriptors of all generated cuboids are concatenated into a final feature vector with a dimensionality of $48 \times 45 = 2160$.

6 Physiological Channels Assessment

This section provides a description of each process involved in the assessment of the recorded physiological data. First, a description of the sensor system used to acquire the data is provided, followed by a description of each recorded physiological channel. Next, each step involved in the preprocessing of the recorded data, as well as in the extraction of descriptors from each specific physiological channel is described.

6.1 Sensor System Description

Physiological data was acquired throughout the experiments using the multi-purpose version of the g.MOBIlab+[3] wireless biosignal acquisition system, equipped with several sensors. All physiological channels were synchronously recorded at a fixed sampling rate of $256\,\text{Hz}$.

Electromyography. EMG measures the electrical activity caused by muscle contractions and propagated through the skin's surface. The intensity of the recorded electrical potential is proportional to the strength of the contractions. For the current experiments, the electrical activity of the upper trapezius muscle (located at the upper back of the human torso) was acquired by using three sintered ($Ag/AgCl$) electrodes (positive, negative, neutral) attached to the surface of the skin. In order to improve the robustness of the recorded signal against noise, a conductive gel was applied on the electrodes before they were attached to the skin. The conductive gel increases the conductivity between the skin and the electrodes and therefore improves the quality of the recorded signals (improved signal to noise ratio). While the difference of electrical potential is measured between the positive and negative electrodes placed on the right upper trapezius muscle, the neutral electrode is used to define a baseline in order to filter out electrical activities propagated through the skin which are unrelated to the muscle activity. Numerous studies [53], [54], [55] report an increase in muscle activity (in particular in the trapezius muscles)

[3] http://www.gtec.at/Products/Hardware-and-Accessories/

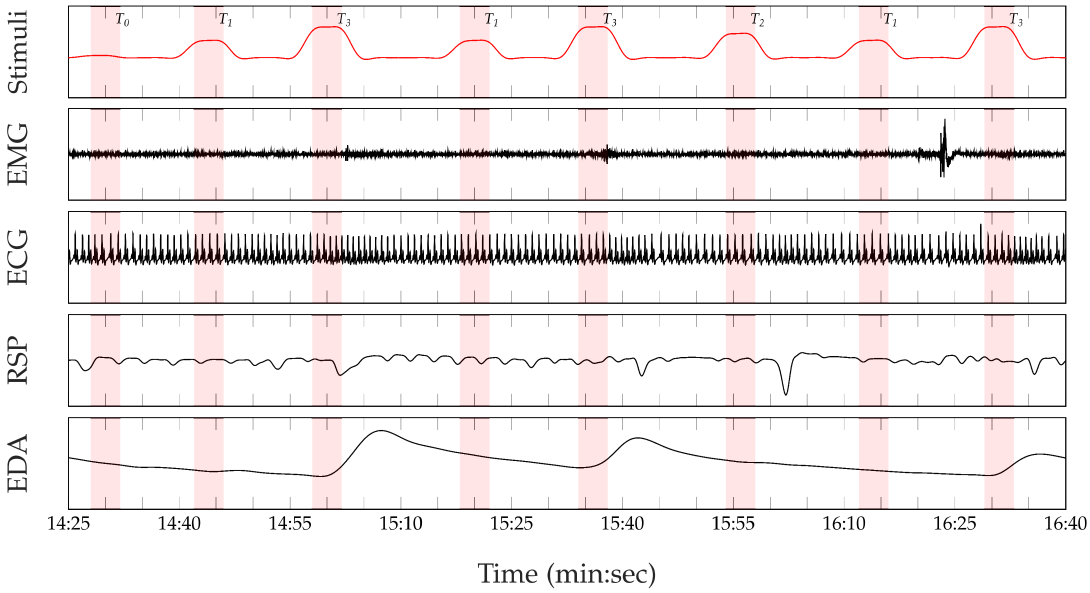


Fig. 5. *Recorded physiological data (preprocessed signals).* From top to bottom: series of artificially induced pain elicitation with the corresponding temperatures (T_0: baseline temperature, T_1: pain threshold temperature, T_2: intermediate elicitation temperature, T_3: pain tolerance temperature), EMG (μV), ECG (μV), RSP (μV), EDA (μS).

concordantly with the experience of stress. The current experiment is based on the assumption that a similar response is to be observed when the participants are subjected to painful stimuli.

Electrocardiography. ECG data was acquired using three sintered electrodes attached to the surface of the skin. Analogously to EMG, a conductive gel was applied on the electrodes prior to their attachment to the skin's surface, in order to perform robust recordings of the electrical activity of the heart muscle. Previous studies [56], [57], [58] have shown that abrupt changes in electrocardiography patterns correspond to physiological arousal as a response to external stimuli, hence the relevance of ECG for the current study.

Respiration (RSP). RSP data was acquired using an elastic belt system worn over clothing around the thorax. The embedded piezoelectric sensor reacts to pressure variations caused by the fluctuation of the thoracic circumference during respiration. Thereby, several respiration patterns (e.g., inhalation and exhalation) can be acquired and recorded. Various studies [59], [60], [61] have investigated the relationship between emotion and respiration, and have shown the existence of a strong correlation between specific emotional states and respiration patterns. This can be observed by a change in breathing patterns when an individual transits from one affective state to another, thus the relevance of RSP for the current study.

Electrodermal Activity. EDA, also referred to as galvanic skin response (GSR) or skin conductance (SC), depicts the change in the electrical resistance of the skin triggered by the activation of sweat glands. The degree of activation of the sweat glands is regulated by the sympathetic nervous system and therefore is sensitive to external stimuli. EDA is considered as a good indicator of physiological arousal [62], [63], [64]. EDA data was acquired by applying a very low constant voltage to the skin through two electrodes fixed respectively at the index finger and ring finger of a participant's right hand. Based on the applied constant voltage and the measured current that flows through the skin of the participant, the skin conductance can be measured and recorded.

6.2 Signal Processing

Prior to the extraction of descriptors from each of the recorded physiological modalities, an individual preprocessing step was undertaken in order to substantially reduce the amount of noise and artefacts within each specific signal. Concerning the EMG signal, a third order bandpass Butterworth filter with a frequency range of $[0.05, 25]$ Hz was applied in order to further isolate the bursts in the signal which carry potentially useful information about the muscles' activity and thus the induced level of pain. The resulting signal was subsequently detrended (by subtracting a least-squares-fit straight line from the filtered signal) in order to focus uniquely on the fluctuations within the filtered signal. Analogously, the ECG signal was first filtered using a third order bandpass Butterworth filter with a frequency range of $[0.1, 25]$ Hz followed by signal detrending. Additionally, the filtered ECG signal was normalised in order to obtain a uniform range of signal values for all involved participants, since a huge inter-individual variance of signal values could be observed during the processing of the recorded ECG data. The RSP signal was smoothed using a third order low-pass Butterworth filter with a cut-off frequency of 0.8 Hz. Finally, the EDA signal was filtered by applying a third order low-pass Butterworth filter with a cut-off frequency set to 0.2 Hz. A sample of the preprocessed signals is depicted in Fig. 5.

6.3 Feature Extraction

Several descriptors from both frequency and temporal domains were extracted from fixed size temporal windows of the preprocessed physiological signals (see Section 8.1 for more details about the conducted temporal window analysis). A common set of 65 features was extracted from each of the involved modalities (EMG, ECG, RSP, EDA). This common set of features includes amongst others the following set of statistical features extracted from the filtered signal, as well as from its first and second temporal derivatives [65]: *mean value of the signal*, *mean value of the normalised signal*, *mean value of the absolute values of the signal*, *mean value of* *the absolute values of the normalised signal* ($3 \times 4 = 12$ features). Moreover, the following additional features from the temporal domain proposed in [18] were extracted uniquely from the filtered signal: *standard deviation of the signal, standard deviation of the normalised signal, skewness, maximum to minimum peak value ratio, kurtosis, peak amplitude (maximum peak value), peak range (difference between maximum and minimum peak values), root mean squared value of the signal, mean value of local maxima, mean value of local minima, temporal slope of the signal* (11 features). Based on [66], [67], the following set of features was also extracted uniquely from the filtered signal: *integrated EMG (IEMG), modified mean absolute values (MMAV1 and MMAV2), slope of mean absolute value (MAVSLP), simple square integral (SSI), signal variance, waveform length, slope sign change (SSC), Willison amplitude (WAMP),* $\nu - Order = \sqrt[\nu]{E\{|x_k|^{\nu}\}}$, *log-Detector* ($logDetect = exp(\frac{1}{N}\sum_i log(|x_i|))$) (11 features). Furthermore, *normalised histogram coefficients* [66] (8 features) as well as *coefficients* resulting from fitting an autoregressive model using the Burg method [68] (5 features) were also extracted.

From the frequency domain, numerous descriptors were also computed including *low frequency to very low frequency ratio* based on Welch's power spectrum density estimation, *zero crossing, frequency mode, bandwidth, central frequency, mean frequency* and *median frequency* (7 features). Additionally, specific features that capture relevant information from the non-stationary nature of the acquired signals were also computed. It comprises *stationary mean, median, area, variance* and *standard deviation* (5 features). Finally, several features were computed in order to capture the irregularities within the recorded signals. These features consist of the following: *Shannon entropy* [69], *approximate entropy (ApEn), sample entropy (SampEn), fuzzy entropy (FuzzyEn)* [70], *spectral entropy* and *Shannon entropy of the peak frequency shifting (SEPFS)* [71] (6 features).

From the ECG modality, an additional set of 58 features was extracted. Most of these features are based on the analysis of the *PQRST* waves of the recorded signals and include several statistical features (*mean, standard deviation, minimum, maximum*) computed from the *amplitudes* and *widths* of the P, Q, R, S and T wavelets, the *temporal delay* between each couple of peaks, as well as the following *angles*: $\angle PQR$, $\angle QRS$ and $\angle RST$ [72]. Subsequently, based on the detected R peaks, the heart rate variability was computed and further features were extracted from the resulting signal, including the *mean* and *root mean square deviation* of the heart rate variability. Moreover, the *slope* of a linear regression fitted to the R peaks occurrences was computed. Additionally, based on [73], wavelet transform decomposition coefficients were also extracted, using a Daubechies wavelet of order 8 at the level 4 applied on the detected and aligned R peaks. The final feature was generated by computing the *mean* of the low frequencies coefficients representing an approximation of the original ECG signal.

