Universität Ulm | 89069 Ulm | Germany

Fakultät für
Ingenieurwissenschaften,
Informatik
und Psychologie
Institut für Neuroinformatik
Direktor: Prof. Dr. Dr. Daniel Braun

Information Fusion Mechanisms for Multi-Modal Affect Recognition

Dissertation zur Erlangung des Doktorgrades
Doktor der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie
der Universität Ulm

vorgelegt von
Patrick Thiam
aus Bafoussam (Kamerun)

Ulm 2020

Amtierender Dekan der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie: Prof. Dr.-Ing. Maurits Ortmanns

Gutachter: PD Dr. Friedhelm Schwenker
Gutachter: Prof. Dr. Mariofanna Milanova
Gutachter: Prof. Dr. Hans Armin Kestler

Tag der Promotion: 15.02.2021

Abstract

Natural human interactions are enabled by the inherent ability of human beings to use a combination of different communication cues stemming from a diverse set of modalities (such as facial expressions, paralinguistic vocalizations, intonation patterns, or changes in body posture) in order to continuously assess the current context of the interaction and respond accordingly. This points out at the fact that multi-modality is rather a natural and crucial element of natural human interactions. In order to enable a similar form of communication in *human-computer interactions*, computer systems need to be able to perceive, assess and successfully aggregate several forms of information (which are usually converted into measurable parameters) across various modalities (e.g. audio, video, bio-physiological signals) in order to continuously model and dynamically adapt to a user's current affective state. This is particularly relevant nowadays, since significant technological advancements in such domains as sensors and data persistence, enable almost every single ubiquitous device with the ability of recording and streaming a huge amount of bio-physiological and biometric data. Therefore, these devices provide abundant and diverse data (which are continuously and systematically collected) that can be used to model and assess a user's affective, physical and psychological state in order to further improve human-computer interactions, as well as enabling new fields of applications, such as remote patient monitoring. Thus, suitable fusion approaches are needed for the extraction, selection and successful combination of relevant information from a set of diverse modalities in order to improve both the robustness as well as the performance of a specific affect recognition system. This thesis introduces novel multi-modal information fusion mechanisms for several pattern recognition tasks, in the domains of *active learning*, *supervised learning* and *deep learning*.

Concerning active learning, the optimization of the training process of an inference model is pursued by first introducing a multiple criteria sample selection approach. The goal of the proposed method is to perform an aggregation of the outputs of multiple heuristics in order to improve the robustness of the sample selection process. Moreover, the method is further extended in order to exploit complementary information stemming from adjacent and correlated input channels for the selection of the most informative samples, which are subsequently used to train an audiovisual events detection model.

Furthermore, each single step involved in the design, optimization and assessment of a multiple classifier system is thoroughly described and assessed in the context of the development of a pain intensity classification system. Manual feature engineering is applied for the design and extraction of several feature representations, based on a multitude of modalities ranging from audio to video signals as well as bio-physiological parameters. The extracted feature representations are subsequently combined using different fixed and trainable information fusion approaches for the optimization of a robust and effective pain intensity classification model.

Lastly, feature learning in the form of deep physiological models and multi-stream attention-based convolutional neural networks are investigated, in order to improve on some of the shortcomings of manual feature engineering for the development of robust pain intensity classification models. The assumption behind the proposed deep fusion architectures is that enabling a specific neural network architecture to autonomously and simultaneously optimize a set of suitable feature representations as well as the corresponding set of fusion parameters can significantly improve the performance of the classification system. The validity of the proposed approaches and methodological contributions is empirically evaluated through an extensive assessment involving custom as well as publicly available datasets.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor PD Dr. Friedhelm Schwenker, for the outstanding guidance and supervision, and also his undeterrable patience throughout the years that led to the completion of the current thesis. Moreover, I would like to thank Prof. Dr. Günther Palm for the various constructive discussions and insightful suggestions. I would also like to thank Prof. Dr. Hans A. Kestler for the provided invaluable support and guidance. I also wish to thank Prof. Dr. Mariofanna Milanova for kindly accepting to review and evaluate the current work.

I can not begin to express my thanks to my family, in particular my Mom, for the never ending and unconditional love, support and encouragement, thought physically distant. Mom, there are no words that can explain how much you mean to me. This achievement would not have been possible without your endless love and never ending support. I will never be able to thank you enough for everything. Also, many thanks to my colleagues and friends, for the countless hours spent together designing, implementing, discussing and writing the publications upon which this dissertation is built. Special thanks to Viktor Kessler, Peter Bellmann and Dr. Ludwig Lausser for the diverse and invaluable inputs and help provided at each step of the completion of the thesis.

Finally, I would like to thank anybody involved directly or indirectly in the process of creation of the present work. Thank you all for the support and insights provided throughout the last years.

Contents

Chapter 1

Introduction

During human interactions, each single individual continuously alternates between assessing the actual context of an interaction, and accordingly choosing a suitable channel of communication in order to convey some form of information. This is made possible by the inherent ability of human beings to display and recognize specific communication cues using a combination of different information channels (or modalities) such as speech, facial expressions, postural shifts and gestures, among others. Therefore, the concept of multi-modality is rather natural and crucial regarding human interactions, since the inability to perceive or communicate social and interpersonal cues negatively affects the dynamics of such interactions. Hence, the field of *Affective Computing* [Picard 1997] aims to significantly improve Human-Computer Interaction (HCI) through the implementation of a similar form of multi-modal communication. Computer systems should therefore be able to perceive and aggregate information stemming from diverse sources in order to assess, model and recognize users' affective dispositions. This will enable computer systems to adapt more naturally to the users' needs and provide suitable responses accordingly, therefore substantially enhancing the user experience [Schwenker, Böck, et al. 2017] and further enabling new fields of application such as in the domains of health monitoring [Dautov et al. 2019] or user-centered content retrieval [Gupta et al. 2016; Rinaldi and Russo 2018].
Furthermore, the concept of multi-modality becomes even more relevant when considering the multifaceted characteristics of affective dispositions. More specifically, emotions are conveyed through a combination of several distinct channels, each of which provides a specific amount of complementary information depicting just a single facet or aspect of the underlying affective state, when observed alone. Hence, the aggregation of complementary information stemming from different sources is more likely to improve the discrimination performance between ambiguous emotional states than relying on the information stemming from a unique source. Moreover, the combination of multiple sources of complementary

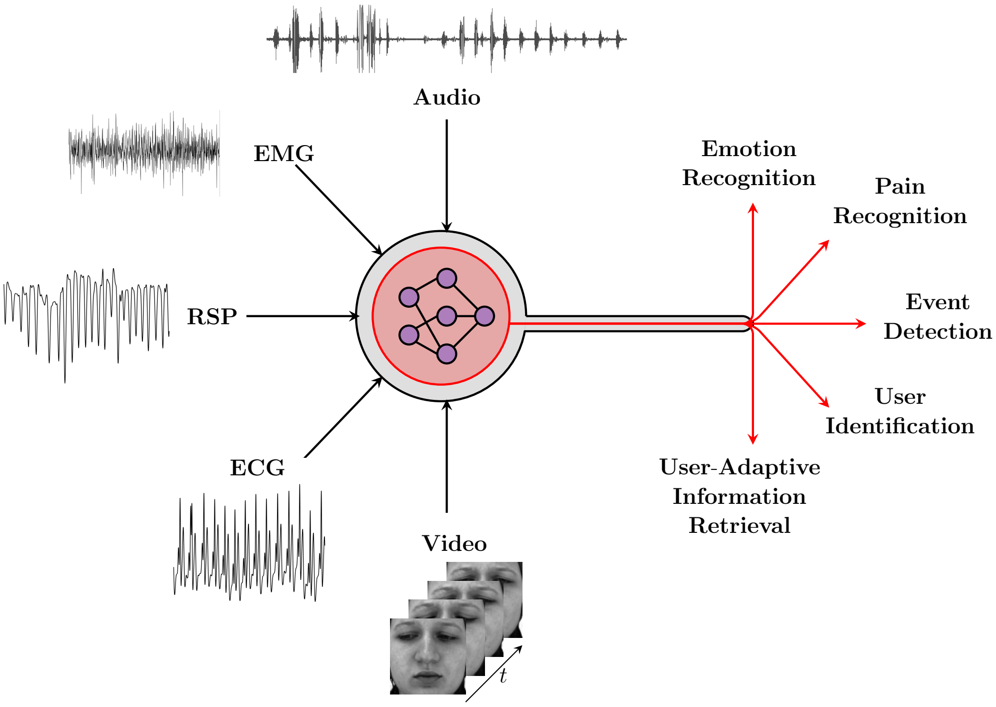


Figure 1.1: *How to combine information stemming from a multitude of sources in order to improve the performance of an inference model, depending on the underlying pattern recognition task?*

information improves the robustness of the system by simultaneously capitalizing on each source's advantages and mitigating the drawbacks of each individual channel.

This specific topic has become very relevant throughout the last decade, since most of the ubiquitous smart devices used nowadays (e.g. phones, watches, tablets, voice assistants, fitness devices) are interconnected and equipped with countless sensors, that continuously record and stream a huge amount of biometric and bio-physiological data such as electrocardiogram (ECG), respiration signal (RSP), electromyography (EMG) or electrodermal activity (EDA). Therefore, researchers are provided with an inexhaustible source of heterogeneous data that can be used in order to assess a user's habits and affective dynamics. However, several challenges arise when it comes to dive into such a huge amount of data in order to extract and combine relevant information for the optimization of an inference model specific to an underlying pattern recognition task (see Figure 1.1). These challenges can be summarized into the following research questions:

Which information is relevant for the underlying pattern recognition task?

Since the performance of an inference model relies primarily on the quality of the training material, the identification and selection of relevant information relative to the underlying pattern recognition task constitutes a crucial and challenging endeavor. This information can be represented either by a set of relevant modalities (e.g. video, audio, EDA), as well as by a set of informative samples or by a specific set of diverse feature representations. This becomes even more relevant in a scenario where a huge amount of unprocessed data is available. Using the totality of available sources of information can be extremely computationally expensive and in most cases inefficient, since there is a high probability of running into a huge amount of redundant, contradicting or noisy information. Therefore, the available data has to be pre-processed and the most relevant information has to be identified and extracted into suitable feature representations.

Hence, the optimization of the training material can occur at different levels of abstraction. The lowest level consists in selecting relevant modalities. This is usually done based on some expert knowledge of the underlying pattern recognition task. The mid-level consists of the selection of relevant modality specific samples. For example, in the case of *supervised learning* pattern recognition tasks, the training material consists of a set of labeled samples. Data labeling is known to be error prone, cumbersome, temporally expensive and costly. Therefore, several approaches stemming from the domains of *active learning* [Settles 2009], and *semi-supervised learning* [Chapelle et al. 2006] are proposed in order to improve the efficiency of the labeling process by substantially reducing the amount of labeled data needed for the optimization of an effective inference model through the combination of both manual and automatic annotation [Thiam, Meudt, Schwenker, et al. 2016]. This is done by assessing the informativeness of the input samples based on designed heuristics and selecting the most informative samples to be either manually labeled by a human annotator or automatically labeled by a pre-trained inference model. The highest level consists of the selection of relevant feature representations (or feature selection): at this level, several techniques ranging from sequential to uni-variate feature selection approaches have been proposed in order to simultaneously reduce the dimensionality of the feature space while improving the performance of the trained inference model [Khalid et al. 2014; Kächele, Zharkov, et al. 2014]. Thus, the selection of relevant information constitutes an essential pre-processing step that contributes to the reduction of computational costs, as well as to the improvement of the performance of an inference model. Moreover, the identification of each source of information that positively or negatively impacts the performance of a specific inference model can provide more insights for a better understanding and interpretation of the underlying process.

When should the processed information be aggregated?

Modality specific characteristics such as temporal granularity (or sampling rate) and signal representation (e.g. one-dimensional signal representations such as bio-physiological signals, two-dimensional signal representations such as images, three-dimensional signal representations such as video signals), as well as feature representations' attributes and data temporal alignment (or synchronization) constitute a set of diverse properties across modalities that have to be taken in consideration while designing a suitable architecture to perform the aggregation of the chosen set of information [Poh and Kittler 2010]. Depending on the level of abstraction at which the aggregation is performed, three main aggregation architectures can be identified, consisting of *early fusion*, *mid-level fusion* and *late fusion* [Schels et al. 2013]. Early fusion consists in aggregating either the preprocessed raw input signals or the extracted feature representations across the modalities into a single high dimensional representation. This representation is subsequently used for the optimization of an inference model. It constitutes one of the simplest aggregation approaches. However it can just be applied on identical signal representations characterized with an identical sampling rate. Moreover, the high dimensional representation resulting from the aggregation at such a low level can be problematic both computationally and also in case the content across the respective representations is highly contradicting, which can negatively impact the performance of the trained model. A more appropriate approach consists in performing the aggregation at a higher level of abstraction such as in the case of mid-level and late fusion techniques.

Mid-level approaches consist of a layered construct of successive processing steps, during which the output of each layer is used as input for the subsequent layer. The conducted processing steps consist usually of generating diverse subsets of the input data and subsequently optimizing an inference model for each specific subset, resulting in the generation of different sets of intermediate representations. This process can be carried out throughout successive layers. The last layer however performs the mapping between the current intermediate representations and the defined output of the underlying pattern recognition task. Late fusion approaches on the other hand consist of training a single inference model on each specific representation and subsequently performing the aggregation of the models' outputs at a subsequent layer [Bellmann et al. 2018]. Both approaches improve the scalability of the trained inference model and reduce the complexity of the underlying task by performing an aggregation of related and less complex sub-tasks. However, a suitable amount of training material is needed in order to effectively optimize the models specific to each level of abstraction.

How should the processed information be aggregated?

Once relevant feature representations have been computed and a specific aggregation architecture has been designed, an aggregation rule has to be defined in order to map the resulting intermediate representations to the predefined outputs of the underlying pattern recognition task. One can distinguish between fixed aggregation rules and trainable aggregation rules [Kuncheva 2004]. Fixed aggregation rules such as Majority Voting, Average Rule, Product Rule are straightforward and do not require any type of further optimization, due to the nonexistence of adaptive parameters to be optimized. On the other hand, trainable aggregation rules such as Weighted Average Rule, Dempster-Shafer Aggregation Rule [Wu et al. 2003], or Decision Templates [Schwenker, Dietrich, et al. 2006] are characterized by a set of trainable parameters to be optimized and subsequently used to perform a weighted aggregation of the outputs of the respective intermediate inference models. The optimized weights should reflect the relative significance or informativeness of each intermediate representation. Moreover, a deep neural network can also be designed and directly applied on the pre-processed raw signals in order to take advantage of its hierarchical construct for the simultaneous and autonomous generation of suitable representations and aggregation parameters [Ramachandram and Taylor 2017]. Even though trainable aggregation rules have proven to be able to significantly outperform fixed rules, such approaches require substantially more training material as well as computational resources for an effective optimization of the aggregation parameters.

The contributions described in the current thesis consist of novel information fusion approaches applied in the domains of *active learning*, *supervised learning* and *deep learning* respectively. The description of the proposed approaches involves providing answers to each of the previously described research questions related to specific pattern recognition tasks.

The work presented in [Thiam, Meudt, Palm, et al. 2018] and summarized in Chapter 2 consists of the description of two novel active learning approaches for the detection of audiovisual events. First, a Multiple Criteria Sample Selection (MCSS) approach is proposed, which is characterized by the aggregation of the results of a diverse set of heuristics for the selection of the most relevant and informative samples that are subsequently used for the optimization of an audiovisual events detection model. Furthermore, the proposed approach is extended into a multi-modal approach in order to further reduce the amount of labeled samples needed for the optimization of a suitable events detection model. Therefore, the proposed extension consists of a combination of active and semi-supervised learning techniques, consisting of using the temporal correlation of events' occurrences in both audio and video channels for the selection and automatic annotation of relevant samples. The assessment of both approaches was performed on the *Ulm University Multimodal Affective Corpus* (UUlmMAC) [Hazer-Rau et al. 2020] and showed that in most cases, an inference model trained with a little less than 30% of the dataset is able to achieve the same performance as a model trained on the totality of the available dataset. Therefore, the application of the proposed approaches in a real world scenario would further improve the efficiency of the cumbersome and costly annotation process, without any loss of performance of the optimized inference model.

In Chapter 3, a summary of the work presented in [Thiam, Kessler, et al. 2019] is provided. The work conducted consists of the design and assessment of a multimodal pain intensity classification system based on the *SenseEmotion Database* [Velana et al. 2017]. The whole design process starting from the pre-processing of each involved single modality, through the extraction and selection of relevant feature representations, until the assessment of the performance of uni-modal classification systems based on each single modality, as well as the assessment of various information fusion approaches combined with different aggregation rules is thoroughly described. The performed assessments show that given a substantial amount of training material, late fusion approaches with trainable aggregation rules significantly outperform other information fusion approaches.

Chapter 4 provides a summary of the works presented in both [Thiam, Bellmann, et al. 2019] and [Thiam, Kestler, et al. 2020b]. Both works consist of deep multi-modal fusion architectures applied on bio-physiological signals of the *BioVid Heat Pain Database* [Walter, Gruss, Ehleiter, et al. 2013] and the video signals of both the *BioVid Heat Pain Database* and *SenseEmotion Database* respectively. In [Thiam, Bellmann, et al. 2019], a multi-modal deep neural network based on a hierarchical construct of modality specific one-dimensional convolutional neural networks and coupled to a weighted aggregation layer is proposed and assessed for the classification of several levels of heat induced nociceptive pain based on bio-physiological signals. In [Thiam, Kestler, et al. 2020b], an end-to-end approach based on attention networks [Zhou et al. 2016] is proposed for the assessment of pain related facial expressions. The proposed approach relies on both temporal and spatial components of video signals for the extraction of specific spatio-temporal representations of the input data in the form of Optical Flow Images (OFIs) [Horn and Schunck 1981] and Motion Histogram Images (MHIs) [Ahad et al. 2012]. These representations are subsequently processed through a hybrid deep neural network involving two-dimensional convolutional neural networks coupled to channel specific attention-based Bidirectional Long Short-Term Memory (BiLSTM) [Hochreiter and Schmidhuber 1997] Recurrent Neural Networks (RNNs). The output of the whole architecture is subsequently computed based on a weighted aggregation of the output of each channel specific attention-based BiLSTM RNN. The performed assessments show in both works that enabling an inference model to autonomously generate not just relevant feature representations of specific input signals but also to optimize a suitable multi-modal aggregation architecture can lead to a significant improvement of the classification performance of the whole system.

This thesis is subsequently concluded with a summary of the main findings as well as a description of potential future works in Chapter 5. The full versions of the summarized works presented in the Chapters 2 to 4 are included in Chapter I in addition to a detailed description of the individual contributions, followed by a summary of major contributions as a list of publications in Chapter II.

Chapter 2

Multi-Modal Active Learning

In this chapter, a summary of the work in [Thiam, Meudt, Palm, et al. 2018] (see Chapter I.1) is provided, including a description of the proposed multi-modal active learning approach for audiovisual events detection. Moreover, a short description of the main findings and results is also provided.

2.1 Introduction and Motivation

Supervised learning approaches rely on a set of labeled samples $\{(x_i, y_i) : i = 1, \ldots, n\}$ (where $y_i \in \mathcal{Y}$ depicts the label or class membership of the sample $x_i \in \mathcal{X} \subset \mathbb{R}^m$), in order to optimize a classifier f_θ (where θ depicts the set of trainable parameters specific to the classifier), which maps each data sample x_i to its predefined class membership y_i as follows:

$$f_\theta : \mathcal{X} \rightarrow \mathcal{Y} \tag{2.1}$$

The main goal of the optimization process of the classifier f_θ is the improvement of its generalization ability, which depicts its capability to properly adapt to and successfully classify unseen samples (samples that do not belong to the set of training samples and were never seen during the optimization process of the classifier). Therefore, the performance of supervised learning approaches is contingent upon both the amount of available labeled samples and the quality of the performed labeling, since noisy labeled samples negatively affect the performance of a trained classifier. Recent advances in sensors have led to a substantial increase of the amount of unlabeled data, since nowadays almost every single ubiquitous device continuously gathers different categories of information related to a user's biometric and bio-physiological characteristics. This data can be used to optimize user-centered pattern recognition systems based on supervised learning approaches, provided that the data is systematically and accurately labeled.

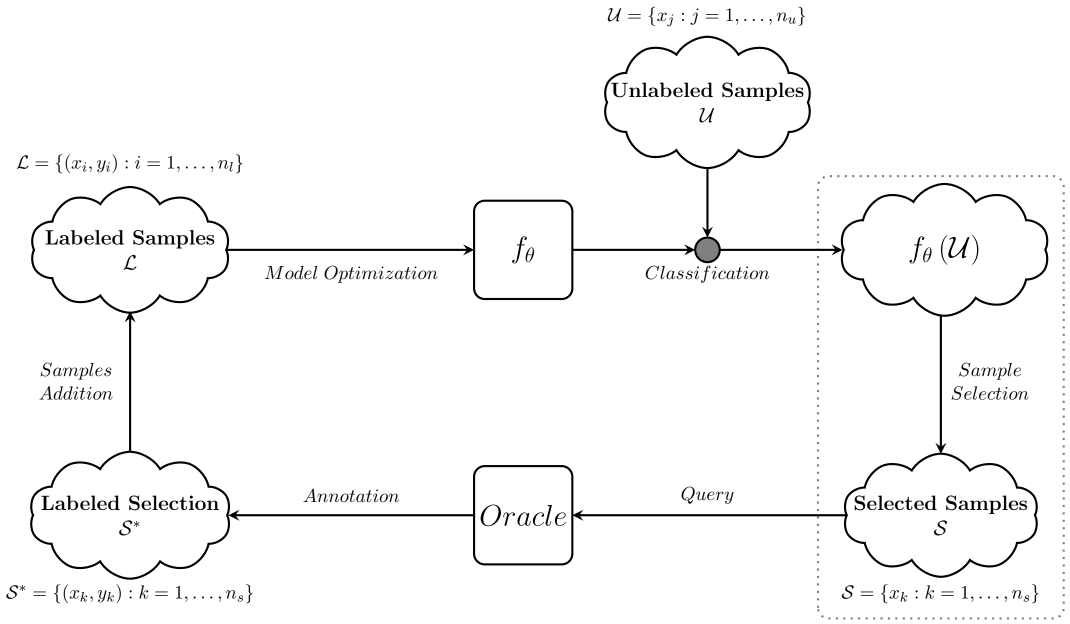

$\mathcal{U} = \{x_j : j = 1, \ldots, n_u\}$
Unlabeled Samples $\mathcal{U}$
$\mathcal{L} = \{(x_i, y_i) : i = 1, \ldots, n_l\}$
Labeled Samples $\mathcal{L}$
Model Optimization
f_θ
Classification
$f_\theta(\mathcal{U})$
Samples Addition
Sample Selection
Labeled Selection $\mathcal{S}^*$
Annotation
Oracle
Query
Selected Samples $\mathcal{S}$
$\mathcal{S}^* = \{(x_k, y_k) : k = 1, \ldots, n_s\}$
$\mathcal{S} = \{x_k : k = 1, \ldots, n_s\}$

Figure 2.1: *Pool-Based Active Learning. The dotted block specifies the area of interest of the current chapter (sample selection approaches).*

However, it is well known that data labeling is a very cumbersome, time consuming and cost expensive process. *Active learning* is a field of pattern recognition specifically defined to address the issue of optimizing an effective classifier under the constraint of scarceness of labeled samples.

Active Learning [Settles 2009] aims to substantially reduce the cost of manual annotation required for the optimization of a supervised learning classification model. Herefore, an iterative process is designed and applied (see Figure 2.1), consisting of the careful selection and annotation of the most informative samples from a large pool of unlabeled samples (pool-based active learning), based on predefined heuristics. The amount of training samples needed to optimize an effective classifier is thereby substantially reduced, without any loss of the generalization ability of the trained classifier. The iterative process usually begins with a relatively small set of labeled samples $\mathcal{L}$ and a comparatively large set of unlabeled samples $\mathcal{U}$ ($|\mathcal{L}| \ll |\mathcal{U}|$). A classifier f_θ is first initialized using the set of labeled samples. The trained classifier is subsequently applied on the pool of unlabeled samples. Based on a predefined heuristic (used to assess the informativeness of each unlabeled sample) and the resulting class distribution relative to the classifier f_θ, the most informative instances of the pool of unlabeled samples are selected and further labeled by an oracle, which is usually a human annotator. These manually labeled samples are subsequently removed from the pool of unlabeled samples, added into the pool of labeled samples, and used to actualize the classifier. This iterative process is repeated until a predefined termination criterion is reached.

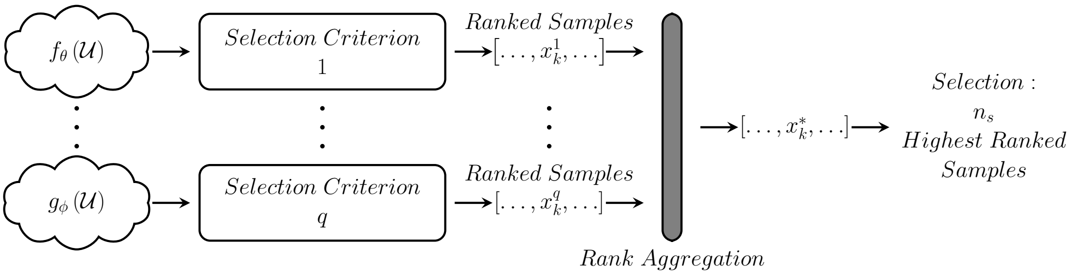


Figure 2.2: *Multiple Criteria Sample Selection (MCSS). Similarly to Multiple Classifier Systems, MCSS consists of aggregating the results of a set of heuristics based on different classifiers combined with different selection criteria in order to perform the selection of the most interesting samples.*

Through the years, several active learning approaches have been proposed and successfully applied in diverse domains such as content-based information retrieval [C. Zhang and Chen 2002; Gosselin and Cord 2008], anomaly and event detection [Pelleg and Moore 2004; Thiam, Meudt, Kächele, et al. 2014; Thiam, Kächele, et al. 2015] and also emotion recognition [Zhao and Ma 2013; Y. Zhang et al. 2015]. Most of the proposed approaches focus on the design and optimization of effective sample selection heuristics in order to significantly reduce the amount of labeled data needed to train an effective classification model. However, these approaches consist of optimizing a single selection heuristic for the underlying pattern recognition task. In Chapter 2.2 a *Multiple Criteria Sample Selection* (MCSS) approach is proposed and consists of the aggregation of the results of different selection heuristics in order to select the most informative samples. Moreover, potentially useful information can be extracted from correlated modalities (when such modalities are available) and subsequently used to further reduce the amount of samples needed for the optimization of an effective inference model. However, most of previous works focus uniquely on a single modality. A multi-modal active learning approach is presented in Chapter 2.2, for the optimization of a model specifically designed for the detection of audiovisual events.

2.2 Multi-Modal Active Learning for Audiovisual Events Detection

In [Thiam, Meudt, Palm, et al. 2018], a Multiple Criteria Sample Selection (MCSS) approach is proposed (see Figure 2.2). Inspired by Multiple Classifier Systems (MCS), a MCSS approach leverages the strength of multiple heuristics through the combination of complementary sample selection criteria, instead of relying on a single heuristic in order to estimate the informativeness of each

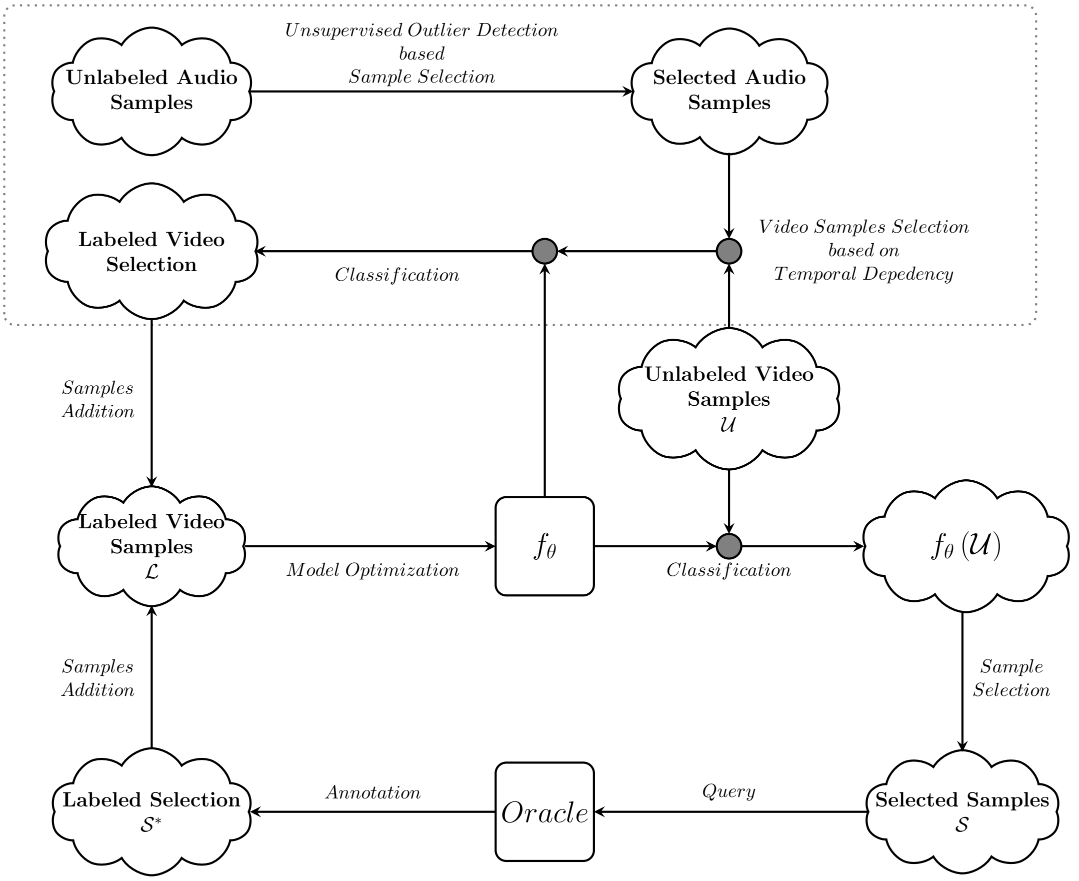

Unlabeled Audio Samples
Unsupervised Outlier Detection based Sample Selection
Selected Audio Samples
Labeled Video Selection
Classification
Video Samples Selection based on Temporal Depedency
Samples Addition
Unlabeled Video Samples $\mathcal{U}$
Labeled Video Samples $\mathcal{L}$
Model Optimization
f_θ
Classification
$f_\theta(\mathcal{U})$
Samples Addition
Sample Selection
Labeled Selection $\mathcal{S}^*$
Annotation
Oracle
Query
Selected Samples $\mathcal{S}$

Figure 2.3: *Bimodal Active Learning. The dotted block corresponds to the semi-supervised learning approach, that is combined with the proposed active learning approach in order to improve the efficiency of the training process of the classifier.*

unlabeled sample. In order to perform the aggregation, a ranking of the unlabeled samples relative to their informativeness is generated based on each involved heuristic (using the informativeness score specific to each heuristic). Subsequently, rank aggregation is performed on the resulting set of rankings and the first $n_s \in \mathbb{N}_{>0}$ samples with the highest aggregated rankings are selected for manual annotation.

