Twofold-Multimodal Pain Recognition with the X-ITE Pain Database

Conference Paper · September 2019

DOI: 10.1109/ACIIW.2019.8925061

4 authors:

Philipp Werner
Otto-von-Guericke-Universität Magdeburg
60 PUBLICATIONS **672** CITATIONS
SEE PROFILE

Ayoub Al-Hamadi
Otto-von-Guericke-Universität Magdeburg
324 PUBLICATIONS **2,585** CITATIONS
SEE PROFILE

Sascha Gruss
Ulm University
57 PUBLICATIONS **599** CITATIONS
SEE PROFILE

Steffen Walter
Ulm University
123 PUBLICATIONS **1,117** CITATIONS
SEE PROFILE

Twofold-Multimodal Pain Recognition with the X-ITE Pain Database

Philipp Werner, Ayoub Al-Hamadi
Neuro-Information Technology group
Otto von Guericke University
Magdeburg, Germany
{Philipp.Werner, Ayoub.Al-hamadi}@ovgu.de

Sascha Gruss, Steffen Walter
Medical Psychology group
University Clinic Ulm
Ulm, Germany
{Sascha.Gruss, Steffen.Walter}@uni-ulm.de

***Abstract*—Automatic pain recognition has great potential to improve pain management. In this work, we investigate multimodality in pain recognition in two regards. First, we compare and combine multiple sensor modalities, which capture both behavioral and physiological pain responses. Second, we compare and distinguish the heat and the electrical pain stimulus modalities in both phasic (short) and tonic (long) variants. Experiments show that (1) pain intensity can be recognized automatically in all stimulus variants, (2) that pain of different qualities (heat and electical stimuli) can be distinguished, (3) that electrodermal activity (EDA) is the best performing single modality, and (4) that fusion with modalities can improve results further.**

***Index Terms*—pain, assessment, recognition, intensity, multimodal, modality, fusion, heat stimuli, electrical stimuli**

I. Introduction

Automatic pain recognition may complement current methods of clinical pain assessment one day, especially for patients who cannot utter on their pain experience, such as infants, adults with cognitive impairment, or unconscious persons [1], [2]. Currently, the pain of those patients is assessed by humans by observing behavior and physiological signals. Automatic systems are promising, because they facilitate continuous monitoring of pain, which is not possible with human observers. Further, they may be more objective than humans, who are influenced by personal factors, such as the relationship to the sufferer [3] and the patient's attractiveness [4].

A. Related Work

Most works on automatic pain recognition focus on facial expression and use the UNBC-McMaster Shoulder Pain Database [5]–[7]. However, many recent results showed that other sensor modalities are very promising as well and that fusion of multiple modalities can improve recognition performance and flexibility (e.g. by relying on other modalities if the face is occluded). Tsai et al. [8] demonstrated that pain intensity can be assessed from audio recorded during triage interviews in an emergency department. In another work, they fused audio and facial expression, outperforming each single modality [9]. Head pose is another behavioral signal that has been shown to be useful for pain recognition [10], [11]. Among the physiological signals, electrodermal activity (EDA), surface electromyography (EMG), and electrocardiogram (ECG) are the most widely used modalities. Werner et al. [12] and Kächele et al. [13]–[15] conducted experiments with these physiological signals, head pose, and facial expression showing some strengths and weaknesses of the single modalities and that multimodal fusion can improve recognition in many cases. Thiam et al. [16] additionally included audio and respiration signals in their work. Aung et al. [17] analyzed facial expression (with a camera) and (separately) body movement (with EMG and mocap) for recognizing pain behaviors of chronic low back pain patients. The body movement based recognition was developed and validated further by Olugbade et al. [18]. A clinical study with infants was conducted by Zamzmi et al. [19], in which pain was recognized using facial expression, body movement, audio, physiological signals, and their fusion.

For comparing and advancing the recognition methods, benchmark datasets are very beneficial. Table I summarizes pain recognition databases that are publicly available now or announced to be published soon. In this work we use the X-ITE Pain Database [25]. Of the databases known to the authors, X-ITE comprises the most sensor modalities (also see Sec. II-E). The SenseEmotion Database [23] comes with a similar set of modalities, but in contrast to X-ITE it does *not* include facial EMG, body movement video, and thermal video. In terms of pain stimulus modalities, X-ITE is the first database that includes stimuli of four different types: phasic (short duration) heat, tonic (long duration) heat, phasic electrical, and tonic electrical stimuli. The other databases only include one type of stimulus (BioVid: phasic heat, BP4D+: tonic cold pressor test, EmoPain: physical exercises in a therapy scenario, SenseEmotion: phasic heat, MIntPAIN: phasic electical). X-ITE also surpasses the other databases regarding the number of samples and subjects.

