Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities

Haque, Mohammad Ahsanul; B. Bautista, Ruben; Noroozi, Fatemeh; Kulkarni, Kaustubh; Laursen, Christian B.; Irani, Ramin; Bellantonio, Marco; Escalera, Sergio; Anbarjafari, Gholamreza; Nasrollahi, Kamal; Andersen, Ole Kæseler; Spaich, Erika Geraldina; Moeslund, Thomas B.

Published in:
Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018

DOI (link to publication from Publisher):
10.1109/FG.2018.00044

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities

Mohammad A. Haque*¶, Ruben B. Bautista†, Fatemeh Noroozi§, Kaustubh Kulkarni†, Christian B. Laursen‡, Ramin Irani*, Marco Bellantonio†, Sergio Escalera†*, Golamreza Anbarjafari§, Kamal Nasrollahi*, Ole K. Andersen‡‖, Erika G. Spaich‡ and Thomas B. Moeslund*
*Visual Analysis of People Lab, Aalborg University, Denmark
†Computer Vision Center (CVC) and Universitat Autnoma de Barcelona, Spain
‡SMI, Dept. of Health Science and Technology, Aalborg University, Denmark
§University of Tartu, Estonia
‖Center for Neuroplasticity and Pain (CNAP), Aalborg University, Denmark
¶Corresponding author's email: mah@create.aau.dk

***Abstract*—Pain is a symptom of many disorders associated with actual or potential tissue damage in human body. Managing pain is not only a duty but also highly cost prone. The most primitive state of pain management is the assessment of pain. Traditionally it was accomplished by self-report or visual inspection by experts. However, automatic pain assessment systems from facial videos are also rapidly evolving due to the need of managing pain in a robust and cost effective way. Among different challenges of automatic pain assessment from facial video data two issues are increasingly prevalent: first, exploiting both spatial and temporal information of the face to assess pain level, and second, incorporating multiple visual modalities to capture complementary face information related to pain. Most works in the literature focus on merely exploiting spatial information on chromatic (RGB) video data on shallow learning scenarios. However, employing deep learning techniques for spatio-temporal analysis considering Depth (D) and Thermal (T) along with RGB has high potential in this area. In this paper, we present the first state-of-the-art publicly available database, 'Multimodal Intensity Pain (MIntPAIN)' database, for RGBDT pain level recognition in sequences. We provide a first baseline results including 5 pain levels recognition by analyzing independent visual modalities and their fusion with CNN and LSTM models. From the experimental evaluation we observe that fusion of modalities helps to enhance recognition performance of pain levels in comparison to isolated ones. In particular, the combination of RGB, D, and T in an early fusion fashion achieved the best recognition rate.**

I. Introduction

International Association for the Study of Pain (IASP) defined 'pain' as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage". It is a prevalent medical problem and managing pain is a moral imperative, a professional responsibility and a duty of medical practitioners [4]. However, the dualistic nature of pain has been recognized throughout history containing both sensory and affective components [29]. This dualistic nature states that pain is both a powerful somatic sensation as well as a powerful behavioral state of mind. To evaluate these dimensions there are many different neuro-physiological tools or techniques which can be used. The widely used technique to measure pain level is 'self-report'. However, self-reported pain level assessment does not always effectively apt in practical scenarios due to inconsistent metric properties across dimensions, efforts at impression management or deception, as well as differences between clinicians' and sufferers' conceptualization of pain [3]. Moreover, it requires cognitive, linguistic and social competencies that make self-report unfeasible to use for young children and patients with limited ability to communicate [17], [22], [26].

Beside 'self-report' of pain, visual pain expression can be revealed in the face and expresses emotion valley regarding to experiencing pain [39]. It can also provide the information about the severity of pain that can be assessed by using the Facial Action Coding System (FACS) of Ekman and Friesen [5], [37]. Prkachin first reported the consistency of facial pain expressions for different pain modalities [32] and then together with Solomon developed a pain metric called Prkachin and Solomon Pain Intensity (PSPI) scale based on FACS in [34]. Although there is a debate about the correlation between self-reported pain and facial pain expression [14], many works found significant relationship between these two [8], [33]. Another of the most widely used scales are the Visual Analogue Scale (VAS) [7]. The VAS is a psychometric response scale, which is often utilized to characterize subjective attitudes which cannot be directly measured, on a continuous line between two end-points [7]. The VAS is capable of characterizing both the level of pain intensity, the level of unpleasantness and is able to shed light on both the somatic component and the affective component in a relatively simplistic way.

The scales, like PSPI or VAS, provide notions to calibrate pain existence and intensity by visual observations from facial images or videos either by a human expert or an automated system. While human observation constitutes the ground truth for the pain level for objective assessment, automated system for pain assessment based on facial image or video analysis tries to provide an effective alternative to self-report or human

978-1-5386-2335-0/18/$31.00

expert for pain assessment. However, automatically assessing pain level from facial image or video is rather challenging. This is not only because of the challenges associated with finding the pain features in the absence of enough visual difference between pain/non-pain facial frames. This is also because of the presence of external factors like 'smiling in pain' phenomenon and/or gender difference (male's vs female's way of experiencing) to pain [23], [24], [40]. These result to a non-linearly wrapped facial emotion levels (due to the presence of pain) in a high dimensional space [36].