Finally, following the decomposition of the EDA signal into its phasic and tonic components based on a convex optimisation algorithm proposed by Greco et al. [74], 7 additional statistical features were extracted from the phasic component (*number of skin conductance responses, mean amplitude of the responses, mean, standard deviation, maximum, range* and *area under the curve* of the phasic component) and 10 more from the tonic component (*mean* and *standard deviation* of the tonic component and its first and second temporal derivatives, *maximum, minimum, range* and *area under curve* of the tonic component).

Feature Vectors. Therefore, both RSP and EMG signals are represented by feature vectors of dimensionality 65. The ECG feature vector consists of the common set of features combined with those extracted using the analysis of the *PQRST* waves and those produced throughout the wavelet decomposition of the signal, which results into a feature vector of dimensionality $65 + 58 = 123$. The EDA feature vector is generated by concatenating the set of common features with those extracted from both phasic and tonic components, resulting in a feature vector of dimensionality $65 + 7 + 10 = 82$.

7 Classification Architectures

This section provides a description of the classification architectures assessed within the scope of the current work.

Each modality is characterised by specific properties which provide valuable and distinctive insights about the level of artificially induced pain. A classification system based on a single modality should then be able to use these insights in order to perform its task to a satisfactory extent. However, the performance of the whole system can be significantly improved by appropriately combining the information provided by several modalities. Multiple classifier systems are able to take advantage of the diversity as well as the complementarity of the information extracted from each of the involved modalities in order to improve the performance of the system. Moreover, single modality classification systems can be unstable due to their reliance on one unique modality, in particular in case of missing data. Multiple classifier systems on the other hand can improve the robustness of the recognition system, since the information used to perform the classification task stems from a variety of modalities. Thus, several multiple classifier system architectures have been designed and assessed. Information fusion is performed at different levels of abstraction, using both trainable and fixed mappings.

The designed fusion architectures use Random Forests classifiers as base classifiers. Proposed by Breiman [75], Random Forests consist of a committee of bagged decision trees which are trained using a combination of both random sub-space and random sub-sampling methods. New samples are classified by applying majority voting to the decisions of the bagged trees. Random Forests are known to be efficient and robust against high dimensional data and do not require lengthy parameter searches for performance optimization in comparison to commonly used classifiers as, for example, SVMs.

The first evaluated fusion architecture consists of an early fusion approach, depicted in Fig. 6a. Early fusion consists of concatenating the descriptors extracted from each of the available modalities into one single high dimensional feature vector. A single Random Forests classifier is subsequently trained on the resulting high dimensional dataset. Some of the most prominent advantages of an early fusion approach are its simplicity and the potential reduction of the complexity of the classification task resulting from the

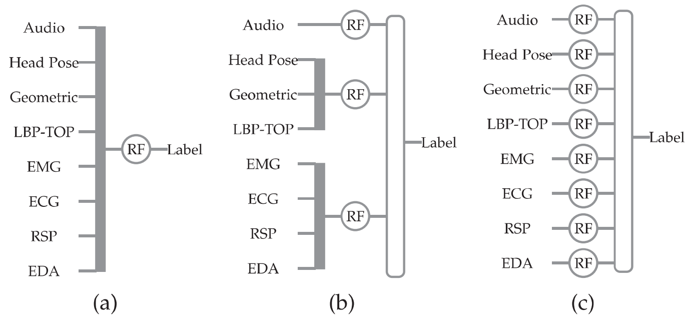


Fig. 6. *Fusion Architectures. (a):* Early Fusion. *(b):* Late Fusion A. *(c)*: Late Fusion B. For both late fusion architectures, two fixed mappings (Mean and Max) and two trainable mappings (LDA and Pseudo-inverse) are evaluated. The mappings are applied on the classification scores of the base classifiers to generate the final label of an unseen sample.

combination of complementary descriptors. Moreover, no additional training and optimisation phase is needed and therefore the whole dataset can be used for the optimisation of the base classifier. However, several drawbacks emerge from the combination of the descriptors at such an early phase. First, the resulting recognition system is inflexible and is unable to deal with missing data, since it relies on the availability of all involved modalities. Moreover, the resulting high dimensional dataset increases the computational requirements. Last, there is a high probability of running into a sub-optimal solution for the classification task due to the so called *curse of dimensionality* [76].

The next fusion architectures consist of late fusion approaches. The fusion strategy in Fig. 6b consists of concatenating the descriptors extracted from the physiological modalities into a single input channel. The same procedure is undertaken with the descriptors extracted from the video modality. Subsequently, a single Random Forests classifier is trained on each of the three input channels (audio, video, physiology), followed by the combination of the resulting scores in an aggregation layer. The last fusion strategy in Fig. 6c consists of training a single Random Forests classifier on each individual set of descriptors, followed by the combination of the base classifiers' outputs in an aggregation layer.

In the current work, several aggregation rules consisting of two fixed mappings (Mean and Max) and two trainable mappings (Linear Discriminant Analysis and Pseudo-inverse) are evaluated. In the following lines, $c \in \mathbb{N}$ represents the number of classes while $n \in \mathbb{N}$ depicts the number of base classifiers. Moreover, $N \in \mathbb{N}$ depicts the size of the testing set and $Tr \in \mathbb{N}$ depicts the size of the training set. The classification output of each base classifier $k \in \{1, \dots, n\}$ is represented by the matrix $C^k = (d^k_{i,j})_{1 \leq i \leq N, 1 \leq j \leq c}$ with $0 \leq d^k_{i,j} \leq 1,\ \forall (i,j) \in [1,N] \times [1,c]$ and $j^* \in \{1, \dots, c\}$ denotes the label assigned to an unseen sample.

Fixed Mappings. Fixed aggregation rules are simple, straightforward and characterised by the non-existence of parameters that have to be optimised in order to proceed with the combination of the base classifiers' outputs. One of the most used fixed mappings is the simple average aggregation rule (Mean). It consists of averaging the classification scores of the base classifiers for each class and subsequently assigning the label of the class with the maximum averaged score:

$$\frac{1}{n}\sum_{k=1}^{n} d^k_{i,j^*} = \max_{1 \leq j \leq c} \left. \frac{1}{n}\sum_{k=1}^{n} d^k_{i,j}\right), \ \forall i \in \{1, \dots, N\}. \tag{1}$$

Another popular fixed mapping is the maximum aggregation rule (Max). Analogously to the average aggregation rule, an unseen sample is assigned the label of the class with the maximum score amongst the outputs of the base classifiers:

$$\max_{1 \leq k \leq n} d^k_{i,j^*} = \max_{1 \leq j \leq c} \left(\max_{1 \leq k \leq n} d^k_{i,j} \right), \ \forall i \in \{1, \dots, N\}. \tag{2}$$

Trainable Mappings. Trainable combination rules are characterised by a second training step following the training of the base classifiers intended to optimise the parameters of the aggregation layer. Therefore, an extra set of data is required (and set aside) in order to proceed with an effective training of the aggregation layer. In the current work, a linear discriminant analysis classifier (LDA) [77] is trained and applied on the outputs of the base classifiers in order to assign a label to an unseen sample. The idea behind a LDA classifier is to consider all involved classes as normally distributed and sharing an identical covariance matrix. Based on these assumptions, each class's conditional probability density function is estimated. The predictions are subsequently undertaken by using the Bayes's rule, and an unseen sample is assigned the label of the class with the maximum conditional probability estimation [78].

A Pseudo-inverse (Pinv) [79] mapping has also been evaluated. The idea behind the Pseudo-inverse aggregation rule is to generate a least-squares linear mapping by computing the pseudo-inverse of the base classifiers horizontally concatenated outputs $C = [C^1, \dots, C^n] \in [0,1]^{Tr \times (cn)}$ ($C^k \in [0,1]^{Tr \times c}$ represents the output of each classifier k for the whole training set) and multiplying it with the corresponding class labels $Y \in \{0,1\}^{Tr \times c}$ accordingly to the data available in the training set:

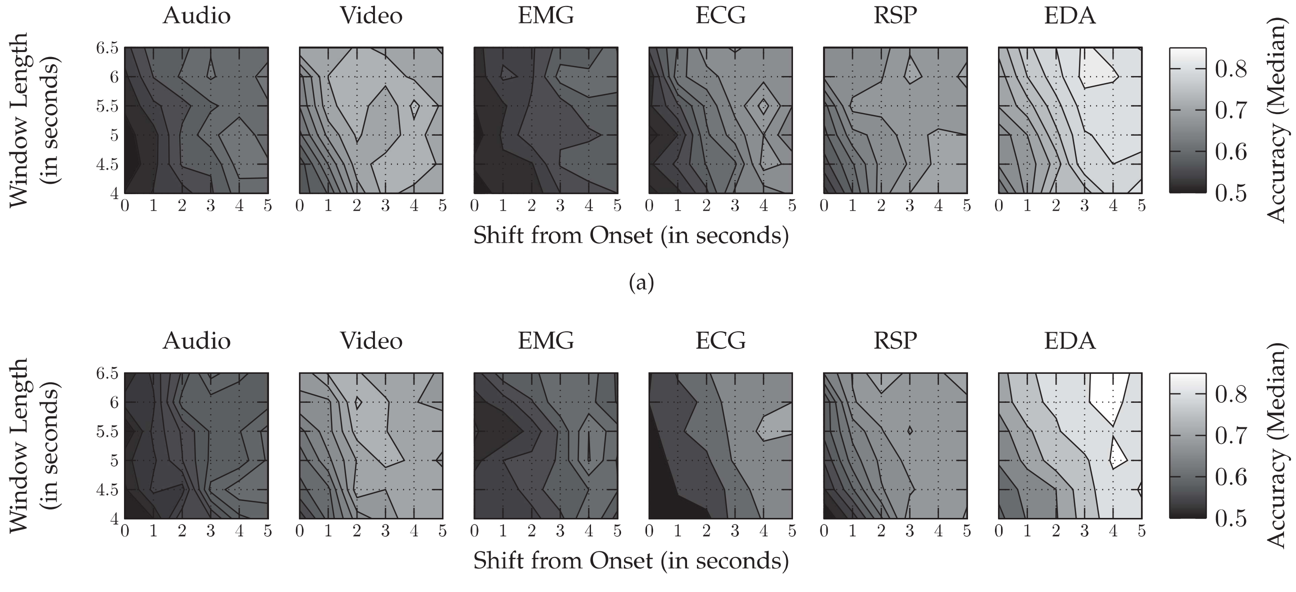
$$M \in \mathbb{R}^{cn \times c} = \lim_{\alpha \to \infty} C^T (CC^T + \alpha I)^{-1} Y. \tag{3}$$