Furthermore, the proposed MCSS approach is extended to a multi-modal active learning approach. The idea is to use the ambiguous temporal correlation of the manifestation of a specific event within several modalities (e.g. audiovisual events), in order to improve the efficiency of the training process of a modality specific classification model. For example, laughter is characterized by expressive facial motions in video sequences which can be accompanied with perceptible vocal utterances in audio sequences. However, the temporal order of occurrence of the manifestations of laughter in each modality is not predictable and unaligned, in most cases. The facial motions can occur before the laughter vocalization

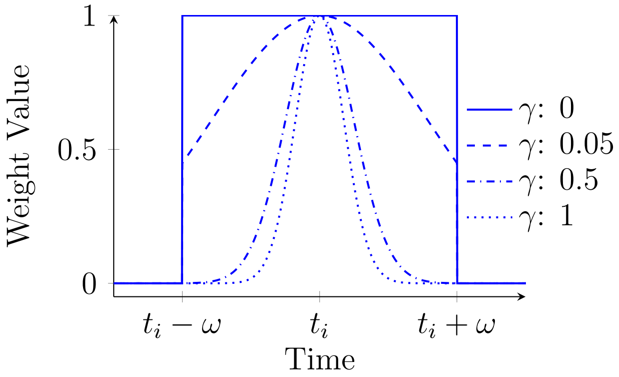


Figure 2.4: *Weight Function. The parameter $\omega \in \mathbb{R}_{\geq 0}$ represents the length of the temporal window, and γ represents the width of the Gaussian kernel. The function penalizes samples of the target pool (video samples) that are located far away from the selected samples of the auxiliary pool (audio samples). Thus, the approach focuses on instances of the target pool that are temporally near to interesting samples in the auxiliary pool. Reprinted by permission from Springer Nature: Springer, Neural Processing Letters, Thiam et al., A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection, © 2018.*

or both phenomena can occur simultaneously. Therefore, in order to efficiently optimize the training process of an event detection model for video sequences, complementary information in the form of co-occurring audio events is included into the active learning approach, as depicted in Figure 2.3. This is achieved by combining the proposed active learning approach applied on video sequences, with a semi-supervised learning approach [Chapelle et al. 2006] consisting of using the temporal dependency between audio and video manifestations of specific events to further select and automatically annotate an additional set of video samples. As usual, a supervised learning classifier f_θ is initialized based on a small set of labeled video data $\mathcal{L}$. The classifier is subsequently applied on the pool of unlabeled video samples $\mathcal{U}$ and based on the defined heuristics, a set of samples is selected, manually annotated and added to the pool of labeled samples. Concurrently, an unsupervised outlier detection approach (e.g. Support Vector Data Description [Tax and Duin 2004]) is applied on the set of audio samples in order to select interesting samples. This is motivated by the fact that samples that substantially deviate from normal observations are susceptible to be lying in regions of low density. These samples are subsequently used to select further video samples from the pool of unlabeled samples $\mathcal{U}$ based on a defined temporal window centered around the time-stamp corresponding to the audio sample, in combination with a Gaussian weighting technique depicted in Figure 2.4. The selected video samples are further automatically labeled by the trained classification model f_θ, and the resulting labeled samples are added to the pool of labeled video samples. The process is repeated iteratively until a predefined stopping

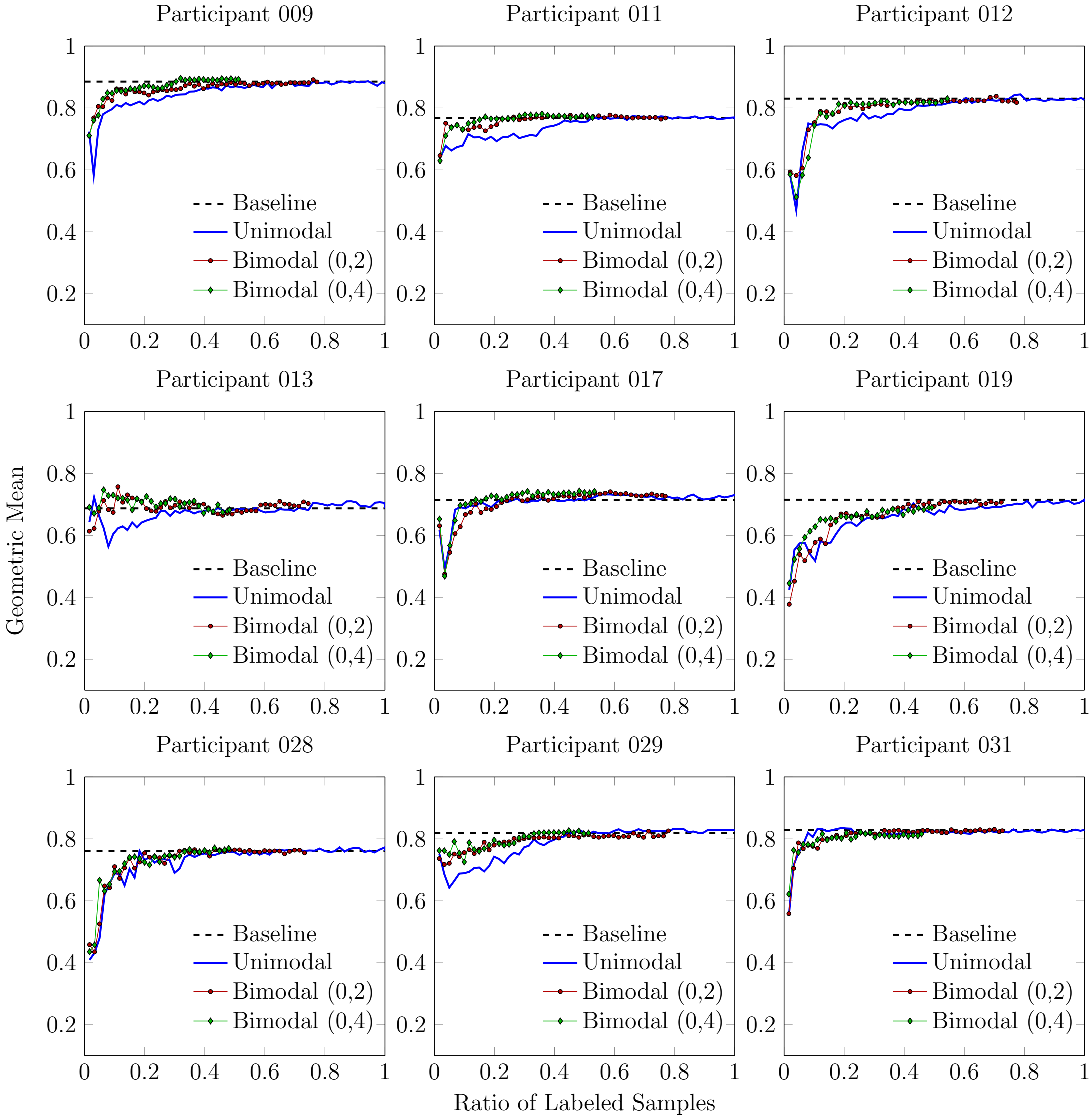


Figure 2.5: *Results based on an approach consisting of a combination of both Support Vector Data Description (SVDD) and Support Vector Machine (SVM) algorithms.* $\gamma = 0$ *and* $\omega \in \{2, 4\}$*. Reprinted by permission from Springer Nature: Springer, Neural Processing Letters, Thiam et al., A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection, © 2018.*

criterion is reached.

Both approaches (MCSS active learning and its multi-modal extension) are assessed on a subset of the *Ulm University Multimodal Affective Corpus* (UUlmMAC) [Hazer-Rau et al. 2020], which consists of several participants taking part to a gamified experimental setup simulating everyday life Human-Computer Interactions. The participants were asked to play a series of games with several levels of difficulties ranging from boring to overwhelming. The demeanor of each participant was synchronously captured using several sensors, including cameras and microphones. The designed active learning approaches were applied in order to efficiently optimize an effective inference model designed for the detection of specific occurrences (events) that substantially deviate from neutral (resp. normal) occurrences, without having to go through the cost expensive process consisting of labeling the entire dataset. The results depicted in Figure 2.5 show the relevance of the proposed approaches, since a classifier trained with a little less than 50% of the entire dataset is able to achieve the same performance as one trained on the entire dataset. Moreover, the multi-modal counterpart of the proposed MCSS active learning approach performs best and outperforms its uni-modal counterpart in almost all settings, since in most cases, a classifier trained on less than 30% of the entire dataset is able to achieve the same performance as one trained on the entire dataset. This shows that integrating complementary information stemming from correlated modalities positively impacts the performance of the designed active learning approach.

Chapter 3

Multiple Classifier Systems and Fusion Mechanisms

In this chapter, a summary of the work in [Thiam, Kessler, et al. 2019] (see Chapter I.2) is provided, including a description of the proposed Multiple Classifier System (MCS) with the corresponding fusion mechanisms. Furthermore, a short description of the main findings and results is also provided.

3.1 Introduction and Motivation

In the last decades, the consensus in the pattern recognition community has been that it is rather ill advised to rely on a single classifier for complex pattern recognition tasks, since such an approach entails several drawbacks which negatively impact both the performance and the robustness of a pattern recognition system [Kittler 2000]. First of all, designing a single classifier that effectively and efficiently exploits the multiple facets and diverse characteristics of a specific classification task is a rather complex and cumbersome endeavor. Moreover, supplementary and potentially complementary information provided by a set of diverse features and classifiers which could substantially reduce the complexity of the underlying pattern recognition task, is also completely ignored and unused. Since a pattern recognition system should benefit from an appropriate combination of complementary information stemming from a heterogeneous ensemble of classifiers, a Multiple Classifier System (MCS) aims at designing a suitable ensemble of classifiers as well as a corresponding and appropriate combination (resp. fusion) approach of the classifiers' outputs, in order to improve the overall performance as well as the robustness of a classification system. Several approaches have therefore been proposed for the optimal design and optimization of a MCS [Kuncheva 2004; Roli 2009; Bellmann et al. 2018].

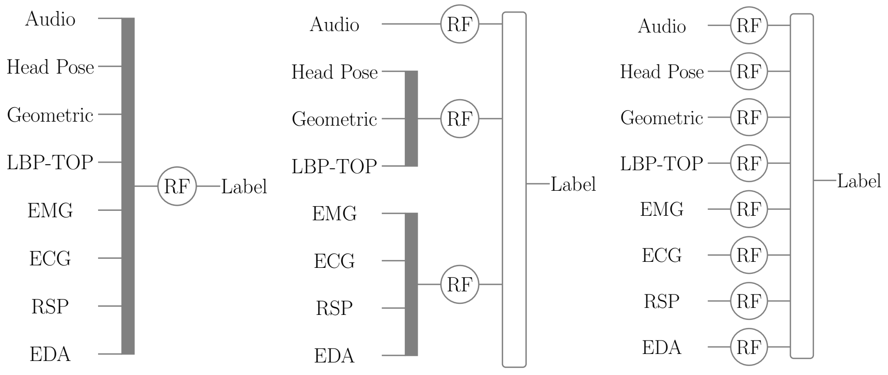
The heterogeneity or diversity of the ensemble of classifiers is necessary in order for the MCS to outperform a system based on a single classifier. Some of the most popular approaches to ensure the diversity of the ensemble are bootstrap aggregation (bagging) [Breiman 1996], boosting [Freund and Schapire 1996] and random subspace modeling [Ho 1998]. Additionally, it is also possible to generate a diverse ensemble by training different classifiers on different sets of features extracted from distinct modalities (e.g. audio, video, bio-physiology). Hence, the performance of a MCS can be further improved through the combination of both mechanisms, which literally consists of applying bagging or boosting approaches on modality specific features and designing a suitable fusion approach for the aggregation of the classifiers' outputs (Multi-modal Classifier Fusion). Such methods have been successfully applied to several pattern recognition tasks and in most cases significantly outperform pattern recognition systems based on a single classifier [Kächele, Thiam, Palm, et al. 2015; Thiam and Schwenker 2017; Bellmann et al. 2019]. Additionally to the diversity of the ensemble, an adequate fusion approach has to be designed in order to successfully combine the information stemming from the different classifiers. Depending on the level of abstraction at which the information from the different classifiers is aggregated, three main categories of information fusion approaches can be distinguished: *early fusion*, *hybrid (mid-level) fusion*, and *late fusion.*

Early fusion approaches consist of concatenating the features extracted from each of the available modalities into one single high dimensional feature vector, which is subsequently used to train a single classifier. The advantages of such an approach are its simplicity and the potential reduction of the complexity of the underlying classification task resulting from the combination of complementary features. However, the major drawback of early fusion approaches is the probability of running into the so called *curse of dimensionality* [Bishop 2006], which negatively affects the overall classification performance. *Hybrid (mid-level) fusion* approaches are characterized by a hierarchical (resp. layered) structure. In each layer, a set of classifiers is trained on different feature sub-spaces stemming from the aggregation of the output of the preceding layer into different groups, based on predefined heuristics. The output of each layer is fed into the next one, where the same procedure takes place. Finally, the last layer uses a specific aggregation rule in order to compute the final output of the classification architecture. *Late fusion* approaches consist of training a diverse set of classifiers based on each modality specific set of features and subsequently using an aggregation rule to combine the outputs of the trained classifiers. The aggregation rules can be categorized into fixed aggregation rules (e.g. majority voting, product rule, averaging rule) [Kuncheva 2002] and trainable aggregation rules (e.g. linear discriminant analysis, decision templates, Pseudo-inverse) [Schwenker, Dietrich, et al. 2006]. Fixed rules are simple, straightforward and are characterized by the non-existence of trainable parameters. Trainable rules on the other hand, are characterized by a set of trainable parameters to be optimized in order to perform the aggregation

of the classifiers' outputs. This inquires that a sufficient amount of labeled data is available for the optimization of the base classifiers, as well as the optimization of the trainable parameters specific to the aggregation layer.
Hence, designing a MCS for a specific pattern recognition task constitutes a manual and an iterative process, where crucial decisions have to be taken at each single step of the design process. First of all, relevant features have to be extracted from each single modality. This is done based on some expert knowledge related to the nature of each specific modality. Next, a specific set of classifiers has to be defined in order to perform the classification experiments at each level of abstraction. Furthermore, a specific fusion approach has to be defined. This also involves defining an aggregation rule in order to perform the fusion of the classifiers' outputs at the level of the aggregation layer. Finally, the whole architecture has to be evaluated depending on the nature of the underlying pattern recognition task. The work presented in Chapter 3.2 depicts exactly such a process, with the underlying task being the recognition of different intensities of pain elicitation, starting from the description of the involved modalities until the evaluation of the designed fusion architecture.

3.2 Multi-Modal Pain Intensity Recognition

In [Thiam, Kessler, et al. 2019] (see Chapter I.2), several experiments are undertaken for the development and assessment of a MCS for the recognition of different intensities of pain elicitation based on the *SenseEmotion Database* [Velana et al. 2017]. Pain is a very challenging phenomenon to assess due to its inherent subjective nature. Therefore, instead of relying uniquely on self-reporting tools, an additional reliable and automatic multi-modal pain assessment system should substantially improve the effectiveness of pain management. Hence, in order to investigate the suitability of a MCS for the detection and assessment of nociceptive pain, a dataset consisting of several healthy individuals submitted to a series of gradually increasing levels of painful thermal stimuli is recorded. Several modalities including audio signals, video signals, electrocardiography (ECG), electromyography (EMG), respiration signal (RSP), electrodermal activity (EDA), are simultaneously recorded during the conducted experiments, before being individually assessed and subsequently combined to identify the underlying set of predefined and gradually increasing intensities of pain elicitation (T_0, T_1, T_2, and T_3). Different sets of features are designed and extracted from each involved modality, before being subsequently assessed on this specific pattern recognition task. Furthermore, several MCS architectures are designed and evaluated, using different fixed and trainable aggregation rules in order to perform the combination of the recorded modalities (see Figure 3.1).
The performed uni-modal assessment shows that the EDA is the best perform-



***(a)** Early Fusion.* ***(b)** Hybrid Fusion.* ***(c)** Late Fusion.*

***Figure 3.1:** MCS Fusion Architectures. For both hybrid and late fusion architectures (resp. Fig.3.1b and Fig.3.1c), 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 (Random Forest (RF)) to generate the final label of an unseen sample. © 2019 IEEE. Reprinted, with permission, from Thiam et al., Multi-Modal Pain Intensity Recognition Based on the SenseEmotion Database, IEEE Transactions On Affective Computing, January/2019.*

ing single modality for this specific experimental setting (heat induced painful stimuli in a controlled environment). EDA significantly outperforms all other modalities for the vast majority of the performed experiments. Moreover, the assessment of the classification architectures shows that given a sufficient amount of training data, late fusion architectures in combination with trainable aggregation rules significantly outperform the other forms of MCS (early and hybrid fusion). The resulting best performing architecture consists of a late fusion approach, with the Pseudo-inverse (Pinv) aggregation rule [Schwenker, Dietrich, et al. 2006] (see Table 3.1). Given a matrix consisting of the horizontally concatenated outputs of the base classifiers $C = \left[C^1 : \ldots : C^k : \ldots : C^n\right] \in [0,1]^{N\times nc}$, where $C^k \in [0,1]^{N\times c}$ (N is the total number of training samples, c is the number of classes of the classification task, n is the number of base classifiers), a least-squares optimal linear mapping $M \in \mathbb{R}^{nc\times c}$ is generated by computing the Moore-Penrose Pseudo-inverse of C and multiplying it with the matrix of the corresponding class labels $Y \in [0,1]^{N\times c}$:

$$M = \lim_{\alpha\to\infty} C^{\intercal}\left(CC^{\intercal} + \alpha I\right)^{-1} Y \tag{3.1}$$

which is subsequently used to perform the aggregation of the base classifiers' outputs for unseen samples.

Moreover, the performed assessment also shows that the classification of lower levels of pain elicitation is very challenging. In this specific case, the aggregation of multiple modalities does not improve the overall performance of the classification system, since the modality (resp. channel) specific classifiers are performing at

Table 3.1: *Classification Results (Mean ± Standard Deviation(in%)). These results have been achieved by merging the data specific to each forearms of the SenseEmotion Database into a single set and performing a Leave One Subject Out (LOSO) 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 (late fusion with the Pseudo-inverse combination approach) and the corresponding best performing single modality. The test has been conducted using a Wilcoxon signed-rank test with a significance level of* 5%. *© 2019 IEEE. Based on Thiam et al., Multi-Modal Pain Intensity Recognition Based on the SenseEmotion Database, IEEE Transactions On Affective Computing, January/2019.*

Task	**T$_0$vs.T$_1$**	**T$_0$vs.T$_2$**	**T$_0$vs.T$_3$**	**T$_0$vs.T$_2$vs.T$_3$**	**T$_0$vs.T$_1$vs.T$_2$ vs.T$_3$**
Random	50.00	50.00	50.00	33.33	25.00
Audio	49.23 ± 4.37	50.19 ± 5.47	64.75 ± 14.27	42.80 ± 8.77	32.35 ± 6.87
Head Pose	51.73 ± 4.71	51.68 ± 5.10	63.05 ± 14.28	42.94 ± 9.48	32.06 ± 7.08
Geometric	**52.58 ± 4.00**	52.87 ± 4.49	66.22 ± 14.48	45.15 ± 10.10	34.22 ± 7.54
LBP-TOP	51.50 ± 4.34	51.73 ± 4.23	62.42 ± 12.18	41.78 ± 8.12	30.87 ± 5.99
EMG	49.97 ± 5.48	50.50 ± 4.99	59.33 ± 10.18	39.39 ± 6.43	29.73 ± 5.30
ECG	50.39 ± 3.58	51.69 ± 5.16	66.28 ± 12.59	44.42 ± 8.41	33.58 ± 6.85
RSP	50.21 ± 4.75	52.04 ± 5.61	67.27 ± 11.17	45.18 ± 8.19	33.89 ± 5.90
EDA	52.14 ± 3.95	**62.96 ± 9.02**	82.23 ± 10.57	57.84 ± 10.51	42.92 ± 7.07
Late Fusion	51.39 ± 4.18	62.28 ± 8.98	**83.39 ± 10.23***	**59.53 ± 9.94***	**43.89 ± 7.61**

the random level. However, the fusion architecture significantly outperforms all modality specific classifiers for higher levels of pain elicitation as well as in the case of multi-class classification tasks, therefore proving the effectiveness of MCS and its relevance for multi-modal pain recognition tasks.

Chapter 4

Deep Multi-Modal Fusion Mechanisms

In this chapter, a summary of the deep multi-modal fusion mechanisms proposed in both [Thiam, Bellmann, et al. 2019] and [Thiam, Kestler, et al. 2020b] is provided, including a short description of each specific approach, with the corresponding main findings and results.

4.1 Introduction and Motivation

A classical supervised learning pipeline involves a set of subsequent and interrelated steps. The very first step consists of the *data pre-processing*, and involves various techniques applied directly on the raw input signal in order to reduce the computational requirements related to the optimization of an inference model. Hence, the data pre-processing step aims at significantly reducing the amount of noise within the input signal, as well as detecting and extracting regions of interest (ROI), which are specific areas of the input signal that are relevant for the task at hand (e.g. facial area for the assessment of facial expressions). Data pre-processing is followed by *feature engineering*. It aims at extracting information from the pre-processed input signal that is most relevant with regards to the underlying inference task. It relies on some expert knowledge in the area of application that is used to manually design a set of measurable descriptors, which is subsequently extracted from the pre-processed signal in the form of feature vectors. The performance as well as the robustness of traditional inference models heavily rely on the discriminative and representative ability of the designed features.

Traditionally, feature engineering is followed by *feature selection*, which aims at improving both the robustness and the efficiency of an inference model by selecting a subset of the most relevant and informative features, while removing irrelevant and redundant ones. Therefore, the goal of feature selection is two-fold and involves a substantial reduction of the dimensionality of the feature space, and at the same time, an improvement of the performance of the corresponding inference model. Lastly, based on the set of selected features in combination with the set of corresponding labels, the *model optimization* phase can be carried out. During this phase, a specific inference architecture is designed, optimized and assessed. It involves different types of inference models such as Support Vector Machines (SVMs) [Abe 2010], Decision Trees [Breiman 1996; Breiman 2001] and Artificial Neural Networks (ANNs), among others.

Besides, when several channels (resp. modalities) are involved in the underlying classification task, the designed architecture should efficiently and effectively aggregate complementary information stemming from each of the channels in order to improve the overall performance and robustness of the classification system (see Chapter 3). Usually in such a case, pre-processing techniques as well as feature extraction and selection approaches, specific to each of the involved channels are separately applied. Subsequently, either a single model is optimized based on the descriptor resulting from the concatenation of the channel specific features (Early Fusion), or several channel specific models have to be optimized and subsequently integrated in order to perform the inference task (Multiple Classifier Systems (MCS)) [Kittler and Roli 2000; Kuncheva 2004].

Hence, classical supervised learning approaches are characterized by a set of interrelated sequential phases. The overall optimization of such systems is therefore an iterative process during which the output of each phase constitutes the unique information shared from one phase to another. The parameters characterizing each phase are either fixed or optimized based on the output of the preceding phase. Consequently, even though such approaches can eventually attain state-of-the-art inference performances, both the robustness as well as the generalization ability of the trained inference models are hampered and constrained due to three principal factors: the manual nature of the optimization pipeline, the reliance on an expert knowledge in the area of application for the design and extraction of relevant features and the isolation of each module characterizing the whole optimization pipeline. This becomes particularly challenging when dealing with a multi-modal classification task, since a modality-specific pre-processing pipeline has to be undertaken, subsequently followed by the optimization of a suitable fusion (resp. aggregation) mechanism involving the information stemming from all the involved modalities.

Meanwhile, several works have shown that this whole manual process can be effectively and efficiently replaced by deep learning approaches [LeCun et al. 2015], which have been outperforming classical supervised learning approaches in such domains as image processing [Szegedy et al. 2015], speech recognition [Hinton et al. 2012] or natural language processing [Costa-jussà 2018]. Deep learning approaches are characterized by a hierarchical construct of successive processing layers, which typically consist of simple and non-linear operations, that enable the whole architecture to learn very complex functions as well as suitable features directly from the pre-processed input signal. Therefore, the work presented in both Chapter 4.2 and Chapter 4.3, aims at enabling an inference model to autonomously generate not just relevant feature representations of specific input signals, but also to optimize a suitable multi-modal aggregation architecture in order to adequately perform the underlying classification task. In both cases, the pattern recognition task consists of the assessment of different levels of nociceptive pain elicitation based respectively on bio-physiological signals (Chapter 4.2) and video sequences (Chapter 4.3).

4.2 Deep Physiological Models for Pain Assessment

In [Thiam, Bellmann, et al. 2019] (see Chapter I.3), the reliance on expert knowledge for the design of competitive feature representations as well as the iterative and time consuming manual model optimization process characterizing classical supervised learning approaches are both avoided by the introduction of deep neural networks. Feature engineering and model optimization are therefore si-

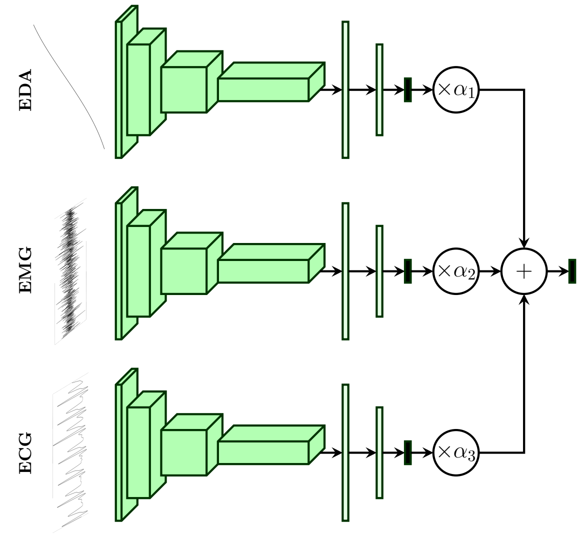


Figure 4.1: *Late Fusion Architecture. The final output is computed by an additional layer consisting of trainable parameters and a linear output function. Therefore, the final output consists of a weighted average of the modality-specific outputs. The whole architecture is trained in an end-to-end manner. Reproduced with permission from [Thiam, Bellmann, et al. 2019]; published by Sensors, 2019, licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).*

Table 4.1: *Fusion performance comparison to early work on the BVDB (Part A) in a Leave One Subject Out (LOSO) cross-validation setting for the classification task* $T_0 vs. T_4$. *The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold. Based on [Thiam, Bellmann, et al. 2019]; published by Sensors, 2019, licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).*

Approach	**Description**	**Performance**
[Werner, Al-Hamadi, Limbrecht-Ecklundt, et al. 2017]	Early Fusion with Random Forests (Head Pose and Facial Activity Descriptors)	72.40
[Werner, Al-Hamadi, Niese, et al. 2014]	Early Fusion with Random Forests (EDA, EMG, ECG, Video)	77.80
[Kächele, Werner, et al. 2015]	Early Fusion with Random Forests (EDA, ECG, Video)	78.90
[Kächele, Thiam, Amirian, et al. 2015]	Late Fusion with Random Forests and Pseudo-inverse (EDA, EMG, ECG, Video)	83.10
Our Approach	**Late Fusion with CNNs (EDA, EMG, ECG)**	$\mathbf{84.40 \pm 14.43}$

multaneously undertaken by a deep artificial neural network consisting of one-dimensional convolutional layers. Furthermore, an aggregation layer is proposed for the fusion of the modality-specific models' outputs (see Figure 4.1). The aggregation layer consists of a set of trainable weights ($\alpha_1, \alpha_2, \alpha_3$), and a linear activation function. Therefore, given the set of class probabilities of each modality-specific model i related to a specific sample j ($\{\theta_{i,j} \in [0,1]^c : 1 \leq i \leq 3\}$, where c is the number of classes of the underlying classification task), the final class probability output is computed as follows:

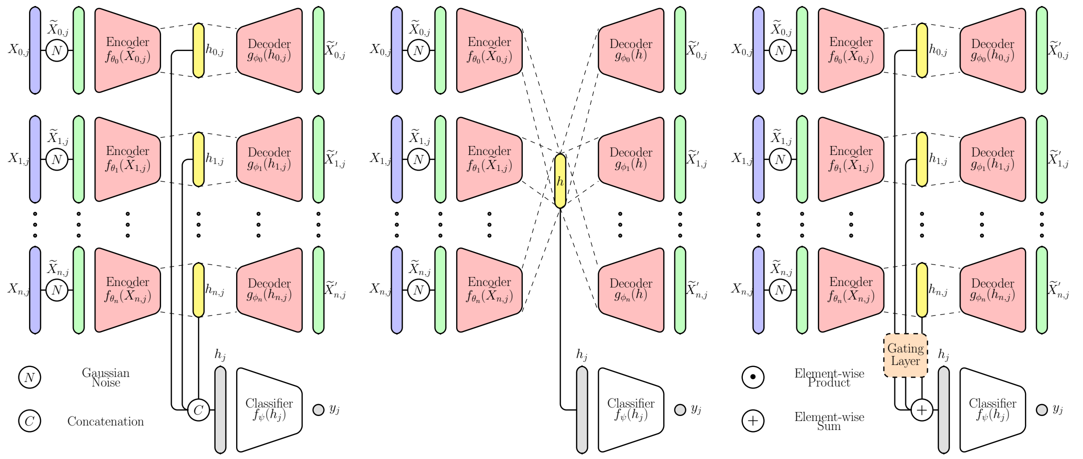
$$\theta_j = \frac{1}{3}\left(\sum_{i=1}^{3} \alpha_i \theta_{i,j}\right) \tag{4.1}$$

with the following constraints: $\forall i, \alpha_i \geq 0$ and $\sum_{i=1}^{3} \alpha_i = 1$. The whole architecture is trained in an end-to-end fashion, which means that the aggregation parameters are simultaneously optimized with the parameters of each modality-specific network. The proposed approach is evaluated on both parts (Part A and Part B) of the *BioVid Heat Pain Database* [Walter, Gruss, Ehleiter, et al. 2013], for the classification of different levels of nociceptive pain elicitation, based on bio-physiological signals (Electrocardiography (ECG), Electromyography (EMG), Electrodermal Activity (EDA)). The performed assessment shows that the designed deep architectures are able to attain new state-of-the-art classification performances in the case of the uni-modal approach based on the EDA, and also in the case of the fusion of all three modalities, while significantly outperforming previous approaches that rely on a set of carefully engineered manual features (see Table 4.1 and Table 4.2). Therefore, domain expert knowledge can potentially be bypassed by enabling a suitable artificial neural network to autonomously learn

Table 4.2: *Fusion performance comparison to early work on the BVDB (Part B) in a Leave One Subject Out (LOSO) cross-validation setting for the classification task $T_0 vs. T_4$. The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold. Based on [Thiam, Bellmann, et al. 2019]; published by Sensors, 2019, licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).*

Approach	**Description**	**Performance**
[Kächele, Werner, et al. 2015]	Late Fusion with SVMs and Mean Aggregation (EMG (zygomaticus), EMG (corrugator), EMG (trapezius), ECG, EDA, Video)	76.60
[Walter, Gruss, Limbrecht-Ecklundt, et al. 2014]	Early Fusion with SVM (EMG (zygomaticus), EMG (corrugator), EMG (trapezius), ECG, EDA)	77.05
Our Approach	**Late Fusion with CNNs (EMG (trapezius), ECG, EDA)**	$\mathbf{79.48 \pm 14.96}$

both relevant representations of the input data as well as the aggregation parameters needed to perform the fusion of several modality-specific models' outputs. Hence, motivated by these findings, more information fusion architectures were assessed as described in [Thiam, Kestler, et al. 2020a], where several architectures consisting of Deep Denoising Convolutional Auto-Encoders (DDCAEs) (see Figure 4.2) were designed and also assessed on the same dataset. The described architectures are characterized by the simultaneous optimization of both a single



(a) *Latent representation concatenation.* ***(b)*** *Shared latent representation.* ***(c)*** *Gated latent representation.*

Figure 4.2: *Fusion architectures based on Deep Denoising Convolutional Auto-Encoders (DDCAEs), trained simultaneously with an additional neural network performing the classification task. Reproduced with permission from [Thiam, Kestler, et al. 2020a], licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).*

representation of all involved modalities through the use of DDCAEs and a feed-forward neural network performing the classification of the different levels of pain elicitation. In this case, the performed assessment points at the fact that using a trainable gating layer for the generation of a weighted representation of the modality-specific latent representations significantly improves the performance of the whole architecture. The gated latent representation architecture (see Figure 4.2c) significantly outperforms previous works based on manually designed features.

Thus, the autonomous representation learning based on deep neural networks constitutes a sound alternative for manual feature engineering. Moreover, a significant performance improvement can be achieved by integrating feature learning, classifier design and classifier aggregation into a single deep neural network architecture.

4.3 Multi-Attention Network for Video Sequence Analysis

In [Thiam, Kestler, et al. 2020b] (see Chapter I.4), an end-to-end approach based on attention networks [Zhou et al. 2016] is proposed, for the analysis of pain related facial expressions in video sequences. The approach consists of the combination of both spatial and temporal aspects of facial expressions in video sequences at both representational (input signal) and structural (inference model architecture) levels of the classification architecture (an overview of the approach is depicted in Figure 4.3). For this purpose, video sequences are first encoded into suitable spatio-temporal representations using Motion Histogram Images (MHIs) [Ahad et al. 2012] and Optical Flow Images (OFIs) [Horn and Schunck 1981]. Given the j^{th} video sequence $\{f_{0,1}, \dots, f_{k,j}, \dots, f_{l,j}\}$, each single representation (OFI or MHI) is computed using a predefined and fixed reference frame ($f_{k^*,j}$) and

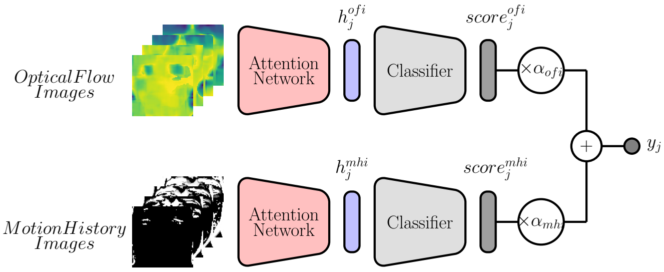


__Figure 4.3:__ Two-Stream Attention Network based on Optical Flow Images (OFIs) and Motion Histogram Images (MHIs) with Weighted Score Aggregation. Reproduced with permission from [Thiam, Kestler, et al. 2020b]; published by Sensors, 2020, licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).

the consecutive frames constituting the video sequence. In the current work, the very first frame of each video sequence is chosen as the reference frame ($k^* = 0$). Therefore, each generated spatio-temporal representation of the j^{th} video sequence consists of a total of l images $\{im_{1,j}, \ldots, im_{l,j}\}$, encoding both long and short term facial motions. These representations are subsequently fed into specific hybrid neural network architectures, each one consisting of a feature embedding Convolutional Neural Network (CNN) coupled to an attention-based Bidirectional Long Short-Term Memory (BiLSTM) [Hochreiter and Schmidhuber 1997] recurrent Neural Network (RNN) (see Figure 4.4).