A vast body of literature was produced in the recent years to automatically measure pain levels from facial color RGB images or videos [8], [33]. On the other hand, recent advances in facial video analysis using deep learning frameworks such as Convolutional Neural Networks (CNN) or Deep Belief Networks (DBN) provide the notion of realizing non-linear high dimensional compositions [35]. Deep learning architectures have been widely used in face recognition [16], [25], [42], facial expression recognition [21], [31], [43] and emotion detection [20], [30], [35]. Pain level estimation using a deep learning framework was also proposed [2], [44]. Employing deep learning framework for pain level assessment from facial video entails two kinds of information processing from facial video sequences: i) spatial information, and ii) temporal information [19]. Spatial information provides pain related information in the facial expressions of a single video frame. On the other hand, temporal information exhibits the relationship between pain expressions revealed in consecutive video frames and it provides a valuable information about the behavioral state of subjects [38]

Besides the spatial and temporal information from facial images, many other factors such as face qualities (e.g. low face resolution or brightness) [2], [9]–[13] and face capturing modalities (e.g. color RGB, depth and/or thermal) play important role in automatic pain assessment. Face quality in pain assessment was investigated in the literature [2] by using super resolved images. However, multimodal pain detection from RGB, Depth and Thermal (RGBDT) imagery is a hardly explored area both in terms of availability of databases and effective methodology for pain level classification. Lack of database in such area is a major issue of concern [22] and employing effective methodology affects the performance [18], [22]. Irani et al. collected a RGB-Depth-Thermal (RGBDT) database by employing pressure pain on the shoulder of healthy subjects and employed Support Vector Machine (SVM) on spatio-temporal features from different modalities to distinguish between different pain levels [19]. However, the database is not publicly available.

In this paper, we present the first state of the art publicly available multimodal pain intensity database for RGBDT pain level recognition in sequences[1]. We employ a hybrid deep learning framework by combining a CNN and a Recurrent Neural Network (RNN) to exploit spatio-temporal information of the collected data for each of the modalities. Then, we employ both early and late fusion strategies between modalities to investigate both the suitability of individual modalities and their complementarity.

The rest of the paper is organized as follows. Section II provides the description of the new database, including data collection, post-processing, and characteristics. Section III describes the proposed methodology in order to provide a first baseline analysis for pain level assessment using multimodal spatio-temporal data with deep learning strategies. Section IV presents the experimental setup and the obtained results. Finally, section V concludes the paper.

II. Multimodal Intensity Pain Database

Two notable databases that are publicly available for video pain experiments are the UNBC-McMaster database [27] and BioVid database [41]. However, none of these are having RGBDT modalities; and to the best of our knowledge, there is no publicly available RGBDT database that focuses on pain analysis from face. By considering such lack of availability of a multimodal RGBDT pain database, the main contribution of this work is creating a multimodal pain intensity database using experimental pain on healthy subjects. It is to be noted that electrical stimulation is a highly reproducible and non-invasive method to elicit experimental pain or discomfort; and both Functional Electrical Stimulation (FES) of muscle nerves and electrical stimulation of the Nociceptive Withdrawal Reflex (NWR) are easy to manipulate using graded stimulation intensities to generate pain [1]. Thus, we have collected the new RGBDT database, named as Multimodal Intensity Pain (MIntPAIN) database, by employing controlled electrical stimulation to generate pain to the subjects' muscles. Table I shows the distinction between the new MIntPAIN database and the other two available pain databases. Though BioVid database is a big database and has a number of non-visual modalities like ECG and EEG, the only visual modality is RGB. This is the case for the small UNBC-McMaster database as well. However, the new MIntPAIN database is a sizable one and includes all RGBDT modalities. The following subsections describe the procedure of data collection and the database structure.

A. Experimental Pain by Electrical Stimulation

Two self-adhesive electrodes (Pals Platinum Round 3.2cm, Axelgaard Ltd., USA) were placed on the muscle belly of the 'extensor digitorum' muscle. The stimulation consisted of a pulse train with a frequency of 30 Hz and square pulse duration of 200s, stimulating for 1.5 seconds, administered by an electrical stimulator (Noxitest IES 230) controlled by a computer. One cathode (Ambu® Neuroline 700) was placed on the anterior surface of the head of the second metacarpal bone and a common anode (7,5cm x 10cm Axelgaard Manufacturing Co., Ltd ® PALS Platinum) was placed on the dorsum of the subject's left hand to elicit the NWR. Each stimulus consisted of a constant current pulse train of five separate 1ms pulses controlled and delivered at 200 Hz by a