The mapping is subsequently applied to the horizontally concatenated outputs of the base classifiers for an unseen sample and the assigned label corresponds to the class with the maximum estimated score:

$$\sum_{k=1}^{n}\sum_{j=1}^{c} d^k_{i,j} M_{l,m^*} = \max_{1 \leq m \leq c} \left. \sum_{k=1}^{n}\sum_{j=1}^{c} d^k_{i,j} M_{l,m} \right) \\ \forall i \in \{1, \dots, N\} \tag{4}$$

with $l = c(k-1) + j$ and $M = \left(M_{l,m}\right)_{1 \leq l \leq cn, 1 \leq m \leq c} \in \mathbb{R}^{cn \times c}$.

Late fusion architectures offer more flexibility in comparison to an early fusion approach since the modalities are grouped in several input channels. Moreover, the probability of running into a sub-optimal solution due to the size of the feature sets is reduced. However, the system still relies on the availability of all recorded modalities and an extra set of data is needed in order to train the aggregation layer in case a trainable combination rule is applied. Thus, a substantial amount of data is needed in order to effectively train not just the base classifiers, but the aggregation mapping as well.



(a)

(b)

Fig. 7. *Temporal Window Analysis.* (a): Left Forearm. (b): Right Forearm. The results depict the median accuracy for each evaluated temporal window, computed for each modality by applying a 10-fold cross validation evaluation in a user specific setting. The features involved in this evaluation are specific to each temporal window. These windows have lengths ranging from 4 sec to 6.5 sec and are temporally shifted in steps of 1 sec, starting from the temperature elevation onset point until a maximum shift of 5 sec.

8 Experiments and Results

In this section, a description of the undertaken experiments and the corresponding results is given. First, the experiments undertaken to proceed with the segmentation of the recorded signals consisting of defining adequate windows for modality specific feature extraction is described. Next, classification experiments and results in both user specific and user independent settings are described. Since the temperature calibration was performed individually and iteratively at the beginning of each session, the assessment is performed for each forearm separately in the Sections 8.1, 8.2 and 8.3. Further experiments with the merged data are performed and described in Section 8.4.

8.1 Temporal Window Analysis

The first experiment consisted of the evaluation of several temporal windows from which the descriptors were extracted for each specific modality in order to proceed with the classification task. This analysis was motivated by the existence of a temporal latency between the moment in time at which an artificial pain elicitation is triggered and the moment at which the reaction of a participant to this specific elicitation is observable in a given signal. Therefore, an approximation of this temporal latency could help in defining the boundaries of the response to the triggered elicitation for each signal individually and thus improve the classification performance of the recognition system.

In [27], the authors show that the level of energy within an audio signal is low during the elicitation phase before it shortly and significantly increases within the phase during which the corresponding temperature is gradually dropped to the baseline temperature. This observation corresponds to a typical demeanour of the participants observed during the experiments and consisting of the participants' breath being held during painful phases, subsequently followed by some deep exhale as soon as the temperatures became bearable and the pain receded (see Fig. 5). This heavy expiration corresponds to the aforementioned peak of audio signal energy. This observation also suggests that potentially valuable insights about the actual level of pain elicitation could be extracted from temporal windows defined within the last seconds of an elicitation.

On the other hand, facial movements as response to a painful stimulation have a lower latency compared to the audio signal. For most of the participants, observable reactions in the facial area were almost instantaneous as soon as the targeted tolerance temperature (T_3) was reached. Furthermore, the response latency in the physiological signals seems to be the highest amongst all recorded modalities. Since these physiological modalities are regulated by the sympathetic nervous system, a certain delay is to be acknowledged between the acquisition of the relayed information (related to an external stimulus) by the central nervous system and the feedback consisting of a specific response to the stimulus.

The temporal window analysis was conducted by performing a grid search, which consists of performing successive classification tasks based on descriptors specific to each modality and extracted from several windows of varying lengths and positions in time. The lengths of the windows vary between 4 sec and 6.5 sec. Each window, was temporally shifted in steps of 1 sec starting from the onset point in time when the temperature starts to increase (see Fig. 1a), with a maximum shift of 5 sec. These ranges were selected in order to avoid extracting ambiguous information from sections in time which are not related to the current pain elicitation. From each specific window, the extracted descriptors were used to perform a 10-fold cross validation evaluation of a binary classification task (T_0 versus T_3) in a user specific setting. For the audio modality, the MFCC-based descriptors are the unique features involved in this evaluation. For the video modality only the descriptors based on the tracked landmarks are involved while all

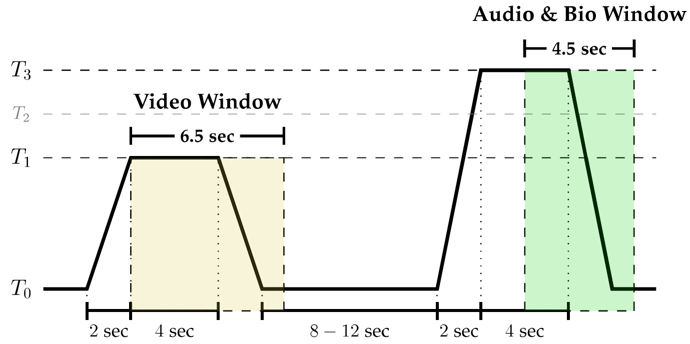


Fig. 8. *Signal segmentation.* The video features are extracted from a window of length 6.5 sec with a temporal shift of 2 sec from the onset. The audio and physiological features are extracted from an identical window of length 4.5 sec with a temporal shift of 4 sec from the onset.

descriptors extracted from each physiological modality are used to proceed with this evaluation.

Figs. 7a and 7b depict the results of the performed grid search for the left forearm and right forearm respectively. The results displayed correspond to the median of the classification accuracy of the user specific 10-fold cross validation evaluation for each specific modality. A first look at these figures points at the similarity of the results for both forearms. At a glance, EDA appears to achieve the best classification performances in comparison to the other modalities. Moreover, both EMG and audio modalities appear to be the worst performing modalities. Still, most of the modalities achieve low classification rates when the descriptors are extracted from windows having a lower boundary located within the first 2 sec of pain elicitation, regardless of the length of the windows.

Furthermore, a substantial improvement in the classification performances can be observed when the temporal windows are starting within a range of 3 sec to 5 sec following the temperature elevation onset, for both audio and physiological modalities. On the other hand, the performance improvements concerning the video modality appear to rely more on the length of the window than on the temporal shift, since relatively good classification performances are depicted for windows extracted within temporal shifts ranging from 1 sec to 5 sec. Thus, the exact combination of window length and temporal shift in order to achieve the best classification performance depends on the nature of each modality which confirms the assumptions stipulated earlier.

Based on these findings, a modality specific signal segmentation is performed as depicted in Fig. 8. Video descriptors are extracted from temporal windows with a length of 6.5 sec and a temporal shift of 2 sec from the onset. The descriptors of the audio and physiological modalities are extracted from identical windows of length 4.5 sec, with a temporal shift of 4 sec.

8.2 User Specific Binary Classification Results

The next experiment consists of the evaluation of the pain recognition system in a user specific setting. For this evaluation, the descriptors extracted from the audio modality are concatenated into a single input channel. The evaluation is performed as a "No Pain" versus "Pain" binary classification problem, consisting of the discrimination between the baseline temperature (T_0) and each of the 3 different temperatures (T_1, T_2, T_3). Therefore, a stratified 10-fold cross validation evaluation is performed on the dataset specific to each single participant. Moreover, the evaluation is performed for each modality individually, followed by the evaluation of the fusion strategies presented in Section 7. The results, consisting of the average classification accuracy and the standard deviation over the entire 40 participants, are depicted in Table 1.