The CNN generates a sequence of l feature embeddings $\{X_{1,j}, \ldots, X_{l,j}\}$ which are fed to the attention-based BiLSTM. The BiLSTM generates hidden representations $\{\overrightarrow{h_{k,j}}\}_{k=1}^{l}$, $\{\overleftarrow{h_{k,j}}\}_{k=1}^{l}$ by processing the input signal backwards and forwards in time. These representations are subsequently concatenated $\{h_{k,j}\}_{k=1}^{l} = \left\{ \left[\overrightarrow{h_{k,j}} : \overleftarrow{h_{k,j}}\right] \right\}_{k=1}^{l}$ and fed into an attention layer that generates a single weighted

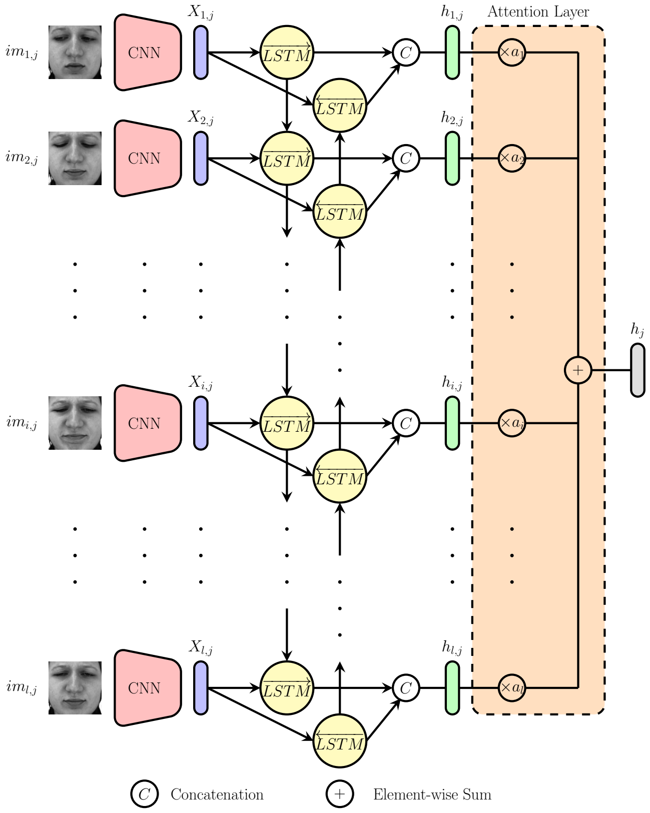


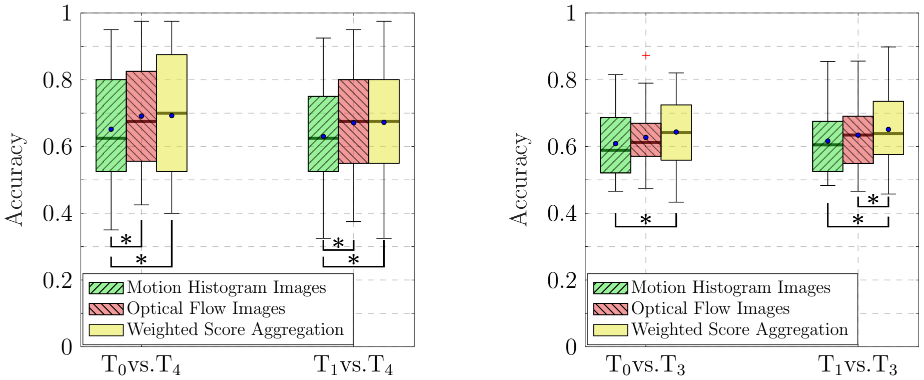
Figure 4.4: *Attention Network. Reproduced with permission from [Thiam, Kestler, et al. 2020b]; published by Sensors, 2020, licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).*

representation h_j as follows:

$$a_k = elu\left(W_k h_{k,j} + b_k\right) \tag{4.2}$$

$$h_j = \frac{\sum\limits_{k=1}^{l} \exp(a_k) h_{k,j}}{\sum\limits_{k=1}^{l} \exp(a_k)} \tag{4.3}$$

where $\{W_k, b_k\}_{k=1}^{l}$ represents the set of trainable parameters of the attention layer and *elu* refers to the *exponential linear unit* activation function [Clevert et al. 2016].
Next, the resulting spatio-temporal representations specific to both OFI sequences (h_j^{ofi}) and MHI sequences (h_j^{mhi}) are fed into specific classifiers (which are Multi-Layer Perceptrons (MLPs) in this case), before the resulting class probabilities are further aggregated using a weighting layer, similar to the one presented in Chapter 4.2. The assessment performed on the video recordings of both the *BioVid Heat Pain Database (Part A)* and the *SenseEmotion Database* [Velana et al. 2017] points at the relevance of the proposed architecture, as its performance is on par with state-of-the-art classification approaches (see Figure 4.5). The proposed approach also outperforms most of the proposed methods in the literature based on manually engineered features, thus pointing at the fact that the integration of the spatial and temporal dimensions of facial expressions in



(a) *BioVid Heat Pain Database.* ***(b)*** *SenseEmotion Database.*

Figure 4.5: *Classification performance (Accuracy). An asterisk (*) indicates a significant performance improvement. The test has been conducted using a Wilcoxon signed-rank test with a significance level of* 5%. *Within each box plot, the mean and the median classification accuracy are depicted respectively with a dot and a horizontal line. Reproduced with permission from [Thiam, Kestler, et al. 2020b]; published by Sensors, 2020, licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).*

both representational and structural levels of a deep architecture significantly improves its overall performance.

Chapter 5

Summary & Conclusion

The work presented in this thesis focused on the development of information fusion mechanisms for the optimization of effective inference models applied in diverse scenarios. The assessment of the proposed approaches was performed on different pattern recognition tasks ranging from the detection of audiovisual events, to the classification of several levels of heat induced painful stimuli. The performed assessment involved different datasets, each consisting of a set of healthy participants submitted to a series of audiovisual or thermal stimuli. The demeanor of the participants involved in the experiments leading to the creation of each dataset was recorded using a diverse set of modalities, ranging from audio and video signals, to bio-physiological signals (electrodermal activity, electrocardiography, electromyography, respiration signal). Accordingly to the underlying pattern recognition task, the works summarized in the Chapters 2, 3 and 4 have shown that a careful integration of a diverse set of information stemming from multiple sources can substantially improve the efficiency of the training process of an inference model, as well as its performance. Furthermore, the design of a suitable fusion architecture not only depends on the underlying pattern recognition task, but more importantly, on the amount, diversity and quality of the training material. These attributes offer more flexibility concerning the design of the fusion architecture which can be exploited to improve the efficiency of the training process as well as the overall performance of the optimized inference model.
In [Thiam, Meudt, Palm, et al. 2018] (summarized in Chapter 2), a uni-modal active learning approach based on a multiple criteria sample selection mechanism as well as its multi-modal counterpart have been proposed for the optimization of the training process of an audiovisual events detection model. The proposed multi-modal approach relies on the temporal correlation of events in both video and audio channels, for the selection and annotation of informative samples, that are subsequently used to continuously actualize the parameters of the inference model. Hence, an effective event detection model can therefore be optimized

on a significantly reduced set of manually annotated samples, without any loss of the generalization performance. The assessment of the proposed approaches performed on the UUlmMAC dataset [Hazer-Rau et al. 2020] has shown that a suitable model can be trained on a little less than 30% of an entire dataset, while still achieving the same performance as a model trained on the entire dataset. This proves that, depending on the underlying pattern recognition task, a huge amount of redundant, inconsistent and noisy data is comprised within the training material. This specific type of data causes the labeling process to be cumbersome, error-prone and temporally expensive. An adequate selection and annotation of a reduced set of samples suffices for the optimization of an effective inference model. Moreover, the proposed multi-modal approach significantly outperforms its uni-modal counterpart in most cases, thus pointing at the validity and relevance of the integration of diverse sources of information for the optimization of a specific inference model. Therefore, potential future works should consist in investigating further heuristics for an effective combination of sample selection approaches stemming from different modalities. Moreover, the integration of more channels into the whole process should also be investigated. In the case of bio-physiological channels, expert knowledge is needed for the localization and interpretation of specific events within the processed signals. A multi-modal active learning approach could be implemented in order to perform the detection of such events within bio-physiological signals based on complementary information stemming from synchronous video and audio modalities and be subsequently assessed by comparing the generated output to the results of the detection process performed manually by an expert. Such an approach would further improve the efficiency of the training process of bio-physiological inference models, since little to no expert knowledge for the annotation of the targeted events would be needed.

In [Thiam, Kessler, et al. 2019] (summarized in Chapter 3), a systematic analysis and assessment of several modalities including audio, video and bio-physiological signals is performed for the classification of several levels of heat induced pain elicitation. Additionally, a thorough assessment of several information fusion architectures (early, hybrid and late fusion) in combination with different fixed and trainable aggregation rules is conducted in order to design an effective multiple classifier system that significantly outperforms each uni-modal classification model. One of the main findings of the performed evaluation of the designed feature representations specific to each modality on the *SenseEmotion Database* [Velana et al. 2017] is the fact that the electrodermal activity (EDA) constitutes the best performing single modality. EDA significantly outperformed all the other modalities in almost the entirety of the performed experiments. However, these results could be biased due to the nature of the stimuli, since a strong correlation between the signals specific to the thermal stimuli and the EDA signals could be observed. Further evaluation of EDA in different experimental settings, involving different types of pain elicitation (e.g. pressure or cold) should be performed in order to validate the aforementioned finding.

This work is also one of the first to assess the audio channel as a potential suitable modality for the assessment of pain. Even though the experimental settings did not include any type of verbal interaction which resulted in audio signals comprising mostly of mourning sounds, the performed assessment proved the relevance of the audio modality, which generally performed better than the electromyography (EMG) (which, by the way, is the worst performing single modality). The assessment of the audio channel in a more realistic setting including verbal interactions should also be performed. Finally, the performed experiments also showed that late fusion architectures with trainable aggregation rules significantly outperformed the other forms of fusion mechanisms. The best performing fusion approach consists of a late fusion architecture with the Pseudo-inverse aggregation rule [Schwenker, Dietrich, et al. 2006]. This specific fusion approach was also able to significantly outperform the best performing single modality (EDA) classification model in almost all conducted experiments. Therefore, information fusion can substantially improve the performance of a pattern recognition system and the choice of a suitable architecture for the aggregation of a diverse set of information also relies on the size of the available training material. Given enough training samples, late fusion approaches with trainable aggregation rules should be preferable to other fusion mechanisms. Lastly, these results were acquired based on a dataset collected in a controlled environment. Therefore, valuable insights would be achieved by implementing and assessing the proposed architecture in a real world scenario.

Based on the findings in [Thiam, Kessler, et al. 2019], specific fusion approaches based on deep neural networks are proposed in both [Thiam, Bellmann, et al. 2019] and [Thiam, Kestler, et al. 2020b] (summarized in Chapter 4). In [Thiam, Bellmann, et al. 2019], a deep neural network based on one-dimensional convolutional layers is proposed for the simultaneous generation and aggregation of suitable high level bio-physiological representations in order to perform the classification of several levels of heat induced painful stimuli. The proposed late fusion architecture is characterized by the weighted aggregation of the outputs of a set of channel-specific deep neural architectures. The whole architecture is subsequently trained in an end-to-end manner. Therefore, the reliance on an expert knowledge in the domain of application for the extraction of suitable feature representations can be bypassed by enabling the network to autonomously engineer suitable representations directly from the pre-processed raw input signals. Furthermore, an end-to-end optimization including the modality-specific weighting parameters, enables the system to automatically identify the most relevant sources of information and adapt the weighting parameters accordingly. The proposed deep fusion approach achieved new state-of-the-art classification performances on the *BioVid Heat Pain Database* [Walter, Gruss, Ehleiter, et al. 2013], and also significantly outperformed related approaches based on handcrafted features. A further integration of recurrent neural networks in order to optimally use the temporal aspect of one-dimensional bio-physiological signals as well as the optimization of a temporal gating layer should be further investigated. In [Thiam, Kestler, et al. 2020b], a two-stream attention network for the analysis and assessment of pain related facial expressions in video sequences is proposed. The architecture also consists of a late fusion approach with trainable weighting parameters, which performs the aggregation of the outputs of two channel-specific attention networks, based respectively on sequences of Motion Histogram Images (MHIs) [Ahad et al. 2012] and Optical flow Images (OFIs) [Horn and Schunck 1981]. The proposed approach integrates both temporal and spatial aspects of facial motion in image sequences at both the representational level (by using OFIs and MHIs as input) and structural level (by using attention-based Bidirectional Long Short-Term Memory (BiLSTM) [Hochreiter and Schmidhuber 1997] Recurrent Neural Networks (RNNs)). Therefore, the network is able to identify the most interesting frames within each video sequence in relation with the corresponding level of pain elicitation based on the optimized attention weights. This information is further integrated into the resulting weighted representations generated by each channel-specific attention network and subsequently used to process with the classification of the video sequences. The outputs of the channel-specific networks are aggregated using a layer characterized by a set of trainable parameters and a linear output function. The whole architecture is trained in an end-to-end manner based on the findings in [Thiam, Bellmann, et al. 2019]. The assessment performed on both the *BioVid Heat Pain Database* and the *SenseEmotion Database* shows that the proposed approach is capable to attain state-of-the-art classification performances and is also on-par with the best performing approaches proposed in the literature. An optimization of the channel-specific attention-based neural networks should be further investigated in order to improve the performance of the whole classification system.

In summary, the work presented in this thesis provides the following contributions regarding each of the three research questions raised in Chapter 1:

Which information is relevant for the underlying pattern recognition task?

This question was addressed by introducing a novel uni-modal active learning approach based on a multiple criteria sample selection mechanism, as well as its multi-modal active learning counterpart in [Thiam, Meudt, Palm, et al. 2018]. The assessment of the proposed approaches has shown that an effective inference model can be efficiently optimized on a significantly small amount of carefully selected training samples. Furthermore, the sample selection and annotation process can be further optimized by exploiting relevant information extracted from auxiliary and temporally correlated input channels. Herewith, the annotation of a huge amount of irrelevant, noisy and redundant information can be effectively avoided and therefore, the cumbersome and temporally expensive labeling

process, needed for the optimization of supervised learning approaches, can be efficiently optimized.
Moreover, a thorough assessment of several modalities regarding the underlying pattern recognition task consisting of the classification of pain intensities was conducted in [Thiam, Kessler, et al. 2019]. Several sets of feature representations based on spatial, spectral and temporal domains were subsequently designed and extracted from each modality, and individually assessed. The results of the performed assessment indicate that the electrodermal activity (EDA) constitutes the most relevant and best performing modality for the underlying and specific experimental settings (involving thermal pain stimuli), followed by the video channel. Furthermore, the audio channel outperforms the electromyography (EMG) channel measured at the level of the trapezius muscles (which, by the way, is the worst performing modality), and therefore constitutes an additional and relevant modality for the classification of pain intensities. A similar overall performance could be observed for both respiration (RSP) and electrocardiography (ECG) channels. Finally, the integration of temporal information at both the structural and representational levels of a deep neural network in combination with a frame attention mechanism for the analysis of pain induced facial expressions in video sequences was proposed and assessed in [Thiam, Kestler, et al. 2020b]. The performed assessment of the proposed approach indicates that the attention mechanism enables the architecture to automatically detect the most relevant frames in a specific video sequence and assign specific weights accordingly, while the integration of the temporal aspect of facial motions significantly improves the overall performance of the pain intensity classification system.

When should the processed information be aggregated?
This question was addressed by proposing and assessing several information fusion architectures characterized by the aggregation of information stemming from multiple modalities at different levels of abstraction (early fusion, hybrid fusion and late fusion) for the classification of several levels of pain elicitation in [Thiam, Kessler, et al. 2019] based on Random Forest base classifiers, and in [Thiam, Bellmann, et al. 2019] based on deep neural networks. In both works, the performed assessment of the proposed architectures points at the fact that the choice of an effective information fusion architecture depends to a large extent on the size of the training data. Late fusion approaches perform consistently better and significantly outperform the other information fusion architectures given that an accordingly high amount of training material is available. The high amount of training material allows a better optimization of single representation inference models, and offers more flexibility concerning the combination of the resulting intermediate representations at a higher level of abstraction, while significantly improving the scalability of the whole architecture.

How should the processed information be aggregated?
Several information aggregation rules ranging from fixed to trainable aggregation mechanisms were designed and assessed in [Thiam, Kessler, et al. 2019] in combination with different information fusion architectures and in [Thiam, Bellmann, et al. 2019; Thiam, Kestler, et al. 2020b] in combination with a late fusion architecture, in order to address this question. The assessment performed in each work points to the fact that trainable aggregation mechanisms significantly outperform fixed aggregation mechanisms given that a sufficient amount of training material is available for an effective optimization of the base classifiers, as well as the aggregation parameters. Moreover, the Moore-Penrose Pseudo-inverse aggregation approach [Schwenker, Dietrich, et al. 2006] significantly outperformed the other forms of trainable and fixed aggregation rules proposed in [Thiam, Kessler, et al. 2019] for the classification of pain intensities, while using Random Forest models as base classifiers. Furthermore, the experiments conducted in both [Thiam, Bellmann, et al. 2019] and [Thiam, Kestler, et al. 2020b] have shown, that enabling an architecture to simultaneously optimize the modality-specific representations as well as the aggregation parameters using a deep neural network, can significantly improve the performance of the classification system. In each work, an end-to-end multi-modal fusion architecture based on several modality-specific deep neural networks combined with a weighting aggregation layer is proposed and assessed. In both cases, the weighted aggregation layer is able to consistently outperform each system based on a specific single modality, by automatically identifying the most relevant input channel and correspondingly assigning specific aggregation weights. Herewith, new state-of-the-art pain intensity classification performances could be achieved on publicly available as well as custom datasets.

In conclusion, this thesis has introduced and thoroughly assessed several new information fusion mechanisms for the optimization of inference models specific to different pattern recognition tasks. The performed experimental evaluation have proven the effectiveness and relevance of the proposed approaches. However, most of the performed experiments were conducted in controlled environments. The application of the proposed approaches in real world scenarios is more than desirable and also relevant, in particular nowadays, when ubiquitous devices provide an unlimited source of diverse data, which can be used to implement and perform different pattern recognition tasks. When performed wisely, this technology can substantially improve human-computer interactions as well as open the door to potentially new fields of application.

Bibliography

Abe, Shigeo (2010). *Support Vector Machines for Pattern Classification.* Springer-Verlag London. DOI: 10.1007/978-1-84996-098-4.

Ahad, Md. Atiqur Rahman, J. K. Tan, H. Kim, and S. Ishikawa (2012). “Motion History Image: its variants and applications”. In: *Machine Vision and Applications* 23, pp. 255–281. DOI: 10.1007/s00138-010-0298-4.

Bellmann, Peter, Patrick Thiam, and Friedhelm Schwenker (2018). “Multi-Classifier-Systems: Architectures, Algorithms and Applications”. In: *Computational Intelligence for Pattern Recognition.* Ed. by Witold Pedrycz and Shyi-Ming Chen. Vol. 777. Cham: Springer International Publishing, pp. 83–113. DOI: 10.1007/978-3-319-89629-8_4.

Bellmann, Peter, Patrick Thiam, and Friedhelm Schwenker (2019). “Using a Quartile-based Data Transformation for Pain Intensity Classification based on the SenseEmotion Database”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 310–316. DOI: 10.1109/ACIIW.2019.8925244.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning.* Springer-Verlag New York.

Breiman, Leo (1996). “Bagging Predictors”. In: *Machine Learning* 24(2), pp. 123–140. DOI: 10.1023/A:1018054314350.

Breiman, Leo (2001). “Random Forests”. In: *Machine Learning* 45(1), pp. 5–32. DOI: 10.1023/A:1010933404324.

Chapelle, Olivier, Bernhard Schölkopf, and Alexander Zien (2006). *Semi-Supervised Learning.* 1. The MIT Press.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter (2016). “Fast and Accurate Deep Neural Network Learning by Exponential Linear Units (ELUs)”. In: *Proceedings of the 4th International Conference on Learning Representations.* Ed. by Yoshua Bengio and Yann LeCun.

Costa-jussà, Marta R. (2018). “From Feature to Paradigm: Deep Learning in Machine Translation”. In: *Journal of Artificial Intelligence Research* 61(1), pp. 947–974. DOI: 10.1613/jair.1.11198.

Dautov, Rustem, Salvatore Distefano, and Rajkumaar Buyya (2019). “Hierarchical Data Fusion for Smart Healthcare”. In: *Journal of Big Data* 6(19). DOI: 10.1186/s40537-019-0183-6.

Freund, Yoav and Robert E. Schapire (1996). “Experiments with a New Boosting Algorithm”. In: *Proceedings of the Thirteenth International Conference on machine Learning*, pp. 148–156.

Gosselin, Philippe-Henri and Matthieu Cord (2008). “Active Learning Methods for Interactive Image Retrieval”. In: *IEEE Transactions on Image Processing* 17(7), pp. 1200–1211. DOI: 10.1109/TIP.2008.924286.

Gupta, Rishabh, Mojtaba Khomami Abadi, Jesús Alejandro Cárdenes Cabré, Fabio Morreale, Tiago H. Falk, and Nicu Sebe (2016). “A Quality Adaptive Multimodal Affect Recognition System for User-Centric Multimedia Indexing”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. New York, NY, USA: Association for Computing Machinery, pp. 317–320. DOI: 10.1145/2911996.2912059.

Hazer-Rau, Dilana, Sascha Meudt, Andreas Daucher, Jennifer Spohrs, et al. (2020). “The uulmMAC Database - A Multimodal Affective Corpus for Affective Computing in Human - Computer Interaction”. In: *Sensors* 20(2308). DOI: 10.3390/s20082308.

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, et al. (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared View of Four Research Groups”. In: *IEEE Signal Processing Magazine* 29(6), pp. 82–97. DOI: 10.1109/MSP.2012.2205597.

Ho, Tin Kam (1998). “The Random Subspace Method for Constructing Decision Forests”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8), pp. 832–844. DOI: 10.1109/34.709601.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long Short-Term Memory”. In: *Neural Computation* 9(8), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Horn, Berthold K. P. and Brian G. Schunck (1981). “Determining Optical Flow”. In: *Artificial Intelligence* 17(1), pp. 185–203. DOI: 10.1016/0004-3702(81)90024-2.

Kächele, Markus, Patrick Thiam, Mohammadreza Amirian, Philipp Werner, et al. (2015). “Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity”. In: *Engineering Applications of Neural Networks, EANN 2015*. Ed. by Lazaros Iliadis and Chrisina Jayne. Vol. 517. Cham: Springer International Publishing, pp. 275–285. DOI: 10.1007/978-3-319-23983-5_26.

Kächele, Markus, Patrick Thiam, Günther Palm, Friedhelm Schwenker, and Martin Schels (2015). “Ensemble Methods for Continuous Affect Recognition: Multi-Modality, Temporality, and Challenges”. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA:

Association for Computing Machinery, pp. 9–16. DOI: 10.1145/2808196.2811637.

Kächele, Markus, Philipp Werner, Steffen Walter, Ayoub Al-Hamadi, and Friedhelm Schwenker (2015). "Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity". In: *Multiple Classifier Systems (MCS).* Ed. by Friedhelm Schwenker, Fabio Roli, and Josef Kittler. Cham: Springer International Publishing, pp. 220–230. DOI: 10.1007/978-3-319-20248-8_19.

Kächele, Markus, Dimitrij Zharkov, Sascha Meudt, and Friedhelm Schwenker (2014). "Prosodic, Spectral and Voice Quality Feature Selection Using a Long-Term Stopping Criterion for Audio-Based Emotion Recognition". In: *2014 22nd International Conference on Pattern Recognition*, pp. 803–808. DOI: 10.1109/ICPR.2014.148.

Khalid, Samina, Tehmina Khalil, and Shamila Nasreen (2014). "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning". In: *2014 Science and Information Conference*, pp. 372–378. DOI: 10.1109/SAI.2014.6918213.

Kittler, Josef (2000). "A Framework for Classifier Fusion: Is It Still Needed ?" In: *Advances in Pattern Recognition.* Ed. by Francesc J. Ferri, José M. Iñesta, Adnan Amin, and Pavel Pudil. Berlin Heidelberg: Springer Berlin Heidelberg, pp. 45–56. DOI: 10.1007/3-540-44522-6_5.

Kittler, Josef and Fabio Roli, eds. (2000). *Multiple Classifier Systems.* Vol. 1857. Springer-Verlag Berlin Heidelberg. DOI: 10.1007/3-540-45014-9.

Kuncheva, Ludmila I. (2002). "A Theoretical Study on Six Classifier Fusion Strategies". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(2), pp. 281–286. DOI: 10.1109/34.982906.

Kuncheva, Ludmila I. (2004). *Combining Pattern Classifiers: Methods and Algorithms.* John Wiley and Sons, Inc.

LeCun, Yann, Yosgua Bengio, and Geoffrey Hinton (2015). "Deep Learning". In: *Nature* 521, pp. 436–444. DOI: 10.1038/nature14539.

Pelleg, Dan and Andrew Moore (2004). "Active Learning for Anomaly and Rare-Category Detection". In: *Proceedings of the 17th International Conference on Neural Information Processing Systems.* Cambridge, MA, USA: MIT Press, pp. 1073–1080.

Picard, Rosalind W. (1997). *Affective Computing.* Cambridge, MA, USA: MIT Press.

Poh, Norman and Josef Kittler (2010). "Chapter 8 - Multimodal Information Fusion". In: *Multimodal Signal Processing.* Ed. by Jean-Philippe Thiran, Ferran Marqués, and Hervé Bourlard. Oxford: Academic Press, pp. 153–169. DOI: 10.1016/B978-0-12-374825-6.00017-4.

Ramachandram, Dhanesh and Graham W. Taylor (2017). "Deep Multimodal Learning". In: *IEEE Signal Processing Magazine* 34(6), pp. 96–108. DOI: 10.1109/MSP.2017.2738401.

Rinaldi, Antonio M. and Cristiano Russo (2018). “User-centered Information Retrieval using Semantic Multimedia Big Data”. In: *2018 IEEE International Conference on Big Data (Big Data)*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 2304–2313. DOI: 10.1109/BigData.2018.8622613.

Roli, Fabio (2009). “Multiple Classifier Systems”. In: *Encyclopedia of Biometrics*. Ed. by Stan Z. Li and Anil Jain. Boston, MA: Springer US, pp. 981–986. DOI: 10.1007/978-0-387-73003-5_148.

Schels, Martin, Michael Glodek, Sascha Meudt, Stefan Scherer, et al. (2013). “Multi-modal Classifier-Fusion for the Recognition of Emotions”. In: *Coverbal Synchrony in Human-Machine Interaction*. CRC Press, pp. 73–97.

Schwenker, Friedhelm, Ronald Böck, Martin Schels, Sascha Meudt, et al. (2017). “Multimodal Affect Recognition in the Context of Human-Computer Interaction for Companion-Systems”. In: *Companion Technology: A Paradigm Shift in Human-Technology Interaction*. Ed. by Susanne Biundo and Andreas Wendemuth. Cham: Springer International Publishing, pp. 387–408. DOI: 10.1007/978-3-319-43665-4_19.

Schwenker, Friedhelm, Christian R. Dietrich, Christian Thiel, and Günther Palm (2006). “Learning of Decision Fusion Mappings for Pattern Recognition”. In: *International Journal on Artificial Intelligence and Machine Learning* 6, pp. 17–21.

Settles, Burr (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report. University of Wisconsin–Madison.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, et al. (2015). “Going Deeper with Convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

Tax, David M.J. and Robert P.W. Duin (2004). “Support Vector Data Description”. In: *Machine Learning* 54, pp. 45–66. DOI: 10.1023/B:MACH.0000008084.60811.49.

Thiam, Patrick, Peter Bellmann, Hans A. Kestler, and Friedhelm Schwenker (2019). “Exploring Deep Physiological Models for Nociceptive Pain Recognition”. In: *Sensors* 4503(20). DOI: 10.3390/s19204503.

Thiam, Patrick, Markus Kächele, Friedhelm Schwenker, and Günther Palm (2015). “Ensembles of Support Vector Data Description for Active Learning Based Annotation of Affective Corpora”. In: *2015 IEEE Symposium Series on Computational Intelligence*, pp. 1801–1807. DOI: 10.1109/SSCI.2015.251.

Thiam, Patrick, Viktor Kessler, Mohammadreza Amirian, Peter Bellmann, et al. (2019). “Multi-Modal Pain Intensity Recognition Based on the SenseEmotion Database”. In: *IEEE Transactions on Affective Computing*. © 2019 IEEE. DOI: 10.1109/TAFFC.2019.2892090.

Thiam, Patrick, Hans A. Kestler, and Friedhelm Schwenker (2020a). “Multimodal Deep Denoising Convolutional Autoencoders for Pain Intensity Classification based on Physiological Signals”. In: *Proceedings of the 9th International Con-*

ference on Pattern Recognition Applications and Methods (ICPRAM). Vol. 1. INSTICC. SciTePress, pp. 289–296. DOI: 10.5220/0008896102890296.

Thiam, Patrick, Hans A. Kestler, and Friedhelm Schwenker (2020b). “Two-Stream Attention Network for Pain Recognition from Video Sequences”. In: *Sensors* 20(839). DOI: 10.3390/s20030839.

Thiam, Patrick, Sascha Meudt, Markus Kächele, Günther Palm, and Friedhelm Schwenker (2014). “Detection of Emotional Events Utilizing Support Vector Methods in an Active Learning HCI Scenario”. In: *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems.* New York, NY, USA: Association for Computing Machinery, pp. 31–36. DOI: 10.1145/2668056.2668062.

Thiam, Patrick, Sascha Meudt, Günther Palm, and Friedhelm Schwenker (2018). “A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection”. In: *Neural Processing Letters* 48(2), pp. 709–732. DOI: 10.1007/s11063-017-9719-y.

Thiam, Patrick, Sascha Meudt, Friedhelm Schwenker, and Günther Palm (2016). “Active Learning for Speech Event Detection in HCI”. In: *Artificial Neural Networks in Pattern Recognition.* Ed. by Friedhelm Schwenker, Hazem M. Abbas, Neamat El Gayar, and Edmondo Trentin. Cham: Springer International Publishing, pp. 285–297. DOI: 10.1007/978-3-319-46182-3_24.

Thiam, Patrick and Friedhelm Schwenker (2017). “Multi-Modal Data Fusion for Pain Intensity Assessement and Classification”. In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6. DOI: 10.1109/IPTA.2017.8310115.

Velana, Maria, Sascha Gruss, Georg Layher, Patrick Thiam, et al. (2017). “The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Pain- and Emotion-Recognition System”. In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction.* Ed. by Friedhelm Schwenker and Stefan Scherer. Cham: Springer International Publishing, pp. 127–139. DOI: 10.1007/978-3-319-59259-6_11.

Walter, Steffen, Sascha Gruss, Hagen Ehleiter, Junwen Tan, et al. (2013). “The BioVid Heat Pain Database: Data for the Advancement and Systematic Validation of an Automated Pain Recognition System”. In: *2013 IEEE International Conference on Cybernetics*, pp. 128–131. DOI: 10.1109/CYBConf.2013.6617456.

Walter, Steffen, Sascha Gruss, Kerstin Limbrecht-Ecklundt, Harald C. Traue, et al. (2014). “Automatic Pain Quantification using Autonomic Parameters”. In: *Psychology and Neuroscience* 7(3), pp. 363–380. DOI: 10.3922/j.psns.2014.041.

Werner, Philipp, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C. Traue (2017). “Automatic Pain Assessment with Facial Activity Descriptors”. In: *IEEE Transactions on Affective Computing* 8(3), pp. 286–299. DOI: 10.1109/TAFFC.2016.2537327.

Werner, Philipp, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue (2014). "Automatic Pain Recognition from Video and Biomedical Signals". In: *2014 22nd International Conference on Pattern Recognition*, pp. 4582–4587. DOI: 10.1109/ICPR.2014.784.

Wu, Huadong, M. Siegel, and S. Ablay (2003). "Sensor Fusion using Dempster-Shafer Theory II: Static Weighting and Kalman Filter-like Dynamic Weighting ". In: *Proceedings of the 20th IEEE Instrumentation Technology Conference.* Vol. 2, pp. 907–912. DOI: 10.1109/IMTC.2003.1207885.

Zhang, Cha and Tsuhan Chen (2002). "An Active Learning Framework for Content Based Information Retrieval". In: *IEEE Transactions on Multimedia* 4(2), pp. 260–268. DOI: 10.1109/TMM.2002.1017738.

Zhang, Yue, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller (2015). "Dynamic Active Learning Based on Agreement and Applied to Emotion Recognition in Spoken Interactions". In: *Proceedings of the 2015 ACM on International Conference on Multimedia Interaction.* New York, NY, USA: Association for Computing Machinery, pp. 275–278. DOI: 10.1145/2818346.2820774.

Zhao, Ziping and Xirong Ma (2013). "Active Learning for Speech Emotion Recognition Using Conditional Random Fields". In: *2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 127–131. DOI: 10.1109/SNPD.2013.102.

Zhou, Peng, Wei Shi, Jun Tian, Zhenyu Qi, et al. (2016). "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.* Berlin, Germany: Association for Computational Linguistics, pp. 207–212. DOI: 10.18653/v1/P16-2034.

I

Articles of the Thesis

I.1 A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection

P. Thiam, S. Meudt, G. Palm, F. Schwenker. A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection. Neural Processing Letters 48(2), 709-732 (2018).
DOI: 10.1007/s11063-017-9719-y.
Status: Published.
Journal: Neural Processing Letters.
Impact Factor: 2.591 (2018).

The final publication is available at `https://link.springer.com/article/10.1007/s11063-017-9719-y`.

Noname manuscript No.
(will be inserted by the editor)

A Temporal Dependency based Multi-modal Active Learning Approach for Audiovisual Event Detection

Patrick Thiam · Sascha Meudt · Günther Palm · Friedhelm Schwenker

Received: date / Accepted: date

Abstract In this work, two novel active learning approaches for the annotation and detection of audiovisual events are proposed. The assumption behind the proposed approaches is that events are susceptible to substantively deviate from the distribution of normal observations and therefore should be lying in regions of low density. Thus, it is believed that an event detection model can be trained more efficiently by focusing on samples that appear to be inconsistent with the majority of the dataset. The first approach is an uni-modal method which consists in using rank aggregation to select informative samples which have previously been ranked using different unsupervised outlier detection techniques in combination with an uncertainty sampling technique. The information used for the sample selection stems from an unique modality (e.g. video channel). Since most active learning approaches focus on one target channel to perform the selection of informative samples and thus do not take advantage of potentially useful and complementary information among correlated modalities, we propose an extension of the previous uni-modal approach to multi-modality. From a target pool of instances belonging to a specific modality, the uni-modal approach is used to select and manually label a set of informative instances. Additionally, a second set of automatically labelled instances of the target pool is generated, based on a transfer of information stemming from an auxiliary modality which is temporally dependent to the target one. Both sets of labelled instances (automatically and manually labelled instances) are used for the semi-supervised training of a classification model to be used in the next active learning iteration. Both methods have been assessed on a set of

Patrick Thiam · Sascha Meudt · Günther Palm · Friedhelm Schwenker
Institute of Neural Information Processing, Ulm University, James-Franck-Ring 89081 Ulm Germany
E-mail: patrick.thiam@uni-ulm.de

Sascha Meudt
E-mail: sascha.meudt@uni-ulm.de

Günther Palm
E-mail: guenther.palm@uni-ulm.de

Friedhelm Schwenker
E-mail: friedhelm.schwenker@uni-ulm.de

participants selected from the UUlmMAC dataset and have proven to be effective in substantially reducing the cost of manual annotation required for the training of a facial event detection model. The assessment is done based on two different methods: Support Vector Data Description (SVDD) and Expected Similarity Estimation (EXPoSE). Furthermore, given an appropriate sampling approach, the multi-modal approach outperforms its uni-modal counterpart in most of the cases.