[1]Database download page: http://www.vap.aau.dk/mintpain-database/

TABLE I
THE NEW MINTPAIN DATABASE AND THE OTHER PUBLIC VISUAL DATABASES FOCUSING ON PAIN

Attribute	UNBC-McMaster database [27]	BioVid database [41]	The new MIntPAIN database
Number of subjects	129 (16 are available)	90 (87 are available)	20 (all available)
Subject's type	Self-identified pain patient	Healthy volunteers	Healthy volunteers
Pain type	Natural shoulder pain	Stimulated heat pain	Stimulated electrical pain
Pain levels	0-16 (PSPI) and 0-10 (VAS)	1-4 (Stimuli)	0-4 (Stimuli)
Available visual modalities	RGB	RGB	RGB, Depth, Thermal
Size of the database	200 variable length videos with 31,571 frames	17300 5s videos with 25fps	9366 variable length videos for all modalities with 1,87,939 frames
Year of publishing	2011	2013	2017

computer controlled electrical stimulator (Noxitest IES 230, Aalborg, Denmark).

The FES motor threshold (M^{th}) was visually identified using a simplistic staircase model with increasing steps of stimulation intensities starting at 5mA with increasing intensities in increments of 1mA until a motor response could be visually identified. The FES pain threshold (P^{th}) and the NWR stimulation P^{th} and reflex threshold (R^{th}) was detected using a more complex staircase model, with increasing and decreasing stimulation intensities. The stimulation intensity starting point was 1mA. From 1mA the intensity was increased with ascending steps of 2mA until the stimuli was considered painful (P^{th}) or a reflex occurred (R^{th}). Whereupon the stimulation intensity was reduced in steps of 2mA until the stimulations was not considered painful or no motor response was visually confirmed. The following two staircases then consisted of increasing and decreasing stimulation intensities in smaller steps of 1mA, until a total of 3 ascending and 3 descending estimations of the pain and motor threshold had been made. From these six intensity values, an average of the last four ascending and descending estimations was used as the final estimate of the thresholds. The R^{th} detection was performed online using customized software based on evaluation of interval peak z-score. A reflex was detected if the EMG signal had an interval peak z-score larger than 12. The EMG signals were rectified and their interval peak z-score was calculated as:

$$Interval\_peak\_zscore = \frac{reflex\_peak - \mu_{baseline}}{\sigma_{baseline}} \quad (1)$$

To grade the stimulation intensities used for FES, we defined four equally spaced Stimulation Intensities (SI_{1-4}) by as follows:

$$SI_{i=1to4} = M^{th} + \frac{1}{3}(M^{th} - P^{th}) * (i-1) \quad (2)$$

To grade the stimulation intensities used to stimulate the NWR, the P^{th} was multiplied with four fixed factors; 0.8, 1.0, 1.25 and 1.5, giving four stimulation intensities. These four stimulation intensities for each cases of FES and NWR were then used to stimulate the subject 10 times in a randomized order with an inter-stimulus interval of 15-20 seconds leaving time for the subjects to rate the pain intensity and unpleasantness of the pain sensation. Before starting the data acquisition, the stimulation intensities were tested to familiarize the subject to the stimulation.

As an experimental protocol, the subjects were seated comfortably in an armchair with the left arm resting. The position of the chair was adjusted to obtain a comfortable resting position and an angle of 100° in the elbow joint of the left arm. The subject was instructed to position the left arm within a marked area to ensure identical positioning throughout the entire experiment.

B. Data Acquisition Setup

EMG was recorded from the subjects left arm, from the Extensor Carpi Ulnaris (ECU) muscle, using three surface electrodes (30 x 22 mm, type 720, Ambu A/S, Denmark) in a tripolar configuration with an inter-electrode distance of 2cm. The EMG signal was pre-amplified and filtered (10-500 Hz) and stored as a 1s recording; 200ms pre- and 800ms post-stimulation. The subjects' perceived pain intensity was rated on a 10cm electronic VAS, anchored by assigning 0 as the perception threshold, 5 as the PTh and 10 as the most intense pain imaginable. Each stimulation was rated by the subject and stored. The subjects' perceived unpleasantness intensity was also rated on a 10cm electronic VAS, anchored by assigning '0' as not unpleasant, 5 as unpleasant and 10 as the most unpleasant imaginable.

To capture the facial pain expression of the subject during the stimulation, the database was collected in three modalities: color RGB, depth and thermal. Color RGB and depth data of frontal facial images are captured by Microsoft Kinect Version2. The thermal data was captured by Axis Q1922 thermal camera. A Logitech camera was also used to determine a light signal to indicate the starting and ending time stamps of electrical stimulation. After collecting the raw data while giving the electrical stimulation to the subjects, we followed some post-processing steps to organize the database. In particular, we synchronized the facial image frames in all three modalities by following the capturing time stamps and annotated them in sequences with different pain levels obtained from the EMG data. Time synchronization of the frames was accomplished by the time stamps in the RGB frames with variable frame rate. Depth frames followed the same time stamps than RGB. However, the thermal frames were captured exactly at 30 fps. Thus, we just discarded thermal frames by keeping only the ones corresponding to the RGB time stamps. We also

(a) RGB (b) Thermal (c) Depth

Fig. 1. Example of captured video frames in different modalities of the new MIntPAIN database. Depth image is histogram equalized for visualization.