Overall, low pain elicitation temperatures (T_1 and T_2) are very difficult to discriminate from the baseline temperature (T_0). The best performance for the classification task T_0

TABLE 1
User Specific Classification Results (**Mean**(**in**%) ± **Standard Deviation**)

Forearm	Left			Right		
Task	$\mathbf{T_0 vs. T_1}$	$\mathbf{T_0 vs. T_2}$	$\mathbf{T_0 vs. T_3}$	$\mathbf{T_0 vs. T_1}$	$\mathbf{T_0 vs. T_2}$	$\mathbf{T_0 vs. T_3}$
Audio	51.91 ± 8.47	52.53 ± 9.38	66.64 ± 17.68	51.29 ± 6.74	52.45 ± 10.81	66.45 ± 16.08
Head Pose	48.08 ± 8.41	52.52 ± 10.13	70.58 ± 14.23	50.60 ± 8.76	56.37 ± 10.04	71.18 ± 13.84
Geometric	49.60 ± 7.31	52.88 ± 9.25	72.44 ± 12.97	50.38 ± 9.06	57.22 ± 9.29	72.51 ± 14.72
LBP-TOP	50.36 ± 7.00	53.72 ± 9.79	72.40 ± 15.76	50.32 ± 8.40	57.35 ± 10.71	75.50 ± 12.41
EMG	48.67 ± 9.14	52.16 ± 9.29	60.15 ± 14.35	48.26 ± 7.94	52.09 ± 8.28	61.53 ± 15.90
ECG	50.37 ± 7.90	53.39 ± 9.08	68.41 ± 13.61	49.23 ± 7.41	53.98 ± 8.22	68.81 ± 15.60
RSP	50.44 ± 9.23	53.94 ± 10.95	70.29 ± 13.16	50.25 ± 8.08	55.32 ± 8.19	70.22 ± 14.02
EDA	**52.74 ± 7.64**	**59.96 ± 12.86**	80.24 ± 13.51	48.84 ± 8.62	59.16 ± 14.31	79.78 ± 16.03
Early Fusion	51.46 ± 7.89	57.40 ± 9.52	81.56 ± 12.12	50.88 ± 8.28	59.45 ± 11.75	82.63 ± 11.56
Late Fusion A (Mean)	50.59 ± 8.94	59.02 ± 10.81	**83.13 ± 12.00**	51.61 ± 7.87	**60.91 ± 12.69**	84.67 ± 11.01
Late Fusion A (Max)	51.06 ± 9.04	59.82 ± 10.08	82.53 ± 12.20	50.67 ± 8.46	60.65 ± 12.24	**84.72 ± 11.09***
Late Fusion A (LDA)	50.90 ± 7.53	58.60 ± 10.53	81.02 ± 12.68	50.00 ± 7.47	56.94 ± 10.15	81.24 ± 12.53
Late Fusion A (Pinv)	50.24 ± 7.43	58.27 ± 10.42	82.16 ± 12.85	49.83 ± 7.04	56.99 ± 10.48	82.30 ± 12.49
Late Fusion B (Mean)	51.36 ± 8.72	58.30 ± 10.60	82.16 ± 12.81	50.94 ± 8.30	59.88 ± 12.42	83.36 ± 11.52
Late Fusion B (Max)	50.19 ± 8.72	58.47 ± 11.74	83.13 ± 12.85	**52.64 ± 8.06**	59.71 ± 13.46	83.19 ± 12.49
Late Fusion B (LDA)	50.11 ± 6.38	57.62 ± 9.83	80.46 ± 13.04	50.14 ± 6.77	57.14 ± 12.15	81.16 ± 14.37
Late Fusion B (Pinv)	49.36 ± 6.61	57.79 ± 9.58	80.42 ± 13.07	51.04 ± 7.46	57.39 ± 12.10	81.33 ± 14.63

The best performance achieved by a single modality is underlined and the best overall performance is depicted in bold. An asterisk (*) indicates a significant performance improvement between the best performing fusion architecture and the best performing single modality. The test has been conducted using a Wilcoxon signed rank test with a significance level of 5%.

TABLE 2
User Independent Classification Results (**Mean**(in%) ± **Standard Deviation**)

Forearm	Left			Right		
Task	$\mathbf{T_0 vs. T_1}$	$\mathbf{T_0 vs. T_2}$	$\mathbf{T_0 vs. T_3}$	$\mathbf{T_0 vs. T_1}$	$\mathbf{T_0 vs. T_2}$	$\mathbf{T_0 vs. T_3}$
Audio	50.80 ± 5.50	52.25 ± 6.24	63.40 ± 15.63	51.04 ± 7.01	49.73 ± 5.84	65.04 ± 13.99
Head Pose	50.42 ± 6.71	50.09 ± 7.08	61.07 ± 15.58	52.55 ± 5.67	51.84 ± 6.75	63.78 ± 16.28
Geometric	51.06 ± 5.36	52.19 ± 6.95	65.84 ± 15.44	52.46 ± 5.09	54.61 ± 6.85	65.39 ± 17.16
LBP-TOP	$\underline{51.69 \pm 6.79}$	51.34 ± 6.21	60.82 ± 13.30	51.43 ± 5.63	51.89 ± 6.11	63.74 ± 13.27
EMG	48.96 ± 6.37	48.00 ± 4.83	56.23 ± 9.30	49.65 ± 5.76	48.46 ± 6.03	62.00 ± 14.01
ECG	51.16 ± 5.91	51.29 ± 5.45	65.04 ± 13.24	49.57 ± 5.73	51.57 ± 6.67	67.26 ± 14.01
RSP	51.35 ± 6.43	50.68 ± 5.49	65.86 ± 15.53	49.24 ± 6.88	49.84 ± 5.47	66.90 ± 14.58
EDA	48.93 ± 5.84	$\underline{\mathbf{62.34} \pm \mathbf{10.50}}$	$\underline{80.43 \pm 13.18}$	$\underline{53.13 \pm 5.82}$	$\underline{62.87 \pm 12.10}$	$\underline{82.16 \pm 13.40}$
Early Fusion	**51.88 ± 5.81**	59.91 ± 8.13	80.79 ± 12.27	**53.86 ± 5.70**	62.37 ± 10.85	80.61 ± 12.33
Late Fusion A (Mean)	50.75 ± 5.28	61.05 ± 9.70	80.86 ± 12.23	52.65 ± 6.83	61.88 ± 10.04	81.58 ± 12.18
Late Fusion A (Max)	51.03 ± 5.71	60.46 ± 9.41	80.70 ± 12.14	52.15 ± 7.04	61.89 ± 10.50	81.58 ± 12.10
Late Fusion A (LDA)	49.88 ± 6.90	58.72 ± 10.96	70.57 ± 12.63	50.87 ± 7.11	62.36 ± 10.88	82.21 ± 13.18
Late Fusion A (Pinv)	49.93 ± 6.51	58.62 ± 10.85	81.04 ± 11.76	49.42 ± 6.63	62.83 ± 11.09	82.81 ± 12.21
Late Fusion B (Mean)	48.91 ± 5.26	58.25 ± 8.52	77.84 ± 14.25	53.20 ± 7.00	61.89 ± 10.16	80.01 ± 13.27
Late Fusion B (Max)	49.74 ± 5.72	58.72 ± 9.33	81.08 ± 12.99	51.73 ± 6.47	61.97 ± 9.64	81.31 ± 11.71
Late Fusion B (LDA)	50.75 ± 7.60	59.40 ± 10.87	81.46 ± 11.95	51.71 ± 6.04	62.33 ± 12.01	83.36 ± 12.75
Late Fusion B (Pinv)	51.00 ± 6.89	59.44 ± 10.28	**81.76 ± 12.08**	51.45 ± 5.74	**62.88 ± 11.02**	**83.95 ± 12.65***

A leave one user out (LOUO) cross validation evaluation is performed. The best performance achieved by a single modality is underlined and the best overall performance is depicted in bold. An asterisk (*) indicates a significant performance improvement between the best performing fusion architecture and the best performing single modality. The test has been conducted using a Wilcoxon signed rank test with a significance level of 5%.

versus T_1 is achieved by the EDA with a performance of 52.74 percent for the left forearm and by the second late fusion architecture in combination with the maximum aggregation rule for the right forearm, with an average accuracy of 52.64 percent. Concerning the classification task T_0 versus T_2, both the EDA and the first late fusion architecture in combination with the average aggregation rule achieve the best performances for the left and right forearm, with average accuracies of 59.96 and 60.91 percent respectively. Although these values are significantly above chance level, only the classification system based on EDA is able to discriminate between those low levels of pain elicitation to an acceptable extent. Meanwhile, each single modality achieves relatively good classification performance for the classification problem T_0 versus T_3. This can be explained by the fact that the stimuli induced with the pain tolerance temperature (T_3) resulted in more observable reactions in each modality. EDA is once again the best performing single modality and significantly outperforms all the other modalities, while the worst performing single modality consists of the trapezius EMG with an average classification accuracy of 60.15 percent and 61.53 percent for the left and right forearm respectively.

Moreover, the best performing fusion architecture is the first late fusion architecture (Late Fusion A) in combination with fixed mappings. The performances of the fixed fusion mappings are quite similar, with the average combination rule performing best in case of the left forearm with an average accuracy of 83.13 percent, and the maximum aggregation rule performing best in case of the right forearm with an average accuracy of 84.72 percent. Additionally, fixed mappings perform significantly better than trainable mappings, regardless of the applied late fusion architecture. This can be explained by the fact that in a user specific setting, the amount of training data is insufficient in order to effectively train the base classifiers and optimise a trainable aggregation layer.

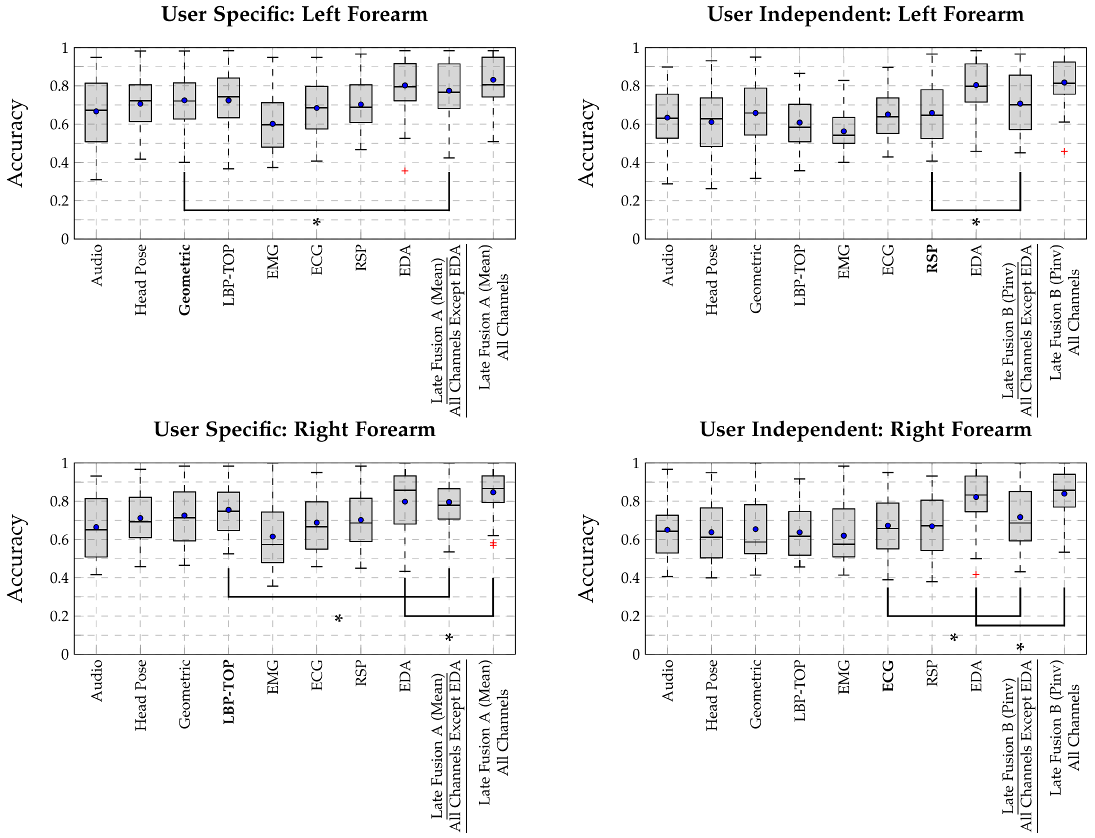
8.3 User Independent Binary Classification Results

The following experiment consists of the evaluation of the generalisation capabilities of the different classification models to unseen users by performing a leave one user out (LOUO) cross validation evaluation with the same binary classification settings as in Section 8.2. The results of the evaluation are depicted in Table 2.