Next to intensity of pain, which is widely covered in literature on automatic pain recognition, pain quality is another relevant characteristic that is commonly assessed in clinical practice [26]. The pain quality refers to the description or type of pain, e.g. whether it is sharp, dull, crushing, burning, tearing, etc. and whether it is intermittent, constant, or throbbing. It has been shown that the reported pain quality differs between

This work was funded by the German Research Foundation (DFG, no. AL 638/3-2) and the Federal Ministry of Education and Research (BMBF, no. 03ZZ0470). The sole responsibility for the content lies with the authors.

P. Werner, A. Al-Hamadi, S. Gruss, S. Walter, "Twofold-Multimodal Pain Recognition with the X-ITE Pain Database", in International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019.
This is the accepted manuscript. The final, published version is available on IEEE Xplore.

TABLE I
Pain Databases

	Sensor Modalities						Stimulus Modalities / Types					Size		
Database	Video	Audio	EMG	ECG	EDA	Other	Thermal	Electrical	Movement	Phasic	Tonic	Subjects	Painful Stimuli	Medical Condition
UNBC-McMaster [5]	✓								✓			25	200	shoulder pain
BioVid [20], [21]	✓		✓	✓	✓		✓			✓		ca. 90	14k	healthy
BP4D+ [22]	✓			✓	✓	✓	✓				✓	140	140	healthy
EmoPain [17]	✓	✓	✓			✓			✓			50	?	chronic low back pain
SenseEmotion [23]	✓	✓	✓	✓	✓	✓	✓			✓		45	8k	healthy
MIntPAIN [24]	✓					✓		✓		✓		20	2k	healthy
X-ITE [25]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	134	25k	healthy

experimental pain models [27] that are widely used in research. For instance, phasic electric pain stimulation was described as pricking, flickering, sharp, and pinching; longer lasting (tonic) ischemia pain as dull, pinching, hot, annoying, spreading, and tight [28]. To the authors' knowledge, distinguishing pain of different qualities has not been addressed in automatic pain recognition yet.

B. Contribution

This work features several novelties: (1) It is the first work reporting results on the new X-ITE Pain Database (which is described in Sec. II). (2) To our best knowledge, it is the first work addressing multimodality in terms of sensors and pain stimulation at the same time. (3) It compares recognition of four types of stimuli that differ in duration (phasic/tonic) and the used experimental pain model (heat/electrical). (4) We report on experiments on using facial video (facial expression and head pose), audio, three EMGs (including facial EMG), EDA, and ECG individually and combining the modalities. (5) Results show that it is possible to distinguish two types of stimuli that are associated with a different quality of pain (heat/electrical). In Sec. III we describe the used recognition method. Sec. IV reports on our experiments and Sec. V discusses the results and draws conclusions.

II. X-ITE Pain Database

In this section we give an overview of the new multimodal Experimentally Induced Thermal and Electrical (X-ITE) Pain Database. For more details, refer to Gruss et al. [25].

A. Participants

A total of 134 healthy adults (67 men and 67 women) aged between 18 and 50 years participated in the experiment. The average age of all subjects was 31.4 (SD = 9.7), of all men = 33.4 (SD = 9.3), and of all women = 32.9 (SD = 10.2) years. None of them suffered from chronic pain, depression, or had a history of psychiatric disorders; none had neurological conditions, headache syndrome, or cardiovascular disease; none had taken pain medication or used painkillers directly before the experiment.

B. Pain Stimulation

Heat pain was stimulated at the participants forearm using the Medoc PATHWAY Model ATS thermal stimulator. Electrical pain was stimulated with electrodes attached to the index and middle finger using the electrical stimulator Digitimer DS7A. Both, heat pain (H) and electrical pain (E) were applied in two variants: phasic (short) stimuli of 5 seconds duration (P) and tonic (long) stimuli of 60 seconds duration (T). Each of the four stimulus types (HP, HT, EP, ET) was stimulated in three intensities.

C. Stimulus Calibration Phase

In order to handle differences in pain sensitivity, each participant underwent a stimulus calibration procedure prior to the main experiment. The calibration comprised four parts, one for each stimulus type. In each part, the stimulus intensity was gradually increased while asking the participant to report the felt pain intensity to determine the person-specific pain threshold and tolerance. Afterwards, the six heat stimulation temperatures (three intensities for phasic and three for tonic stimulation) and six electrical stimulation currents (three phasic and three tonic) were calculated based on the pain thresholds and tolerances.

D. Main Stimulation Phase

After the calibration procedure, the participant laid down on an examination couch and underwent the main stimulation phase, which took about 90 minutes. The phasic stimuli of each modality (heat and electrical pain) and intensity were repeated 30 times in randomized order with pauses of 8-12 seconds. The 1-minute tonic stimuli were only applied once per intensity, i.e. there were six tonic stimuli per participant, each followed by a pause of five minutes. They were applied in three phases: The highest intensity tonic heat and electrical pain stimuli were applied at the end of the experiment. The other two phases, which each contained one tonic heat and one tonic electrical stimulus of lower intensity, were randomly started during the phasic stimulation period.