Keywords Active Learning · Unsupervised Outlier Detection · Support Vector Data Description · Expected Similarity Estimation

1 Introduction

In recent years, there has been a growing interest of the affective computing community towards the analysis of the affective state of users in different fields of application like the development of personal assistance systems, health care applications, robotics or entertainment industry. Potential applications of the analysis of an individual's affective state span from personal assistance to elderly care monitoring and improved human computer interaction (HCI). All these fields of application have in common that a classification system has to be trained on a preferably huge set of annotated training data. In the past years this data started to become multi-modal, containing at least audio and video information, often additional physiological data or infrared and depth information [1]. Furthermore, the content of the data moved from small datasets of acted emotions with condensed strong expressive emotions to more realistic everyday life situation data. As a consequence, an exponential growth of unannotated realistic datasets characterised by a scarceness of emotional content can be observed. This is a specific characteristic of human behaviour, as we do not express our emotions, or more general, our affective state all the time. So these datasets are much harder to classify and contain lots of useless and redundant data from a classifier's point of view [2,3].
Given the fact that nowadays, almost every single device is equipped with sensors that continuously gather and stream valuable information about the user's daily habits, physical shape and mental state, it is highly probable that this amount of data will continue growing. But before a classifier can be trained, the available data or at least a part of it, needs to be annotated first. This labelling is so far done mostly by human experts. Thus, the problem does not reside in the availability of data but instead in the labelling of such a huge amount of data. It is well known that data annotation is a very cumbersome and cost expensive process. Therefore, several machine learning methods spanning from semi-supervised learning [4] to active learning [5] have been proposed to substantially reduce the cost of manual annotation without any significant degradation of the performance of the trained classifier [6].
Moreover, the combination of multiple modalities can be advantageous, because emotions are expressed differently depending on the modality. Furthermore, some signs of emotion only appear in specific modalities [7,8]. For instance, the first signs of happiness are shown by smiling and laughter. This information would be lost if the laughter is not vocalized and the only available modality is the audio channel. On the other hand, a frustrated user may show his emotion by using some strong language. This would not be visible on a video recording. Therefore,

a multi-modal approach in affective computing is often preferred to analyse the overall emotional state of an individual. However multi-modal scenarios are difficult to handle, because each modality has specific characteristics (e.g. distinct feature extraction methods are needed for each modality). Thus, feature vectors are calculated from different points of time and cover different time intervals. In combination with emotion detection, a well designed strategy is needed to successfully master the task of multi-modal emotion detection.

In the following work, we first present a uni-modal pool-based active learning approach for audiovisual event detection. The approach consists in using rank aggregation to select informative samples which have previously been ranked using different outlier detection techniques in combination with an uncertainty sampling technique. Moreover, an extension of the uni-modal approach to multi-modality is proposed. The multi-modal approach uses complementary information from an auxiliary modality to select samples from a target modality, which are subsequently used together with manually annotated samples to train an event detection model in a semi-supervised manner. The approach exploits the temporal dependency between the modalities to select instances from the target modality that are located at defined temporal neighbourhoods of interesting samples of the auxiliary modality, which have previously been selected using unsupervised outlier detection techniques. The goal here is to further improve the efficiency of the training process of an event detection model on a specific target modality, by further reducing the cost of manual annotation through the exploitation of complementary information extracted from auxiliary modalities. The proposed approaches are assessed on a subset of the Ulm University Multimodal Affective Corpus (UUlmMAC) database.

The remainder of this work is organized as follows. In Section 2, an overview of several active learning approaches in a multitude of domain spanning from content-based information retrieval to event detection and emotion recognition in human computer interaction is provided. In Section 3 a thorough description of the proposed approaches is given. The dataset utilized to assess the proposed approaches including its annotation process is described in Section 4. In Section 5, a description of the experimental validation and results assessment is provided. Finally in Section 6 we conclude the present work and offer ideas for future work in the current direction.

2 Related Works

In recent years, research in active learning has expanded from content-based information retrieval [9–11], to anomaly and rare category detection [12–15] and to emotion recognition in Human Computer Interaction (HCI) [16–18] as well as physiological signal processing [19–21]. Content-based information retrieval has gained more interest lately due to its relevance in many multimedia applications as web search and audiovisual content management. Rare category detection consists in identifying instances from rare classes within unlabelled datasets. Görnitz et al. [22] proposed an active learning method based on Support Vector Data Description (SVDD) [23] for anomaly detection in network traffic. The proposed approach incorporates unlabelled and labelled data in a semi-supervised training of a decision boundary (ActiveSVDD). Subsequently, the instances that are to be manually annotated are selected using a combined strategy, consisting of querying those instances that are close to the decision boundary and that also lie in potentially anomalous clusters detected using a k-nearest neighbour graph. By doing so, the cost of annotation required for a proper calibration and validation of an intrusion detection model is substantially reduced.

Pelleg et al. [12] proposed a pool based active learning approach based on semi-supervised Gaussian Mixture Model (GMM) for anomaly detection in the case of extremely unbalanced datasets. The proposed querying strategy (Interleave method) consists of generating ranked lists of unlabelled instances by using the probability distribution function of each component of the mixture. These lists are subsequently merged into one final ranked list by cycling through each of them and picking the top instance that has not already been put in the final list. The picked instance is subsequently placed at the next position in the final list. Thereafter, the top n instances of the final list are selected for manual annotation. The approach was tested on synthetic as well as real datasets and proved to be effective in discovering rare classes for extremely unbalanced datasets while reducing the cost of manual annotation. In [13], He and Carbonell proposed two nearest-neighbour based rare category detection algorithms, one for the binary case (NNDB) and the other for the multiple classes case (NNDM) respectively, by performing local density differential sampling. Both methods rely on the prior(s) of the minority class(es). Based on the latter, an estimate of the number of instances belonging to the minority class(es) is computed. Thereafter, the latter is used in combination with a nearest neighbour clustering algorithm to measure the density around each unlabelled sample. The querying method consists of measuring the change in local density for each sample and selecting those with the maximum change. The approach was tested on both synthetic and real datasets and proved to be significantly better than the Interleave method proposed by Pelleg et al. [12].

Furthermore, He et al. in [24] proposed a graph-based rare category detection algorithm (GRADE) which is a generalised form of the NNDB algorithm proposed in [13]. The proposed approach also relies on the prior of the minority class and utilizes a global similarity matrix to detect compact clusters susceptible to correspond to the minority class. The querying strategy then consists of selecting instances from regions where the local density changes the most (regions where the probability of selecting samples of the minority class is high). In [14], the authors proposed a pool based active learning approach to jointly perform rare classes discovery and classification. The approach consists of adaptively selecting an appropriate query strategy online, amongst several query strategies based on their performance at discovering new classes and also their classification performance. Moreover, a combination of a generative model based on Gaussian mixture model and a discriminative model based on Support Vector Machine (SVM) [25] is proposed, since generative models have proven to be effective in rare category detection tasks, while discriminative models have proven to be effective in classification tasks.

Pichara and Soto [15] proposed a semi-supervised anomaly detection method that first uses a Bayesian network (BN) [26] to build an initial set of anomalous samples based on the generated probabilistic model. Thereafter, a subspace clustering method is used to identify relevant subspaces and micro clusters within the previously selected samples. Finally, a probabilistic active learning scheme (a hierarchical Bayesian generative model in this case) based on the earlier identified subspaces and micro clusters, is used to select and annotate the most relevant samples. How-

ever, the main focus of the proposed approach does not reside in improving the generalization performance of a classifier while significantly reducing the cost of annotation. Instead, the main goal is to quickly learn to identify relevant anomalous instances based on the feedback provided by the expert while minimizing the amount of query. Yan et al. [27] proposed a multi-class active learning approach for video data annotation, with query strategies based on the minimization of the expected generalization risk.
Concerning emotion recognition, Zhao and Ma in [16] proposed an active learning method consisting of applying Conditional Random Fields (CRF) [28] with a combination of uncertainty sampling and density measure for speech emotion recognition. Zhang et al. in [17] proposed the Dynamic Active Learning (DAL) for emotion recognition in spoken interactions in order to reduce the cost of human annotation by adaptively deciding for each instance if it should be labelled automatically by the trained model, or manually. The method also decides how many human annotators are required for an effective annotation of an instance. The method utilizes the medium certainty query strategy [29] to select the samples to be labelled. Subsequently, instead of using a majority voting model amongst all available annotators, an agreement level based selection scheme is used to select a subset of the annotators for the labelling of the selected samples. In [30], the authors use active learning for the detection and annotation of facial action units in webcam videos. The annotated facial action units are used subsequently for the classification of the facial expressions displayed in the videos. The proposed approach consists in using pre-trained action unit classifiers to provide probabilities of the presence of the action units in each frame of a video segment. Thereafter, video segments are chosen based on a specific threshold and ranked using the average values of the classifier outputs over the segment. The highest ranked segments are subsequently selected for manual annotation.
Regarding physiological signal processing, Xia et al. [18] applied simple margin active learning based on SVMs for skin conductance responses detection, as well as artefacts detection from electro-dermal activity (EDA) data. In [19, 20], the authors utilize active learning to improve the efficiency of the annotation process of patient-specific electrocardiogram (ECG) signals, for the discrimination between ventricular ectopic beats (VEBs) and non-VEBs, and also for the detection of ECG abnormalities. The querying strategy applied in this case is a combination of uncertainty sampling and hierarchical clustering. In [21], the authors apply active learning in order to improve the scalability of personalised long-term physiological monitoring in the context of epileptic seizure onset detection, based on the analysis of Electroencephalogram (EEG) signals.

3 Proposed Approach

We propose a pool-based active learning method which is a further iteration of the methods presented by Thiam et al. in [31, 32]. The method consists in combining outlier detection [33–35] with uncertainty sampling [5] to select informative samples that are labelled by an oracle and subsequently used to train a classification model. This approach has been previously applied in [36] and has proven to be effective in substantially reducing the cost of manual annotation required for the training of a good speech event detection model without any degradation of

performance in comparison to a fully supervised trained model. In this work, the approach is assessed on the task of facial event detection.
Furthermore, the uni-modal approach is extended to a multi-modal approach. The idea is to use complementary information from an auxiliary modality to perform the sample selection from a target modality and improve the efficiency as well as the performance of the generated classification model. Both target and auxiliary modalities depict some strong temporal dependency and are partially correlated. The partial correlation is defined as the frequency of events occurring at the same time in both modalities. The approach is assessed on the same task of facial event detection, where the target modality is the video channel and the auxiliary modality is the audio channel.
The assumption behind the proposed uni-modal approach is that events are susceptible to deviate substantively from normal observations and thus should be lying in regions of low density. Therefore, it is believed that an event detection model can be effectively trained by providing labels to the samples which appear to be inconsistent with the majority of the dataset. Furthermore, uncertainty sampling is also performed to select samples that would further refine the decision boundary between both classes (event class vs normal class). The approach relies on rankings of the unlabelled data samples relative to their informativeness. These rankings are generated using unsupervised outlier detection techniques and uncertainty sampling. Subsequently, rank aggregation is performed to combine the resulting rankings and the k samples with the highest aggregated rankings are selected for manual annotation. Two outlier detection techniques have been assessed in the present work.

3.1 Support Vector Data Description (SVDD)

Introduced by Tax and Duin [23], Support Vector Data Description (SVDD) is inspired by the Support Vector Classifier [37]. Given a set of observations $\{x_i\}_{i=1}^N = \mathcal{X} \subset \mathbb{R}^n$ with $x_i \in \mathbb{R}^n \; \forall i$, the SVDD generates a closed boundary around the data, characterised by a center $a \in \mathbb{R}^n$ and a radius $R > 0$. The optimization problem to be solved consists in minimizing the volume of the hypersphere with the constraint that all observations are located within the boundary. To allow the possibility of outliers in the training set, slack variables $\xi_i \geq 0$ are introduced to relax the strictness of the constraint, as well as a parameter $C \geq 0$ which controls the trade-off between the volume of the closed boundary and the amount of miss classifications. The optimization problem translates into:

$$\begin{aligned} &\min_{R,a,\xi_i} F(R,a) = R^2 + C\sum_i \xi_i \\ &\text{subject to } \langle x_i - a, x_i - a\rangle \leq R^2 + \xi_i,\; \xi_i \geq 0,\; i = 1,\dots,N \end{aligned} \tag{1}$$

The optimization problem is solved by quadratic programming and a new observation z is classified as an outlier if the distance from z to the center of the hypersphere a is greater than the radius of the hypersphere R: $\langle z - a, z - a\rangle = \|z - a\|^2 > R^2$. Similarly to the Support Vector Classifier, a more flexible hypersphere can be obtained by using the kernel trick and replacing the inner product in Equation 1 by an appropriate kernel function. In the present work, the Gaussian kernel function ($K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, $\gamma > 0$) is applied, since it produces a more flexibel decision boundary than a model based on either a linear or a polynomial kernel.

3.2 Expected Similarity Estimation (EXPoSE)

Proposed by Schneider et al. [38], Expected Similarity Estimation (EXPoSE) is an outlier detection method which utilizes a similarity function to compute a score ($\eta(z)$, z being an observation), which represents the likelihood of an observation belonging to the distribution of regular data $\mathcal{P}$. The lower the score ($\eta(z)$) the higher the likelihood that the corresponding observation is an outlier. Let $\mathcal{X} \subset \mathbb{R}^n$ be the input space. Given a kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which measures the similarity between observations of the input space $\mathcal{X}$, the expected similarity of an observation $z \in \mathcal{X}$ under the probability distribution $\mathcal{P}$ is defined as follows:

$$\eta(z) = \mathbb{E}\left[k(z, \cdot)\right] = \int\limits_{\mathcal{X}} k(z, x) d\mathcal{P}(x) \tag{2}$$

Given a finite set of observations $\{x_i\}_{i=1}^{N}$ drawn independently from $\mathcal{P}$, the expected similarity estimation can be approximated by the following scalar product:

$$\eta(z) = \left\langle \phi(z), \frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \right\rangle \tag{3}$$

where ϕ is an approximated feature map such that $k(z, x) \approx \langle \phi(z), \phi(x) \rangle$. In the present work, we use the Gaussian RBF kernel to measure the similarity between the observations and we apply the Nyström method [39, 40] for the computation of the approximated feature map ϕ.

3.3 Unsupervised Outlier Detection

Throughout the active learning iterations, outlier detection is performed in an unsupervised manner since the labels of the observations are unknown. The absence of labels hinders the optimization of the parameters for a single SVDD (respectively EXPoSE) model. Therefore, the unsupervised outlier detection is conducted with an ensemble of SVDD (respectively EXPoSE) models.
Based on the work of Chang et al. [41], $\mu \times \nu$ SVDD models are generated by choosing μ values for the trade-off parameter C, equally spaced within the interval $\left[\frac{1}{N}, 1\right]$ where N is the number of unlabelled instances, and ν values for the Gaussian RBF kernel parameter γ, equally spaced within the interval $\left[\frac{1}{f}, 1\right]$ where f is the dimensionality of the feature vector. In this way, the grid of possible values for both parameters is covered and the diversity in the ensemble is ensured. Furthermore a threshold is used to prune the generated ensemble, based on the reclassification results of each generated model. This threshold specifies the maximum ratio of observations that have to be classified as outliers. Models with outlier classification rates higher than the specified threshold are discarded.

Subsequently, the unlabelled samples are ranked in descending order of the voting count of the generated ensemble.
Concerning the EXPoSE method, the Gaussian RBF kernel parameter γ constitutes the unique parameter that has to be optimized. At each iteration, a random subset constituting 20% of the unlabelled instances is chosen for the approximation of the feature map ϕ. This value was selected emperically. The ensemble is generated by choosing several equally spaced γ values within the interval $(0, 1]$. Subsequently, the unlabelled samples are ranked in ascending order of the averaged scores of the generated EXPoSE models.

3.4 Pool-based Active Learning: Uni-modal Sample Selection

An overview of the proposed approach can be seen in Figure 1. The approach is applied to an artificial dataset. Unlabelled instances are represented by black dots. Labelled instances of the normal class are represented by filled blue dots, while labelled instances of the event class are represented by filled red diamonds. The samples selected using the outlier detection methods (see Figure 1(b)) and the uncertainty sampling method (see Figure 1(c)) are enclosed in yellow filled circles. The samples selected for manual annotation are enclosed within yellow filled diamonds (see Figure 1(d)). The decision boundary of the model trained after each iteration with the labelled instances is depicted with a full black line and the corresponding hyperplane is depicted in green.

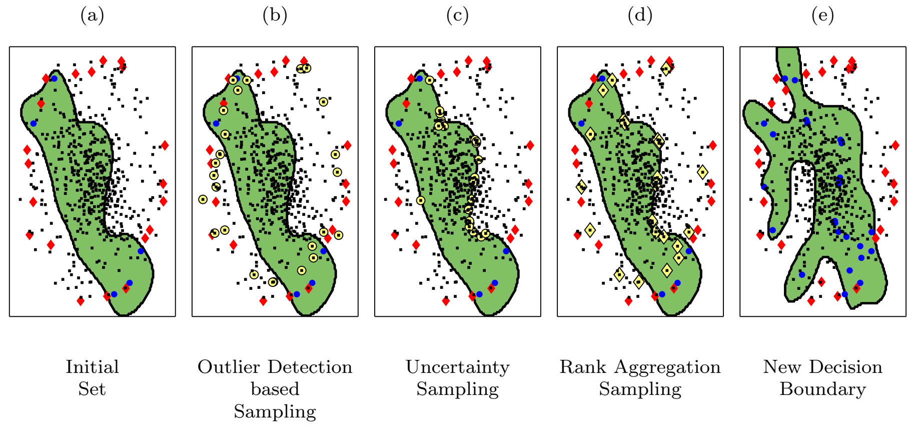


Fig. 1: **Proposed pool-based active learning approach.**

Initially, a set of samples is selected using an unsupervised outlier detection method (SVDD, EXPoSE) and subsequently manually labelled. From the labelled instances, a binary Support Vector Machine (SVM) model is trained (see Figure 1(a)). In the next iteration, besides the set of samples selected using the unsupervised outlier detection method (see Figure 1(b)), an additional set of observations is selected using the SVM model trained in the previous iteration by using an uncertainty sampling method. In this work, the applied uncertainty sampling method is based on the distance from the unlabelled instances to the decision boundary (see Figure 1(c)). Thereafter, both sets of selected instances are merged using Borda's geometric mean rank aggregation [42]. The k samples with the highest aggregated rankings are selected for manual annotation (see Figure 1(d)) before the supervised SVM model is updated (see Figure 1(e)). A detailed description of the proposed approach can be seen in Algorithm 1.

Algorithm 1 Pool-based Active Learning: Uni-modal Sample Selection

```
Require:
    $U = \{s_i | s_i \in \mathbb{R}^n \ \forall i\}$                ▷ Pool of unlabelled instances
    $k \in \mathbb{N}_{>0}$                                          ▷ Query size
$U_0 \leftarrow U$                                                   ▷ Unlabelled set initialisation
$L_0 \leftarrow \emptyset$                                           ▷ Labelled set initialisation
$M_0 \leftarrow \emptyset$                                           ▷ Classification model
$t \leftarrow 1$
while $U_{t-1} \neq \emptyset$ do
    Outlier Detection based Sampling on $U_{t-1}$: $S_{unsupervised}$
    if $M_{t-1} \neq \emptyset$ then
        Uncertainty Sampling by applying $M_{t-1}$ on $U_{t-1}$: $S_{supervised}$
        Rank Aggregation Sampling based on $S_{unsupervised}$ and $S_{supervised}$: $S_{final}$
    else
        $S_{final} \leftarrow S_{unsupervised}$
    end if
    Selection $S_{final}$ Annotation
    $L_t \leftarrow L_{t-1} \cup S_{final}$
    $U_t \leftarrow U_{t-1} \setminus S_{final}$
    $P \leftarrow \{(s_j, l_j) | s_j \in L_t \wedge l_j > 0\}$      ▷ Labelled samples belonging to the majority class
    $N \leftarrow \{(s_j, l_j) | s_j \in L_t \wedge l_j < 0\}$      ▷ Labelled samples belonging to the minority class
    if $P \neq \emptyset \wedge N \neq \emptyset$ then
        Train a supervised classification model using $L_t$: $M_t$
    end if
    $t \leftarrow t + 1$
end while
```

3.5 Pool-based Active Learning: Bimodal Sample Selection

In the previous section (see Section 3.4), the unlabelled instances belong to a single modality (e.g. audio or video). The information used to select the most informative instances is restricted to that specific modality. In the following section we describe an active learning method that takes advantage of the temporal dependency and the complementary information between two modalities, in combination with semi-supervised learning [4] to further reduce the cost of manual annotation and improve the performance of a model on a target modality.
It has to be noted that the proposed approach is completely different from multi view active learning [43–45]. In the case of multi view active learning, the different views constitute disjoint subsets of features, each of which is sufficient to learn an adequate decision boundary. Furthermore, multi view active learning approaches rely on the assumptions that each observation is identically labelled in each view

and the feature vectors of each observation in each view are independent.
The proposed approach is based on the assumption that the instances, as well as the corresponding labels in both modalities are independent and partially correlated (i.e. the temporal occurrences of the target events are partially correlated in both modalities). For instance, if the task at hand is laughter detection, an unvoiced laughter in a video sequence is labelled as an event due to the presence of facial activities. But in the audio channel, the corresponding temporal segment is labelled as normal given the absence of laughter vocalization. There will be a correlation in both modalities in the case of vocalized laughter. Still, the correlation will be partial because of the length of the vocalization in the audio segment compared to the length of the facial activity display.
Although the task at hand remains identical to the one described in Section 3.4, the initial setting is however different: additionaly to a specific target pool of observations (e.g. video), a secondary pool of observations (auxiliary pool), which is temporally linked to the target pool (e.g. audio), is also available. The goal is to exploit information from the auxiliary pool to improve the efficiency and the performance of a model generated and applied on the target pool by reducing the cost of annotation. This transfer of information is realised by exploiting the temporal dependencies between the observations in each pool in combination with semi-supervised learning. Outlier detection methods are applied on the auxiliary pool to select interesting samples. Instances from the target pool lying in a specific temporal neighbourhood of the previously chosen samples are selected and automatically labelled by the classification model of the previous active learning iteration. The samples automatically labelled by the trained model in combination with those which have been manually labelled are subsequently used to update the classification model.
In the following lines, the target pool is refered to as $A \subset \mathbb{R}^m$ and the auxiliary pool as $B \subset \mathbb{R}^n$. The features describing the observations ($s_j \in A$, $s_i \in B$) are specific to the modalities and independent in each pool. The timestamps of the occurrences of the observations in both pools are also provided ($t_j \in A$, $t_i \in B$). Therefore, the initial setting consists of two pools of unlabelled instances: one target pool $A = \{(t_j, s_j)\,|t_j \in \mathbb{R}_{\geq 0}, s_j \in \mathbb{R}^m, \forall j = 1, \ldots, p\}$ and one auxiliary pool $B = \{(t_i, s_i)\,|t_i \in \mathbb{R}_{\geq 0}, s_i \in \mathbb{R}^n, \forall i = 1, \ldots, q\}$. The labels of the observations are set separately in each modality. At each iteration, the method presented in Section 3.4 is applied on the target pool for the selection and annotation of instances. The classification model is thereafter actualised using the new set of labelled instances. Following, the same outlier detection method is applied on the auxiliary pool, but the selected samples are not annotated. Based on a fixed temporal window $\omega \in \mathbb{R}_{\geq 0}$ and for each selected sample i from the auxiliary pool, all samples from the target pool that are located within the temporal window $[t_i - \omega, t_i + \omega]$ are selected (t_i is the time stamp of the selected observation in the auxiliary pool) and weighted using the following Gaussian function:

$$\delta_{ij} = \exp(-\gamma\,|t_i - t_j|^2) \tag{4}$$

where t_j is the time stamp of an observation in the target pool that falls within the specified window. This specific weight is computed to penalise samples that are located far away from the observation in the auxiliary pool (see Figure 2). Thereafter, the selected samples from the auxiliary pool are discarded. As soon as the auxiliary pool is empty, the uni-modal approach is further applied to the

Algorithm 2 Bimodal Sample Selection

```
procedure BiModalSelection($A$, $B$, $L$, $M$, $\omega$, $\gamma$, $k$)
    $A = \{(t_j, s_j) \mid t_j \in \mathbb{R}_{\geq 0}\ \forall j,\ s_j \in \mathbb{R}^m\ \forall j\}$        ▷ Target pool
    $B = \{(t_i, s_i) \mid t_i \in \mathbb{R}_{\geq 0}\ \forall i,\ s_i \in \mathbb{R}^n\ \forall i\}$        ▷ Auxiliary pool
    $L$                                                     ▷ Labelled set
    $M$                                                     ▷ Classification model
    $\omega \in \mathbb{R}_{\geq 0}$                                          ▷ Temporal window
    $\gamma \in [0, 1]$                                     ▷ Gaussian function parameter
    $k \in \mathbb{N}_{>0}$                                          ▷ Query size
    Outlier Detection based Sampling on $B$: $S$
    $S_{auxiliary} \leftarrow \bigcup_{(t_i,s_i)\in S} \{(t_j, s_j, \delta_{ij}) \mid (t_j, s_j) \in A \land t_i - \omega \leq t_j \leq t_i + \omega\}$
    Apply $M$ on $S_{auxiliary}$ and calculate the weighted scores $W_{scores}$ based on $\delta_{ij}$
    Calculate the threshold $th_{bimodal}$ based on $W_{scores}$
    $S_{auxiliary} \leftarrow \{(t_j, s_j, l_j^*) \mid (t_j, s_j) \in S_{auxiliary} \land W_{scores}(s_j) \geq th_{bimodal}\}$   ▷ $l_j^*$ is the label generated by the model $M$
    $B \leftarrow B \setminus S$
    $A \leftarrow A \setminus S_{auxiliary}$
    $L \leftarrow L \cup S_{auxiliary}$
    Train a classification model using $L$: $M$
end procedure
```

target pool until a stopping criteria is attained (in the present work, until the target pool is empty). The final set of selected samples from the target pool using the time dependency between both pools can be expressed as:

$$S = \bigcup_{(t_i,s_i)} \{(t_j, s_j, \delta_{ij}) \mid (t_j, s_j) \in A \land t_i - w \leq t_j \leq t_i + w\} \tag{5}$$

where s_j is the selected observation in the target pool.

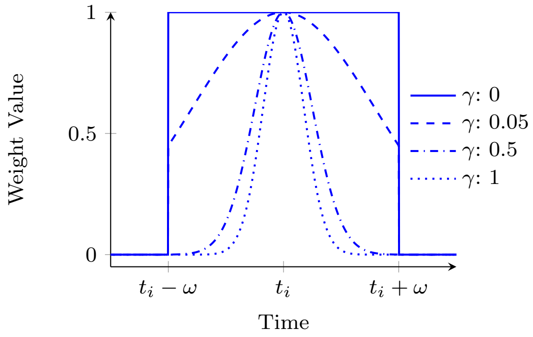


Fig. 2: **Weight Function.** The function penalises samples of the target pool that are located far away from the selected sample of the auxiliary pool. Thus, the approach focuses on instances of the target pool that are temporally near to interesting samples in the auxiliary pool.

Subsequently, the classification model is applied on the samples selected using the temporal dependency constraint (S). The classification confidence of each observation s_j is weighted using δ_{ij} and the observations with a weighted confidence above a specific threshold $th_{bimodal}$ are selected and added with the corresponding machine generated label to the labelled set. Those samples are also discarded from the target pool. Following that, the classification model is actualised using both the manually annotated and the machine annotated samples for the next iteration. A detailed description of the proposed approach can be seen in both Algorithms 2 and 3.

Algorithm 3 Pool-based Active Learning with Bimodal Sample Selection

```
Require:
    $A = \{(t_j, s_j) \mid t_j \in \mathbb{R}_{\geq 0}\ \forall j,\ s_j \in \mathbb{R}^m\ \forall j\}$    ▷ Target pool
    $B = \{(t_i, s_i) \mid t_i \in \mathbb{R}_{\geq 0}\ \forall i,\ s_i \in \mathbb{R}^n\ \forall i\}$    ▷ Auxiliary pool
    $\omega \in \mathbb{R}_{\geq 0}$    ▷ Temporal window
    $\gamma \in [0, 1]$    ▷ Gaussian function parameter
    $k \in \mathbb{N}_{>0}$    ▷ Query size
$A_0 \leftarrow A$    ▷ Unlabelled set initialisation (target pool)
$B_0 \leftarrow B$    ▷ Unlabelled set initialisation (auxiliary pool)
$L_0 \leftarrow \emptyset$    ▷ Labelled set initialisation
$M_0 \leftarrow \emptyset$    ▷ Classification model
$t \leftarrow 1$
while $A_{t-1} \neq \emptyset$ do
    Outlier Detection based Sampling on $A_{t-1}$: $S_{unsupervised}$
    if $M_{t-1} \neq \emptyset$ then
        Uncertainty Sampling by applying $M_{t-1}$ on $A_{t-1}$: $S_{supervised}$
        Rank Aggregation Sampling based on $S_{supervised}$ and $S_{unsupervised}$: $S_{target}$
        Selection $S_{target}$ Annotation
        $L_t \leftarrow L_{t-1} \cup S_{target}$
        $A_t \leftarrow A_{t-1} \setminus S_{target}$
        Train a classification model using $L_t$: $M_t$
        if $B_{t-1} \neq \emptyset$ then
            BiModalSelection($A_t$, $B_{t-1}$, $L_t$, $M_t$, $\omega$, $\gamma$, $k$)
        end if
    else
        $S_{target} \leftarrow S_{unsupervised}$
        Selection $S_{target}$ Annotation
        $L_t \leftarrow L_{t-1} \cup S_{target}$
        $A_t \leftarrow A_{t-1} \setminus S_{target}$
        $P \leftarrow \{(t_j, s_j, l_j) | s_j \in L_t \wedge l_j > 0\}$
        $N \leftarrow \{(t_j, s_j, l_j) | s_j \in L_t \wedge l_j < 0\}$
        if $P \neq \emptyset \wedge N \neq \emptyset$ then
            Train a classification model using $L_t$: $M_t$
        end if
    end if
    $t \leftarrow t + 1$
end while
```

4 Dataset Description

The dataset utilised in the present work is a subset of the Ulm University Multimodal Affective Corpus (UUlmMAC) which was recorded at Ulm university. 60 participants were recorded for this dataset in 100 recording sessions. Each session lasted for about 45 minutes. The participant's tasks were designed referring to

Schuessel et al. [46] who proposed a gamified experimental setup close to everyday life human computer interactions. Each recorded session consisted of a series of puzzle games during which the participant had to find the unique symbol (unique in shape and color) in a grid of different symbols (see Figure 3).

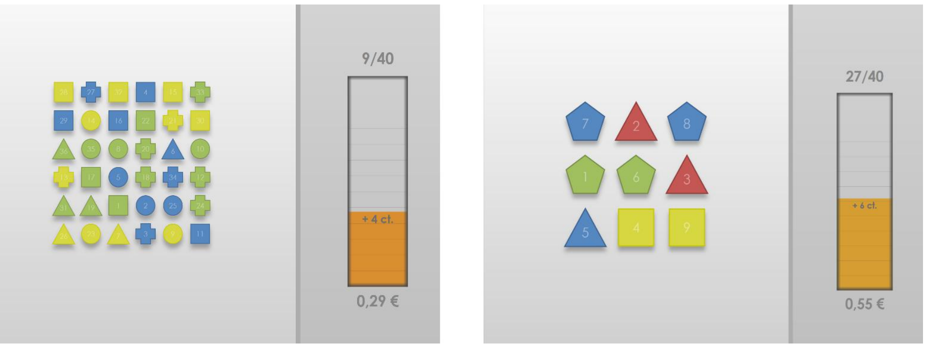


Fig. 3: **Screenshot of the difficult level (left) with target element 6 and the easyest level (right) with target element 5.**

For each correct answer, the participant was paid an additional amount of money. The already earned amount of money was displayed on the right side of the gaming field. The longer the participant took to provide an answer, the smaller the amount of money earned was. If the participant ran out of time, no extra reward was given at all.

A recording session consisted of a total of six sequences. Each sequence was composed of a multitude of puzzles and was designed to induce cognitive under- or overload by tuning the levels of difficulty of the presented puzzles. The levels of difficulty were tuned by changing the size of the grid of puzzle and by adjusting the answering time. The first sequence was an introduction to the gaming mechanics. The following sequences gradually decreased in difficulty. In the last sequence, the game logged in wrong answers on purpose to induce frustration. A complete overview of the sequence's settings can be seen in Table 1.

Table 1: **Gaming sequences of the recording session for UUlmMAC database.**

Sequence	**Gridsize**	**Time**	**Difficulty**	**Additional info**
0	3x3, 4x4	20 sec	easy	Introduction session, not used
1	6x6	6sec	overload	
2	4x4	10sec	high difficulty	
3	3x3	10sec	medium difficulty	
4	3x3	100sec	underload	
5	3x3	10sec	frustration	game logged in wrong answers

After each sequence the participants had to answer a survey designed to assess the level of valence, arousal and dominance in the latest sequence [47]. First, the participants had to describe in their own words how they felt during the game sequence. In the following questions, the participants had to choose the matching values on three scales to their experience (Self Assessment Manikin Scale (SAM)) [48]. Hihn et al. [49,50] show that the reported (V, A, D) values differ significantly between the mental overload and underload sequences. Valence is higher during mental underload sequences, while arousal is higher during mental overload sequences. Finally, dominance is higher during the mental underload sequences. This suggests that mental overload and underload can be expressed using the (V, A, D) space as overload: (V-, A+, D-) and underload: (V+, A-, D+).

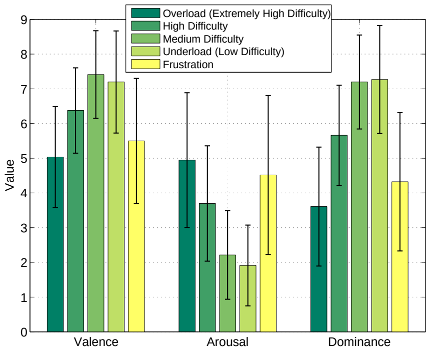


Fig. 4: **Mean VAD over all sequences.** It can be seen, that the reported VAD varies significantly between the sequences, especially between sequences 1 and 4.

Different modalities were recorded (see Figure 5): three video streams (frontal HD, frontal webcam, rear webcam), three audio lines (headset, ambiance and directional microphone), Microsoft Kinect 2 (depth, infrared, video, audio, posture) and biophysiology (EMG, ECG, SCL, respiration and temperature).
For this work, nine participants were selected out of the total 60 based on their grade of expressiveness. This subgroup consists of four male and five female participants with age ranging from 20 to 27 years. Table 2 shows the participants' information. This work uses video and audio data from the recorded sessions. Audio information was taken from the participant's headset. For video information, the frontal HD camera stream was used. A precedent analysis of the recorded data shows that all of the participants reacted emotionally after providing an incorrect answer. This observation is most noticeable in the overload sequences when the task is nearly impossible to solve in the given time. The present work focuses on the detection and discrimination of such emotional reactions like laugher, heavy breathing, idiomatic expressions as a signal of boredom or frustration, from non-emotional (or neutral) reactions like the participant giving an answer to the task. Emotional reactions are referred to as events and neutral reactions as normal. To assess the performance of the proposed approaches, the recorded data of the nine selected participants was manually annotated. Table 2 shows the duration of the manual annotation of the dataset specific to each participant for each modality.