At a glance, there is a significant drop of performance for the video modality in comparison to the results computed in a user specific setting (see Table 1). This can be explained by the diversity of expressiveness of pain perception due to user specific attributes. This drop of performance can also be seen in the other modalities, except for the EDA which seems not to be affected by individual characteristics. As a matter of fact, the performances of the EDA are quite similar, and in some cases better than those yielded in a user specific setting. Analogously to the user specific results, EDA significantly outperforms the other modalities.

The second late fusion architecture (Late Fusion B) performs in most cases better than the first late fusion architecture (Late Fusion A) in this setting. Moreover, in contrast to the results yielded in a user specific setting, trainable mappings perform in most cases better than fixed mappings. The amount of training data available in a LOUO cross validation seems to be sufficient to effectively train the base classifiers and the trainable fusion layer. The best classification performances are yielded for the classification task T_0 versus T_3 and for each forearm by the second late fusion architecture in combination with the pseudo-inverse fusion layer, with performances of 81.76 and 83.95 percent for the left and right forearm respectively.

For some further assessment of the proposed fusion approaches, an additional experiment is carried out using all previously described channels except the EDA. This experiment is motivated by the previous results (see Tables 1 and 2) which depict a very high correlation between the performance of the fusion architectures and the performance of



(a)

(b)

Fig. 9. $\mathbf{T_0}$ *versus* $\mathbf{T_3}$ *Results Comparison. (a):* User Specific (10-fold cross validation). *(b):* User Independent (LOUO cross validation). An asterisk (*) indicates a significant performance improvement between the fusion architecture and the corresponding best performing single modality. The test has been conducted using a Wilcoxon signed rank test with a significance level of 5 percent. Within each box plot, the mean and median classification accuracy across all 40 participants are depicted respectively with a dot and a horizontal line.

EDA. Although the fusion approaches outperform the best performing single modality, the benefit of the combination of the information stemming from different sources is overshadowed by the performance of EDA. Therefore, the best performing fusion architectures in each evaluation setting (Late Fusion A with the average aggregation rule for a user specific evaluation and Late Fusion B with the pseudo-inverse aggregation rule for a user independent evaluation) are used to perform the fusion of all involved channels except EDA. A summary of the results for the classification problem T_0 versus T_3 is depicted in Fig. 9.

In the absence of EDA, the best performing modality in a user specific evaluation setting is the video modality. The best performing single channel consists of the geometric and LBP-TOP features with classification rates of 72.44 and 75.50 percent for the left and the right forearm respectively. The fusion approach (Late Fusion A (Mean)) significantly outperforms the video modality for both sessions with classification rates of 77.46 and 79.54 percent for the left and the right forearm respectively. In a user independent setting, both RSP and ECG modalities perform best with similar classification performances. RSP performs slightly better with an average accuracy of 65.86 percent for the left forearm, and ECG performs better with a performance of 67.26 percent for the right forearm. The fusion approach (Late Fusion B (Pinv)) significantly outperforms both channels with classification performances of 70.71 percent for the left forearm, and 71.66 percent for the right forearm. In both evaluation settings and for both forearms, there is a significant drop of performance when the information stemming from EDA is excluded. However, the fusion approaches are still able to significantly outperform the best performing single modality in all cases by combining the information provided by the remaining sources.

Altogether, in both user specific and user independent settings, the discrimination between the different levels of pain becomes more challenging the lower the level of pain elicitation gets. Each single modality provides valuable insights for the recognition of the different pain intensities, whereby some of them seem to be more appropriate for the current experimental settings (thermal pain elicitation). Although the recorded audio material comprises substantially paralinguistic vocalisations, the performance of the audio modality is significantly better than chance for the classification task T_0 versus T_3 in both user specific and user independent settings. The audio channel also outperforms the trapezius EMG in all classification tasks and settings. Moreover, the sensor used to perform the audio recordings is less invasive than physiological sensors and audio data is also much cheaper to acquire. Furthermore, the recorded audio signal does not require any substantial processing step (except for the usual signal filtering and denoising steps) like the localisation of the facial area for the video signal as an example. Finally, the audio channel is less affected than the video channel by the inter-individual differences in pain perception and pain expressions (see Fig. 9). Therefore, the audio signal is a promising and relevant modality for the development of a pain intensity recognition system.

TABLE 3
Multi-Class Classification Results
(**Mean(in%) ± Standard Deviation)**

Dataset	Left Forearm	Right Forearm	Both Forearms
Audio	31.99 ± 7.66	31.87 ± 7.65	32.35 ± 6.87
Head Pose	30.75 ± 7.53	33.31 ± 7.87	32.06 ± 7.08
Geometric	33.76 ± 7.61	34.37 ± 9.57	34.22 ± 7.54
LBP-TOP	30.97 ± 6.34	31.80 ± 7.94	30.87 ± 5.99
EMG	28.34 ± 4.99	30.82 ± 7.67	29.73 ± 5.30
ECG	31.83 ± 6.76	33.62 ± 7.17	33.58 ± 6.85
RSP	33.16 ± 7.83	33.62 ± 7.61	33.89 ± 5.90
EDA	$\underline{42.17 \pm 9.11}$	$\underline{41.63 \pm 9.89}$	$\underline{42.92 \pm 7.07}$
Late Fusion B (Pinv)	$\mathbf{42.48 \pm 8.35}$	$\mathbf{43.11 \pm 7.93}$	$\mathbf{43.89 \pm 7.61}$

The results correspond to a 4-class classification task (T_0 versus T_1 versus T_2 versus T_3). The random performance for a 4-class classification task is 25%. The evaluation is performed in a LOUO setting. The best performance achieved by a single modality is underlined and the best overall performance is depicted in bold.

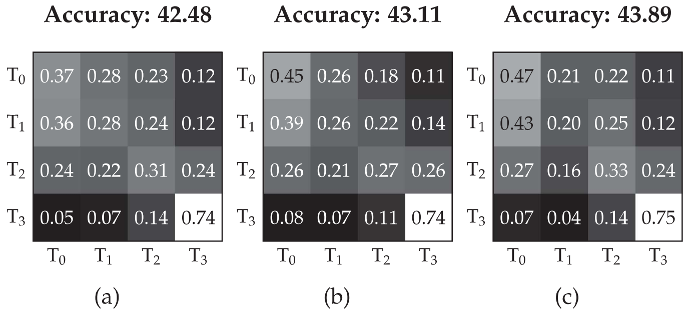
The significant drop of performance of the video modality in a user independent evaluation points to the negative effect of generalisation on a recognition system based uniquely on the video modality. A personalisation scheme is needed in this case in order to improve the classification performance of the system. The worst performing modality so far has been the EMG of the trapezius muscle. While both RSP and ECG perform similarly in both user specific and user independent settings, EDA has proven to be the best performing single modality in all evaluated settings. EDA not only significantly outperforms all the other modalities but also does not seem to be affected by the variety of inter-individual responses to pain. However, this observation is susceptible to be biased by the current experimental settings which consist of an isolated and controlled laboratory environment combined with pain elicitation through thermal stimuli. Further evaluations with diverse experiments covering different types of pain (chronic and acute pain) in both experimental and clinical settings, have to be carried out in order to better assess the relevance of EDA for pain assessment.

Finally, for the task T_0 versus T_3 in both user specific and independent settings, the proposed fusion architectures are able to improve the performance of the recognition system by combining the insights provided by each specific modality. The performance of each fusion approach depends substantially on the amount of data available for the training phase. Given enough training data, trainable mappings are able to outperform fixed mappings.

8.4 Multi-Class Classification Experiments in a User Independent Setting

In the previous experiments, the data specific to each forearm was assessed separately. This was motivated by the fact that the calibration of the temperatures was performed individually at the beginning of each session, resulting in different ranges of temperature for each forearm. However, the results depicted so far are quite similar, which hints at the similarity of the responses, regardless of the forearm on which the elicitations are performed. Based on this observation, further experiments, involving the merged data of both sessions, are conducted.

In Table 3, the results of a 4-class classification task in a user independent setting are depicted. A comparison between the performance of the pain intensity classification system is addressed when the data specific to each session is assessed separately and when it is combined in a single dataset. Similarly to the results depicted so far, EDA significantly outperforms the other modalities and the overall performance of the classification system is improved by the fusion architecture. The yielded classification rates are quite similar in all three cases. This can also be seen in the corresponding confusion matrices of the late fusion classification approach depicted in Fig. 10. The lower temperatures T_1 and T_2 are mostly confused with the baseline temperature T_0, while the pain tolerance temperature can be effectively classified.