E. Data Recording

Synchronously to the applied pain stimulation, several sensors were used to collect multimodal pain response data:

- RGB video of the face (frontal and side view) for analyzing facial expression and head pose,
- audio signal for analyzing para-linguistic responses,
- electrocardiogram (ECG) for analyzing heart rate and its variability,

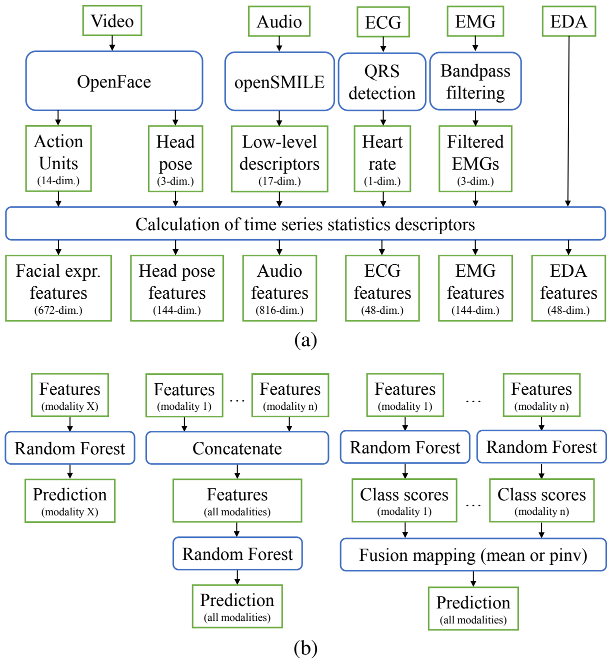

Fig. 1. Recognition method comprising (a) feature extraction and (b) classification of single modality (left), multimodal feature fusion (middle), and multimodal decision fusion (right).

- surface electromyography (EMG) to measure activity of three muscles, the trapezius (neck/shoulder), corrugator supercilii (close to eyebrows), and zygomaticus major (at the cheeks),
- electrodermal activity (EDA) to measure sweating,
- video of the body to analyze body movements, and
- thermal video of the face.

III. Recognition Approach

This section describes the recognition approach used for the experiments. It was adopted from prior work, because this work focuses on providing and analyzing baseline results for the new X-ITE Pain Database rather than proposing a new recognition approach. Fig. 1 gives an overview of the used method. We processed facial RGB video, audio, ECG, EMG, and EDA data as detailed in the subsections. This yielded one or multiple time series per modality, which were condensed in time-series statistics descriptors. The descriptors were used as features for recognizing the pain stimulus intensity and stimulus modality using Random Forest classification. We applied the classification for each sensor modality individually and also fused all modalities using three approaches: feature fusion, decision fusion with mean-score mapping, and decision fusion with pseudoinverse mapping.

A. Processing of Sensor Signals

Video: The frontal facial RGB video was processed with OpenFace [29] to extract facial expression and head pose information. As frame-level expression features we used the 14 FACS Action Unit intensities provided by OpenFace without post-processing, i.e. the outputs were *not* clipped to the AU intensity range of 0 to 5. The head pose was described using the provided yaw, pitch, and roll head orientation angles.

Audio: The audio signal was processed with openSMILE [30] using frames of 25 ms length extracted in steps of 10 ms. For each frame we extracted a 17-dimensional low-level descriptor comprising the logarithmic signal energy, the voicing probability, the pitch (F_0), the F_0 envelope, and 13 Mel Frequency Cepstral Coefficients (MFCCs). The low-level descriptor time series were smoothed with a moving average filter over three frames.

ECG: We applied the QRS-detection algorithm by Hamilton and Tompkins [31] in order to find the R-peaks in the ECG signal. Afterwards, the heart rate was calculated from the R-to-R intervals. Finally, we interpolated the heart rate signal linearly to match the sampling of the EMG and EDA (500 Hz).

EMG: The three EMG channels were preprocessed with a zero-phase 3rd-order Butterworth band-pass filter with cut-off frequencies of 20 and 250 Hz. Further, they were downsampled from 1,000 Hz to 500 Hz to speed-up processing.

EDA: The 1-dimensional EDA time series was downsampled to 500 Hz as the other biopotentials, but was not processed any further.

B. Time Series Statistics Descriptor

The time series of the above-mentioned sensor modalities were segmented based on the applied stimulation, i.e. the samples were temporally aligned with the pain stimulation. For the phasic stimuli, we extracted time windows with a duration of 6 s, each starting with the stimulus. As phasic baseline samples we selected 6 s time windows following phasic heat stimuli of lowest intensity. For the tonic stimuli, time windows of 60 s were extracted, temporally matching the applied pain stimulus. The respective baseline samples of the same length were cut from the pause following the tonic heat stimuli of lowest intensity.

We calculated a feature vector for each sample and sensor modality as proposed by Werner et al. [11] in the context of facial activity. Each time series was summarized by several statistics of the time series itself and of its first and second derivative, including mean, maximum, range, time of maximum, and others, yielding a 48-dimensional descriptor per time series. The smoothing proposed by Werner et al. (1 Hz first-order Butterworth low-pass filtering) was only applied for the video time series, because further noise reduction was not necessary for the other modalities. We applied a person-specific standardization of the features [11] in order to focus this work on the within-subject response variation rather than the differences between subjects.