TABLE II
KEY ATTRIBUTES OF THE NEW MULTIMODAL INTENSITY PAIN (MINTPAIN) DATABASE OBTAINED BY ELECTRICAL STIMULATION

Attribute	Value and comments
No. of subjects	20 healthy volunteers
Age range	(22-42)y with mean 29.8y
Height range	(1.60-2.00)m with mean 1.79m
Weight range	(50.0-110.0)kg with mean 81.20kg
Pain levels	(0-4), 0 for no-pain and (1-4) for four pain levels
Pain type	Electrical stimulation (including both FES and NWR in two trial) for (1-10)sec in each sweep
Self-report	Using VAS ranging (0-10)
Visual Modalities	RGB resolution 1920x1080 with fps<30 Depth resolution 512x424 with fps<30 Thermal resolution 640x480 with fps=30
Sequence details	Total 9366 videos (50-50 pain/non-pain) Average frames in each sequence = 20.07 (for RGB) Duration of the sequences [1-10]sec

provide an approximate image registration across modalities by means of homography estimation. For this, we calculated homography matrices from RGB to D and T using [15].

C. Structure of the Database

The new MIntPAIN database has multimodal pain data obtained by giving electrical stimulation in five different levels (Level0 to Level4, where 0 implies no stimulation and 4 implies the highest degree of stimulation) to 20 healthy subjects. After prior ethical approval for the data collection, the subjects were invited to be volunteer. They were adequately informed about the electrical pain stimulation and overall data recording procedure. Each subject exhibited two trials during the data capturing session and each trial has 40 sweeps of pain stimulation. In each sweeps we captured two data: one for no pain (Label0) and the other one for one of the four pain levels (Level1-Level4). As a whole each trial has 80 videos (50-50 pain/non-pain ratio) for 40 sweeps. Among these, some sweeps are missing for few subjects. This is due to the unexpected noise in the EMG reading of one subject, talking by one subject during data capturing, and lack of VAS scale by two experimental subjects. Fig. 1 shows some full-frame samples of a recorded subject for the three different modalities. Fig. 2 shows examples of cropped database faces for the different annotated pain levels and modalities. Note the clear difficulty in performing visual assessment of this complex multi-class problem, particularly for the second subject (at the right) in the figure. A list of key attributes of the database is shown in Table II.

III. DEEP MULTIMODAL PAIN DETECTION

In this section, we describe the methodological proposal in order to perform a baseline analysis for the 5-level pain recognition on the presented database. We test standard deep approaches to the three provided modalities. We then fuse the modalities by employing both early and late fusion techniques. The pain scores are measured by employing CNN (to exploit spatial characteristics) and a combination of CNN and RNN (to exploit spatio-temporal characteristics) on both individual and fused modalities. In this section, we first describe the preprocessing steps and the architecture of the deep learning strategies considered. Finally, we discuss the different fusion strategies that have been applied.

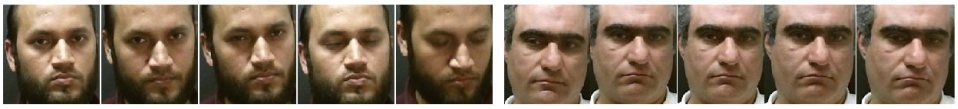
A. Preprocessing

Fig. 1 shows that the original RGBDT video frames present a large portion of the subject body in the space of the acquisition room. For the experimental evaluation we just focus on face-based pain recognition. Thus, on the synchronized data modalities, we applied face detection using the method of [28] on RGB modality and cropped associated faces on D and T modalities by using computed homographs. The procedure is shown in Fig. 3 and described below. We are providing time synchronization and homography matrices codes together with the database. Time and space calibration steps are described in database section. The cropped faces are then fed to deep learning frameworks for individual and fusion performance analysis.

B. Baseline evaluation

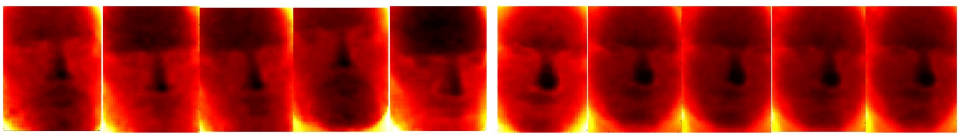
In order to provide a baseline results on the presented database we use a standard two step deep approach. First we apply a 2D-CNN for frame wise feature extraction and pain recognition. Secondly, an implementation of RNN called Long-Short Term Memory (LSTM) [2] is used to estimate the temporal relations between the frames and to perform sequence level pain recognition.

We fine-tuned the VGG-FACE model [2] pre-trained with faces. The model was fine-tuned against different modalities, specifically RGB, D and T, creating three different models used for feature extraction. Given the lack of existing pre-trained models of faces on depth and thermal modalities and considering the moderate amount of data of our database to train it with VGG-FACE from scratch, we considered to use the same pre-trained model (RGB) for the fine-tuning of all three modalities. This allowed us to make important



(a) RGB faces

(b) Thermal faces



(c) Depth faces

Fig. 2. Faces from two subjects for all 5 pain levels (Level0 to Level4 from left to right) for all different modalities. The depth images are depict by editing the colormap for visualization purpose.