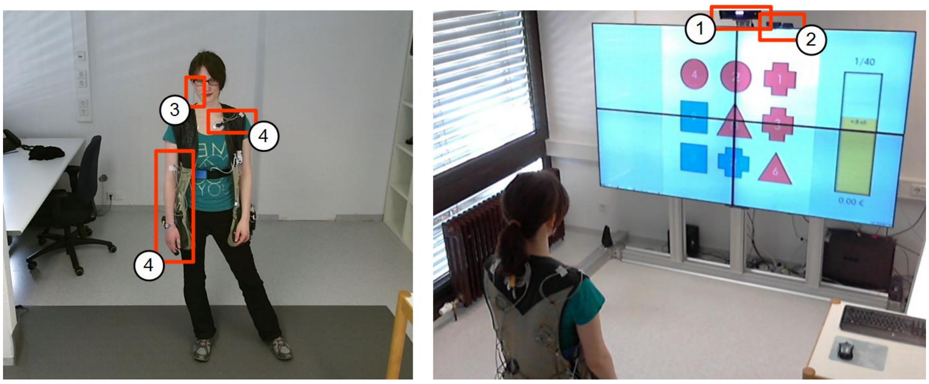


Fig. 5: **Overview of the experimental setting with sensors:** (1) MS Kinect v2, (2) frontal webcam, (3) wireless headset, (4) GTec g.MOBIlab+ physiologic sensor with sensors attached to the users body.

4.1 Data Annotation

Annotating the recorded data is a necessary step to create ground truth labels which will be used as the knowledge base of the oracle in active learning tasks. Audio and video recordings were annotated separately by two different persons. Audio annotation was preprocessed by an automatic segmentation of speech in voice active and voice inactive segments. This step is necessary since the task at hand relies uniquely on the voice active segments. Thus, the voice inactive segments are irrelevant. Because the recording sessions took place in a noise free environment, a simple unsupervised voice activity detection [51] based on the energy of the speech signal was used to distinguish between speech and silence. After preprocessing, each speech segment was manually annotated either as event or as normal utterance. Additionally the label's temporal boundaries were adjusted for a perfect segmentation and for further noise reduction. The annotation process was performed with ATLAS [52, 53] and took in average 90 minutes per participant (see Table 2).

Similar to audio annotation, the video track was annotated manually. No preprocessing was used, every label was set by hand. As in the case of the audio data, there are no segments that could create a third class to the two class problem of differentiating events from normal behaviour. So, only labels for events were set and all other video information was interpreted as normal behaviour. The annotator focused solely on the participant's face while annotating. Movements of the arms or legs were ignored. Like the audio annotation, the video annotation process was performed with ATLAS. The time needed to manually annotate one participant ranged from 4 to 7 hours. More precise information about the annotation times can be found in Table 2.

After the manual annotation, a first assessment of the dataset was undertaken. Tables 3 and 4 show the amount of labelled observations. For each modality, the *normal* observations clearly outnumber the *event* observations. Regarding the audio channel, there is an average of 1537 *normal* observations to 693 *event* ob-

Table 2: **Amount of time needed for the manual annotation of each participant for each modality.** The amount of time is specified in minutes. In average, the length of the audio recording for each participant is 2.7 minutes which results in a total annotation time of 95.6 minutes. Concerning the video recordings, the annotator needs to review 36.6 minutes which results in an annotation time of 320.6 minutes. The resulting large factors of 37.5 for audio and 9 for video clearly indicate the cost-intensity of the annotation process.

			audio			video		
ID	sex	age	rec. duration	ann. time	$\frac{ann.}{rec.}$	rec. duration	ann. time	$\frac{ann.}{rec.}$
09	f	21	2	85	42.5	36	340	9.4
11	f	27	5	150	30.0	31	315	10.2
12	f	20	3	105	35.0	32	405	12.7
13	m	21	2	70	35.0	35	330	9.4
17	f	22	3	∼90	30.0	32	275	9.2
19	m	20	2	∼90	45.0	32	285	6.3
28	m	25	3	∼90	30.0	33	245	8.2
29	f	23	2	∼90	45.0	34	325	7.2
31	m	23	2	∼90	45.0	35	365	8.1
avr:		22.4	2.7	95.6	37.5	36.6	320.6	9.0

servations per participant. Concerning the video channel, there is an average of 3284 *normal* observations and 401 *event* observations per participant. This imbalance is a very important characteristic of this dataset and is typical in realistic HCI scenarios. The amount of audio samples from one participant to the next is very similar, except for the participant 011, who was extremely expressive. The mentioned imbalance is a very typical characteristic of a dataset derived from a human computer interaction scenario. Usually, a person remains most of the time in a neutral state. For example, when using a companion technology [54], the user only interacts with the companion device if he has a question or receives a notification. But most of the time, the device waits for an event like a keyword to start interacting. Of course, the whole time the device's sensors need to be active to catch such a keyword.

4.2 Audio Feature Extraction

Following the annotation of the audio signal, a fixed window of 215 milliseconds was selected for the segmentation of the voice active samples. The window was shifted with a fixed offset of 65 milliseconds. The resulting data distribution for each participant can be seen in Table 3.

The following set of features was subsequently extracted from the resulting segments with fixed frames of 25 milliseconds, sampled at a rate of 10 milliseconds: 8 linear predictive coding coefficients (LPC) [55]; 5 perceptual linear prediction cepstral coefficients (PLP-CC) [56], each with delta and acceleration coefficients; 12 Mel frequency cepstral coefficients (MFCC) [57], each with delta and acceleration coefficients; fundamental frequency (F0); voicing probability; loudness contour; log-energy with its delta and acceleration [58]. Thus, each frame is represented by a 65 dimensional feature vector. Consequently, each labelled speech segment is represented by a 20×65 feature matrix. The features were extracted using the

Table 3: **Audio Data Distribution**: window size set at 215 milliseconds

Participant ID	**009**	**011**	**012**	**013**	**017**	**019**	**028**	**029**	**031**
Events	350	2206	890	195	1036	297	439	632	196
Normal	1033	2162	1574	1464	1330	1363	1851	1503	1555
Total	1383	4368	2464	1659	2366	1660	2290	2135	1751

openSMILE feature extraction tool [59].
Subsequently, in order to compute the features at the segment level, 14 statistical functions (mean, median, standard deviation, maximum, minimum, range, skewness, kurtosis, first and second quartile, inter quartile, 1%-percentile and 99%-percentile, range of 1% and 99% percentile), were applied on each of the extracted frame level feature resulting in a 910 dimensional segment level feature vector. Principal Component Analysis (PCA) was subsequently applied on the extracted features to reduce the dimensionality of the feature space. The first 10 components represented 98% of the variance and were selected as segment level feature vectors.

4.3 Video Feature Extraction

Concerning the video signal, a window size of 1 second (30 frames) with a shift of 0.5 seconds (15 frames) was chosen. The resulting data distribution for each participant can be seen in Table 4.
Using the facial behaviour analysis toolkit OpenFace [60], the facial area of each participant was automatically detected and tracked from one frame to the next through out the entire video. Subsequently, Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [61] features were extracted based on the detected facial regions. Within each video segment, each facial region was divided in a 4×4 grid of cells with a 25% overlap from one cell to the next. From each resulting cuboid a uniform LBP-TOP feature vector was extracted. These feature vectors were subsequently concatenated, resulting in a 720 dimensional segment level feature vector. Additionally, Pyramid Histogram of Oriented Gradients (PHOG) [62] features were also extracted from each detected facial region. Within each segment, a 3 levels PHOG feature vector with 20 bins was extracted from the detected facial

Table 4: **Video Data Distribution**: window size set at 1 second (30 frames)

Participant ID	**009**	**011**	**012**	**013**	**017**	**019**	**028**	**029**	**031**
Events	231	307	610	207	277	283	140	217	1337
Normal	3874	3007	2488	3736	3350	3346	3615	3552	2592
Total	4105	3314	3098	3943	3627	3629	3755	3769	3929

region in each frame. The feature for the whole segment was subsequently generated by performing a max pooling from the frame level feature vectors resulting in a 210 dimensional segment level feature.

For further assessment of the dataset, both PHOG and LBP-TOP features were concatenated to form a 930 dimensional segment level feature vector. Thereafter, PCA was applied as well to reduce the dimensionality of the feature space. As a result, the first 5 components, which accounted for 95% of the variance, were selected for the assessment of the proposed methods.

5 Evaluation

In the following section, the proposed approaches are assessed on the earlier described dataset (see Section 4). The pool of video observations is set as target pool and the pool of audio observations is set as auxiliary pool. More specifically, the uni-modal method described in Section 3.4 is applied on the video modality. The multi-modal method described in Section 3.5 is assessed by using the pool of video instances as target pool and the pool of audio instances as auxiliary pool. The goal is to improve the performance of the uni-modal method applied on the video modality by transferring information acquired from the audio modality using the temporal dependency between both modalities.

The proposed approaches are assessed by performing a 5-fold blocked cross validation [63]. During a blocked cross validation the dataset is partitioned sequentially (not randomly) into several subsets. Each single subset is subsequently used within each cross validation iteration as a test set while the remaining sets are used as training sets. During each active learning iteration, the baseline result is first computed by training a fully supervised SVM model with Gaussian RBF kernel on the entirely labelled training set and applying the generated model on the test set. Next, the proposed approaches are applied on the unlabelled training set. At each active learning iteration, the updated supervised (respectively semi-supervised) SVM model is applied on the test set.

Data imbalance is known to affect the performance of a SVM model, since it is biased into classifying almost every samples as belonging to the majority class. Therefore, the Synthetic Minority Over-Sampling Technique (SMOTE) [64] is applied on the labelled set to balance the data before the supervised (respectively semi-supervised) classification model is trained. Undersampling the majority class to deal with data imbalance results in losing a huge amount of potentially useful information in the form of the discarded data samples. By oversampling the minority class this information is preserved. However, oversampling can lead to overfitting when the samples of the minority class are simply duplicated. Instead, by generating synthetic samples of the minority class, the overfitting effect is substantively reduced.

The performance of the generated model is expected to be high on both majority and minority classes simultaneously. Therefore, the geometric mean (gmean) [65, 66] defined in Equation 6 is used as performance metric for the assessment of the developed methods:

$$gmean = \sqrt{acc^{+} \times acc^{-}} \tag{6}$$

where acc^{+} stands for the accuracy on the minority class and acc^{-} stands for the accuracy on the majority class. It depicts the balance of classification performances

of the trained model on both majority and minority classes. The SVM models are trained using the libsvm library [67].
Concerning the uni-modal approach, a fixed number of 25 SVDD models ($\mu = 5$, $\nu = 5$) is generated for the unsupervised outlier detection ensemble. The size of the ensemble was set empirically and the influence of the size of the committee on the performance of the system has not been explored at this point. The same committee size is also generated for the EXPoSE based method. At each active learning iteration, a fixed size of 50 samples is selected for the manual annotation. Regarding the multi-modal approach, the uni-modal sample selection performed once on the target pool and once on the auxiliary pool is performed using the same settings as previously described. The experiments are conducted using two specific window sizes of 2 and 4 seconds ($\omega \in \{2, 4\}$). The impact of the weighting function was evaluated using four specific γ values: $\gamma \in \{0, 0.05, 0.5, 1\}$. The threshold $th_{bimodal}$ for the sample selection in the target pool is dynamically set to be the 65%-quantile of the weighted confidences of the samples located within the specified windows. Figures 6 and 7 show the assessment results for the SVDD-based approaches for two different participants.

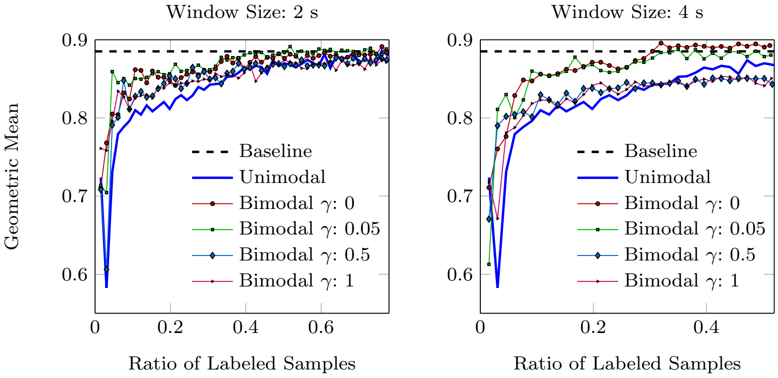


Fig. 6: **Participant 009: SVDD based approaches.** The bimodal approaches converge faster than the uni-modal approach. However, higher values of the parameter γ (0.5, 1) converge to sub-optimal models that are worse than the uni-modal approach, while lower values (0, 0.05) outperform the uni-modal approach.

Figure 6 proves the effectiveness of the bimodal approach, in particular during the early stages of the active learning iterations. For both windows ($\omega \in \{2, 4\}$), the performance curves of the bimodal approaches rise and converge faster than the performance curve of the uni-modal approach, confirming that the bimodal approaches outperform the uni-modal approach with a reduced set of labelled instances. Moreover, lower values of the γ parameter $(0, 0.05)$ outperform higher values $(0.5, 1)$ as it can be clearly seen in the 4 seconds window figure. Given the same amount of labelled instances, higher values of the weighting parameter γ might converge to sub-optimal classification models which are outperformed by the uni-modal generated model, while lower γ values still lead to models that outperform the uni-modal generated model.

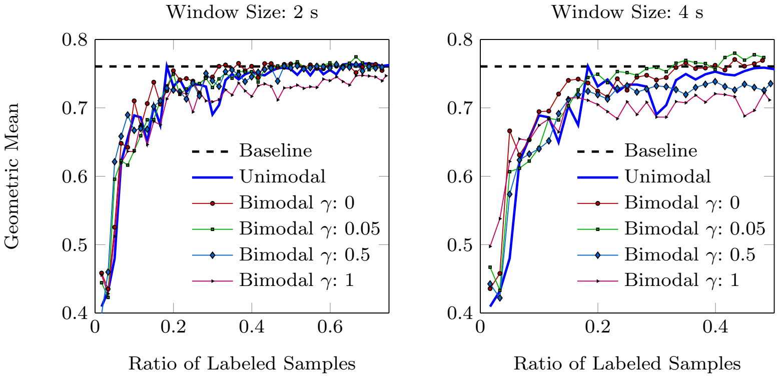


Fig. 7: **Participant 028: SVDD based approaches.** The performances of both approaches are quite similar. In this case, the bimodal approach does not outperform the uni-modal approach. Given the correct parameter γ, the multi-modal approach converges to a model with similar performances as a uni-modal generated model.

Figure 7 shows that the bimodal approach can also be unnecessary, since the performance of both bimodal and uni-modal approaches are quite similar. Given the wrong weighting parameter (γ) the bimodal approach might lead to a sub-optimal classification model. But given an optimized weighting parameter, the bimodal approach will lead to a model which is at least as good as a model trained by the uni-modal approach. The EXPoSE variant of the approaches yielded similar performance results.

For the next assessment, the weighting parameter is set to 0, which means literally that all video samples situated within the define window $[t_i - \omega, t_i + \omega]$ receive the same weight equal to 1. For the sake of clarity, the variance of the computed results is not displayed. Figures 8 and 9 depict the results of the approaches for the selected participants, based respectively on the SVDD outlier detection method and the EXPoSE outlier detection method. The uni-modal SVDD approach proves to be effective since the baseline performance is attained for almost all participants (except for the participants 019 and 031), by manually labelling in average less than 60% of the entire dataset. The uni-modal EXPoSE approach depicts similar results, but converges slower than the uni-modal SVDD approach.

In both settings (SVDD, EXPoSE), the performance curves for the bimodal approaches stop earlier than the performance curves of the uni-modal approaches due to the fact that a larger amount of instances are labelled at each iteration through the semi-supervised labelling process. Concerning the SVDD based approaches, the bimodal methods converge quicker than the uni-modal methods for almost all participants, except for the participants 028 and 031, where the performances of all approaches are very similar. The bimodal SVDD approaches outperform the uni-modal approach in most of the cases. Regarding the EXPoSE based approaches, similar results can be reported. Except for the participant 012, the bimodal approaches outperform the uni-modal approach and converge faster than the uni-modal approach. The size of the window for the selection of samples in

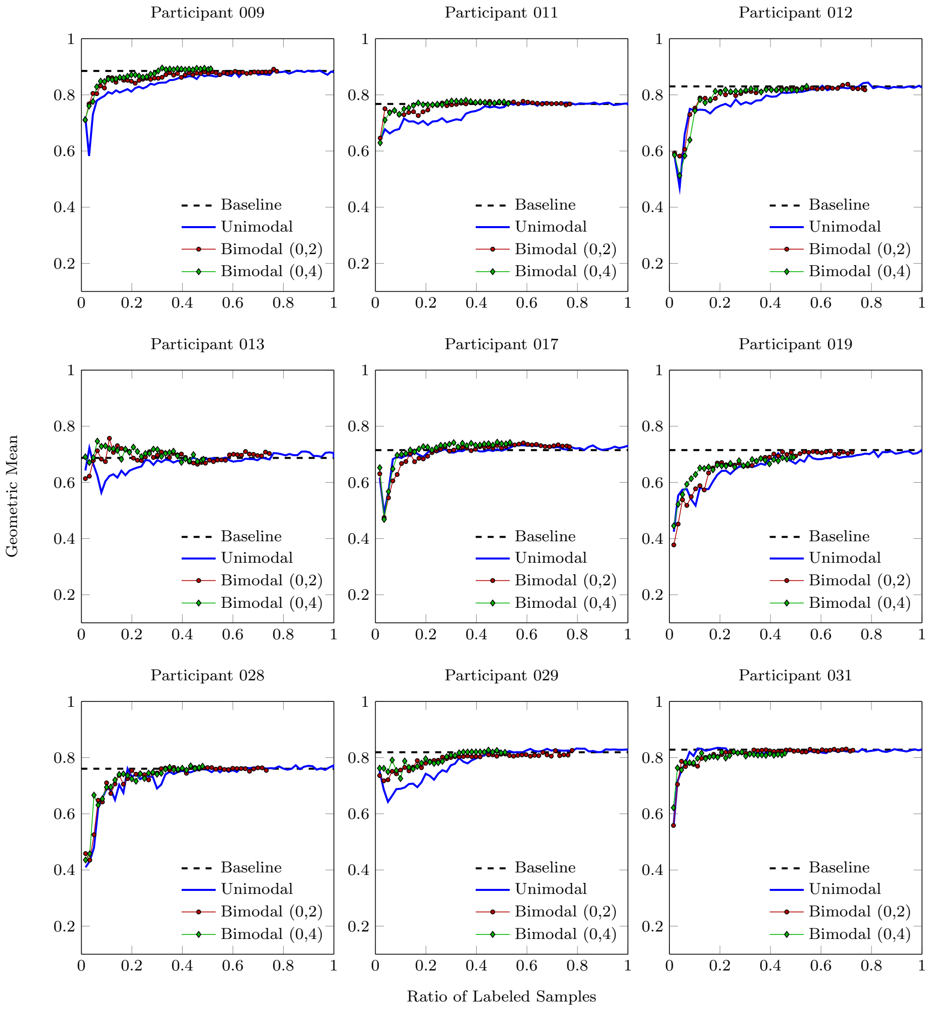


Fig. 8: **Results of the SVDD based approaches.** $\gamma = 0$ and $\omega \in \{2, 4\}$.

the target pool seems to only affect the amount of instances that are automatically labelled. The performances of both windows are very similar and a further assessment is needed here to have a better understanding of the impact of the size of the window.

In order to further assess the performances of the proposed approaches, the averaged quantity of instances from the target pool that has to be labelled in order to attain 95% of the average performance of the fully supervised SVM model (baseline) is computed for each approach. The results are depicted in Table 5.

The result for the EXPoSE based method specific to the participant 012, with a weighting parameter $\gamma = 0$ and a window $\omega = 4$ seconds is not available due to its poor performance as can be seen in Figure 9. The corresponding performance curve could never reach the fully supervised performance curve (baseline).

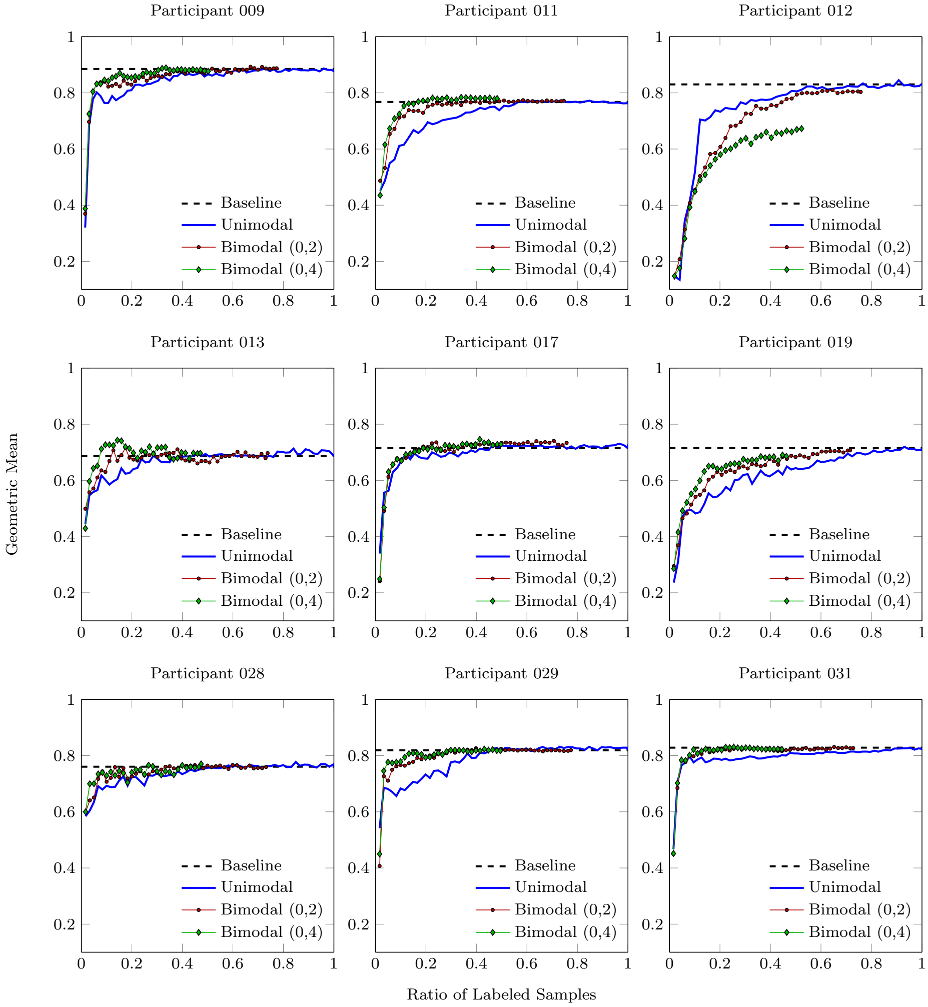


Fig. 9: **Results of the EXPoSE based approaches.** $\gamma = 0$ and $\omega \in \{2, 4\}$. Except for the participant 012, the bimodal approaches outperform the uni-modal approach.

However, the bimodal SVDD based approach with a weighting parameter $\gamma = 0$ and a window $\omega = 4$ seconds outperforms the other active learning approaches in most cases. Moreover, both SVDD based bimodal approaches outperform the uni-modal approaches in all cases except for the participants 017 and 031. This proves the effectiveness of the proposed bimodal approaches. Given a set of optimized parameters (γ, ω), the bimodal approaches converge faster than the uni-modal approaches, thus the cost of manual annotation is considerably reduced. Moreover, given the same amount of labelled instances, the bimodal SVDD approaches are for the hardest classification tasks at least as good as the uni-modal approach.

Table 5: **Active learning performance assessment.** Ratio of manually labelled samples in order to attain at least 95% of the baseline performance.

Participant ID	**009**	**011**	**012**	**013**	**017**	**019**	**028**	**029**	**031**
Unimodal SVDD	0.32	0.38	0.38	0.22	**0.12**	0.41	0.20	0.33	**0.08**
Unimodal EXPoSE	0.30	0.38	0.44	0.24	0.16	0.69	0.28	0.33	0.25
Bimodal SVDD $(\mathbf{0},\mathbf{2})$	**0.11**	**0.09**	**0.18**	0.05	0.16	**0.36**	0.18	0.20	0.13
Bimodal EXPoSE $(\mathbf{0},\mathbf{2})$	0.21	0.15	0.52	0.11	**0.12**	0.48	0.12	0.15	0.10
Bimodal SVDD $(\mathbf{0},\mathbf{4})$	**0.11**	**0.09**	**0.18**	**0.02**	**0.12**	0.41	0.17	0.20	0.10
Bimodal EXPoSE $(\mathbf{0},\mathbf{4})$	**0.11**	0.11	N.A.	0.08	**0.12**	0.38	**0.08**	**0.10**	0.10

6 Conclusion

In this work, two novel active learning approaches for the annotation and detection of audiovisual events have been proposed. The first approach is a uni-modal method which consists of a combination of unsupervised outlier detection techniques and uncertainty sampling throughout rank aggregation for the selection of informative samples. The pool of unlabelled samples belongs to one specific modality and the method exploits uniquely the information from this modality to perform the selection of informative samples. The method has been previously applied for speech event detection in [36] and has shown its effectiveness by substantially reducing the cost of manual annotation required for the training of an effective speech event detection model. In this work, the method was further assessed and applied in the domain of facial events detection. The results also offer evidence of the effectiveness of the method for facial events detection. Furthermore, a new outlier detection technique (EXPoSE [38]) was also assessed within the proposed active learning approach and has also yielded satisfactory results.
The second approach is a multi-modal method which consists of the exploitation of the temporal dependency between two modalities, in combination with semi-supervised learning in order to further reduce the cost of annotation, while improving the performance of the generated model. In this work, the method was also assessed in the domain of facial events detection, by using both audio and video modalities. Informative samples are detected within the pools of audio and video instances using the uni-modal method described earlier. The selected video samples are manually labelled while further video samples are selected within a temporal neighbourhood of the selected audio samples and classified using the model trained in the previous iteration. The corresponding scores of the selected video samples are subsequently weighted based on a temporal Gaussian function and the samples with weighted scores above a specific threshold are discarded from

the pool of video instances and placed in the pool of labelled instances with the machine generated labels. The labelled instances are used to train a semi-supervised classification model that is used in the following iteration. The assessment results show that the bimodal approach outperforms its uni-modal counterpart in most of the cases by achieving better classification performances with less manually labelled instances.

For future work, the proposed bimodal approach has to be assessed on more multimodal datasets and also in different classification tasks. Furthermore, several methods for the transfer of information between the modalities have to be developed and assessed. The assessment of combination methods for the selection of informative samples has also to be undertaken.

Acknowledgements This paper is based on work done within the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems funded by the German Research Foundation (DFG). Patrick Thiam is supported by the Federal Ministry of Education and Research (BMBF) within the project *SenseEmotion*. The work of Sascha Meudt and Friedhelm Schwenker is supported by the SFB/TRR 62. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Markus Kächele, Martin Schels, Sascha Meudt, Viktor Kessler, Michael Glodek, Patrick Thiam, Stephan Tschechne, Günther Palm, and Friedhelm Schwenker. On annotation and evaluation of multi-modal corpora in affective human-computer interaction. In *International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, pages 35–44, 2014.
2. Markus Kächele, Martin Schels, Sascha Meudt, Günther Palm, and Friedhelm Schwenker. Revisiting the emotiw challenge: how wild is it really? *Journal on Multimodal User Interfaces*, pages 151–162, 2016.
3. Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2016.
4. Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning.* The MIT Press, 2006.
5. Burr Settles. Active learning literature survey. Computer sciences technical report, University of Wisconsin–Madison, 2009.
6. Friedhelm Schwenker and Edmonto Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, pages 4–14, 2014.
7. Sascha Meudt, Miriam Schmidt-Wack, Frank Honold, Felix Schüssel, Michael Weber, Friedhelm Schwenker, and Günther Palm. Going further in affective computing: how emotion recognition can improve adaptive user interaction. In *Toward Robotic Socially Believable Behaving Systems-Volume I*, pages 73–103. Springer, 2016.
8. Martin Schels, Michael Glodek, Sascha Meudt, Stefan Scherer, Miriam Schmidt, Georg Layher, Stephan Tschechne, Tobias Brosch, David Hrabal, Steffen Walter, et al. Multimodal classifier-fusion for the recognition of emotions. *Coverbal synchrony in Human-Machine Interaction*, pages 73–97, 2013.
9. Cha Zhang and Tsuhan Chen. An active learning framework for content based information retrieval. *IEEE Transactions on Multimedia*, pages 260–268, 2002.
10. Philippe Henri Gosselin and Matthieu Cord. Active learning methods for interactive image retrieval. *IEEE Transactions on Image Processing*, pages 1200–1211, 2008.
11. Meng Wang and Xian-Sheng Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology*, pages 1–21, 2011.

12. Dan Pelleg and Andrew Moore. Active learning for anomaly and rare-category detection. In *Advances in Neural Information Processing Systems 17*, pages 1073–1080. MIT Press, 2004.
13. Jingrui He and Jaime Carbonell. Nearest-neighbor-based active learning for rare category detection. In *Advances in Neural Information Processing Systems*, pages 633–640, 2007.
14. Timothy M. Hospedales, Shaogang Gong, and Tao Xiang. Finding rare classes: Active learning with generative and discriminative models. In *Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 296–308, 2011.
15. Karim Pichara and Alvaro Soto. Active learning and subspace clustering for anomaly detection. *Intelligent Data Analysis*, pages 151–171, 2011.
16. Ziping Zhao and Xirong Ma. Active learning for speech emotion recognition using conditional random fields. In *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 127–131, 2013.
17. Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller. Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In *Proceedings of the 2015 ACM on International Conference on Multimedia Interaction*, pages 275–278, 2015.
18. Victoria Xia, Natasha Jaques, Sara Taylor, Szymon Fedor, and Rosalind Picard. Active learning for electrodermal activity classification. In *2015 IEEE Signal Processing in Medicine and Biology Symposium*, pages 1–6, 2015.
19. Jenna Wiens and John V. Guttag. Patient-adaptive ectopic beat classification using active learning. In *2010 Computing in Cardiology*, pages 109–112, 2010.
20. Jenna Wiens and John V. Guttag. Active learning applied to patient-adaptive heartbeat classification. In *Advances in Neural Information Processing Systems 23*, pages 2442–2450. 2010.
21. Guha Balakrishnan and Zeeshan Syed. Scalable personalization of long-term physiological monitoring: Active learning methodologies for epileptic seizure onset detection. *Journal of Machine Learning Research*, pages 73–81, 2012.
22. Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Active learning for network intrusion detection. In *Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence*, pages 47–54, 2009.
23. David M.J. Tax and Robert P.W. Duin. Support Vector Data Description. *Machine Learning*, pages 45–66, 2004.
24. Jingrui He, Yan Liu, and Richard Lawrence. Graph-based rare category detection. In *Eight IEEE International Conference on Data Mining*, pages 833–838, 2008.
25. Shigeo Abe. *Support Vector Machines for Pattern Classification.* Springer, 2005.
26. Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, pages 131–163, 1997.
27. Rong Yan, Jie Yang, and Alexander Hauptmann. Automatically labeling video data using multi-class active learning. In *Proceedings of the ninth IEEE International Conference on Computer Vision*, pages 516–523, 2003.
28. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.
29. Zixing Zhang and Björn Schuller. Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *Confidence in Acoustic Emotion in Proceedings Interspeech 2012*, pages 362–365, 2012.
30. Thibaud Senechal, Daniel McDuff, and Rana el Kaliouby. Facial action unit detection using active learning and an efficient non-linear kernel approximation. In *2015 IEEE International Conference on Computer Vision Workshop*, pages 10–18, 2015.
31. Patrick Thiam, Sascha Meudt, Markus Kächele, Günther Palm, and Friedhelm Schwenker. Detection of emotional events utilizing Support Vector Methods in an active learning HCI scenario. In *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems*, pages 31–36, 2014.
32. Patrick Thiam, Markus Kächele, Friedhelm Schwenker, and Günther Palm. Ensembles of Support Vector Data Description for Active Learning Based Annotation of Affective Corpora. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 1801–1807, 2015.

33. Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, pages 85–126, 2004.
34. Varun Chandola, Arindam Baerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, pages 1–58, 2009.
35. Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, pages 215–249, 2014.
36. Patrick Thiam, Sascha Meudt, Friedhelm Schwenker, and Günther Palm. Active learning for speech event detection in HCI. In *Proceedings of the 7th IAPR TC3 Workshop, Artificial Neural Networks in Pattern Recognition, ANNPR 2016*, pages 285–297, 2016.
37. Vladimir Naoumovitch Vapnik. *Methods of Pattern Recognition*, pages 123–170. Springer, 2013.
38. Markus Schneider, Wolfgang Ertel, and Fabio Ramos. Expected similarity estimation for large-scale batch streaming anomaly detection. *Machine Learning*, pages 305–333, 2016.
39. Christopher Williams and Mathias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688, 2001.
40. Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, pages 2153–2175, 2005.
41. Wei-Cheng Chang, Ching-Pei Lee, and Chih-Jen Lin. A Revisit to Support Vector Data Description (SVDD). In *Technical Reports*, 2013.
42. Shili Lin. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, pages 555–570, 2010.
43. Ion Muslea, Steven Minton, and Craig A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the 19th International Conference of Machine Learning*, pages 435–442, 2002.
44. Craig A. Knoblock, Steven Minton, and Ion Muslea. Active learning with multiple view. *Journal of Artificial Intelligence Research*, pages 203–233, 2006.
45. Wei Wang and Zhi-Hua Zhou. On multi-view active learning and the combination with semi-supervised learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1152–1159, 2008.
46. Felix Schüssel, Frank Honold, Nikola Bubalo, Anke Huckauf, Harald Traue, and Dilana Hazer-Rau. In-depth analysis of multimodal interaction: An explorative paradigm. In *International Conference on Human-Computer Interaction*, pages 233–240, 2016.
47. James A Russell. Emotion, core affect and psychological construction. *Cognition and Emotion*, pages 1259–1283, 2009.
48. Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, pages 49–59, 1994.
49. Heinke Hihn, Sascha Meudt, and Friedhelm Schwenker. Inferring mental overload based on postural behavior and gestures. In *Proceedings of the 2nd workshop on Emotion Representations and Modelling for Companion Systems*, pages 1–4, 2016.
50. Heinke Hihn, Sascha Meudt, and Friedhelm Schwenker. On gestures and postural behavior as a modality in ensemble methods. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 312–323, 2016.
51. J Alam, Patrick Kenny, Pierre Ouellet, Themos Stafylakis, and Pierre Dumouchel. Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus. In *Odyssey Speaker and Language Recognition Workshop*, 2014.
52. Sascha Meudt, Lutz Bigalke, and Friedhelm Schwenker. ATLAS–an annotation tool for HCI data utilizing machine learning methods. *Advances in Affective and Pleasurable Design*, pages 5347–5352, 2012.
53. Sascha Meudt, Lutz Bigalke, and Friedhelm Schwenker. ATLAS-annotation tool using partially supervised learning and multi-view co-learning in human-computer-interaction scenarios. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 1309–1312, 2012.
54. Susanne Biundo, Daniel Höller, and Pascal Schattenberg. Companion-technology: An overview. *KI - Künstliche Intelligenz*, pages 11–20, 2016.
55. Sreenivasa Rao Krothapalli and Shashidhar G. Koolagudi. *Emotion Recognition using Speech Features*, chapter Emotion Recognition Using Vocal Tract Information, pages 67–78. 2013.

56. Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of Acoustical Society of America*, pages 1738–1752, 1990.
57. Bhadragiri Jagan Mohan and Ramesh Badu N. Speech recognition using mfcc and dtw. In *International Conference on Advances in Electrical Engineering (ICAEE)*, pages 1–4, 2014.
58. Sreenivasa Rao Krothapalli and Shashidhar G. Koolagudi. *Emotion Recognition using Speech Features*, chapter Speech Emotion Recognition: A Review, pages 15–34. 2013.
59. Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *ACM Multimedia (MM)*, pages 835–838, 2013.
60. Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016.
61. Guoying Zhao and Matti Pietikaeinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 915–928, 2007.
62. Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007.
63. Christoph Bergmeir and José M. Benìtez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, pages 192–213, 2012.
64. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, pages 321–357, 2002.
65. Qiong Gu, Li Zhu, and Zhihua Cai. *Computational Intelligence and Intelligent Systems*, chapter Evaluation Measures of the Classification Performance of Imbalanced Data Sets, pages 461–471. 2009.
66. Victoria Lòpez, Alberto Fernàndez, Salvador Garcìa, Vasile Palade, and Francisco Herrera. Strategies for learning in class imbalance problems. *Pattern Recognition*, pages 849–851, 2003.
67. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, pages 1–27, 2011.

I.2 Multi-modal Pain Intensity Recognition based on the SenseEmotion Database

P. Thiam, V. Kessler, M. Amirian, P. Bellmann, G. Layher, Y. Zhang, M. Velana, S. Gruss, S. Walter, H. C. Traue, J. Kim, D. Schork, E. André, H. Neumann, F. Schwenker. Multi-modal Pain Intensity Recognition based on the *SenseEmotion Database*. IEEE Transactions on Affective Computing. 2019.
DOI: 10.1109/TAFFC.2019.2892090.
Status: Published.
Journal: IEEE Transactions on Affective Computing.
Impact Factor: 6.288 (2019).

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

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 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 assess-

- *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, 89081 Ulm, 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, Switzerland.*
E-mail: amir@zhaw.ch
- *M. Velana, S. Gruss, S. Walter, H. C. Traue are with the University Clinic for Psychosomatic Medicine and Psychotherapy, Medical Psychology, Ulm University, Frauensteige 6, 89075 Ulm, 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, 86159 Augsburg, Germany*
E-mail: {schork, andre}@informatik.uni-augsburg.de
- *J. Kim is with the Department of Information and Communication Technology, Cheju Halla University, Korea*
E-mail: kim@ieee.org

ment 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 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 biovisual 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 Figure 1(a)).

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 $32\,^{\circ}C$ and a maximum of $50.5\,^{\circ}C$. 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 Figure 1(b)). The baseline temperature (T_0) corresponding to no pain stimulation was set to $32\,^{\circ}C$ for all participants. Each temperature 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 Figure 1(c) 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

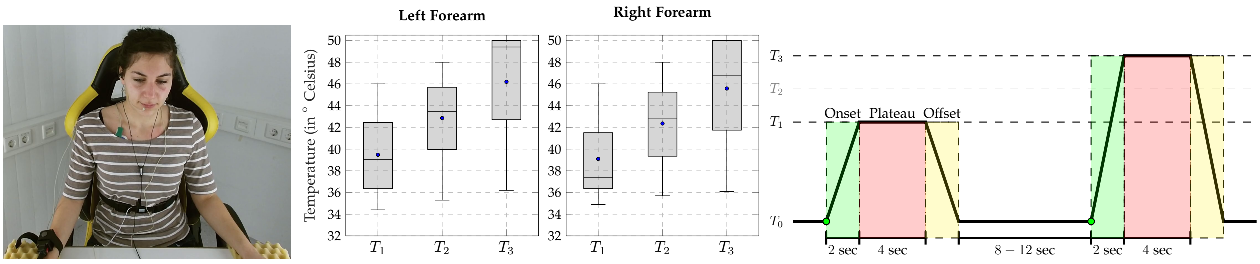
The following section provides a description of the experimental settings specific to the audio channel. A description of each single step involved in the assessment of the data is also provided.

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

[1] http://medoc-weg.com/products/pathway-model-ats/

[2] Ethics Committee Approval: 196/10-UBB/bal



(a) (b) (c)

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 ($32\,°C$); 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.

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 subsections, 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*.

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 $45°$ angle (see Figure 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.

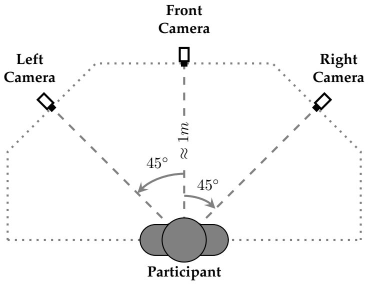


Fig. 2. **Multiple view camera set-up.** The multiple view camera set-up consists of one front camera placed at approximately 1 meter from the participant and two additional cameras placed each in a $45°$ angle at the left and right hand-side of the participant. Hence, the facial area can still be recorded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

5.2 Signal Processing

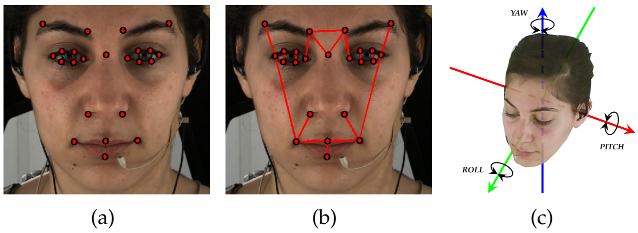
Prior to the assessment of the recorded data, all recordings were first converted into full color videos using demosaicking [46], since the recordings were performed using a Bayer pattern color filter array (CFA). Then, the full color videos were compressed using the codec H.264. Missing frames were 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 Figure 3(a)), characterising each of the defined regions of interest, are detected and tracked from one video frame to the next.



(a) (b) (c)

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.

Based on these landmarks, a set of 17 Euclidean distances are computed at the frame level (see Figure 3(b)). 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 Figure 3(c)), 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 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 Figure 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.

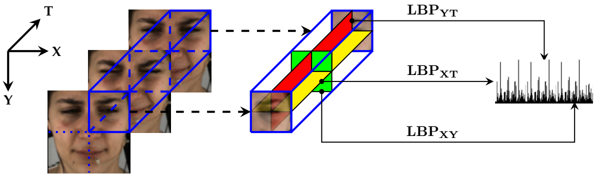


Fig. 4. **Local binary patterns from three orthogonal planes (*LBP-TOP*).** Given a fixed size video sequence, a cuboid consisting of a specific region of interest is extracted. LBP-TOP are subsequently computed based on the cuboid by combining local binary patterns (*LBP*) extracted from the spatial 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.

In this work, each detected facial region within a fixed temporal window is divided into a 4×4 grid of cells with a 25% overlap from one cell to the next. Furthermore, the temporal window is divided in 3 temporal blocks with a 20% 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 Hz.

Electromyography (EMG). 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

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

muscles) 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). 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). 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 Figure 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*,

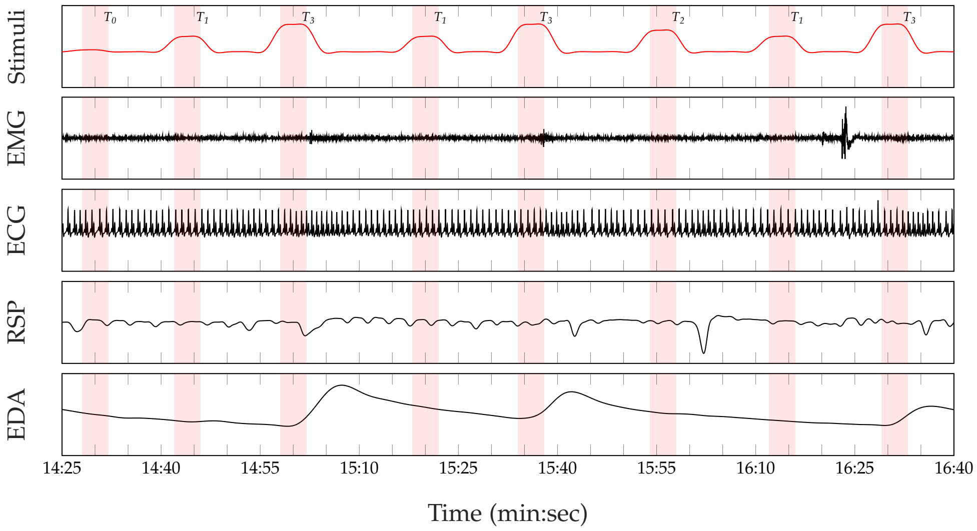


Fig. 5. **Recorded physiological data (preprocessed signals).** From top to bottom: series of artificially induced pain elicitation with the corresponding temperatures (T_0: baseline temperature, T_1: pain threshold temperature, T_2: intermediate elicitation temperature, T_3: pain tolerance temperature),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

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 subspace 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 Figure 6(a). Early fusion con-

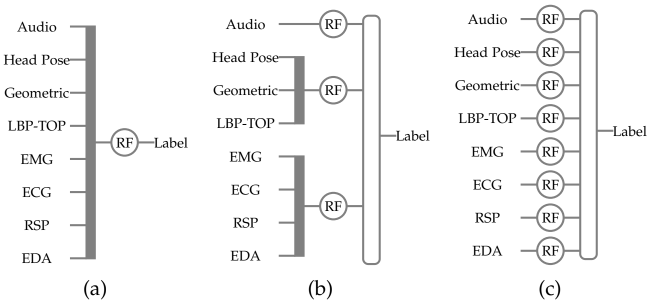


Fig. 6. **Fusion Architectures. (a):** Early Fusion. **(b):** Late Fusion A. **(c):** Late Fusion B. For both late fusion architectures, two fixed mappings (Mean and Max) and two trainable mappings (LDA and Pseudo-inverse) are evaluated. The mappings are applied on the classification scores of the base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

sists 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 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 Figure 6(b) 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 Figure 6(c) 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:

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

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:

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

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 $C = \left[C^1, \dots, C^n\right] \in [0,1]^{Tr \times (cn)}$ ($C^k \in [0,1]^{Tr \times c}$ 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:

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

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:

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

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

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.

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 Figure 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 Figure 1(c)), 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 vs. 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 descriptors extracted from each physiological modality are used to proceed with this evaluation.

Figure 7(a) and Figure 7(b) 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 Figure 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.

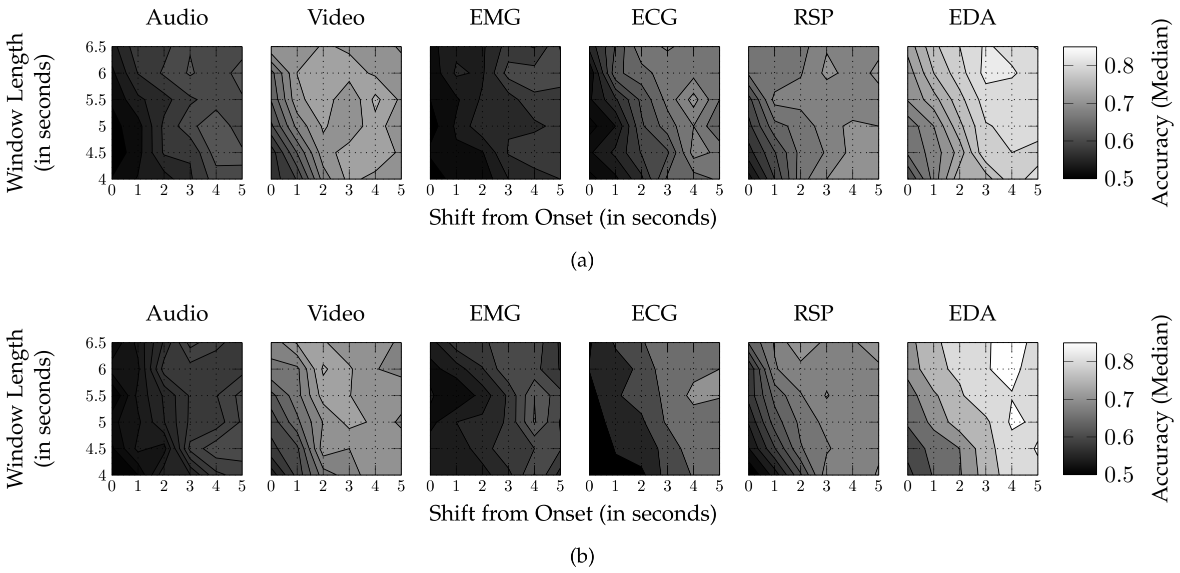


Fig. 7. **Temporal Window Analysis.** (a): Left Forearm. (b): Right Forearm. The results depict the median accuracy for each evaluated temporal window, computed for each modality by applying a 10-fold cross validation evaluation in a user specific setting. The features involved in this evaluation are 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.

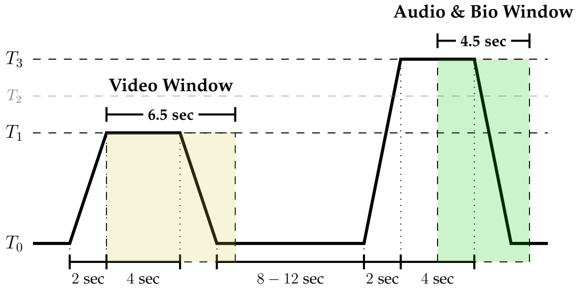


Fig. 8. **Signal segmentation.** The video features are extracted from a window of length 6.5 sec with a temporal shift of 2 sec from the onset. The audio and physiological features are extracted from an identical window of length 4.5 sec with a temporal shift of 4 sec from the onset.

TABLE 1
User specific classification results
(Mean(in %) ± Standard Deviation). 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%.

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	$\mathbf{52.74 \pm 7.64}$	$\mathbf{59.96 \pm 12.86}$	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	$\mathbf{83.13 \pm 12.00}$	51.61 ± 7.87	$\mathbf{60.91 \pm 12.69}$	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	$\mathbf{84.72 \pm 11.09}$*
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	$\mathbf{52.64 \pm 8.06}$	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

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" vs. "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 vs. T_1 is achieved by the EDA with a performance of 52.74% 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%. Concerning the classification task T_0 vs. 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% 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 vs. 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% and 61.53% 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%, and the maximum aggregation rule performing best in case of the right forearm with an average accuracy of 84.72%. 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.

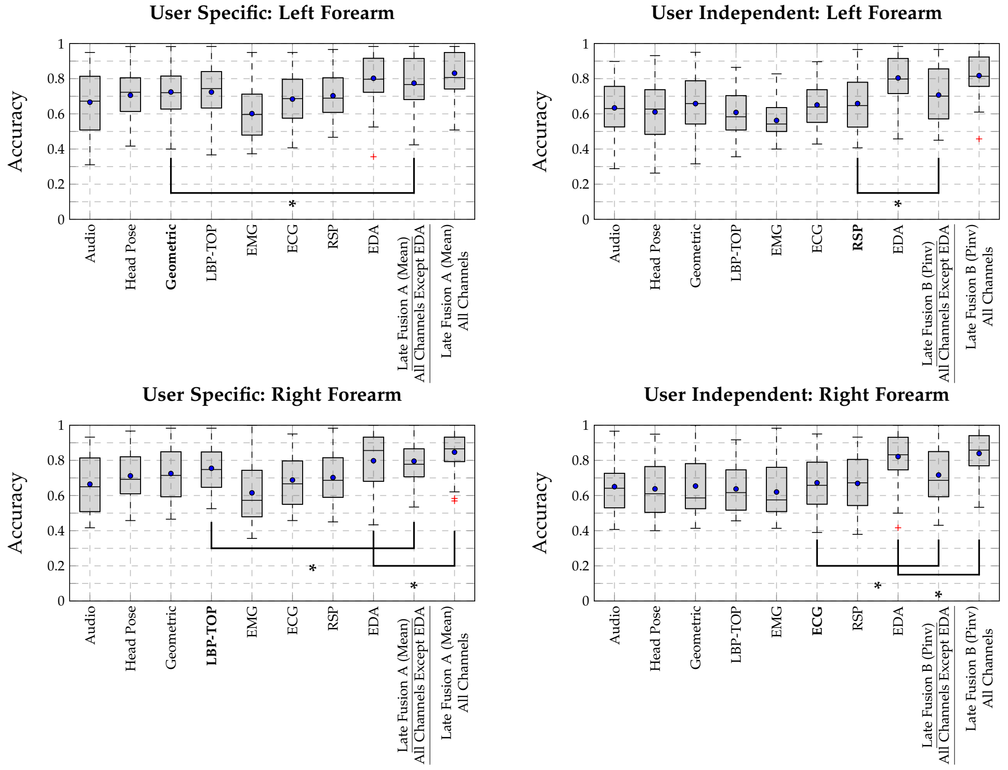
The second late fusion architecture (Late Fusion B) performs in most cases better than the first late fusion architecture (Late Fusion A) in this setting. Moreover, in contrast to the results yielded in a user specific setting, trainable mappings perform in most cases better than fixed mappings. The 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 vs. 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% for the left and right forearm respectively.

TABLE 2
User independent classification results (Mean(in %) ± Standard Deviation). 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%.

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	$\underline{\mathbf{62.34 \pm 10.50}}$	$\underline{80.43 \pm 13.18}$	$\underline{53.13 \pm 5.82}$	$\underline{62.87 \pm 12.10}$	$\underline{82.16 \pm 13.40}$
Early Fusion	$\mathbf{51.88 \pm 5.81}$	59.91 ± 8.13	80.79 ± 12.27	$\mathbf{53.86 \pm 5.70}$	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	$\mathbf{81.76 \pm 12.08}$	51.45 ± 5.74	$\mathbf{62.88 \pm 11.02}$	$\mathbf{83.95 \pm 12.65}$*

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 Table 1 and Table 2) which depict a very high correlation between the performance of the fusion architectures and the performance of 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 vs. T_3 is depicted in Figure 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% 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% 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% for the left forearm, and



(a)

(b)

Fig. 9. $\mathbf{T_0}$ **vs.** $\mathbf{T_3}$ **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%. Within each box plot, the mean and median classification accuracy across all 40 participants are depicted respectively with a dot and a horizontal line.

ECG performs better with a performance of 67.26% for the right forearm. The fusion approach (Late Fusion B (Pinv)) significantly outperforms both channels with classification performances of 70.71% for the left forearm, and 71.66% 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 vs. 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 Figure 9). Therefore, the audio signal is a promising and relevant modality for the development of a pain intensity recognition system.

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 vs. 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 Figure 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.

TABLE 3
Multi-class classification results
(Mean(in %) ± Standard Deviation). The results correspond to a 4-class classification task (T_0 vs. T_1 vs. T_2 vs. 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.

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)	$\mathbf{42.48 \pm 8.35}$	$\mathbf{43.11 \pm 7.93}$	$\mathbf{43.89 \pm 7.61}$

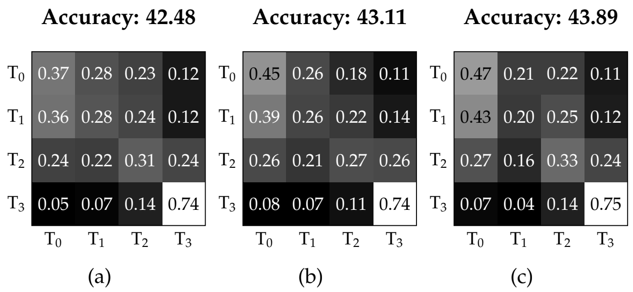


Fig. 10. **4-class Classification Task Confusion Matrices (Late Fusion B (Pinv) in a LOUO setting).** (a): Left Forearm. (b): Right Forearm. (c): Both Forearms. The rows correspond to the ground truth, while the columns correspond to the predictions.

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 Figure 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.

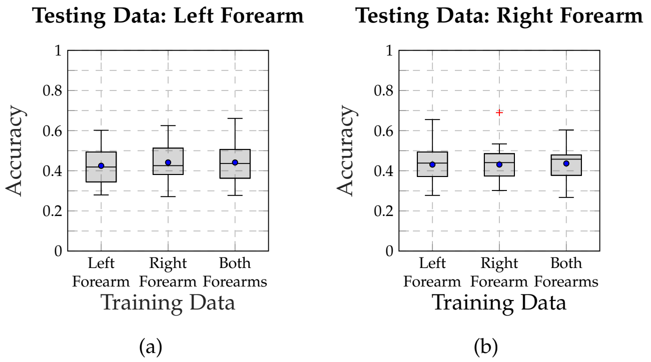


Fig. 11. **4-class Classification Task Results Comparison (Late Fusion B (Pinv)) in a LOUO setting.** (a): The test evaluation is performed on the data specific to the left forearm. (b): The test evaluation is performed on the data specific to the right forearm.

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 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 vs. 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 vs. T_3 and T_0 vs. T_2 vs. T_3 classification tasks respectively. By taking into account that the class labels are

TABLE 4

Classification results (Mean(in %) ± Standard Deviation). 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%.

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	$\mathbf{52.58 \pm 4.00}$	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	$\mathbf{62.96 \pm 9.02}$	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	82.23 ± 10.57	$\mathbf{83.39 \pm 10.23}$*
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	57.84 ± 10.51	$\mathbf{59.53 \pm 9.94}$*
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	42.92 ± 7.07	$\mathbf{43.89 \pm 7.61}$

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.

Acknowledgements

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,” *Proceedings of the National Academy of Sciences of the United States of America*, 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,” *Journal of Palliative Medicine*, 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 and Alternative Communication*, 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 and Research*, 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," *Journal of Pain Management & Medecine*, vol. 2, no. 3, p. 118, 2016.

[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 *Multimodal Pattern Recognition of Social Signals in Human-Computer-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 *Face and Gesture*, 2011, pp. 57–64.

[10] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and 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 *2013 IEEE International Conference on Cybernetics*, 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 Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, 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 *Proceedings of the 25th European Symposium of Artificial Neural Networks, Computational Intelligence and Machine Learning*, 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 Transactions on Affective Computing*, vol. 8, no. 3, pp. 286–299, 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 Transactions on Cybernetics*, pp. 1–11, 2017.

[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 Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, pp. 854–864, 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 Systems*, 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 and 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*. 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 Computing and Applications*, 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 Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[25] M. Kächele, M. Schels, P. Thiam, and F. Schwenker, "Fusion mappings for multimodal affect recognition," in *IEEE Symposium Series on Computational Intelligence*, 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 *Proceedings of the 5th International Workshop on 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 *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, 2017, pp. 110–126.

[28] V. Kessler, P. Thiam, M. Amirian, and F. Schwenker, "Pain recognition with camera photoplethysmography," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 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 *Proceedings of the 7th Annual Workshop on 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 *In Proceedings of the International Conference on Pattern Recognition (ICPR)*, 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*. 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 *Proceedings of the International Workshop on Multiple Classifier Systems (MCS)*, 2015, pp. 220–230.

[33] P. Thiam and F. Schwenker, "Multi-modal data fusion for pain intensity assessement and classification," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 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," *Journal on 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 *Proceedings of the 15th ACM on International Conference on 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 *Neural Information Processing*, 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 *Interspeech 2016*, 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 *Proceedings of the 21st ACM International Conference on 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 *ACM Multimedia (MM)*, 2013, pp. 835–838.

[40] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[41] B. Jagan Mohan and R. Badu N., "Speech recognition using mfcc and dtw," in *International Conference on Advances in Electrical Engineering (ICAEE)*, 2014, pp. 1–4.

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

[43] A. Neerja, "Automatic speech recognition system: A review," *International Journal of Computer Applications*, vol. 151, no. 1, pp. 24–28, 2016.

[44] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992, pp. 121–124.

[45] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, 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 Processing Magazine*, vol. 22, no. 1, pp. 44–554, 2005.

[47] 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*, 2016, pp. 1–10.

[48] ——, "Constrained Local Neural Fields for robust facial landmark detection in the wild," in *IEEE International Conference on Computer Vision 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[52] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, 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 *Wireless Health 2010*, 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 *Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives*. Information Science Reference, 2009, pp. 281–307.

[57] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG pattern analysis for emotion detection," *IEEE Transactions of Affective Computing*, vol. 3, no. 1, pp. 102–115, 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 in Human Neuroscience*, vol. 7, p. 295, 2013.

[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," *Experimental 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 Transactions on Affective Computing*, vol. 3, no. 4, pp. 482–495, 2012.

[62] Y. Chu, X. Zhao, J. Han, and Y. Su, "Physiological signal-based method for measurement of pain intensity," *Frontiers in Neuroscience*, vol. 11, p. 279, 2017.

[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," *Journal of Intelligent Manufacturing*, 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," *Journal of Physiological Anthropology*, vol. 34, no. 1, p. 25, 2015.

[65] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[66] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "A novel feature extraction for robust EMG pattern recognition," *Journal of Computing*, 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," *Journal of Neuroingineering and Rehabilitation*, vol. 7, no. 1, p. 21, 2010.

[68] J. P. Burg, *Modern Spectrum Analysis*. New York: 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," *Journal of Rehabilitation Research & Development*, 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," *Medical Engineering & Physics*, 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 journal of the International Federation of 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," *International Journal of Multimedia and Ubiquitious Engineering*, 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 *International Conference on Neural Networks and 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 Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2016.

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

[76] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

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

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

[79] F. Schwenker, C. R. 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.

[80] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *Ninth International Conference on Intelligent Systems Design and Applications*, 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 Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.

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

Patrick Thiam received the M.Sc. degree in computer science from Ulm University, Ulm, Germany, in 2014. He is currently working toward the Ph.D. 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 M.Sc. degree in computer science from Ulm University, Ulm, Germany, in 2015. After his graduation he started working as a research assistant at the University of Ulm, Germany, as part of the SFB/TRR 62. He is currently working toward the Ph.D. 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 degree in Communications Technology from Ulm University, Ulm, Germany, in 2017. He is currently working as a researcher at the Institute of Applied Information Technology (InIT) of Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland, and simultaneously pursuing a Ph.D. 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 pursuing the Ph.D. 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 Georg Layher graduated in computer science from the University of Ulm, Germany in 2009. After his graduation he started working as a research assistant at 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 at the Institute of Neural Information Processing, Ulm University, Ulm, Germany. Before, he studied in Saarland University and worked at the Max Planck Institute of Informatics, Computer Science Department of Saarland University and at 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 her Bachelors in Psychology from the Panteion University of Social and Political Sciences, Athens, Greece, in 2008. She received her MSc degree (Hons) in Applied Public Health, National School of Public Health in collaboration with Technological Educational Institute of Athens, Greece, 2010. From 2011 to 2012, she was a research associate in the Department of Public Health (Public & Administrative Health), National School of Public Health, Athens, Greece. In 2016, she received her MA (Hons) in Physical Activity and Health, Institute of Sport Science and Sport, University of Erlangen-Nuremberg, Germany. Currently, she is working toward her PhD degree in Medical Psychology, University Clinic of Psychosomatic Medicine and Psychotherapy, Ulm University, Germany. 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 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 Harald C. Traue studied Electrical Engineering, Computer Sciences, Cybernetics, Communication and Social Sciences at Universities in Berlin, Lemgo and Bremen. PhD graduation in Human Biology 1978, Senior Researcher and Lecturer at Ulm University, Visiting Professor at the University of Calgary/Canada. Since 1993 Professor of Medical Psychology at 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 his 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 Jonghwa Kim is Full Professor of the Department of Information and Communication Technology at the Cheju Halla University, Korea. In 2016-2018, he was Professor of Computer Software at the University of Science and Technology (UST), Korea. He received the BS and MS degree in Electronic Engineering from the Gachon University, Korea, in 1992 and 1994 respectively. He received the Ph.D. degree in Communication Engineering from the Technical University of Berlin, Germany, 2003, and the Habilitation Dr. degree (venia legendi) in Computer Science in 2010 from the University of Augsburg, Germany, where he worked as Professor of Applied Computer Science until 2016. 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. His current research interests include affective artificial intelligence, pain and emotion recognition, and the deep learning for cognitive systems.

Elisabeth André is a full professor of Computer Science and Founding Chair of Human-Centered Multimedia at Augsburg University in Germany where she has been since 2001. She has multiple degrees in computer science from Saarland University, including a doctorate. Previously, she was a principal researcher at the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken. Elisabeth André has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing and social signal processing. In 2010, Elisabeth André 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 a doctoral degree in computer science at the University of Hamburg in 1988. He received the Habilitation degree in 1995 and was appointed as professor of computer science in the Institute of Neural Information Processing at Ulm University in 1995. While at 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) at Ulm University. His research interests include neural modelling in computational and cognitive neuroscience, biologically inspired computational vision, and neuromorphic computation.

Friedhelm Schwenker received the Ph.D. 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.

I.3 Exploring Deep Physiological Models for Nociceptive Pain Recognition

P. Thiam, P. Bellmann, H. A. Kestler, F. Schwenker. Exploring Deep Physiological Models for Nociceptive Pain Recognition. Sensors 2019, 19, 4503.
DOI: 10.3390/s19204503.
Status: Published.
Journal: Sensors.
Impact Factor: 3.031 (2018).

sensors

MDPI

Article

Exploring Deep Physiological Models for Nociceptive Pain Recognition

Patrick Thiam [1,2,*], Peter Bellmann [2], Hans A. Kestler [1] and Friedhelm Schwenker [2]

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany; hans.kestler@uni-ulm.de

[2] Institute of Neural Information Processing, Ulm University, James-Franck-Ring, 89081 Ulm, Germany; peter.bellmann@uni-ulm.de (P.B.); friedhelm.schwenker@uni-ulm.de (F.S.)

* Correspondence: patrick.thiam@uni-ulm.de

Received: 5 September 2019; Accepted: 14 October 2019; Published: 17 October 2019

check for updates

Abstract: Standard feature engineering involves manually designing measurable descriptors based on some expert knowledge in the domain of application, followed by the selection of the best performing set of designed features for the subsequent optimisation of an inference model. Several studies have shown that this whole manual process can be efficiently replaced by deep learning approaches which are characterised by the integration of feature engineering, feature selection and inference model optimisation into a single learning process. In the following work, deep learning architectures are designed for the assessment of measurable physiological channels in order to perform an accurate classification of different levels of artificially induced nociceptive pain. In contrast to previous works, which rely on carefully designed sets of hand-crafted features, the current work aims at building competitive pain intensity inference models through autonomous feature learning, based on deep neural networks. The assessment of the designed deep learning architectures is based on the *BioVid Heat Pain Database (Part A)* and experimental validation demonstrates that the proposed uni-modal architecture for the electrodermal activity (EDA) and the deep fusion approaches significantly outperform previous methods reported in the literature, with respective average performances of 84.57% and 84.40% for the binary classification experiment consisting of the discrimination between the baseline and the pain tolerance level (T_0 vs. T_4) in a *Leave-One-Subject-Out* (LOSO) cross-validation evaluation setting. Moreover, the experimental results clearly show the relevance of the proposed approaches, which also offer more flexibility in the case of transfer learning due to the modular nature of deep neural networks.

Keywords: convolutional neural networks; signal processing; information fusion; pain intensity classification

1. Introduction

Conventional machine learning approaches are built upon a set of carefully engineered representations, which consist of measurable parameters extracted from raw data. Based on some expert knowledge in the domain of application, a feature extractor is designed and used to extract relevant information in the form of a feature vector from the preprocessed raw data. This high level representation of the input data is subsequently used to optimise an inference model. Although such approaches have proven to be very effective and can potentially lead to state-of-the-art results (given that the set of extracted descriptors is suitable for the task at hand), the corresponding performance and generalisation capability is limited by the reliance on expert knowledge as well as the inability of the designed model to process raw data directly and to dynamically adapt to related new tasks.

Meanwhile, deep learning approaches [1] automatically generate suitable representations by applying a succession of simple and non-linear transformations on the raw data. A deep learning architecture consists of a hierarchical construct of several processing layers. Each processing layer is characterised by a set of parameters that are used to transform its input (which is the representation generated by the previous layer) into a new and more abstract representation. This specific hierarchical combination of several non-linear transformations enables deep learning architectures to learn very complex functions as well as abstract descriptive (or discriminative) representations directly from raw data [2]. Moreover, the hierarchical construct characterising deep learning architectures offers more flexibility when it comes to adapting such approaches to new and related tasks. Hence, deep learning approaches have been outperforming previous state-of-the-art machine learning approaches, especially in the field of image processing [3–7]. Similar performances have been achieved in the field of speech recognition [8,9] and natural language processing [10,11].