C. Classification and Fusion

The samples were classified using Random Forests (RF) [32] with 100 trees and a maximum depth of 10 nodes. To compare the usefulness of the sensor modalities alone, we first trained and tested RFs using the features of each modality individually. Second, we concatenated the feature vectors of all

TABLE II

CROSS-VALIDATION ACCURACY OF **PHASIC STIMULI 2-CLASS TASKS** (COLUMNS) AND MODALITIES / FUSION APPROACHES (ROWS).

	No Pain vs H. Pain			No Pain vs E. Pain			H. Pain vs H. Pain			E. Pain vs E. Pain			H. Pain vs E. Pain			Mean	
	B/H1	B/H2	B/H3	B/E1	B/E2	B/E3	H1/H2	H2/H3	H1/H3	E1/E2	E2/E3	E1/E3	H1/E1	H2/E2	H3/E3	modality	category
Facial expression	52.2	58.2	75.2	52.8	59.1	75.9	55.0	68.7	73.9	55.2	71.2	75.4	52.1	59.3	73.4	63.9	61.4
Head pose	51.2	54.9	70.7	51.8	56.6	70.0	53.9	67.1	70.9	54.5	67.4	71.0	51.4	55.5	67.4	61.0	
Audio	50.5	53.5	67.8	50.2	54.5	69.8	54.0	65.1	68.1	53.3	67.8	70.3	51.0	52.4	62.2	59.4	
ECG	50.4	53.9	65.6	52.2	60.4	72.9	52.5	62.9	65.0	58.9	66.4	73.5	51.4	60.6	68.6	61.0	67.9
EMG	51.4	59.0	78.0	64.3	74.9	87.8	57.0	71.1	77.3	63.3	75.1	82.2	63.4	74.6	82.7	70.8	
EDA	59.3	64.1	79.1	66.6	79.9	91.1	56.8	68.6	72.9	64.8	75.0	83.6	61.5	71.7	81.6	71.8	
Feature fusion	58.4	63.8	82.2	68.0	81.0	92.4	57.1	72.1	78.6	66.4	79.0	86.9	63.3	75.5	85.4	74.0	74.8
Decision fusion (mean)	58.6	63.4	79.7	68.6	78.7	88.2	58.3	71.8	77.9	66.2	76.9	84.6	63.1	77.5	86.3	73.3	
Decision fusion (pinv)	55.7	65.9	83.3	72.3	83.7	94.3	60.3	72.7	80.4	70.7	82.2	90.3	71.7	81.1	90.5	77.0	
Mean	54.2	59.6	75.8	60.8	69.9	82.5	56.1	68.9	73.9	61.5	73.4	79.8	58.8	67.6	77.6	68.0	
Mean (category)	63.2			71.0			66.3			71.6			68.0				

B: Baseline (no pain) H: Heat (pain) E: Electrical (pain) Hx: Heat pain of intensity x (1 = lowest, 3 = highest) Ex: Electrical pain of ...

TABLE III

CROSS-VALIDATION ACCURACY OF **TONIC STIMULI 2-CLASS TASKS** (COLUMNS) AND MODALITIES / FUSION APPROACHES (ROWS).

	No Pain vs H. Pain			No Pain vs E. Pain			H. Pain vs H. Pain			E. Pain vs E. Pain			H. Pain vs E. Pain			Mean	
	B/H1	B/H2	B/H3	B/E1	B/E2	B/E3	H1/H2	H2/H3	H1/H3	E1/E2	E2/E3	E1/E3	H1/E1	H2/E2	H3/E3	modality	category
Facial expression	59.6	58.0	69.2	59.8	62.4	71.3	58.1	66.5	73.4	55.9	69.0	70.7	52.6	58.1	65.6	63.3	62.0
Head pose	58.8	60.0	68.4	65.0	60.0	65.6	61.8	65.3	73.8	50.6	63.3	69.5	55.9	52.8	59.2	62.0	
Audio	58.8	55.9	68.0	56.9	53.9	59.9	60.6	64.1	70.2	57.9	62.9	68.7	58.3	51.6	62.0	60.6	
ECG	53.5	59.6	72.1	61.0	56.3	75.3	63.0	67.3	76.2	56.7	73.4	69.9	56.7	65.9	63.2	64.7	70.7
EMG	58.8	67.3	78.5	67.1	73.5	87.0	61.0	70.2	78.6	65.2	72.2	78.7	70.0	69.9	76.4	71.6	
EDA	55.5	67.8	86.6	73.2	79.6	87.0	68.7	79.4	86.3	61.5	77.0	84.3	72.9	79.7	79.6	75.9	
Feature fusion	64.5	71.4	84.6	69.5	81.6	89.5	65.9	74.2	82.7	59.5	74.6	82.7	68.0	69.9	78.0	74.4	72.7
Decision fusion (mean)	66.5	69.8	84.2	74.8	82.0	88.3	69.9	75.0	86.3	64.8	78.6	83.9	73.3	72.8	81.2	76.8	
Decision fusion (pinv)	56.3	66.1	69.2	72.0	72.2	68.8	63.8	60.5	71.0	56.3	71.0	67.9	63.2	71.1	72.0	66.8	
Mean	59.1	64.0	75.7	66.6	69.1	77.0	63.6	69.2	77.6	58.7	71.3	75.1	63.4	65.8	70.8	68.5	
Mean (category)	66.3			70.9			70.1			68.4			66.7				