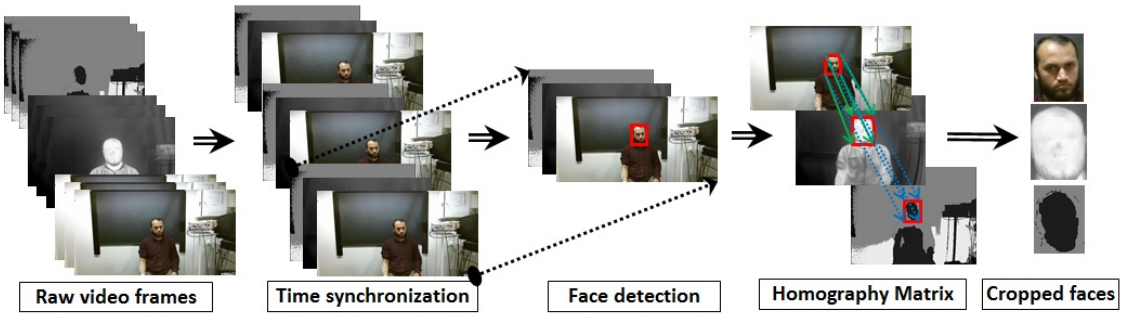


Fig. 3. Preprocessing steps employed on the raw video frames of different modalities to crop facial regions before deep learning of pain levels.

contributions by asking: whether a model pre-trained against RGB data can also be used with similar data captured in the other modalities like D and T, and whether what the network learned in RGB is still meaningful in the other modalities.

A 2D-CNN is unable to estimate long term temporal relations between frames. Therefore, a LSTM is used to learn these temporal relations. The hybrid CNN+LSTM pain detection framework is depicted in Fig. 4 along with different fusion strategies. First, we extract facial features for the frames. We obtain the features of the `fc7` layer of the fine-tuned VGG-FACE and use those as input to the LSTM to exhibit hybrid deep learning performance. Pain levels (labelled from 0-4) are predicted sequence-wise, *i.e.* given an unknown sequence of n frames $f_i \in \{f_1, ..., f_n\}$, the target prediction is the pain level of the f_n frame. Thus, training is set so that the information contained in the past frames is used in order to predict the current pain level.

C. Fusion strategies

In order to analyze the potential of the different visual modalities, in the experimental section we evaluate both early and late fusion methods as described below.

1) Early Fusion: For the early fusion analysis, different modalities are integrated into the single combined data stream. Then classifiers are fed using these large multimodal input vectors for training. The combination of multiple individual modalities into a single one implicates that the joint representations are projected to the same space using all the modalities as input. In our specific case, D and T data have been stacked together as new channels with the RGB data. The fusing process generates a 5-dimensional matrix for each of the video frames (more specifically, cropped facial region), considered as the new input for the CNN.

2) Late Fusion: We consider late fusion by integrating outputs of individual classifiers (deep models in our case) as an input feature vector for a second stacked classifier. This second classifier is the one which is in charge of producing final classification [6].

IV. Experimental results

In order to present the results, first we discuss the experimental setup. In terms of experiments, we evaluate the database for CNN and LSTM 5-class pain recognition both at frame and sequence levels for the different modalities, and early and late fusion strategies.

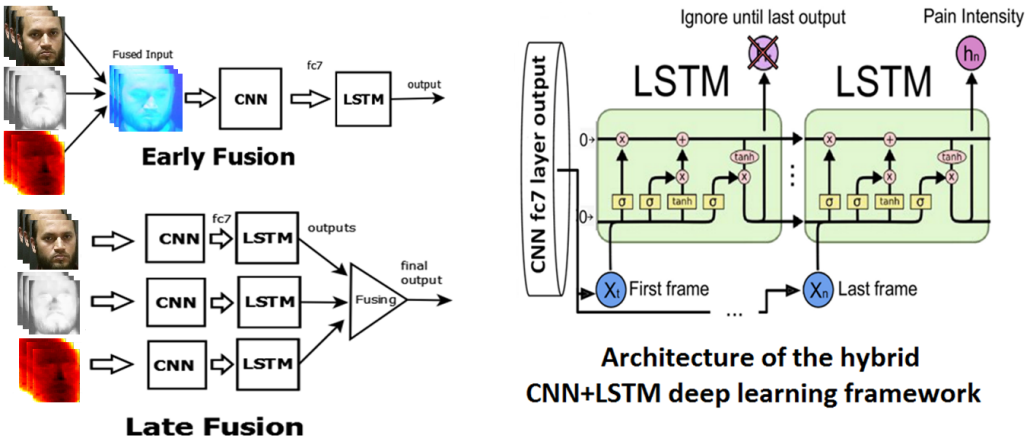


Fig. 4. The block diagram of fusion strategies along with the deep hybrid classification framework based on a combination of CNN and LSTM.