Accuracy: 42.48

	T_0	T_1	T_2	T_3
T_0	0.37	0.28	0.23	0.12
T_1	0.36	0.28	0.24	0.12
T_2	0.24	0.22	0.31	0.24
T_3	0.05	0.07	0.14	0.74

(a)

Accuracy: 43.11

	T_0	T_1	T_2	T_3
T_0	0.45	0.26	0.18	0.11
T_1	0.39	0.26	0.22	0.14
T_2	0.26	0.21	0.27	0.26
T_3	0.08	0.07	0.11	0.74

(b)

Accuracy: 43.89

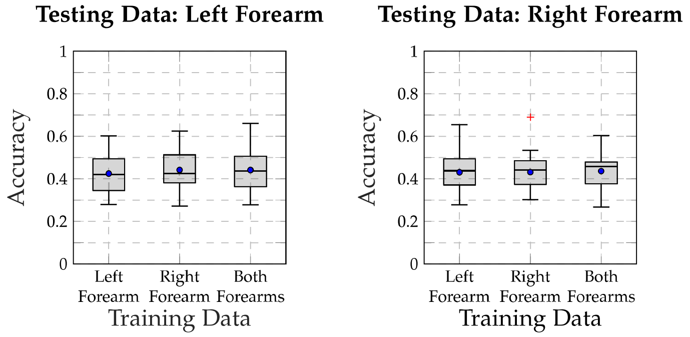
	T_0	T_1	T_2	T_3
T_0	0.47	0.21	0.22	0.11
T_1	0.43	0.20	0.25	0.12
T_2	0.27	0.16	0.33	0.24
T_3	0.07	0.04	0.14	0.75

(c)

Fig. 10. 4*-class Classification Task Confusion Matrices (Late Fusion B (Pinv) in a LOUO setting).* (a): Left Forearm. (b): Right Forearm. (c): Both Forearms. The rows correspond to the ground truth, while the columns correspond to the predictions.

Furthermore, an experiment is performed by training the classification architecture on either datasets separately and also on the combined dataset, and subsequently performing the evaluation on the data specific to either the left or the right forearm. The results of the evaluation are depicted in Fig. 11. The similarity of the depicted results regardless of the data used to train the classification architecture suggests that there is no significant difference between the data specific to both forearms.

Therefore, the previously conducted experiments (see Section 8.3) are reiterated, but this time based on the combined data of both sessions. Additionally, a 3-class classification task involving the baseline temperature, and both temperatures T_2 and T_3 is conducted. This is motivated by the fact that T_1 is mostly confused to T_0 and can not be considered as an effective pain elicitation temperature. The elicitations performed with this specific temperature could not



(a)

(b)

Fig. 11. 4*-class Classification Task Results Comparison (Late Fusion B (Pinv)) in a LOUO setting.* (a): The test evaluation is performed on the data specific to the left forearm. (b): The test evaluation is performed on the data specific to the right forearm.

TABLE 4
Classification Results (**Mean(in%) ± Standard Deviation**)

Task	Random	Audio	Head Pose	Geometric	LBP-TOP	EMG	ECG	RSP	EDA	Late Fusion B (Pinv)
T_0 vs. T_1	50.00	49.23 ± 4.37	51.73 ± 4.71	$\underline{\mathbf{52.58 \pm 4.00}}$	51.50 ± 4.34	49.97 ± 5.48	50.39 ± 3.58	50.21 ± 4.75	52.14 ± 3.95	51.39 ± 4.18
T_0 vs. T_2	50.00	50.19 ± 5.47	51.68 ± 5.10	52.87 ± 4.49	51.73 ± 4.23	50.50 ± 4.99	51.69 ± 5.16	52.04 ± 5.61	$\underline{\mathbf{62.96 \pm 9.02}}$	62.28 ± 8.98
T_0 vs. T_3	50.00	64.75 ± 14.27	63.05 ± 14.28	66.22 ± 14.48	62.42 ± 12.18	59.33 ± 10.18	66.28 ± 12.59	67.27 ± 11.17	$\underline{82.23 \pm 10.57}$	**83.39 ± 10.23***
T_0 vs. T_2 vs. T_3	33.33	42.80 ± 8.77	42.94 ± 9.48	45.15 ± 10.10	41.78 ± 8.12	39.39 ± 6.43	44.42 ± 8.41	45.18 ± 8.19	$\underline{57.84 \pm 10.51}$	**59.53 ± 9.94***
T_0 vs. T_1 vs. T_2 vs. T_3	25.00	32.35 ± 6.87	32.06 ± 7.08	34.22 ± 7.54	30.87 ± 5.99	29.73 ± 5.30	33.58 ± 6.85	33.89 ± 5.90	$\underline{42.92 \pm 7.07}$	**43.89 ± 7.61**

These results have been achieved by merging the data specific to each forearms into a single set and performing a LOUO cross validation evaluation. The best performance achieved by a single modality is underlined and the best overall performance is depicted in bold. An asterisk (*) indicates a significant performance improvement between the fusion architecture and the corresponding best performing single modality. The test has been conducted using a Wilcoxon signed rank test with a significance level of 5%.

trigger any significant reaction in any of the recorded modalities. The results of the evaluation are depicted in Table 4. The depicted results are in conformity with the previous findings, derived from individual forearms. The fusion architecture outperforms the best performing modality in both multi-class classification tasks and for the binary classification task T_0 versus T_3. The improvement is significant with classification rates of 83.39% (p-value: 1.1%) and 59.53% (p-value: 2.2%) for both T_0 versus T_3 and T_0 versus T_2 versus T_3 classification tasks respectively. By taking into account that the class labels are ordinal scaled, the average deviation in absolute value of the predicted class from the true one (*MAE*) [80], [81] for both classification tasks are respectively 0.468 and 0.811. The observed agreement based on linear (respectively quadratic) weights [82] is respectively 0.750 (0.826) and 0.728 (0.844) for each of both classification tasks.

9 Conclusion

In this work, several classifier fusion strategies have been evaluated within the scope of the development of a multimodal pain recognition system. The assessment of the proposed approaches is performed on the recently recorded *SenseEmotion Database*, which consists of several individuals subjected to three gradually increasing levels of artificially induced pain stimuli. The authors suggest for the first time a combination of three distinctive modalities (Audio, Video, Physiology) for the recognition of artificially induced pain intensities. The fusion approaches consist of a combination of modality specific descriptors at several levels of abstraction with different aggregation rules (fixed and trainable mappings). EDA has proven to be the best performing single modality regardless of the classification setting, and seems not to be affected by the individual characteristics of each participant.

Furthermore, the experimental results have proven the effectiveness of the proposed fusion approaches for these specific experimental settings. Late fusion architectures in combination with fixed mappings are able to outperform the best performing single modality in a user specific classification setting. Moreover, late fusion architectures combined with trainable mappings perform better than those combined with fixed mappings in a user independent setting, and improve the performance of a classification system based uniquely on the best performing single modality. These findings suggest that the amount of data available at the training phase plays a crucial role in the selection of an appropriate fusion strategy which can substantially improve the performance of a pain recognition system.

Still, the assessment and recognition of pain intensities remains very challenging. Furthermore, the data used for the current assessment stems from an experimental setting in a controlled environment. Therefore, the current assessment does not reflect the conditions of a clinical setting. In order to realise a reliable online pain recognition system, more realistic data are to be gathered and evaluated. Several challenges have to be addressed, beginning with the sensor system to be used in a realistic context in order to reliably record the data. This also concerns the actual real time implementation of several data pre-processing steps as well as the design and implementation of the classification architectures. In the future iterations of the current work, fusion approaches which are robust against missing and erroneous data as well as feature selection for dimensionality reduction should be addressed. Also, deep learning fusion architectures have shown promising results in different fields of application and are therefore believed to be able to significantly improve the performance as well as the robustness of a pain recognition system. Furthermore, the extent to which the designed approaches can be applied for the discrimination between pain intensities and different types of emotional states resulting from the combination of different levels of arousal and valence (e.g., stress, disgust, anger) has not been addressed and therefore constitutes an interesting extension of the current work.

Acknowledgments

This paper is based on work done within the project *SenseEmotion* funded by the Federal Ministry of Education and Research (BMBF). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] R. C. Coghill, J. G. McHaffie, and Y.-F. Yen, “Neural correlates of interindividual differences in the subjective experience of pain,” *Proc. Nat. Acad. Sci. United States Amer.*, vol. 100, no. 14, pp. 8538–8542, 2003.

[2] C. S. Nielsen, A. Stubhaug, D. D. Price, O. Vassend, N. Czajkowski, and J. R. Harris, “Individual differences in pain sensitivity: Genetic and environment contributions,” *Pain*, vol. 136, no. 1, pp. 21–29, 2008.

[3] R. C. Coghill, “Individual differences in the subjective experience of pain: New insights into mechanisms and models,” *Headache*, vol. 50, no. 9, pp. 1531–1535, 2010.

[4] E. B. Kim, H.-S. Han, J. H. Chung, B. R. Park, S.-N. Lim, K. H. Yim, Y. D. Shin, K. H. Lee, W.-J. Kim, and S. T. Kim, “The effectiveness of a self-reporting bedside pain assessment tool for oncology inpatients,” *J. Palliative Med.*, vol. 15, no. 11, pp. 1222–1233, 2012.

[5] N. C. De Knegt, F. Lobbezoo, C. Schuengel, H. M. Evenhuis, and E. J. A. Scherder, “Self-reporting tool on pain in people with intellectual disabilities (STOP-ID!): A usability study,” *Augmentative Alternative Commun.*, vol. 32, no. 1, pp. 1–11, 2016.

[6] G. A. Hawker, S. Mian, T. Kendzerska, and M. French, “Measures of adult pain: Visual analog scale for pain (VAS pain), numeric rating scale for pain (NRS pain), McGill pain questionnaire (MPQ), short-form McGill pain questionnaire (SF-MPQ), chronic pain grade scale (CPGS), short form-36 bodily pain scale (SF-36 BPS), and measure of intermittent and constant osteoarthritis pain (ICOAP),” *Arthritis Care Res.*, vol. 63, no. S11, pp. S240–S252, 2011.