B: Baseline (no pain) H: Heat (pain) E: Electrical (pain) Hx: Heat pain of intensity x (1 = lowest, 3 = highest) Ex: Electrical pain of ...

modalities and trained and tested RFs on the resulting 1,872-dimensional feature space. This approach is called *Feature Fusion*. Third, we applied two types of *Decision Fusion*, in which an individual RF is trained for each modality. In this approach, each RF not only yields the class that the majority of its trees predicted, but it provides a score for each possible class (the proportion of trees that predicted this class). There are many ways how to aggregate these classifier scores into a final decision. Here we used (1) a fixed mapping approach calculating the *mean* of all RF scores per class and selecting the class with the highest score, and (2) a trained mapping approach that learns class-score weights for each modality by calculating the pseudoinverse (pinv) matrix [33].

IV. EXPERIMENTS

Dataset: In this work we experimented with stimulus-aligned samples, which had been cut out from the continuous recording of the main stimulation phase of the X-ITE study (about 90 minutes per participant). See Sec. III-B for details about the time windows. The phasic (short) and tonic (long) stimuli were evaluated separately, because the number of repetitions is very different (one for tonic vs 30 for phasic). Combining both in one recognition framework is possible but challenging. Thus, we leave this for future work. Due to technical problems, some sensor modalities and stimuli are not available. We used the intersection set of the samples, i.e. we only included samples for which there was no failure for any of the used sensors (frontal RGB camera, audio, ECG, EMG, EDA). This way we were able to use data of 127 subjects. The tonic dataset contained 865 samples in total; the phasic dataset contained 26,308 samples. The classes were approximately balanced in both the tonic and the phasic dataset.

Validation: We applied leave-one-subject out cross validation to estimate the performance of various models on unseen subjects. The performance is reported with the accuracy measure[1] (in percent), which is intuitive and well-suited for the balanced datasets we consider here. In the following we report experimental results of several classification tasks: pairwise binary classification of phasic and tonic pain, 4-class intensity estimation of phasic and tonic pain, and the full 7-class pain recognition task. We considered the single sensor modalities and their fusion (three approaches), heat and electrical pain (which are two different stimulus modalities associated with

[1] The accuracy is defined as the quotient of the number of all correctly classified samples and the total number of tested samples.

TABLE IV
CROSS-VALIDATION ACCURACY OF **4-CLASS TASKS** (COLUMNS) AND MODALITIES / FUSION APPROACHES (ROWS). CHANCE: 25%.

	Phasic Pain		Tonic Pain		
	Heat	Eletrical	Heat	Eletrical	Mean
Facial expression	37.8	39.4	40.2	38.5	38.9
Head pose	35.8	36.6	39.4	35.4	36.8
Audio	34.3	35.5	38.9	34.0	35.7
ECG	33.1	38.6	41.4	38.5	37.9
EMG	40.0	50.5	44.0	52.6	46.8
EDA	41.7	53.1	52.7	55.7	50.8
Mean (single modalities)	37.1	42.3	42.8	42.4	41.1
Feature fusion	43.2	56.3	47.7	55.7	50.7
Decision fusion (mean)	42.9	55.3	51.5	54.7	51.1
Decision fusion (pinv)	45.6	61.4	36.9	46.0	47.5
Mean (fusion)	43.9	57.6	45.4	52.1	49.8
Mean (all)	39.1	46.9	43.5	45.3	43.7

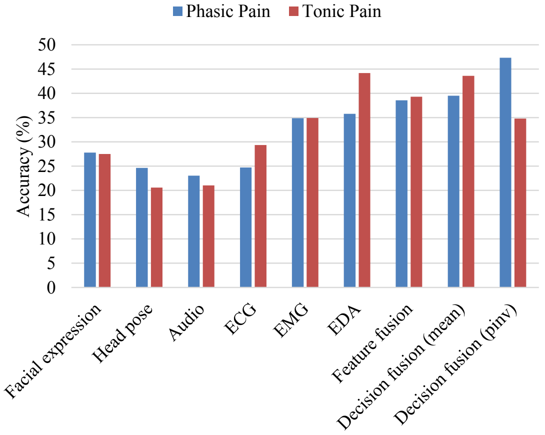

Fig. 2. Cross-validation accuracy of **7-class tasks**, phasic (red) and tonic (blue), and modalities / fusion approaches (bar groups). Chance: 14%.

different pain qualities), and the four pain intensities (ranging from no pain to high pain). No pain (baseline, pain intensity 0) is denoted with the letter *B*, heat pain with *H*, and electrical pain with *E*. The letters *H* and *E* are followed by a number between 1 and 3 denoting the stimulus intensity.