A steadily growing amount of work has been exploring the application of deep learning approaches on physiological signals. Martinéz et al. [12] were able to significantly outperform standard approaches built upon hand-crafted features by using a deep learning algorithm for affect modelling based on physiological signals (two physiological signals consisting of Skin Conductance (SC) and Blood Volume Pulse (BVP) were used in this specific work). The designed approach consisted of a multi-layer Convolutional Neural Network (CNN) [13] combined with a single-layer perceptron (SLP). The parameters of the CNN were trained in an unsupervised manner using denoising auto-encoders [14]. The SLP was subsequently trained in a supervised manner using backpropagation [15] to map the outputs of the CNN to the target affective states. In [16], the authors proposed a multiple-fusion-layer based ensemble classifier of stacked auto-encoder (MESAE) for emotion recognition based on physiological data. A physiological-data-driven approach was proposed in order to identify the structure of the ensemble. The architecture was able to significantly outperform the existing state-of-the-art performance. A deep CNN was also successfully applied in [17] for arousal and valence classification based on both electrocardiogram (ECG) and Galvanic Skin Response (GSR) signals. In [18], a hybrid approach using CNN and Long Short-Term Memory (LSTM) [19] Recurrent Neural Network (RNN) was designed to automatically extract and merge relevant information from several data streams stemming from different modalities (physiological signals, environmental and location data) for emotion classification. Moreover, deep learning approaches have been applied on electromyogram (EMG) signals for gesture recognition [20,21] or hand movement classification [22,23]. Most of the reported approaches consist of first transforming the processed EMG signal into a two dimensional (time-frequency) visual representation (such as a spectrogram or a scalogram) and subsequently using a deep CNN architecture to proceed with the classification. A similar procedure was used in [24] for the analysis of electroencephalogram (EEG) signals. These are just some examples of an increasingly growing field of experimentation for deep neural networks. A better overview of deep learning approaches applied to physiological signals can be found in [25,26]. However, there are few related works that focus specifically on the application of deep neural networks on physiological signals for pain recognition. The authors of [27] recently proposed a classification architecture based on Deep Belief Networks (DBNs) for the assessment of patients' pain level during surgery, using photoplethysmography (PPG). The proposed architecture consists of a bagged ensemble of DBNs, built upon a set of manually engineered features, extracted from the recorded and preprocessed PPG signals. It is important to note that, in this specific study, the ensemble of bagged DBNs was trained on a set of carefully designed hand-crafted features. Therefore, an expert knowledge in this specific area of application is still needed in order to generate a set of relevant descriptors, since the whole classification process is not performed in an end-to-end manner.

Nonetheless, there is a constantly growing amount of works that focus specifically on pain recognition based on physiological signals, and categorised by the nature of the pain elicitations. There is a huge variety of statistical methods that have been proposed, most of them based on more traditional machine learning approaches such as decision trees or Support Vector Machines (SVMs) [28]. In [29], the authors proposed a continuous pain monitoring method using an Artificial Neural Network (ANN), based on hand-crafted features (wavelength (WL) and root mean square (RMS) features) extracted from several physiological signals consisting of heart rate (HR), breath rate (BR), galvanic skin response (GSR) and facial surface electromyogram (sEMG). The proposed approach was assessed on a dataset collected by inducing both thermal and electrical pain stimuli. In [30], the authors proposed a pain detection approach based on EEG signals. Relevant features are extracted from the EEG signals using the Choi–Williams quadratic time–frequency distribution and subsequently used to train a SVM in order to perform the classification task. Pain in this specific work is elicited throughout tonic cold. Most recently, Thiam et al. [31,32] provided the results for a row of pain intensity classification experiments based on the *SenseEmotion Database* (SEDB) [33], by using several fusion architectures to merge hand-crafted features extracted from different modalities, including physiological, audio and video channels. Thereby, the combination of the features extracted from the recorded signals was compared for different fusion approaches, namely the fusion at feature level, the fusion at the classifiers' output level and the fusion at an intermediate level. Random Forests [34] were used as the base classifiers. In [35], the authors combined camera PPG input signals with ECG and EMG signals in order to proceed with a user-independent pain intensity classification using the same dataset. The authors used a fusion architecture at the feature level with Random Forests and SVMs as base classifiers.

In [36–38], the authors performed different pain intensity classification experiments based on the *BioVid Heat Pain Database* [39] (BVDB). All the conducted experiments were based on a carefully selected set of features extracted from both physiological and video channels. The classification was also performed using either Random Forests or SVMs. In [40], Kächele et al. performed a user-independent pain intensity classification evaluation based on physiological input signals, using the same dataset. The authors used the whole data from all recorded pain levels in a classification, as well as a regression setting with Random Forests as the base classifiers. Several personalisation techniques were designed and validated, based on meta information from the test subjects, distance measures and machine learning techniques. The same authors proposed an adaptive confidence learning approach for personalised pain estimation in [41] based on both physiological and video modalities. Thereby, the authors applied the fusion at feature level. The whole pain intensity estimation task was analysed as a regression problem. Random Forests were used as the base regression models. Moreover, a multi-layer perceptron (MLP) was applied to compute the confidence for an additional personalisation step. One recent work included the physiological signals of both datasets (SEDB and BVDB) [42]. The authors analysed different fusion approaches with fixed aggregating rules based on their merging level for the person-independent multi-class scenario using all available pain levels. Thereby, three of the most popular decision tree based classifier systems, i.e., Bagging [43], Boosting [44] and Random Forests, were compared.

The current work focuses on the application of deep learning approaches for nociceptive heat-induced pain recognition based on physiological signals (EMG, ECG and electrodermal activity (EDA)). Several deep learning architectures are proposed for the assessment of measurable physiological parameters in order to perform an end-to-end classification of different levels of artificially induced nociceptive pain. The current work aims at achieving state-of-the-art classification performances based on feature learning (the designed architecture autonomously extracts relevant features from the preprocessed raw signals in an end-to-end manner), therefore removing the reliance on expert knowledge for the design and optimisation of reliable pain intensity classification models (since most of the previous works on pain intensity classification involving autonomic parameters rely on a carefully designed set of hand-crafted features). The remainder of the work is organised as follows. The proposed deep learning approaches as well as the dataset used for the validation of the approaches are described in Section 2. Subsequently, a description of the results corresponding to the conducted assessments specific to each presented approach is provided in Section 3. Finally, the findings of the conducted experiments are discussed in Section 4, followed by the description of potential future works and a conclusion.

2. Materials and Methods

2.1. BioVid Heat Pain Database (BVDB)

The *BioVid Heat Pain Database* [39] (BVDB) was collected at Ulm University. It includes multi modal data recordings from healthy subjects subjected to different levels of artificially induced pain stimuli under strictly controlled conditions. The pain elicitation in the form of heat was conducted through the professionally designed PATHWAY (http://www.medoc-web/products/pathway) thermode attached to the participants' right forearm. Before the data were recorded, a personalised calibration step was undertaken for each participant to determine individual levels for the pain threshold, as well as the tolerance threshold. Therefore, starting at a temperature of 32 °C (global pain free level T_0 for all participants), the temperature was slowly increased until, first, the participant felt a change from heat to pain (pain threshold T_1), and, second, the pain became hardly bearable (tolerance threshold T_4). In addition, two in-between pain elicitation levels T_2 and T_3 were calculated, making the four individual pain levels T_1, T_2, T_3, T_4 equidistant. After the initial calibration steps, starting at the baseline temperature T_0, each of the four individual pain levels was applied randomly 20 times. Each of the pain levels was held for a total of 4 s. Each pain stimulation was followed by a rest period during which the baseline temperature was held for a random duration of 8–12 s. Ninety subjects were recruited for the experiments. The participants covered three age groups, i.e., 18–35 years, 36–50 years and 51–65 years. Each group was equally distributed, including 15 male and 15 female subjects. In the current study, the designed approaches were assessed on the *BioVid Heat Pain Database (Part A)* since most of the related works were conducted based on this specific database. The database is publicly available and consists of a total of 87 participants. Due to technical issues during the recordings, some of the data specific to three participants are missing [36]. Those participants were therefore discarded and the remaining 87 participants, for which all data are available, constitute the *BioVid Heat Pain Database (Part A)*.

During the experiments, three different physiological signals were recorded, namely electrodermal activity (EDA), electrocardiogram (ECG) and electromyogram (EMG) (a sample of the recorded physiological signals is depicted in Figure 1). The EDA is an indicator of the skin conductance level and was measured at both, the participants' index and ring fingers. The ECG signals measure the participants' heart activity, such as the heart rate, the interbeat interval and the heart rate variability. The EMG signal is an indicator of the muscle activity. The EMG signal of the current dataset consists of the muscle activities of the trapezius muscles, which are located at the back, in the shoulder area. In addition to the biopotentials, different video signals were recorded. Since in the current work we only consider the physiological signals, interested readers are referred to [39] to get further details on the whole dataset. Having 20 elicitations for each level of pain elicitation, every subject is represented by a total of $20 \times 5 = 100$ sequences of numerical data points (time series). Therefore, the unprocessed dataset consists of $87 \times 100 = 8700$ samples, each labelled with its corresponding level of nociceptive pain elicitation (T_0, T_1, T_2, T_3 or T_4).

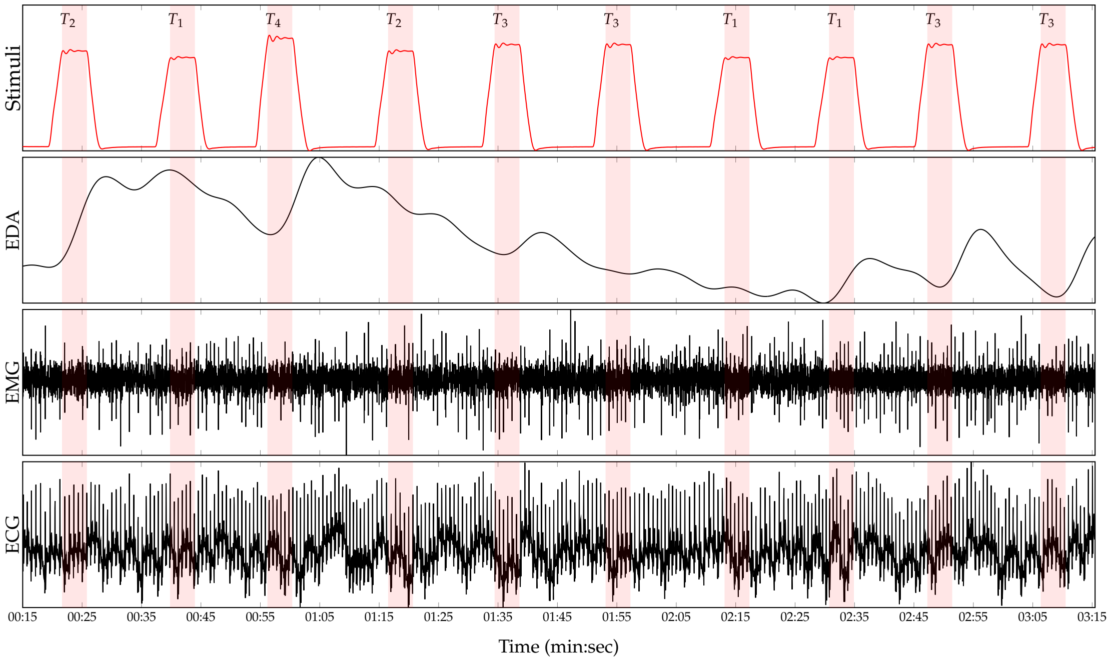
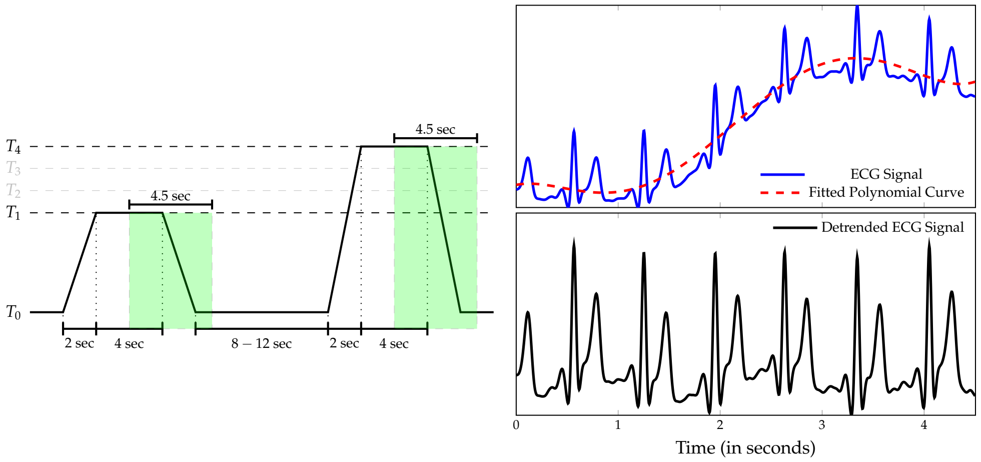


Figure 1. Recorded physiological data. From top to bottom: Series of artificially induced pain elicitation (T_1, pain threshold temperature; T_2, first intermediate elicitation temperature; T_3, second intermediate elicitation temperature; T_4, pain tolerance temperature); EDA (µS); EMG (µV); and ECG (µV).

2.2. Data Preprocessing

Prior to the classification experiments, the sampling rate of the recorded physiological modalities was reduced to 256 Hz, in order to reduce the computational requirements. Subsequently, the amount of noise and artefacts within the recorded data was significantly reduced by applying different signal preprocessing techniques on each specific modality. A third-order low-pass Butterworth filter with a cut-off frequency of 0.2 Hz was applied on the EDA signals. The EMG signals were filtered by applying a fourth-order bandpass Butterworth filter with a frequency range of $[20, 250]$ Hz. Finally, a third-order bandpass Butterworth filter with a frequency range of $[0.1, 250]$ Hz was applied on the ECG signals. Furthermore, the data were segmented as proposed in [37], but rather than using 5.5 s windows with a shift of 3 s from the elicitations' onset, the preprocessed signals were segmented into windows of length 4.5 s, with a shift from the elicitations' onset of 4 s (see Figure 2a), as recently proposed in [31]. Each signal extracted within this window constitutes a 1D array of size $4.5 \times 256 = 1152$ and was later used in combination with the corresponding level of nociceptive pain elicitation to optimise and assess the designed deep classification architectures. Thus, each physiological modality specific to each single participant is represented by a tensor with the dimensionality $100 \times 1152 \times 1$. After some close analysis of the preprocessed physiological signals, a clear baseline wandering of the ECG signal, which is characterised by a strong correlation with the shape of the EDA signal, was observed (see Figure 2b). Therefore, the segmented ECG signals were additionally detrended by subtracting a fifth-degree polynomial least-squares fit from the filtered signals. This step was carried out to remove the aforementioned artefacts from the ECG signals, since these artefacts could potentially bias the classification performance of the corresponding deep classification model (instead of using information stemming uniquely from the ECG signal, the designed system would end up extracting information stemming from a non-linear combination of both the ECG signal and a noisy signal related to the EDA signal). Finally, data augmentation was performed by shifting the 4.5 s window of segmentation backward and forward in time with small shifts of length 250 ms and a maximal total window shift of 1 s in each direction, starting from the initial position of the window depicted in Figure 2a. The signals extracted within these windows were subsequently used as training material for the optimisation of the classification architectures.



(a) (b)

Figure 2. Data preprocessing. (**a**) Signal Segmentation. The classification experiments were performed on windows of length 4.5 s with a temporal shift of 4 s from the elicitations' onset. (**b**) The ECG signal was further detrended by subtracting a least-squares polynomial fit from the preprocessed signal.

2.3. Uni-modal Deep Model Description

As mentioned above, the goal of the current work is to apply feature learning to alleviate the reliance on domain specific expert knowledge that occurs when relevant and adequate features are to be manually designed (hand-crafted features) in order to achieve state-of-the-art classification performances. Therefore, multi-layer CNNs were designed and fed with the preprocessed physiological signals in order to automatically compute relevant signal representations and at the same time optimise the classification architectures. In the following sections, c depicts the number of classes of the classification task.

CNNs [45,46] constitute a distinct category of biologically inspired neural networks, which are characterised by a hierarchical structure of several processing layers. The input to a CNN is sequentially and progressively transformed by each specific layer and the back-propagated information stemming from the error computed between the network's output and the expected output (ground-truth) is used to optimise the whole structure of the architecture in order to efficiently and effectively solve a classification or regression task. The basic processing layers of CNNs are *convolutional layers*, *pooling layers* and *fully connected layers*. *Convolutional layers* are characterised by a set of neurons (or kernels), whereby each specific neuron extracts a specific pattern of information from a patch of the layer's input. Each neuron consists of a set of trainable weights, the size of which is determined by the patch's size (or kernel size). The output of each neuron is calculated by applying a non-linear activation function (e.g., sigmoid function) on the weighted sum of the neuron's input. Each neuron scans the layer's input sequentially and the aggregation of the resulting local information extracted at each specific patch constitutes a feature map. Thus, the output of a convolutional layer is a set of feature maps generated by the convolution of each neuron across the layer's input. *Pooling layers* reduce the spatial resolution of the generated feature maps by merging semantically similar features. *Max Pooling* is a commonly used pooling approach and consists of computing the maximum value of a defined local patch (the size of the patch related to a specific pooling layer is referred in the current work as "pool size") of each feature map. *Fully connected layers* are basically single-layer feed-forward networks that perform the classification or regression task based on the learned deep representations.

Several challenges emerge when it comes to optimising such architectures. One of those challenges is the so-called *vanishing* or *exploding gradients* problem, which is caused by the *internal covariate shift* (constant fluctuations in layers' input distributions) occurring in deep architectures during the training process. In [47], the authors proposed a technique called *Batch Normalisation* to address this specific issue. Batch Normalisation consists of automatically learning the optimal scaling and shifting parameters of each layer's input, so that each layer's input is dynamically normalised, thus significantly reducing the effects of the internal covariate shift and therefore stabilising the training process. Another common challenge occurring when training CNNs is the *overfitting* problem caused by the large amount of parameters that have to be consistently and effectively optimised. Applying regularisation techniques can help to significantly reduce this issue. The authors of [48] introduced the *dropout* approach, which is one of the most commonly used regularisation techniques for deep neural networks. The *dropout* approach consists of randomly and temporarily removing a set of neurons (or units) from the neural network during each training step, each neuron having a fixed probability $p \in [0, 1]$ of being retained. The resulting model is therefore more robust against overfitting and generalises better.

In the current work, the designed architectures are regularised using both techniques and the dropout rate is fixed at 25%. Moreover, the *Exponential Linear Unit* (ELU) function [49] defined in Equation (1)

$$elu_{\alpha}(x) = \begin{cases} \alpha\left(\exp(x) - 1\right) & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{1}$$

is used as activation function for both convolutional layers and fully connected layers (with $\alpha = 1$), except for the last fully connected layer of each architecture where a *softmax* function defined in Equation (2)

$$s(y_i) = \frac{\exp(y_i)}{\sum\limits_{j} \exp(y_j)} \tag{2}$$

is used as activation function, where $y_i = elu_{\alpha}\left(\sum\limits_{k=1}^{n} w_{i,k}x_k + b_i\right)$ ($\{w_{i,k}\}_{k=1}^{n}$ represents the set of weights of the ith neuron, b_i represents the bias term of the ith neuron and $x = (x_1, \ldots, x_k, \ldots, x_n)$ represents the output of the precedent fully connected layer). The designed architectures for each physiological signal are based on 1D convolutional layers and are described in Table 1. The architectures are similar and were inspired by the architecture presented in [50] for the classification of ECG signals. The unique difference between the architectures is the usage of a dropout layer after each convolutional layer in the architecture specific to both modalities EMG and ECG.

Table 1. Deep classification architectures for each of the recorded physiological modality.

EDA				EMG & ECG			
Layer Name	**No. Kernels (Units)**	**Kernel (Pool) Size**	**Stride**	**Layer Name**	**No. Kernels (Units)**	**Kernel (Pool) Size**	**Stride**
Convolution	16	3	1	Convolution	16	11	1
Batch Normalisation	-	-	-	Batch Normalisation	-	-	-
Max Pooling	-	2	2	Max Pooling	-	2	2
Convolution	16	3	1	Dropout	-	-	-
Batch Normalisation	-	-	-	Convolution	16	11	1
Max Pooling	-	2	2	Batch Normalisation	-	-	-
Convolution	32	3	1	Max Pooling	-	2	2
Batch Normalisation	-	-	-	Dropout	-	-	-
Max Pooling	-	2	2	Convolution	32	11	1
Convolution	32	3	1	Batch Normalisation	-	-	-
Batch Normalisation	-	-	-	Max Pooling	-	2	2
Max Pooling	-	2	2	Dropout	-	-	-
Convolution	64	3	1	Convolution	32	11	1
Batch Normalisation	-	-	-	Batch Normalisation	-	-	-
Max Pooling	-	2	2	Max Pooling	-	2	2
Convolution	64	3	1	Dropout	-	-	-
Batch Normalisation	-	-	-	Convolution	64	11	1
Max Pooling	-	2	2	Batch Normalisation	-	-	-
Convolution	128	3	1	Max Pooling	-	2	2
Batch Normalisation	-	-	-	Dropout	-	-	-
Max Pooling	-	2	2	Convolution	64	11	1
Flatten	-	-	-	Batch Normalisation	-	-	-
Fully Connected	1024	-	-	Max Pooling	-	2	2
Dropout	-	-	-	Dropout	-	-	-
Fully Connected	512	-	-	Convolution	128	11	1
Dropout	-	-	-	Batch Normalisation	-	-	-
Fully Connected	c	-	-	Max Pooling	-	2	2
				Flatten	-	-	-
				Dropout	-	-	-
				Fully Connected	1024	-	-
				Dropout	-	-	-
				Fully Connected	512	-	-
				Dropout	-	-	-
				Fully Connected	c	-	-

ELU is used as activation function in both convolutional and fully connected layers, except for the last fully connected layer where a *softmax* activation function is used. The networks are further regularised by using *dropout* layers with a fixed dropout rate of 25%.

2.4. Multi-Modal Deep Model Description

To further investigate the compatibility of the recorded physiological data, several fusion approaches based on CNNs are proposed. The information stemming from each modality is aggregated at different levels of abstraction.

The first approach depicted in Figure 3 consists of an early fusion method, where the aggregation is done at the lowest level of abstraction, which consists of the preprocessed raw signals (input data). A 2D representation of the input data is generated by concatenating the three physiological modalities along the temporal axis, resulting in a tensor with the dimensionality $3 \times 1152 \times 1$. The resulting data are subsequently fed into a network consisting of 2D convolutional layers. The motivation behind such an approach is to enable the architecture to dynamically learn an appropriate set of weights, which will generate feature maps consisting of relevant and compatible information extracted simultaneously from the recorded modalities, when applied to the 2D data representation. The designed fusion architecture is described in Table 2.

Table 2. Early fusion deep CNN architecture.

Layer Name	No. Kernels (Units)	Kernel (Pool) Size	Stride
Convolution	16	2 × 11	1 × 1
Convolution	16	2 × 11	1 × 1
Batch Normalisation	-	-	-
Max Pooling	-	1 × 2	1 × 2
Dropout	-	-	-
Convolution	32	1 × 11	1 × 1
Batch Normalisation	-	-	-
Max Pooling	-	1 × 2	1 × 2
Dropout	-	-	-
Convolution	32	1 × 11	1 × 1
Batch Normalisation	-	-	-
Max Pooling	-	1 × 2	1 × 2
Dropout	-	-	-
Convolution	64	1 × 11	1 × 1
Batch Normalisation	-	-	-
Max Pooling	-	1 × 2	1 × 2
Dropout	-	-	-
Convolution	64	1 × 11	1 × 1
Batch Normalisation	-	-	-
Max Pooling	-	1 × 2	1 × 2
Flatten	-	-	-
Dropout	-	-	-
Fully Connected	1024	-	-
Dropout	-	-	-
Fully Connected	512	-	-
Dropout	-	-	-
Fully Connected	c	-	-

The architecture is based on 2D convolutional layers. A 2D representation of the input data is generated by concatenating the three physiological modalities resulting in a tensor with the dimensionality $3 \times 1152 \times 1$. Similar to the previous architectures (see Table 1), ELU is used as activation function in both convolutional and fully connected layers, except for the last fully connected layer where a *softmax* activation function is used. The network is further regularised by using *dropout* layers with a fixed dropout rate of 25%.

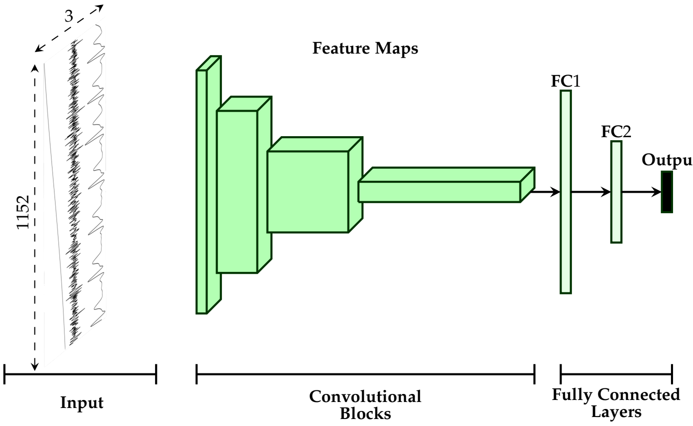


Figure 3. Early Fusion Architecture. A 2D representation of the input data is generated by concatenating the three physiological modalities and is subsequently fed into the designed deep architecture.

Furthermore, two additional late fusion approaches are proposed (see Figure 4). Both approaches are based on the uni-modal CNN architectures described earlier (see Section 2.3). The first approach described in Figure 4a performs the aggregation of the information at the mid-level since it involves using intermediate representations of the input data. It consists of concatenating the outputs of the second fully connected layer of each modality specific architecture and feeding the resulting representation to an output layer with a *softmax* activation function. The second approach depicted in Figure 4b performs the aggregation at the highest level of abstraction, since it involves using the respective *softmax* layers' outputs of each modality specific architecture. An additional layer consisting of a set of trainable positive parameters $(\alpha_1, \alpha_2, \alpha_3) \in \mathbb{R}^3_{\geq 0}$ with a *linear* activation function is directly connected to the outputs of each uni-modal architecture.

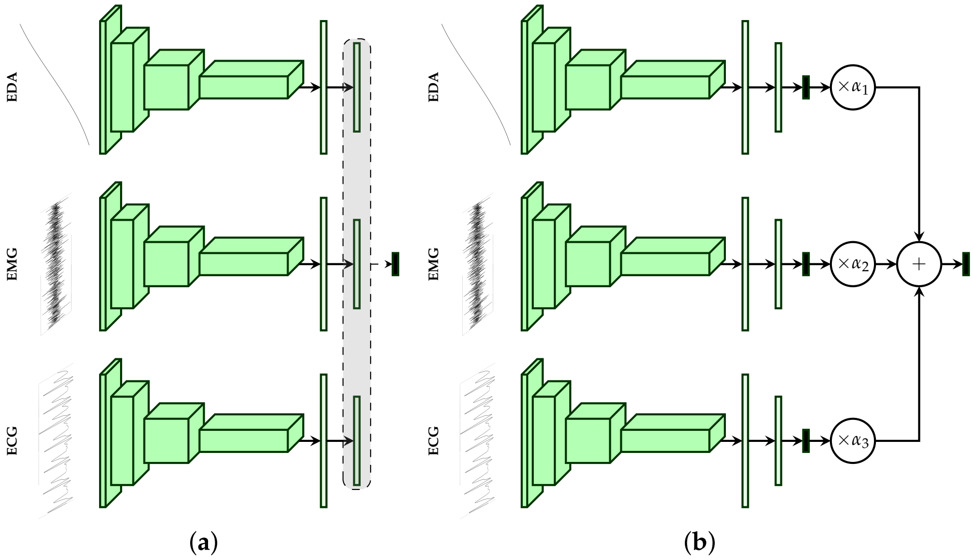


Figure 4. Late Fusion Architectures. (**a**) The features extracted by the second fully connected layer are concatenated and fed into the output layer. (**b**) The final output consists of a weighted average of the outputs of each uni-modal model.

For each modality specific architecture $i \in \{1,2,3\}$ (since we are dealing with three physiological modalities), let $\{\theta_{i,j} \in [0,1] : 1 \leq j \leq \text{c}\}$ be the output values of the respective *softmax* layers. The output of the aggregation layer is computed by using the following formulas:

$$e_j = \frac{1}{3}\left(\sum_{i=1}^{3} \alpha_i \theta_{i,j}\right), \text{ with the constraint: } \sum_{i=1}^{3} \alpha_i = 1 \tag{3}$$

$$s(e_j) = e_j \tag{4}$$

First, a weighted average output of the class probabilities stemming from the uni-modal architectures is computed (see Equation (3)), and the corresponding class probabilities of the fusion architecture are subsequently deducted by applying a *linear* activation function on the previously computed scores (see Equation (4)). Furthermore, the whole architecture is trained by using the loss function defined in Equation (5),

$$L = \sum_{i=1}^{3} \lambda_i L_i + \lambda_{agg} L_{agg} \tag{5}$$

where L_1, L_2 and L_3 are the loss functions of each modality specific architecture and L_{agg} is the loss function of the aggregation layer. The parameters λ_1, λ_2, λ_3 and λ_{agg} are the corresponding weights for each of the loss functions. Once the architecture has been trained, unseen samples are classified based uniquely on the output of the aggregation layer. All described fusion approaches are subsequently trained in an *end-to-end* manner, which means that the fusion parameters are optimised at the same time as the parameters of each modality specific classification architecture. Furthermore, the parameters of each described architecture (uni-modal as well as multi-modal) are optimised using the cross entropy loss function defined in Equation (6),

$$Loss = -\sum_{j=1}^{c} y_j \log(\hat{y}_j) \tag{6}$$

where y_j is the ground-truth value of the jth class and $\hat{y}_j$ is the jth output value of the *softmax* function. Concerning the second late fusion architecture, the cross entropy loss function is used for each uni-modal architecture as well as for the aggregation layer ($L_1 = L_2 = L_3 = L_{agg} = Loss$).

3. Results

All previously described deep architectures are trained using the Adaptive Moment estimation (*Adam*) [51] optimisation algorithm with a fixed learning rate set empirically to 10^{-5}. The training process consisted of 100 epochs with the batch size set to 100. The weights of the loss function for the second late fusion architecture (see Figure 4b) were empirically set as follows: $\lambda_1 = \lambda_2 = \lambda_3 = 0.2$, $\lambda_{agg} = 0.4$. The weight corresponding to the aggregation layer (λ_{agg}) was set higher than the others to push the network to focus on the weighted combination of the single modality architectures' outputs, and therefore to evaluate an optimal set of the weighting parameters $\{\alpha_1(EDA), \alpha_2(EMG), \alpha_3(ECG)\}$. The implementation and evaluation of the described algorithms was done with the libraries Keras [52], Tensorflow [53] and Scikit-learn [54]. The evaluation of the architectures was performed in a *Leave-One-Subject-Out* (LOSO) cross-validation setting, which means that 87 experiments were conducted. During each experiment, the data specific to a single participant were used to evaluate the performance of the trained deep model and were never seen during the optimisation of this specific deep model. The data specific to each single participant were therefore used once as an unseen test set, and the results depicted in this section consist of averaged performance metrics from a set of 87 performance values.

A performance evaluation of the designed architectures in a binary classification task consisting of the discrimination between the baseline temperature T_0 and the pain tolerance temperature T_4 is reported in Table 3. The achieved results based on CNNs are also compared to the state-of-the-art results reported in previous works. At a glance, the designed deep learning architectures outperform the state-of-the-art results in every setting, except for the ECG modality. Regarding the aggregation of all physiological modalities, the second late fusion architecture performs best and sets a new state-of-the-art fusion performance with an average accuracy of 84.40%, which even outperforms the best fusion results reported in [41], where the authors could achieve an average classification performance of 83.1% by using both physiological and video features.

Table 3. Performance comparison to early work on the BVDB (Part A) for the classification task T_0 vs. T_4 in a LOSO cross-validation setting.

Method	ECG	EMG	EDA	Fusion
Werner et al. [36]	**62.00**	57.90	73.80	74.10
Kächele et al. [40,41]	53.90	58.51	81.10	82.73
Our Approaches (CNNs)	57.04 ± 11.58	**58.65 ± 13.82**	**84.57 ± 14.13**	Early Fusion: 82.79 ± 15.22 Late Fusion (a): 83.39 ± 15.54 **Late Fusion (b): 84.40 ± 14.43**

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation (the standard deviation of the cross-validation results for the proposed approaches is also provided). The best performing approach for each modality and the aggregation of all modalities is depicted in bold.

The deep architecture based on the EDA modality significantly outperforms all previously reported classification results with an average accuracy of 84.57%.

Based on these findings, further classification experiments were conducted, based on each physiological modality and also the best performing fusion architecture (Late Fusion (b)). The performance evaluation of the conducted experiments consisting of several binary classification experiments and a multi-class classification experiment is summarised in Table 4.

EDA significantly outperforms both EMG and ECG in all conducted classification experiments and constitutes the best performing single modality, which is consistent with the results reported in previous works. Both EMG and ECG depict similar classification performances and also perform poorly for almost all classification experiments. The discrimination between the baseline temperature T_0 and the pain threshold temperature T_1, as well as the two intermediate temperatures T_2 and T_3, constitute very difficult classification experiments that both modalities are unable to perform successfully. However, the classification performances of both modalities for the classification tasks T_0 vs. T_4 and T_1 vs. T_4 are significantly above chance level, which shows that higher temperatures of elicitation cause observable and measurable responses in the recorded physiological signals, that can be used to perform the classification tasks at a certain degree of satisfaction. However, the overall performance of the fusion architecture is greatly affected by the significantly poor performance of both ECG and EMG in comparison to EDA. As can be seen in Table 4, the EDA classification architecture outperforms the fusion architecture in almost all classification experiments (but not significantly), except for the classification task T_1 vs. T_4 and the multi-class classification task (the performance improvement of the fusion architecture is however not significant).

Table 4. CNN Classification performance on the BVDB (Part A) in a LOSO cross-validation setting (the multi-class classification task corresponds to the five-class classification task T_0 vs. T_1 vs. T_2 vs. T_3 vs. T_4).

Task	ECG	EMG	EDA	Late Fusion (b)
T_0 vs. T_1	49.71 ± 06.90	49.71 ± 02.77	**61.67 ± 12.54** †	61.15 ± 12.22 ^{a,b}
T_0 vs. T_2	50.72 ± 07.30	50.29 ± 03.60	**66.93 ± 16.19** †	66.81 ± 15.92 ^{a,b}
T_0 vs. T_3	52.87 ± 09.32	53.25 ± 08.93	**76.38 ± 14.70** †	76.29 ± 14.62 ^{a,b}
T_0 vs. T_4	57.04 ± 11.58	58.65 ± 13.82	**84.57 ± 14.13** †	84.40 ± 14.43 ^{a,b}
T_1 vs. T_4	58.07 ± 12.36	58.79 ± 12.08	76.61 ± 15.38	**76.72 ± 15.02** ^{a,b,†}
Multi-Class	23.23 ± 05.62	22.85 ± 05.65	36.25 ± 09.01	**36.54 ± 08.55** ^{a,b,†}

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation (the standard deviation of the cross-validation results is also provided). The best performing approach for each classification task is depicted in bold. We also performed a significance test between the fusion approach and each single modality, using a Wilcoxon signed rank test with a significance level of 5%: $(^{a})$ indicates a significant performance improvement between EMG and the fusion approach; $(^{b})$ indicates a significant performance improvement between ECG and the fusion approach; and $(^{\dagger})$ indicates no significant improvement between EDA and the fusion approach.