[7] C. Eckard, C. Asbury, B. Bolduc, C. Camerlengo, J. Gotthardt, L. Healy, L. Waialar, C. Zeigler, J. Childers, and J. Horzempa, “The integration of technology into treatment programs to aid in the reduction of chronic pain,” *J. Pain Manage. Med.*, vol. 2, no. 3, 2016, Art. no. 118.

[8] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Gruss, H. Neumann, J. Kim, F. Schwenker, E. André, H. C. Traue, and S. Walter, “The SenseEmotion Database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system,” in *Proc. Multimodal Pattern Recognit. Social Signals Human-Comput.-Interaction*, 2017, pp. 127–139.

[9] 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 *Proc. Face and Gesture*, 2011, pp. 57–64.

[10] P. Ekman and W. V. Friesen, “Measuring facial movement,” *Environ. Psychology Nonverbal Behavior*, vol. 1, no. 1, pp. 56–75, 1976.

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

[12] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, S. Crawcour, P. Werner, A. Al-Hamadi, and A. Andrade, “The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system,” in *Proc. IEEE Int. Conf. Cybern.*, 2013, pp. 128–131.

[13] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. C. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. J. Watson, A. C. Williams, M. Pantic, and N. Bianchi-Berthouze, “The automatic detection of chronic pain-related expression: Requirements, challenges and multimodal dataset,” *IEEE Trans. Affective Comput.*, vol. 7, no. 4, pp. 435–451, Oct.-Dec. 2016.

[14] 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*, vol. 136, no. 1, pp. e124–e131, 2015.

[15] P. Thiam, V. Kessler, and F. Schwenker, “Hierarchical combination of video features for personalised pain level recognition,” in *Proc. 25th Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn.*, 2017, pp. 465–470.

[16] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, “Automatic pain assessment with facial activity descriptors,” *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 286–299, Jul.-Sep. 2017.

[17] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, “Deep pain: Exploiting long short-term memory networks for facial expression classification,” *IEEE Trans. Cybern.*, 2017, doi: 10.1109/TCYB.2017.2662199.

[18] 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, pp. 1–14, 2015.

[19] M. Kächele, P. Thiam, M. Amirian, F. Schwenker, and G. Palm, “Methods for person-centered continuous pain intensity assessment from bio-physiological channels,” *IEEE J. Select. Topics Signal Process.*, vol. 10, no. 5, pp. 854–864, Aug. 2016.

[20] M. Kächele, M. Amirian, P. Thiam, P. Werner, S. Walter, G. Palm, and F. Schwenker, “Adaptive confidence learning for the personalization of pain intensity estimation systems,” *Evolving Syst.*, vol. 8, no. 1, pp. 1–13, 2016.

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

[22] M. Schels, M. Glodek, S. Meudt, S. Scherer, M. Schmidt, G. Layher, S. Tschechne, T. Brosch, D. Hrabal, S. Walter, H. C. Traue, G. Palm, F. Schwenker, M. Rojc, and N. Campbell, “Multi-modal classifier-fusion for the recognition of emotions,” in *Coverbal Synchrony in Human-Machine Interaction*. Boca Raton, FL, USA: CRC Press, 2013, pp. 73–97.

[23] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, “Pattern classification with missing data: A review,” *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, 2010.

[24] J. Wagner, E. André, F. Lingenfelser, and J. Kim, “Exploring fusion methods for multimodal emotion recognition with missing data,” *IEEE Trans. Affective Comput.*, vol. 2, no. 4, pp. 206–218, Oct.-Dec. 2011.

[25] M. Kächele, M. Schels, P. Thiam, and F. Schwenker, “Fusion mappings for multimodal affect recognition,” in *Proc. IEEE Symp. Series Comput. Intell.*, 2015, pp. 207–313.

[26] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, “Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges,” in *Proc. 5th Int. Workshop Audio/Visual Emotion Challenge*, 2015, pp. 9–16.

[27] P. Thiam, V. Kessler, S. Walter, G. Palm, and F. Scwenker, “Audio-visual recognition of pain intensity,” in *Proc. Multimodal Pattern Recognit. Social Signals Human-Comput.-Interaction*, 2017, pp. 110–126.

[28] V. Kessler, P. Thiam, M. Amirian, and F. Schwenker, “Pain recognition with camera photoplethysmography,” in *Proc. 7th Int. Conf. Image Process. Theory Tools Appl.*, 2017, pp. 1–5.

[29] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, “AVEC 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[30] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, “Automatic pain recognition from video and biomedical signals,” in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 4582–4587.

[31] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, *Engineering Applications of Neural Networks*. New York, NY, USA: Springer International Publishing, 2015, ch. Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity, pp. 275–285.

[32] M. Kächele, P. Werner, S. Walter, A. Al-Hamadi, and F. Schwenker, “Bio-visual fusion for person-independent recognition of pain intensity,” in *Proc. Int. Workshop Multiple Classifier Syst.*, 2015, pp. 220–230.

[33] P. Thiam and F. Schwenker, “Multi-modal data fusion for pain intensity assessement and classification,” in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl.*, 2017, pp. 1–6.

[34] B. Sun, L. Li, X. Wu, T. Zuo, Y. Chen, G. Zhou, J. He, and X. Zhu, “Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild,” *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 125–137, 2016.

[35] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, “Multiple kernel learning for emotion recognition in the wild,” in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 517–524.

[36] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, “Multimodal emotion recognition using deep neural networks,” in *Proc. Neural Inf. Process.*, 2017, pp. 811–819.

[37] F.-S. Tsai, Y.-L. Hsu, W.-C. Chen, Y.-M. Weng, C.-J. Ng, and C.-C. Lee, “Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions,” in *Proc. Interspeech*, 2016, pp. 92–96.

[38] J. Wagner, T. Lingenfelser, Florian abd Baur, I. Damian, F. Kistler, and E. André, “The social signal iterpretation (SSI) framework: Multimodal signal processing and recognition in real-time,” in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 831–834.

[39] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in openSMILE, the munich open-source multimedia feature extractor,” in *Proc. ACM Multimedia*, 2013, pp. 835–838.

[40] S. B. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[41] B. Jagan Mohan and R. Badu N., “Speech recognition using mfcc and dtw,” in *Proc. Int. Conf. Adv. Electr. Eng.*, 2014, pp. 1–4.

[42] S. R. Krothapalli and S. G. Koolagudi, *Emotion Recognition Using Speech Features*. New York, NY, USA: Springer, 2013, ch. Speech Emotion Recognition: A Review, pp. 15–34.

[43] A. Neerja, “Automatic speech recognition system: A review,” *Int. J. Comput. Appl.*, vol. 151, no. 1, pp. 24–28, 2016.

[44] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, “Rasta-plp speech analysis technique,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1992, pp. 121–124.

[45] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *J. Acoustical Society Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[46] B. K. Gunturk, J. Glotzbach, Y. Altunbasak, R. W. Schafer, and R. M. Mersereau, “Demosaicking: Color filter array interpolation,” *IEEE Signal Process. Mag.*, vol. 22, no. 1, pp. 44–554, Jan. 2005.

[47] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[48] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “Constrained local neural fields for robust facial landmark detection in the wild,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 354–361.

[49] K. M. Prkachin, “The consistency of facial expressions of pain: A comparison across modalities,” *PAIN*, vol. 51, no. 3, pp. 297–306, 1992.

[50] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, and H. C. Traue, “Head movements and postures as pain behavior,” *PLOS ONE*, vol. 13, no. 2, pp. 1–17, 2018.

[51] G. Zhao and M. Pietikaeinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[52] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on feature distributions,” *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[53] D. Bansevicius, R. H. Westgaard, and C. Jensen, “Mental stress of long duration: EMG activity, perceived tension, fatigue and pain development in pain-free subjects,” *Headache*, vol. 37, no. 8, pp. 499–510, 1997.

[54] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens, “Trapezius muscle EMG as predictor of mental stress,” in *Proc. Wireless Health*, 2010, pp. 155–163.

[55] R. Luijcks, H. J. Hermens, L. Bodar, C. J. Vossen, J. v. Os, and R. Lousberg, “Experimentally induced stress validated by EMG activity,” *PLOS ONE*, vol. 9, no. 4, pp. 1–8, 2014.

[56] J. Morie, M. Seif El-Nasr, and A. Drachen, “A scientific look at the design of aesthetically and emotionally engaging interactive entertainment experiences,” in *Proc. Affective Comput. Interaction: Psychological, Cognitive Neuroscientific Perspectives*, 2009, pp. 281–307.

[57] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, “ECG pattern analysis for emotion detection,” *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 102–115, Jan.-Mar. 2012.

[58] A.-M. Brouwer, N. Van Wouwe, C. Mühl, J. Van Erp, and A. Toet, “Perceiving blocks of emotional pictures and sounds: effects on physiological variables,” *Frontiers Human Neuroscience*, vol. 7, 2013, Art. no. 295.

[59] F. A. Boiten, “The effects of emotional behaviour on components of the respiratory cycle,” *Biological Psychology*, vol. 49, no. 1, pp. 29–51, 1998.

[60] I. Homma and Y. Masaoka, “Breathing rhythms and emotions,” *Exp. Physiology*, vol. 93, no. 9, pp. 1011–1021, 2008.

[61] C.-K. Wu, P.-C. Chung, and C.-J. Wang, “Representative segment-based emotion analysis and classification with automatic respiration signal segmentation,” *IEEE Trans. Affective Comput.*, vol. 3, no. 4, pp. 482–495, Oct.-Dec. 2012.

[62] Y. Chu, X. Zhao, J. Han, and Y. Su, “Physiological signal-based method for measurement of pain intensity,” *Frontiers Neuroscience*, vol. 11, 2017, Art. no. 279.

[63] S. Balters and M. Steinert, “Capturing emotion reactivity through physiology measurement as a foundation for effective engineering in engineering design science and engineering practices,” *J. Intell. Manuf.*, vol. 28, no. 7, pp. 1585–1607, 2015.