A. Binary Classification of Phasic Stimuli

Table II shows the results of various two-class tasks of phasic pain stimulation (columns). Higher pain intensities and higher differences between considered pain intensities are associated with higher performance, as observed in several other works [11], [12], [16]. To a large extent, low performances are probably caused by the lack of pain responses observed in many participants of experimental pain studies [12], [34]. On average, recognition results of electrical pain are significantly better than those of heat pain, e.g. compare the mean of the no-vs-heat-pain experiments (63.2%) and the mean of the no-vs-electrical-pain experiments (71.0%). Note the results of the heat-vs-electrical-pain experiments (average 68.0%, maximum 90.5%), which show that it is possible to distinguish stimulated pain of different qualities. If we compare the single modalities (rows), EDA and EMG quite consistently perform best (mean: 71.8% and 70.8%), followed by facial expression (63.9%), head pose and ECG (both 61.0%), and audio (59.4%). Fusion of the sensor modalities on average improves performance significantly compared to the best single modality (71.8% for EDA vs 74.8% for mean of fusion approaches and 77.0% for decision fusion with pseudoinverse).

B. Binary Classification of Tonic Stimuli

The two-class tonic stimulation results (see Table III) are similar to the phasic results in many regards. Higher pain intensities and higher differences between considered pain intensities are associated with higher performance. However, the performances for lower intensity tasks tend to be better than for the corresponding phasic cases and performances for higher intensity tasks tend to be worse than for the phasic. Further, the heat pain recognition tasks quite consistently perform better in tonic pain than in phasic pain (e.g. tonic vs phasic category means of 66.3% vs 63.2% and 70.1% vs 66.3%). For the electrical pain recognition task it is vice versa, but to a lesser extent (e.g. tonic vs phasic category means in E-vs-E pain of 68.4% vs 71.6%). On average there is a clear performance gain for the contact-based sensor modalities ECG, EMG, and EDA (tonic 70.7% vs phasic 67.9%). The ranking of the single modality results is similar as in phasic, but the ECG is the third best modality in tonic pain experiments, on average performing better than the contact-free sensor modalities. The best-performing single modality (EDA) is only outperformed by one of the fusion approaches: decision fusion with the mean-score aggregation rule with mean values of 75.9% vs 76.8%, but there is no gain in several of the underlying single classification rates. Decision fusion with pseudoinverse performs poorly in the tonic case.

C. 4-Class Pain Intensity Recognition

The pain intensity recognition involves four classes for each combination of pain duration (phasic/tonic) and stimulation modality (heat/electrical), e.g. in phasic heat: B, H1, H2, and H3. The cross-validation results can be found in Table IV. All results are significantly above the chance level (25%). Consistently with the binary classification, the performance of heat pain recognition is worse than that of electrical pain for phasic stimulation. Heat pain is assessed more reliably in tonic stimuli compared to phasic. In electrical pain it is vice versa, i.e. phasic recognition rates are better than tonic. The best performing single sensor modality is EDA, followed by EMG. But they are outperformed by all fusion approaches in the phasic experiments. The overall best results in this context are achieved with decision fusion with the learned pseudoinverse aggregation: 61.4% for electical and 45.6% for heat pain. In the tonic experiments, fusion fails to outperform EDA (which achieves 55.7% for electical and 52.7% for heat pain).

D. 7-Class Pain Recognition

In this part we evaluate the discrimination of all 7 available classes, combining the pain intensity (from 0=B to 3) and

TABLE V

Confusion matrix: Decision fusion (pinv) on phasic 7-class task (26,308 samples, 1,872 features used in total)

	Prediction							
	B	H1	E1	H2	E2	H3	E3	recall
B	1774	867	409	267	169	191	59	47.5%
H1	1233	790	467	495	310	299	147	21.1%
E1	512	314	1580	215	863	123	174	41.8%
H2	731	601	463	736	343	736	145	19.6%
E2	187	143	461	99	2306	115	474	60.9%
H3	206	195	164	329	200	2315	318	62.1%
E3	28	26	75	38	519	146	2951	78.0%
precision	38.0%	26.9%	43.7%	33.8%	49.0%	59.0%	69.1%	(47.3%)

TABLE VI

Confusion matrix: EDA on tonic 7-class task (865 samples, 48 features)