A. Experimental setup

We divided the 20 subjects of the proposed database in 5 disjoint sets and run 5-fold cross validation. Thus, each partition corresponds to 16 subjects for training and 4 not previously observed subjects for testing. Results are reported as mean per frame and sequence accuracy over all five classes. Sequence evaluation is addressed by majority voting of individual frames predicted labels of the sequence. Since the examples of class 0 i.e. *no pain* are several times more than the other four classes, we augment the samples belonging to the 4 classes to balance the number of training examples. The data augmentation is performed by rotating the cropped faces five degrees to the right and left. Thus, giving us three times more training examples.

B. CNN independent modality evaluation

In this section, we discuss the challenges of finetuning a CNN on our datasebase. We finetune the VGG-Face network independently by each modality with a base learning rate of 0.00001 and momentum 0.1. We train all the layers of the VGG-FACE network. We train the fully connected layers 10 times faster than the convolutional layers. The results are compiled in Table III. One can observe that the accuracy achieved by independent modalities is near guess prediction given the inherent complexity of the pain level recognition problem as well as the presence of new subjects at test stage. Some samples of subjects of the database in Fig. 2 (more particularly, the second subject in the right) show the clear difficulty to perform pain level assessment by human observation. On the other hand, we observed that finetuning D and T channels from a pretained RGB network provides some meaningful information. As we will see below in the case of fusion of modalities, it will be demonstrated by the fact that the fusion of these fine-tuned modalities enhance final performance in relation to isolated CNN models performance.

C. CNN-LSTM independent modality evaluation

For learning the temporal relationships between frames we implement an LSTM. The input to the LSTM are the per frame feature vectors extracted from the *fc7* layer of the fine-tuned VGG-Face CNN. We implement the LSTM framework for each modality. While training LSTM we vary the hidden states between 64 and 256. We also try a single layer to a 3 layers deep LSTM. The learning rate is 0.001 and we trained the network until 50 epochs. Results are shown in tab III. We concluded that there are two main reasons for the low performance of the hybrid CNN+LSTM system. First, the low performance of the independent CNN based features signifies that the per frame feature vectors are not discriminative enough to allow LSTM for a better generalization. Secondly, we have limited the number of sequences to train the LSTM. Although the influence of temporal information is clearly motivated in the literature for pain assessment, we found that a simple state of the art baseline based on LSTM with standard CNN features is not enough to provide good generalization capabilities in this scenario and based on the amount of provided data.

D. Fusion of modalities

In order to analyze if different modalities can complement each other to enhance pain level recognition performance, we ran early and late fusion analyses on all four possible combinations of fusions against the three modalities. Early fusion of modalities is used to fine-tune the VGG-Face network. On the other hand, while doing late fusion the confidence scores of classes from different modalities are combined with a Random Forest classifier. The training parameters are the same to the training parameters for fine-tuning the VGG-Face network as in the case of the independent modalities experiment. In tab. IV we show the results of early fusion (EF) and late fusion (LF) for all combinations of modalities. From this table, we can observe that the best result is achieved by the early fusion of all three modalities. The confusion matrix w.r.t. this result is shown in fig. 5. It is also apparent from the table that both early fusion and late fusion strategies are more discriminative than individual modalities. The sequence level accuracy is slightly higher mainly because the majority voting may help in some cases to recover from isolated frame miss-classifications because of the usage of majority voting proce-

TABLE III
VGG-FACE CNN AND LSTM RESULTS ON INDEPENDENT MODALITIES.THE TOP ROW IS PER FRAME ACCURACY AND THE BOTTOM ROW IS PER SEQUENCE ACCURACY. BEST SCORES IN BOLD

Modalities	CNN-RGB	CNN-T	CNN-D	LSTM-RGB	LSTM-D	LSTM-T
Mean Frame(%)	**18.17**	18.08	16.71	15.36	14.72	13.13
Mean Sequence (%)	**18.55**	18.33	17.41	15.36	14.72	13.13

TABLE IV
EARLY FUSION (EF) AND LATE FUSION (LF) RESULTS FOR DIFFERENT COMBINATIONS OF THE MODALITIES. TOP ROW IS PER FRAME ACCURACY AND THE BOTTOM ROW IS PER SEQUENCE ACCURACY. BEST SCORES IN BOLD

Fusion	EF-RGB-T	EF-RGB-D	EF-D-T	EF-RGB-DT	LF-RGB-T	LF-RGB-D	LF-D-T	LF RGB-D-T
Mean Frame (%)	23.85	24.62	23.12	**32.40**	21.80	23.20	22.50	25.20
Mean Sequence(%)	30.77	27.92	25.30	**36.55**	22.10	22.30	22.70	25.40