[64] E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn, “Analysis of physiological signals for recognition of boredom, pain, and suprise emotions,” *J. Physiological Anthropology*, vol. 34, no. 1, 2015, Art. no. 25.

[65] J. Kim and E. André, “Emotion recognition based on physiological changes in music listening,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[66] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, “A novel feature extraction for robust EMG pattern recognition,” *J. Comput.*, vol. 1, no. 1, pp. 71–80, 2009.

[67] D. Tkach, H. Huang, and T. A. Kuiken, “Study of stability of time-domain features for electromyographic pattern recognition,” *J. Neuroingineering Rehabil.*, vol. 7, no. 1, 2010, Art. no. 21.

[68] J. P. Burg, *Modern Spectrum Analysis*. New York, NY, USA: IEEE Press, 1978, ch. A new Analysis Technique for Time Series Data, pp. 42–49.

[69] T.-R. Lee, Y. H. Kim, and P. S. Sung, “Spectral and entropy changes for back muscle fatigability following spinal stabilisation exercises,” *J. Rehabil. Res. Develop.*, vol. 47, no. 2, pp. 133–142, 2010.

[70] W. Chen, J. Zhuang, W. Yu, and Z. Wang, “Measuring complexity using FuzzyEn, ApEn, and SampEn,” *Med. Eng. Phys.*, vol. 31, no. 1, pp. 61–68, 2009.

[71] C. Cao and S. Slobounov, “Application of a novel measure of EEG non-stationarity as Shannon-entropy of the peak frequancy shifting for detecting residual abnormalities in concussed individuals,” *Clinical Neurophysiology: Official J. Int. Federation Clinical Neurophysiology*, vol. 122, no. 7, pp. 1314–1321, 2011.

[72] X. Tang and L. Shu, “Classification of electrocardiogram signals with RS and quantum neural networks,” *Int. J. Multimedia Ubiquitious Eng.*, vol. 9, no. 2, pp. 363–372, 2014.

[73] Q. Zhao and L. Zhang, “ECG feature extraction and classification using wavelet transform and support vector machines,” in *Proc. Int. Conf. Neural Netw. Brain*, 2005, pp. 1089–1092.

[74] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, “cvxeda: A convex optimization approach to electrodermal activity processing,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.

[75] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[76] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[77] R. A. Fisher, “The use of multiple measurement in taxonomic problems,” *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[78] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[79] F. Schwenker, C. R. Dietrich, C. Thiel, and G. Palm, “Learning of decision fusion mappings for pattern recognition,” *Int. J. Artif. Intell. Mach. Learn.*, vol. 6, pp. 17–21, 2006.

[80] S. Baccianella, A. Esuli, and F. Sebastiani, “Evaluation measures for ordinal regression,” in *Proc. 9th Int. Conf. Intell. Syst. Des. Appl.*, 2009, pp. 283–287.

[81] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, “Ordinal regression methods: Survey and experimental study,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.

[82] K. L. Gwet, *Handbook of Inter-Rater Reliability*, 4th ed. Advanced Analytics, LLC, Gaithersburg, MD, 2014.

Patrick Thiam received the MSc degree in computer science from Ulm University, Ulm, Germany, in 2014. He is currently working toward the PhD degree in computer science from the Neural Information Processing Institute, Ulm University. He is active in a joint project funded by the German Federal Ministry of Education and Research (BMBF) called SenseEmotion. His research interests include multi-modal supervised and semi-supervised learning, active learning and machine learning algorithms for the recognition of affective states in human computer interaction.

Viktor Kessler received the MSc degree in computer science from Ulm University, Ulm, Germany, in 2015. After his graduation he started working as a research assistant with the University of Ulm, Germany, as part of the SFB/TRR 62. He is currently working toward the PhD degree in computer science from the Neural Information Processing Institute, Ulm University. His research interests include multi-modal sensor fusion and transductive learning for the recognition of affective states in human centered signals.

Mohammadreza Amirian received the master’s degree in communications technology from Ulm University, Ulm, Germany, in 2017. He is currently working as a researcher with the Institute of Applied Information Technology (InIT), Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland, and simultaneously working toward the PhD degree at Ulm University. Besides his research interests in bio-physiological signal processing for person-centered medical and affective pattern recognition, his current research focuses on deep learning algorithms for industrial applications in quality assessment and learning to learn.

Peter Bellmann received the degree in mathematics from Ulm University, Ulm, Germany, in 2016. He is currently working toward the PhD degree in computer science from the Neural Information Processing Department, Ulm University. He is supported by a scholarship of the Landesgraduiertenförderung Baden-Württemberg at Ulm University. His research interests include multiple classifier systems, multi-modal fusion architectures, and machine learning techniques for the recognition of affective states in human centered signals.

Georg Layher graduated in computer science from the University of Ulm, Germany, in 2009. After his graduation he started working as a research assistant with the University of Ulm, Germany, as part of the SFB/TRR 62. His research covers computer vision and visual perception. Recent studies focus on biological and articulated motion analysis, as well as unsupervised learning mechanisms in biologically motivated neural models.

Yan Zhang is currently working with the Institute of Neural Information Processing, Ulm University, Ulm, Germany. Before, he studied in Saarland University and worked with the Max Planck Institute of Informatics, Computer Science Department, Saarland University and with the German Cancer Research Center in Heidelberg. He has interests in image analysis, computer vision and machine learning, as well as their applications in biomedical engineering and healthcare.

Maria Velana received the bachelor's in psychology from the Panteion University of Social and Political Sciences, Athens, Greece, in 2008, the MSc degree (Hons) in applied public health from National School of Public Health in collaboration with Technological Educational Institute of Athens, Greece, 2010, and the MA (Hons) degree in physical activity and health from the Institute of Sport Science and Sport, University of Erlangen-Nuremberg, Germany, in 2016. Currently, she is working toward the PhD degree in medical psychology at the University Clinic of Psychosomatic Medicine and Psychotherapy, Ulm University, Germany. From 2011 to 2012, she was a research associate with the Department of Public Health (Public & Administrative Health), National School of Public Health, Athens, Greece. She is primarily interested in multi-modal automatic pain recognition, affective computing, and the study of neural sources of human emotions by neurophysiological signals and how these are affected by addictive behaviours.

Sascha Gruss received the PhD degree in human biology from the University of Ulm, Germany, in 2015 for the recognition of pain via psychophysiological signals using machine learning algorithms. Currently, he is involved as a researcher in many projects concerning the automatic recognition of pain via bio, video and audio signals, the development of pain assessment tools and companion technologies for cognitive technical systems. His main research interests include pain pattern recognition, bio signal processing, machine learning techniques, assistive companion technology and health technology assessment.

Steffen Walter is the head of the Department of Medical Psychology, Clinic of Psychosomatic and Psychotherapy, University Clinic of Ulm, Germany. His research focus is multi-modal automatic pain recognition and trans-situational experiments in affective computing.

Harald C. Traue studied electrical engineering, computer sciences, cybernetics, communication and social sciences from Universities in Berlin, Lemgo and Bremen, and received the PhD degree in human biology, in 1978. He is a senior researcher and lecturer with Ulm University and a visiting professor with the University of Calgary/Canada. Since 1993, he a professor of medical psychology with the University of Ulm (Medical School). His areas of research are psychosocial pain theory, cognitive science and emotion, affective computing, behavioral medicine and e-learning.

Daniel Schork completed the master thesis on the analysis of eye tracking signals at Augsburg University, Germany, in 2015. After that, he joined the Lab on Human-Centered Multimedia as a research scientist to work on the analysis on biosignals within the SenseEmotion project.

Jonghwa Kim received the BS and MS degree in electronic engineering from the Gachon University, Korea, in 1992 and 1994, respectively, the PhD degree in communication engineering from the Technical University of Berlin, Germany, 2003, and the Habilitation Dr. degree (venia legendi) in computer science from the University of Augsburg, Germany, in 2010, where he worked as professor of applied computer science until 2016. He is a full professor with the Department of Information and Communication Technology, Cheju Halla University, Korea. In 2016-2018, he was a professor of computer software with the University of Science and Technology (UST), Korea. His current research interests include affective artificial intelligence, pain and emotion recognition, and the deep learning for cognitive systems. He has served in a number of program committees of Affective Computing conferences and in editorial boards of related journals. He was involved as leader of emotion research teams in various European IST projects such as HUMAINE, CALLAS, CEEDS, and METABO.

Elisabeth André received the degrees in computer science from Saarland University, including a doctorate. She is a full professor of computer science and founding chair of Human-Centered Multimedia with Augsburg University in Germany where she has been since 2001. Previously, she was a principal researcher with the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken. She has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing and social signal processing. In 2010, she was elected a member of the prestigious Academy of Europe, the German Academy of Sciences Leopoldina, and AcademiaNet. To honor her achievements in bringing Artificial Intelligence techniques to HCI, she was awarded a EurAI fellowship (European Coordinating Committee for Artificial Intelligence) in 2013. Most recently, she was elected to the CHI Academy, an honorary group of leaders in the field of human-computer interaction.

Heiko Neumann studied computer science at Technical University of Berlin and received the doctoral degree in computer science from the University of Hamburg, in 1988, and the habilitation degree in 1995 and was appointed as professor of computer science with the Institute of Neural Information Processing, Ulm University, in 1995. While with the University of Hamburg, he was a member of the Graduate School of Cognitive Systems. He spent several research sabbaticals at the Center for Adaptive Systems, Department for Cognitive and Neural Systems, at Boston University. He is co-founder of the competence center for Perception and Interactive Technologies (PIT), Ulm University. His research interests include neural modelling in computational and cognitive neuroscience, biologically inspired computational vision, and neuromorphic computation.

Friedhelm Schwenker received the PhD degree in mathematics from the University of Osnabrück, Osnabrück, Germany, in 1988. From 1989 to 1992, he was a postdoc with the Vogt-Institute for Brain Research, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. Since 1992, he is a researcher and a senior lecturer with the Institute of Neural Information Processing, Ulm University, Ulm, Germany. His research interests include artificial neural networks, machine learning, data mining, pattern recognition, applied statistics and affective computing.