	Prediction							
	B	H1	E1	H2	E2	H3	E3	recall
B	35	32	14	27	5	8	1	28.7%
H1	21	53	16	20	6	3	4	43.1%
E1	12	25	34	15	28	4	6	27.4%
H2	7	29	9	50	10	13	5	40.7%
E2	5	10	22	8	53	7	18	43.1%
H3	3	2	5	15	9	73	18	58.4%
E3	1	3	5	4	18	10	84	67.2%
precision	41.7%	34.4%	32.4%	36.0%	41.1%	61.9%	61.8%	(44.2%)

the stimulus modality (heat and electrical pain) recognition tasks: B, H1, E1, H2, E2, H3, and E3. The results are depicted in Fig. 2. They qualitatively match the observations made in the previous subsections to a great extent. In the phasic pain dataset, the best result is obtained with decision fusion (pinv): 47.3%. In the tonic pain data, EDA alone gives best performance (44.2%). The accumulated confusion matrices of these best models are given in Table V and VI. In both, confusions occur mostly close to the main diagonal, i.e. between similar intensities, but there are more confusions among intensities of the same stimulus modality (H/E) than between the H and E (leading to a chessboard-like pattern partially). When comparing phasic and tonic results (recall and precision), we again see that heat pain recognition performance tends to be better in the tonic pain and electrical pain recognition tends to work better in phasic pain.

V. Discussion and Conclusion

We conducted experiments in recognizing pain with multiple sensor modalities and stimulation modalities. The results in both phasic and tonic datasets show that it is possible to recognize the pain intensity, which is in line with results of many prior works. Most confusions occur between intensity levels that are close to each other. Further, the results show that stimulated pain of different qualities (heat and electical stimuli) can be distinguished with automatic methods, which is a new finding. We observed that responses to heat and electrical pain differ regarding speed, intensity, and the overall pattern in the multimodal signals. Responses to electrical stimuli tend to be more intense and start earlier and more rapidly than responses to heat, because the electrical pain is instantly felt with full intensity whereas the heat pain is building up slower. This is probably also why the recognition of electrial pain stimuli works better than recognition of heat pain stimuli in the phasic dataset, which uses quite short time-windows to avoid overlap of pain and baseline samples. Further, heat pain is assessed more reliably in tonic stimuli compared to phasic. This leads to the hypothesis that heat pain recognition may benefit from longer time windows, which should be investigated in future work. In electrical pain it is vice versa, i.e. phasic recognition rates are better than tonic in many cases. However, this may be a sample size issue: If we compare tonic pain (865 samples) to phasic pain (26k samples), some high-dimensional modalities such as audio (816-dim. features space) perform worse and some low-dimensional modalities such as EDA (48-dim.) perform better. Also decision fusion with pseudoinverse performs poorly in the tonic case, probably because the training datasets are too small (and this approach requires an additional subdivision of the training set).

Among the single sensor modalities, EDA performs best. This is in line with prior works on the BioVid and SenseEmotion databases [14]–[16]. EDA is less person-specific than other modalities [12], [16] and very sensitive to pain. However, EDA responses are not specific to pain, but indicate psychological or physiological arousal in general. To better address this issue, future work should include distinguishing pain and other affective states, such as anxiety or anticipation. On the BioVid data, facial expression (EMG and camera-based) outperforms EDA in some experiments [12], [13]. In our results, EMG is the second best modality, probably due to the facial EMG channels, whereas video-based facial expression performs much worse. Kächele et al. [13], who used a more specialized video analysis, found a smaller performance gap between facial EMG and facial video results. This indicates that using specialized video-based expression recognition (rather than the generic OpenFace software) may facilitate improvements on the X-ITE results as well. Head pose has been identified as an auxiliary pain indicator in prior works with people in upright position [10], [11]. Our results show that head pose is similarly useful in lying position. Audio is the worst performing modality in our X-ITE experiments, which is consistent with the quite poor performance of audio in the SenseEmotion database, but in contrast to the results achieved by Tsai et al. [9]. Tsai et al. used an interview scenario yielding more audio material with potentially discriminative information, whereas moaning and pain-related breathing patterns (which were observed on SenseEmotion and X-ITE) may occur less consistently than other pain responses.

Future work should compare fusion results achievable with different subsets of modalities to find good options for lower-cost pain monitoring (not requiring all sensors). Further, evaluation protocols for the X-ITE database need to be developed in order to improve comparability of results among future papers. They should cover different tasks that are close to prospective clinical applications, including a continuous monitoring task.

References

[1] K. Herr, P. J. Coyne, M. McCaffery, R. Manworren, and S. Merkel, “Pain assessment in the patient unable to self-report: position statement with clinical practice recommendations.” *Pain management nursing : official journal of the American Society of Pain Management Nurses*, vol. 12, no. 4, pp. 230–50, dec 2011.

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

[3] K. D. Craig, “The social communication model of pain.” *Canadian Psychology*, vol. 50, no. 1, p. 22, 2009.

[4] ——, “The facial expression of pain Better than a thousand words?” *APS Journal*, vol. 1, no. 3, pp. 153–162, 1992.

[5] 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 *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 2011, pp. 57–64.

[6] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. W. Picard, “Automatic Recognition Methods Supporting Pain Assessment: A Survey,” *IEEE Transactions on Affective Computing*, vol. PP, 2019.