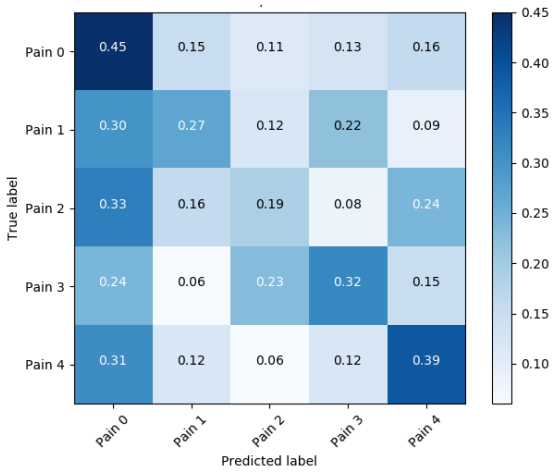


Fig. 5. The confusion matrix corresponding to the early fusion of all three modalities.

dure. We also experimented with features extracted from early fused data to train an LSTM. Our preliminary experiments showed that the performance was not comparable to the early fusion experiments just with CNN, being in correlation to the results obtained by LSTM in the case of isolated modalities evaluation.

V. CONCLUSION

This work presented the first public available state of the art database for pain assessment from RGB-Depth-Thermal sequences. The new database includes 20 subjects and has been annotated at frame level with 5 different levels of pain. We also provided a first baseline based on standard CNN and LSTM deep learning strategies. Furthermore we performed both early and late fusion of modalities in order to evaluate their complementary in order to enhance the recognition performance of pain levels. From our evaluations, we observed that fusion of modalities are more discriminative than training the classifiers with independent ones for this task. The early fusion of all three modalities provided the highest performance. These results support the usability of the different visual data sources provided in the database. We also observed that the usage of LSTM to learn long term dependencies in our data achieves poor performance for the considered input fine-tuned VGG-features. Further work includes the analysis of alternative appearance and temporal features from the different modalities, different models for spatio-temporal inference, as well as fusion strategies in order to provide further insights about the complementary of the three visual modalities.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work has also been partially supported by Estonian Research Council Grant PUT638, the Scientific and Technological Research Council of Turkey (TBTAK) 1001 Project (116E097), and the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund. This study was funded by the Danish National Research Foundation (DNRF121) as well. This project has also received funding from the European Unions Horizon 2020 research and innovation program under Marie Skodowska-Curie grant agreement No 6655919.

REFERENCES

[1] Ole K Andersen, Finn A Sonnenborg, and Lars Arendt-Nielsen. Reflex receptive fields for human withdrawal reflexes elicited by non-painful and painful electrical stimulation of the foot sole. *Clinical Neurophysiology*, 112(4):641 – 649, 2001.

[2] Marco Bellantonio, Mohammad A Haque, Pau Rodriguez, Kamal Nasrollahi, Taisi Telve, Sergio Escarela, Jordi Gonzalez, Thomas B Moeslund, Pejman Rasti, and Gholamreza Anbarjafari. Spatio-temporal pain recognition in cnn-based super-resolved facial images. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 151–162. Springer, 2016.

[3] A. C. de C Williams, H. T. Davies, and Y. Chadury. Simple pain rating scales hide complex idiosyncratic meanings. *Pain*, 85(3):457–463, Apr 2000.

[4] D. J. Debono, L. J. Hoeksema, and R. D. Hobbs. Caring for patients with chronic pain: pearls and pitfalls. *J Am Osteopath Assoc*, 113(8):620–627, Aug 2013.

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

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[7] Ken Steffen Frahm, Carsten Dahl Mørch, Warren M. Grill, and Ole Kæseler Andersen. Experimental and model-based analysis of differences in perception of cutaneous electrical stimulation across the sole of the foot. *Medical & Biological Engineering & Computing*, 51(9):999–1009, Sep 2013.

[8] T. Hadjistavropoulos, D. L. LaChapelle, F. K. MacLeod, B. Snider, and K. D. Craig. Measuring movement-exacerbated pain in cognitively impaired frail elders. *Clin J Pain*, 16(1):54–63, Mar 2000.

[9] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund. Facial video-based detection of physical fatigue for maximal muscle activity. *IET Computer Vision*, 10(4):323–329, 2016.

[10] M. A. Haque, R. Irani, K. Nasrollahi, and T. B. Moeslund. Heartbeat rate measurement from facial video. *IEEE Intelligent Systems*, 31(3):40–48, May 2016.

[11] M. A. Haque, K. Nasrollahi, and T. B. Moeslund. Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera. In *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 443–448, Aug 2013.

[12] M. A. Haque, K. Nasrollahi, and T. B. Moeslund. Constructing facial expression log from video sequences using face quality assessment. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 517–525, Jan 2014.

[13] M. A. Haque, K. Nasrollahi, and T. B. Moeslund. Quality-aware estimation of facial landmarks in video sequences. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 678–685, Jan 2015.

[14] Mohammad A Haque, Kamal Nasrollahi, and Thomas B Moeslund. Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain? In *Identity, Security and Behavior Analysis (ISBA), 2017 IEEE International Conference on*, pages 1–8. IEEE, 2017.

[15] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[16] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning. *Pattern Recognition*, 48(10):3113–3124, 2015.

[17] Chaudhary Muhammad Aqdus Ilyas, Kamal Nasrollahi, Thomas B. Moeslund, Matthias Rehm, and Mohammad Ahsanul Haque. *Facial Expression Recognition for Traumatic Brain Injured Patients*, volume 4, page 1. SCITEPRESS Digital Library, 2018.

[18] R. Irani, K. Nasrollahi, and T. B. Moeslund. Pain recognition using spatiotemporal oriented energy of facial muscles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 80–87, June 2015.

[19] R. Irani, K. Nasrollahi, M. O. Simon, C. A. Corneanu, S. Escalera, C. Bahnsen, D. H. Lundtoft, T. B. Moeslund, T. L. Pedersen, M. L. Klitgaard, and L. Petrini. Spatiotemporal analysis of rgb-d-t facial images for multimodal pain level recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 88–95, June 2015.

[20] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, Raul Chandias Ferrari, Mehdi Mirza, David Warde-Farley, Aaron Courville, Pascal Vincent, Roland Memisevic, Christopher Pal, and Yoshua Bengio. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[21] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2):173–189, 2016.

[22] Juris Klonovs, Mohammad A. Haque, Volker Krueger, Kamal Nasrollahi, Karen Andersen-Ranberg, Thomas B. Moeslund, and Erika G. Spaich. *Distributed Computing and Monitoring Technologies for Older Patients*. SpringerBriefs in Computer Science. Springer International Publishing, Cham, 2016.

[23] Miriam Kunz, Andreas Gruber, and Stefan Lautenbacher. Sex differences in facial encoding of pain. *The Journal of Pain*, 7(12):915 – 928, 2006.

[24] Miriam Kunz, Kenneth Prkachin, and Stefan Lautenbacher. Smiling in Pain: Explorations of Its Social Motives. *Pain Research and Treatment*, 2013:e128093, August 2013.

[25] Haoxiang Li and Gang Hua. Hierarchical-pep model for real-world face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '15, pages 4055–4064, 2015.

[26] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Trans Syst Man Cybern B Cybern*, 41(3):664–674, Jun 2011.

[27] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011.

[28] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

[29] S.B. McMahon, M. Koltzenburg, I. Tracey, and D.C. Turk. *Wall & Melzack's Textbook of Pain,Expert Consult - Online and Print,6: Wall & Melzack's Textbook of Pain*. ClinicalKey 2012. Elsevier/Saunders, 2013.

[30] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, (99):1–1, 2017.

[31] Ikechukwu Ofodile, Kaustubh Kulkarni, Ciprian Adrian Corneanu, Sergio Escalera, Xavier Baro, Sylwia Hyniewska, Juri Allik, and Gholamreza Anbarjafari. Automatic recognition of deceptive facial expressions of emotion. *arXiv preprint arXiv:1707.04061*, 2017.

[32] K. M. Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51(3):297–306, Dec 1992.

[33] K. M. Prkachin, I. Schultz, J. Berkowitz, E. Hughes, and D. Hunt. Assessing pain behaviour of low-back pain patients in real time: concurrent validity and examiner sensitivity. *Behav Res Ther*, 40(5):595–607, May 2002.

[34] K. M. Prkachin and P. E. Solomon. The structure, reliability and validity of pain expression: evidence from patients with shoulder pain. *Pain*, 139(2):267–274, Oct 2008.

[35] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. *Multimodal emotion recognition using deep learning architectures*. Institute of Electrical and Electronics Engineers Inc., United States, 5 2016.

[36] A. Sikdar, S. K. Behera, and D. P. Dogra. Computer vision guided human pulse rate estimation: A review. *IEEE Reviews in Biomedical Engineering*, PP(99):1–1, 2016.

[37] Karan Sikka, Alex A. Ahmed, Damaris Diaz, Matthew S. Goodwin, Kenneth D. Craig, Marian S. Bartlett, and Jeannie S. Huang. Automated assessment of children's postoperative pain using computer vision. *Pediatrics*, 2015.

[38] Daniel Simonsen, Ramin Irani, Kamal Nasrollahi, John Hansen, Erika Geraldina Spaich, Thomas B. Moeslund, and Ole Kæseler Andersen. *Validation and Test of a Closed-Loop Tele-rehabilitation System Based on Functional Electrical Stimulation and Computer Vision for Analysing Facial Expressions in Stroke Patients*, pages 741–750. Springer International Publishing, Cham, 2014.

[39] Dennis C Turk and Ronald Melzack. The Facial Expression of Pain. *Handbook of Pain Assessment*, pages 117–133, 2011.

[40] April Hazard Vallerand and Rosemary C. Polomano. The relationship of gender to pain. *Pain Management Nursing*, 1(3, Supplement 1):8–15, September 2000.

[41] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *Cybernetics (CYBCONF), 2013 IEEE International Conference on*, pages 128–131. IEEE, 2013.

[42] Jiaolong Yang, Peiran Ren, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. *arXiv preprint arXiv:1603.05474*, 2016.

[43] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 435–442, New York, NY, USA, 2015. ACM.

[44] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. *arXiv preprint arXiv:1605.00894*, 2016.