[7] Z. Hammal and J. F. Cohn, “Automatic, objective, and efficient measurement of pain using automated face analysis,” in *Social and Interpersonal Dynamics in Pain: We Don’t Suffer Alone*, T. Vervoort, K. Karos, Z. Trost, and K. M. Prkachin, Eds. Springer, 2018, pp. 121–146.

[8] F.-S. Tsai, Y.-M. Weng, C.-J. Ng, and C.-C. Lee, “Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage,” in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, oct 2017, pp. 313–318.

[9] 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 *Interspeech*, sep 2016, pp. 92–96.

[10] 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, p. e0192767, feb 2018.

[11] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, “Automatic Pain Assessment with Facial Activity Descriptors,” *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 286–299, 2017.

[12] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. Traue, “Automatic pain recognition from video and biomedical signals,” in *International Conference on Pattern Recognition (ICPR)*, 2014.

[13] M. Kächele, P. Werner, A. Al-Hamadi, G. Palm, S. Walter, and F. Schwenker, “Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity,” in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, F. Schwenker, F. Roli, and J. Kittler, Eds. Springer International Publishing, 2015, pp. 220–230.

[14] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, “Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity,” in *Engineering Applications of Neural Networks*, ser. Communications in Computer and Information Science, L. Iliadis and C. Jayne, Eds. Springer International Publishing, 2015, pp. 275–285.

[15] 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 Systems*, vol. 8, no. 1, pp. 71–83, 2017.

[16] P. Thiam and F. Schwenker, “Multi-modal data fusion for pain intensity assessment and classification,” in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, nov 2017, pp. 1–6.

[17] M. S. 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. C. Williams, M. Pantic, and N. Bianchi-Berthouze, “The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset,” *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, 2016.

[18] T. A. Olugbade, A. Singh, N. Bianchi-Berthouze, N. Marquardt, M. S. H. Aung, and A. C. D. C. Williams, “How can affect be detected and represented in technological support for physical rehabilitation?” *ACM Trans. Comput.-Hum. Interact.*, vol. 26, no. 1, pp. 1:1–1:29, Jan. 2019.

[19] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, “An approach for automated multimodal analysis of infants’ pain,” in *International Conference on Pattern Recognition (ICPR)*. IEEE, dec 2016, pp. 4148–4153.

[20] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. Traue, S. Crawcour, P. Werner, A. Al-Hamadi, A. Andrade, and G. Da Silva, “The BioVid Heat Pain Database: Data for the advancement and systematic validation of an automated pain recognition,” in *2013 IEEE International Conference on Cybernetics, CYBCONF 2013*, 2013.

[21] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, “Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges,” in *British Machine Vision Conference (BMVC)*. BMVA Press, 2013, pp. 111–119.

[22] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, “Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3438–3446.

[23] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, 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 *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*. Springer, Cham, 2016, pp. 127–139.

[24] M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O. K. Andersen, E. G. Spaich, and T. B. Moeslund, “Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, may 2018, pp. 250–257.

[25] S. Gruss, M. Geiger, P. Werner, O. Wilhelm, H. C. Traue, A. Al-Hamadi, and S. Walter, “Multi-Modal Signals for Analyzing Pain Responses to Thermal and Electrical Stimuli,” *Journal of Visualized Experiments*, no. 146, apr 2019.

[26] Registered Nurses’ Association of Ontario, “Practice Recommendations,” in *Assessment and Management of Pain*, 3rd ed. Toronto: Registered Nurses’ Association of Ontario, 2013, pp. 19–40.

[27] K. S. kumar Reddy, M. U. R. Naidu, P. U. Rani, and T. R. K. Rao, “Human experimental pain models: A review of standardized methods in drug development,” *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, vol. 17, no. 6, pp. 587–595, 2012.

[28] A. C. N. Chen and R.-D. Treede, “The McGill pain questionnaire in the assessment of phasic and tonic experimental pain: behavioral evaluation of the ’Pain Inhibiting Pain’ Effect,” *Pain*, vol. 22, no. 1, pp. 67–79, 1985.

[29] T. Baltrusaitis, P. Robinson, and L. P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. IEEE, 2016, pp. 1–10.

[30] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the international conference on Multimedia - MM ’10*. New York, New York, USA: ACM Press, 2010, pp. 1459–1462.

[31] P. S. Hamilton and W. J. Tompkins, “Quantitative investigation of QRS detection rules using the MIT\BIH arrhythmia database,” *Biomedical Engineering, IEEE Transactions on*, vol. 12, pp. 1157–1165, 1986.

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

[33] F. Schwenker, C. Dietrich, C. Thiel, and G. Palm, “Learning of decision fusion mappings for pattern recognition,” *International Journal on Artificial Intelligence and Machine Learning (AIML)*, vol. 6, pp. 17–21, 2006.

[34] P. Werner, A. Al-Hamadi, and S. Walter, “Analysis of facial expressiveness during experimentally induced heat pain,” in *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 176–180.