The information stemming from both modalities EMG and ECG harms the optimisation process of the fusion architecture due to its inconsistency. However, it can be seen in Figure 5 that the fusion architecture is able to detect the sources of inconsistent information and dynamically adapt by systematically assigning higher weight values to EDA, while both ECG and EMG are assigned significantly lower weight values for all conducted classification tasks, and therefore improving the generalisation ability of the fusion architecture.

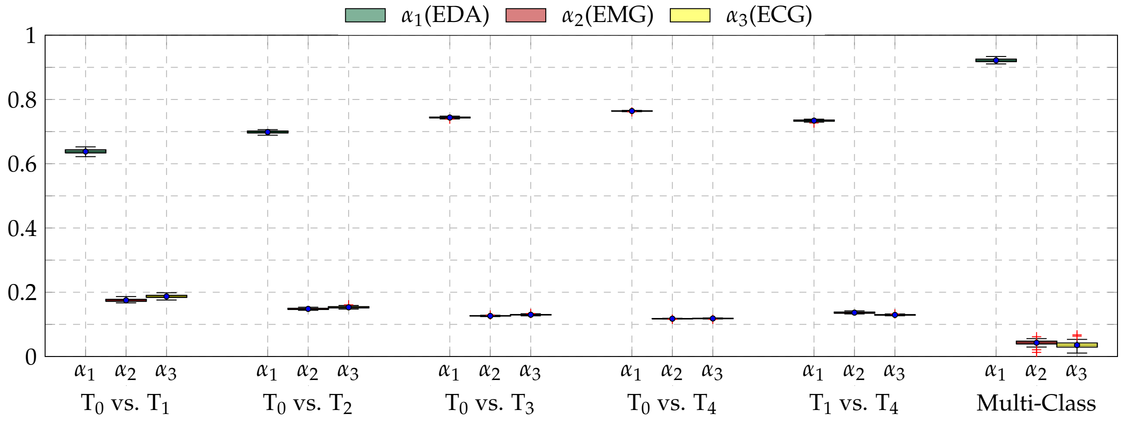


Figure 5. Box plots of the weighting parameters α_1, α_2 and α_3 for the late fusion architecture (Late Fusion (b)), computed during the LOSO cross-validation evaluation of each conducted classification experiment. Within each box plot, the mean and median values of the performed LOSO cross-validation evaluation are depicted with a dot and a horizontal line, respectively.

Subsequently, the performance of both EDA and late fusion architectures were further evaluated using different performance measures. In the case of binary classification experiments, *true positives* (*tp*) correspond to the number of correct acceptances, *false positives* (*fp*) correspond to the number of false acceptances, *true negatives* (*tn*) correspond to the number of correct rejections and *false negatives* (*fn*) correspond to the number of false rejections. These four values stem from the confusion matrix of an evaluated inference model and are used to define different performance measures. Those used for the current evaluation of the designed classification architectures are defined in Table 5.

The performance evaluation of the EDA architecture is depicted in Figure 6, while the performance evaluation of the fusion architecture is depicted in Figure 7. Considering binary classification experiments, both architectures are able to consistently discriminate between the baseline temperature T_0 and the other temperatures of pain elicitation. However, the performance of both architectures with regards to the five-class classification experiment suggests that the discrimination between all five levels of pain elicitation is a very challenging classification task. While the overall accuracy of each architecture is significantly above random performance (which is 20% in the case of a five-class classification task), the discrimination of the intermediate levels of pain elicitation remains very difficult, as can be seen in Figure 8. Both baseline and pain tolerance temperatures T_0 and T_4 can be classified with a relatively good performance. The classification performance of T_2 is barely above random performance and both T_1 and T_3 are mostly confused with T_0 and T_4, respectively. These results are however consistent with previous works on the same dataset.

Table 5. Classification performance measures.

Measure	Binary Classification	Multi-Class Classification
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	$\frac{1}{c}\sum_{i=1}^{c}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}$
Precision	$\frac{tp}{tp+fp}$	$\frac{1}{c}\sum_{i=1}^{c}\frac{tp_i}{tp_i+fp_i}$
Recall	$\frac{tp}{tp+fn}$	$\frac{1}{c}\sum_{i=1}^{c}\frac{tp_i}{tp_i+fn_i}$
F1 score	$\frac{2\times Precision\times Recall}{Precison+Recall}$	

In the case of multi-class classification experiments: tp_i corresponds to true positives, tn_i corresponds to true negatives, fp_i corresponds to false positives and fn_i corresponds to false negatives in the confusion matrix associated with the *i*th class. Furthermore, since the dataset used for the evaluation of the performance of the designed architectures is balanced, we use the macro-averaged F1 score in the case of multi-class classification.

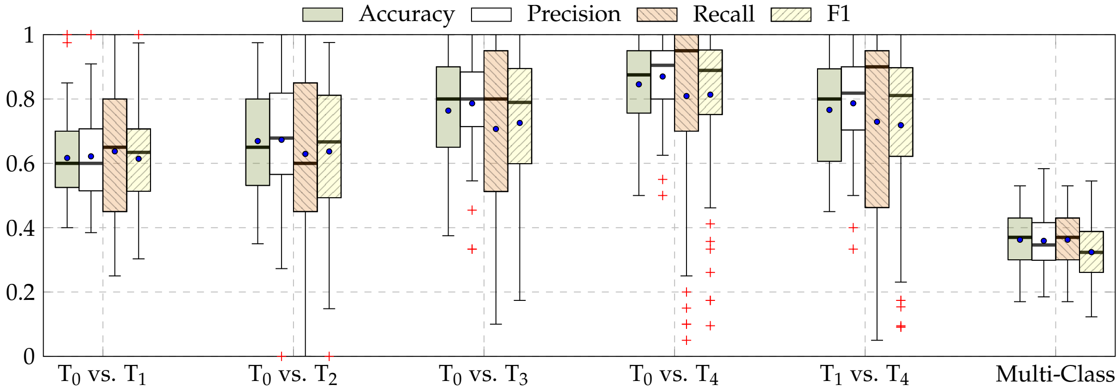


Figure 6. EDA classification performance. Within each box plot, the mean and median values of the respective performance evaluation metrics are depicted with a dot and a horizontal line, respectively.

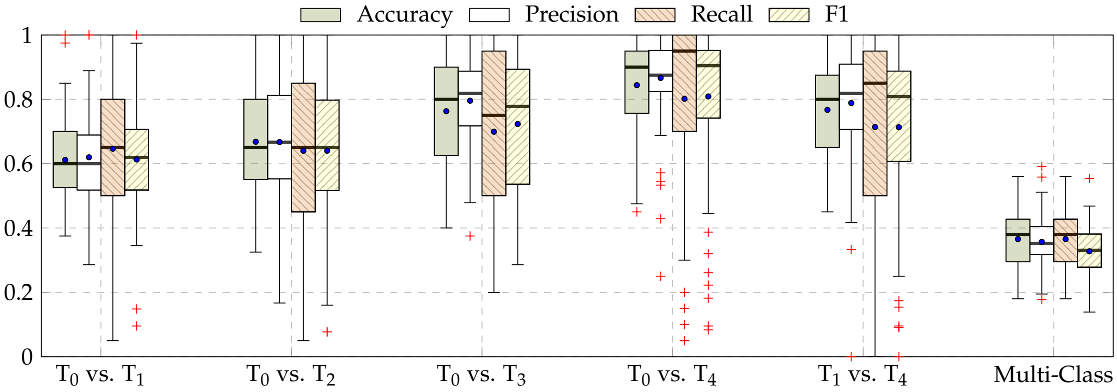
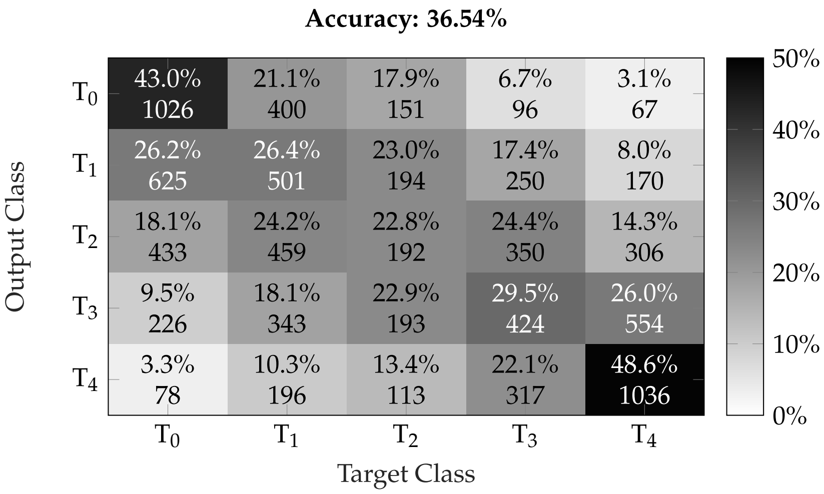


Figure 7. Late fusion classification performance (Late Fusion (b)). Within each box plot, the mean and median values of the respective performance evaluation metrics are depicted with a dot and a horizontal line, respectively.



	T_0	T_1	T_2	T_3	T_4
T_0	43.0% 1026	21.1% 400	17.9% 151	6.7% 96	3.1% 67
T_1	26.2% 625	26.4% 501	23.0% 194	17.4% 250	8.0% 170
T_2	18.1% 433	24.2% 459	22.8% 192	24.4% 350	14.3% 306
T_3	9.5% 226	18.1% 343	22.9% 193	29.5% 424	26.0% 554
T_4	3.3% 78	10.3% 196	13.4% 113	22.1% 317	48.6% 1036

Figure 8. Multi-class classification performance (confusion matrix) of the fusion architecture (Late Fusion (b)). The darker the color the higher the corresponding performance.

Table 6. EDA performance comparison to early work on the BVDB (Part A) in a LOSO cross-validation setting.

Method	**T_0 vs. T_1**	**T_0 vs. T_2**	**T_0 vs. T_3**	**T_0 vs. T_4**
Werner et al. [36]	55.40	60.20	65.90	73.80
Lopez-Martinez et al. [55]	56.44	59.40	66.00	74.21
Our Approach (CNN)	**61.67 ± 12.54**	**66.93 ± 16.19**	**76.38** ± 14.70	**84.57 ± 14.13**

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach for each classification task is depicted in bold.

We therefore compared the performance of the EDA and proposed late fusion approach to early works. For the sake of fairness, we considered the related works performed on the exact same dataset, using the exact same evaluation settings (LOSO with all 87 participants). The results depicted in Table 6 clearly show that the designed CNN architecture specific to EDA is able to consistently and significantly outperform previous approaches in all binary classification settings. Moreover, the authors of [56,57] reported overall accuracy performances of, respectively, 74.40% and 81.30% for the binary classification task T_0 vs. T_4 based uniquely on EDA. These approaches are also based on carefully designed hand-crafted features and are also significantly outperformed by the proposed CNN architecture specific to EDA.

Furthermore, we also compared the proposed late fusion approach with other fusion approaches proposed in early works. The results depicted in Table 7 show that the proposed fusion approach outperforms previous approaches for the binary classification task T_0 vs. T_4. Concerning the multi-class classification task, the proposed fusion approach also outperforms early approaches with an overall accuracy of 36.54%. The authors of [41] reported an overall accuracy of 33% with a classification model based on physiological modalities, while Werner et al. [58] reported an overall accuracy of 30.8% with a classification model based on head pose and facial activity descriptors.

Table 7. Fusion performance comparison to early work on the BVDB (Part A) in a LOSO cross-validation setting for the classification task T_0 vs. T_4.

Approach	Description	Performance
Werner et al. [58]	Early Fusion with Random Forests (Head Pose and Facial Activity Descriptors)	72.40
Werner et al. [36]	Early Fusion with Random Forests (EDA, EMG, ECG, Video)	77.80
Kächele et al. [56]	Early Fusion with Random Forests (EDA, ECG, Video)	78.90
Kächele et al. [57]	Late Fusion with Random Forests and Pseudo-inverse (EDA, EMG, ECG, Video)	83.10
Our Approach (CNN)	**Late Fusion (b) with CNNs (EDA, EMG, ECG)**	**84.40 ± 14.43**

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold.

Moreover, the designed fusion architecture was tested on the *BioVid Heat Pain Database (Part B)*. The database was generated using the same exact procedure as *Part A*. However, it consists of 86 participants and two additional EMG signals (from the corrugator and the zygomaticus muscles) were recorded. In this evaluation, we used the same signals as in *Part A* (EMG of the trapezius muscle, ECG and EDA), and used the same fusion architecture (Late fusion (b) depicted in Figure 4b). The computed results were subsequently compared with those of previous works. The corresponding results are depicted in Table 8.

Table 8. Fusion performance comparison to early work on the BVDB (Part B) in a LOSO cross-validation setting for the classification task T_0 vs. T_4.

Approach	Description	Performance
Kächele et al. [56]	Late Fusion with SVMs and Mean Aggregation (EMG (zygomaticus), EMG (corrugator), EMG (trapezius), ECG, EDA, Video)	76.60
Walter et al. [37]	Early Fusion with SVM (EMG (zygomaticus), EMG (corrugator), EMG (trapezius), ECG, EDA)	77.05
Our Approach (CNN)	**Late Fusion (b) with CNNs (EMG (trapezius), ECG, EDA)**	**79.48 ± 14.96**

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold.

The methods reported in previous works consist of fusion approaches involving all the recorded signals and based on hand-crafted features [37,56]. Although the fusion approach proposed in the current work (late fusion (b)) is based only on three of the recorded physiological signals, it is still able to outperform the previously proposed approaches, as depicted in Table 8. Therefore, it is believed that the performance of the architecture can be further improved by including the remaining signals (EMG corrugator, EMG zygomaticus, and Video) in the proposed architecture.

4. Discussion and Conclusions

This work explored the application of deep neural networks for pain intensity classification based on physiological data including ECG, EMG and EDA. Several CNN architectures, based on 1D and 2D convolutional layers, were designed and assessed based on the *BioVid Heat Pain Database (Part A)*. Furthermore, several deep fusion architectures were also proposed for the aggregation of relevant information stemming from all involved physiological modalities. The proposed architecture specific to EDA significantly outperformed the results presented in previous works in all classification settings. For the classification task T_0 vs. T_4, EDA achieved a state-of-the-art average accuracy of 84.57%. The proposed late fusion approach based on a weighted average of each modality specific model's output also achieved state-of-the-art performances (average accuracy of 84.40% for the classification task T_0 vs. T_4), but was unable to significantly outperform the deep model based uniquely on EDA.

Moreover, all designed architectures were trained in an *end-to-end* manner. Therefore, it is believed that the pre-training and fine tuning at different levels of abstraction of the CNN architectures, as well as the combination with recurrent neural networks (in order to include the temporal aspect of the physiological signals in the inference model), could potentially improve the performance of the current system, since such approaches have been successfully applied in other domains of application such as facial expression recognition [59–61]. Finally, the recorded video data provide an additional channel that can be integrated into the fusion architecture in order to improve the performance of the whole system. Therefore, the video modality should also be evaluated and assessed in combination with the physiological modalities.

In summary, the performed assessment suggests that deep learning approaches are relevant for the inference of pain intensity based on 1D physiological data, and such methods are able to significantly outperform traditional approaches based on hand-crafted features. Domain expert knowledge could be bypassed by enabling the designed deep architecture to learn relevant features from the data. In the future iterations of the current work, approaches consisting of combining both learned and hand-crafted features should be addressed. In addition, the designed architectures should be also assessed by replacing the classification experiments by regression experiments. Additionally, several data transformation approaches applied to the recorded 1D physiological data in order to generate 2D visual representations (e.g., spectrograms) should also be investigated in combination with established deep neural network approaches, specifically designed for this type of data representation.

Author Contributions: Conceptualisation, P.T. and F.S.; Methodology, P.T.; Software, P.T.; Validation, P.T. and P.B.; Formal Analysis, P.T.; Investigation, P.T. and F.S.; Writing—Original Draft Preparation, P.T. and P.B.; Writing—Review and Editing, P.T., P.B., H.A.K. and F.S.; Visualisation, P.T.; Supervision, H.A.K. and F.S.; Project Administration, H.A.K. and F.S.; and Funding Acquisition, H.A.K. and F.S.

Funding: The research leading to these results received funding from the Federal Ministry of Education and Research (BMBF, SenseEmotion) to F.S. (BMBF, e:Med, CONFIRM, ID 01ZX1708C) to H.A.K., and the Ministry of Science and Education Baden-Württemberg (Project ZIV) to H.A.K.

Acknowledgments: Peter Bellmann is supported by a scholarship of the Landesgraduiertenförderung Baden-Württemberg at Ulm University. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
4. Simonyan, K.; Zisserman, A. Very Deep Convolution Networks for Large-Scale Image Recognition. In Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 730–734.
5. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan Dumitru abd Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
6. He, K.; Zhang, X.; Ren, S.; Sun, J.A. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
7. Mohd Kamarufin, J.A.; Abdullah, A.; Sallehuddin, R. A Review of Deep Learning Architectures and Their Application. In *Modeling, Design and Simulation of Systems*; Springer: Singapore, 2017; pp. 83–94.
8. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohammed, A.r.; Jaitly, N.; Senior, A.; Nguyen, P.; Sainath, T.N.; Kingsbury, B. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared View of Four Research Groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
9. Zhang, Z.; Geiger, J.; Pohjalainen, J.; Mousa, A.E.D.; Jin, W.; Schuller, B. Deep Learning for Environmentallly Robust Speech Recognition: An Overview of Recent Developments. *ACM Trans. Intell. Syst. Technol.* **2018**, *9*, 49:1–49:28. [CrossRef]
10. Costa-jussà, M.R. From Feature to Paradigm: Deep Learning in Machine Translation. *J. Artif. Intell. Res.* **2018**, *61*, 947–974. [CrossRef]
11. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
12. Martinez, H.P.; Bengio, Y.; Yannakakis, G.N. Learning Deep Physiological Models of Affect. *IEEE Comput. Intell. Mag.* **2013**, *8*, 20–33. [CrossRef]
13. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and application in vision. In Proceedings of the IEEE International Symposium on Circuits and Systems, Paris, France, 30 May–2 June 2010; pp. 253–256.
14. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and Composing Robust features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
15. Hecht-Nielsen, R. Theory of the backpropagation neural network. In Proceedings of the International Joint Conference on Neural Networks, Washington, DC, USA, 1989; pp. 593–605.
16. Zhong, Y.; Mengyuan, Z.; Yongxiong, W.; Jingdong, Y.; Jianhua, Z. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Comput. Methods Programs Biomed.* **2017**, *140*, 93–110.
17. Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-González, G.; Abdulhay, E.; Arunkumar, N. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access* **2018**, *7*, 57–67. [CrossRef]
18. Kanjo, E.; Younis, E.M.; Ang, C.S. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Inf. Fusion* **2019**, *49*, 46–56. [CrossRef]
19. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
20. Geng, W.; Du, Y.; Jin, W.; Wei, W.; Hu, Y.; Li, J. Gesture recognition by instantaneous surface EMG images. *Sci. Rep.* **2016**, *6*, 36571. [CrossRef] [PubMed]

21. Xing, K.; Ding, Z.; Jiang, S.; Ma, X.; Yang, K.; Yang, C.; Li, X.; Jiang, F. Hand Gesture Recognition Based on Deep Learning Method. In Proceedings of the IEEE Third International Conference on Data Science in Cyberspace, Guangzhou, China, 18–21 June 2018; pp. 542–546.
22. Zhai, X.; Jelfs, B.; Chan, R.H.M.; Tin, C. Self-Recalibrating Surface EMG Pattern Recognition for Neuroprosthesis Control Based on Convolutional Neural Network. *Front. Neurosci.* **2017**, *11*, 379. [CrossRef] [PubMed]
23. Zia ur Rehman, M.; Waris, A.; Gilani, S.O.; Jochumsen, M.; Niazi, I.K.; HJamil, M.; Farina, D.; Kamavuako, E.N. Multiday EMG-Based Classification of Hand Motions with Deep Learning Techniques. *Sensors* **2018**, *18*, 2497. [CrossRef]
24. Tayeb, Z.; Fedjaev, J.; Ghaboosi, N.; Richter, C.; Everding, L.; Qu, X.; Wu, Y.; Cheng, G.; Conradt, J. Validating Deep Neural Networks for Online Decoding of Motor Imagery Movements from EEG Signals. *Sensors* **2019**, *19*, 210. [CrossRef]
25. Faust, O.; Hagiwara, Y.; Hong, T.J.; Lih, O.S.; Acharya, U.R. Deep Learning for Healthcare Applications based on Physiological Signals: A Review. *Comput. Methods Programs Biomed.* **2018**, *161*, 1–13. [CrossRef]
26. Ganapathy, N.; Swaminathan, R.; Deserno, T.M. Deep Learning on 1-D Biosignals: A Taxonomy-based Survey. *Yearb. Med. Inform.* **2018**, *27*, 98–109. [CrossRef]
27. Lim, H.; Kim, B.; Noh, G.J.; Yoo, S.K. A Deep Neural Network-Based Pain Classifier Using a Photoplethysmography Signal. *Sensors* **2019**, *19*, 384. [CrossRef]
28. Abe, S. *Support Vector Machines for Pattern Classification*; Springer: London, UK, 2005.
29. Jiang, M.; Mieronkoski, R.; Syrjälä, E.; Anzanpour, A.; Terävä, V.; Rahmani, A.M.; Salanterä, S.; Aantaa, R.; Hagelberg, N.; Liljeberg, P. Acute pain intensity monitoring with the classification of multiple physiological parameters. *J. Clin. Monit. Comput.* **2019**, *33*, 493–507. [CrossRef]
30. Alazrai, R.; AL-Rawi, S.; Alwanni, H.; Daoud, M.I. Tonic Cold Pain Detection Using Choi-Williams Time-Frequency Distribution Analysis of EEG Signals: A Feasibility Study. *Appl. Sci.* **2019**, *9*, 3433. [CrossRef]
31. Thiam, P.; Kessler, V.; Amirian, M.; Bellmann, P.; Layher, G.; Zhang, Y.; Velana, M.; Gruss, S.; Walter, S.; Traue, H.C.; et al. Multi-modal Pain Intensity Recognition based on the SenseEmotion Database. *IEEE Trans. Affect. Comput.* **2019**, 1. [CrossRef]
32. Thiam, P.; Schwenker, F. Multi-modal data fusion for pain intensity assessement and classification. In Proceedings of the 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), Montreal, QC, Canada, 28 November–1 December 2017; pp. 1–6.
33. Velana, M.; Gruss, S.; Layher, G.; Thiam, P.; Zhang, Y.; Schork, D.; Kessler, V.; Gruss, S.; Neumann, H.; Kim, J.; et al. The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Pain- and Emotion-Recognition System. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*; Springer International Publishing: Cham, Switzerland, 2017; pp. 127–139.
34. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
35. Kessler, V.; Thiam, P.; Amirian, M.; Schwenker, F. Pain recognition with camera photoplethysmography. In Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), Montreal, QC, Canada, 28 November–1 December 2017; pp. 1–5.
36. Werner, P.; Al-Hamadi, A.; Niese, R.; Walter, S.; Gruss, S.; Traue, H.C. Automatic Pain Recognition from Video and Biomedical Signals. In Proceedings of the International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24–28 August 2014; pp. 4582–4587.
37. Walter, S.; Gruss, S.; Limbrecht-Ecklundt, K.; Traue, H.C.; Werner, P.; Al-Hamadi, A.; Diniz, N.; Silva, G.M.; Andrade, A.O. Automatic Pain Quantification using Autonomic Parameters. *Psychol. Neurosci.* **2014**, *7*, 363–380. [CrossRef]
38. Gruss, S.; Treister, R.; Werner, P.; C. Traue, H.; Crawcour, S.; Andrade, A.; Walter, S. Pain Intensity Recognition Rates via Biopotential Feature Patterns with Support Vector Machines. *PLoS ONE* **2015**, *10*, 1–14. [CrossRef]
39. Walter, S.; Gruss, S.; Ehleiter, H.; Tan, J.; Traue, H.C.; Crawcour, S.; Werner, P.; Al-Hamadi, A.; Andrade, A. The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In Proceedings of the IEEE International Conference on Cybernetics, Lausanne, Switzerland, 13–15 June 2013; pp. 128–131.

40. Kächele, M.; Thiam, P.; Amirian, M.; Schwenker, F.; Palm, G. Methods for Person-Centered Continuous Pain Intensity Assessment From Bio-Physiological Channels. *IEEE J. Sel. Top. Signal Process.* **2016**, *10*, 854–864. [CrossRef]
41. Kächele, M.; Amirian, M.; Thiam, P.; Werner, P.; Walter, S.; Palm, G.; Schwenker, F. Adaptive Confidence Learning for the Personalization of Pain Intensity Estimation Systems. *Evol. Syst.* **2016**, *8*, 1–13. [CrossRef]
42. Bellmann, P.; Thiam, P.; Schwenker, F. Chapter Multi-classifier-Systems: Architectures, Algorithms and Applications. In *Computational Intelligence for Pattern Recognition*; Pedrycz, W., Chen, S.M., Eds.; Springer: Cham, Switzerland, 2018; Volume 777, pp. 83–113.
43. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
44. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
45. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Handwritten Digit Recognition with a Back-propagation Network. In Proceedings of the 2nd International Conference on Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1990; pp. 396–404.
46. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
47. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
48. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
49. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Neural Network Learning by Exponential Linear Units (ELUs). In Proceedings of the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
50. Pyakillya, B.; Kazachenko, N.; Mikhailovsky, N. Deep Learning for ECG Classification. *J. Phys. Conf. Ser.* **2017**, *913*, 012004. [CrossRef]
51. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
52. Keras: The Python Deep Learning Library. 2015. Available online: https://keras.io (accessed on 16 October 2019).
53. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, C.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org/ (accessed on 16 October 2019).
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Lopez-Martinez, D.; Picard, R. Continuous Pain Intensity Estimation from Autonomic Signals with Recurrent Neural Networks. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 5624–5627.
56. Kächele, M.; Werner, P.; Walter, S.; Al-Hamadi, A.; Schwenker, F. Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity. In *Multiple Classifier Systems (MCS)*; Springer: Cham, Switzerland, 2015; pp. 220–230.
57. Kächele, M.; Thiam, P.; Amirian, M.; Werner, P.; Walter, S.; Schwenker, F.; Palm, G. Engineering Applications of Neural Networks. In *Engineering Applications of Neural Networks, EANN 2015*; Chapter Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity; Iliadis, L., Jayne, C., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 275–285.
58. Werner, P.; Al-Hamadi, A.; Limbrecht-Ecklundt, K.; Walter, S.; Gruss, S.; Traue, H.C. Automatic Pain Assessment with Facial Activity Descriptors. *IEEE Trans. Affect. Comput.* **2017**, *8*, 286–299. [CrossRef]

59. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
60. Rodriguez, P.; Cucurull, G.; Gonzàlez, J.; Gonfaus, J.M.; Nasrollahi, K.; Moeslund, T.B.; Roca, F.X. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Trans. Cybern.* **2018**, 1–11. [CrossRef]
61. Yan, J.; Zheng, W.; Vui, Z.; Song, P. A Joint Convolutional Bidirectional LSTM Framework for Facial Expression Recognition. *IEICE Trans. Inf. Syst.* **2018**, *E101.D*, 1217–1220. [CrossRef]

I.4 Two-stream Attention Network for Pain Recognition from Video Sequences

P. Thiam, H. A. Kestler, F. Schwenker. Two-stream Attention Network for Pain Recognition from Video Sequences. Sensors 2020, 20, 839.
DOI: 10.3390/s20030839.
Status: Published.
Journal: Sensors.
Impact Factor: 3.031 (2018).
Reprinted with permission from P. Thiam, H. A. Kestler, F. Schwenker, Two-stream Attention Network for Pain Recognition from Video Sequences, Sensors, 2020, licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0/.

sensors

Article

Two-Stream Attention Network for Pain Recognition from Video Sequences

Patrick Thiam [1,2], Hans A. Kestler [1,†] and Friedhelm Schwenker [2,†,*]

[1] Institute of Medical Systems Biology, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany; patrick.thiam@uni-ulm.de (P.T.); hans.kestler@uni-ulm.de (H.A.K.)

[2] Institute of Neural Information Processing, Ulm University, James-Frank-Ring, 89081 Ulm, Germany

* Correspondence: friedhelm.schwenker@uni-ulm.de

† Equally contributing senior authors.

Received: 23 January 2020; Accepted: 2 February 2020; Published: 4 February 2020

Abstract: Several approaches have been proposed for the analysis of pain-related facial expressions. These approaches range from common classification architectures based on a set of carefully designed handcrafted features, to deep neural networks characterised by an autonomous extraction of relevant facial descriptors and simultaneous optimisation of a classification architecture. In the current work, an end-to-end approach based on attention networks for the analysis and recognition of pain-related facial expressions is proposed. The method combines both spatial and temporal aspects of facial expressions through a weighted aggregation of attention-based neural networks' outputs, based on sequences of Motion History Images (MHIs) and Optical Flow Images (OFIs). Each input stream is fed into a specific attention network consisting of a Convolutional Neural Network (CNN) coupled to a Bidirectional Long Short-Term Memory (BiLSTM) Recurrent Neural Network (RNN). An attention mechanism generates a single weighted representation of each input stream (MHI sequence and OFI sequence), which is subsequently used to perform specific classification tasks. Simultaneously, a weighted aggregation of the classification scores specific to each input stream is performed to generate a final classification output. The assessment conducted on both the *BioVid Heat Pain Database (Part A)* and *SenseEmotion Database* points at the relevance of the proposed approach, as its classification performance is on par with state-of-the-art classification approaches proposed in the literature.

Keywords: convolutional neural networks, long short-term memory recurrent neural networks, information fusion, pain recognition

1. Introduction

An individual's affective disposition is often expressed throughout facial expressions. Human beings are therefore able to assess someone's current mood or state of mind by observing his or her facial demeanour. Therefore, an analysis of facial expressions can provide some valuable insight about one's emotional and psychological state. Thus, facial expression recognition (FER) has been attracting a lot of interest from the research community in the recent decades and constitutes a steadily growing area of research, particularly in the domains of machine learning and computer vision. The current work focuses on the analysis of facial expressions for the assessment and recognition of pain in video sequences. More specifically, a two-stream attention network is designed, with the objective of combining both temporal and spatial aspects of facial expressions, based on sequences of motion history images [1] and optical flow images [2], to accurately discriminate between neutral, low, and high levels of nociceptive pain. The current work is organised as follows. An overview of pain recognition approaches based on facial expressions is provided in Section 2, followed by a thorough description of the proposed approach in Section 3. In Section 4, a description of the datasets used for the assessment of the proposed approach as well as the performed experiments is provided, followed by a description of the corresponding results. The current work is subsequently concluded in Section 5 with a short discussion and description of potential future works.

2. Related Work

Recent advances in both domains of computer vision and machine learning, combined with the release of several datasets designed specifically for pain-related research (e.g., *UNBC-McMaster Shouder Pain Expression Archive Database* [3], *BioVid Heat Pain Database* [4], *Multimodal EmoPain Database* [5] and *SenseEmotion Database* [6]), have fostered the development of a multitude of automatic pain assessment and classification approaches. These methods range from unimodal approaches, characterised by the optimisation of an inference model based on one unique and specific input signal (e.g., video sequences [7,8], audio signals [9,10] and bio-physiological signals [11–13]), to multimodal approaches that are characterised by the optimisation of an information fusion architecture based on parameters stemming from a set of distinctive input signals [14–16].

Regarding pain assessment based on facial expressions, several approaches have been proposed, ranging from conventional supervised learning techniques based on specific sets of handcrafted features, to deep learning techniques. These approaches rely on an effective preprocessing of the input signal (which in this case consists of a set of images or video sequences) and involves the localisation, alignment and normalisation of the facial area in each input frame. Moreover, further preprocessing techniques include the localisation and extraction of several fiducial points characterising specific facial landmarks, and in some cases, the continuous extraction of facial Action Units (AUs) [17,18]. The preprocessed input signal, as well as the extracted parameters, are subsequently used to optimise a specific inference model based on different methods. In [19], the authors use an ensemble of linear Support Vector Machines (SVMs) [20] (each trained on a specific AU), in which inference scores are subsequently combined using Logistical Linear Regression (LLR) [21] for the detection of pain at a frame-by-frame level. The authors in [22] apply a *k*-Nearest Neighbours (*k*-NN) [23] model on geometric features extracted from a specific set of facial landmarks for the recognition of AUs. Subsequently, the pain intensity in a particular frame is evaluated based on the detected AUs by using a pain evaluation scale provided by Prkachin and Solomon [24]. Most recently, the authors in [25] improve the performance of a pain detection system based on automatically detected AUs by applying a transfer learning approach based on neural networks to map automated AU codings to a subspace of manual AU codings. The encoded AUs are subsequently used to perform pain classification, using an Artificial Neural Network (ANN) [26].

In addition to AU-based pain assessment approaches, several techniques based on either facial texture, shape, appearance and geometry or on a combination of several of such facial descriptors have been proposed. Yang et al. [27] assess several appearance-based facial descriptors by comparing the pain classification performance of each feature with its spatio-temporal counterpart using SVMs. The assessed spatial descriptors consist of Local Binary Patterns (LBP) [28], Local Phase Quantization (LPQ) [29], Binarized Statistical Image Features (BSIF) [30] as well as each descriptor's spatio-temporal counterpart extracted from video sequences on three orthogonal planes (LBP-TOP, LPQ-TOP and BSIF-TOP). In [8,31], the authors propose several sets of spatio-temporal facial action descriptors based on both appearance- and geometry-based features extracted from both the facial area, as well as the head pose. Those descriptors are further used to perform the classification of several levels of pain intensities using a Random Forest (RF) [32] model. Similarly, the authors in [7,14,15,33], propose several spatio-temporal descriptors extracted either from the localised facial area or from the estimated head pose, including, among others, Pyramid Histograms of Oriented Gradients (PHOG) [34] and Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [35], to perform the classification of several levels of nociceptive pain. The classification experiments are also performed with RF models and ANNs.

Alongside handcrafted feature-based approaches, several techniques based on deep neural networks have also been proposed for the assessment of pain induced facial expressions. Such approaches are characterised by the joint extraction of relevant descriptors (from the preprocessed raw input data) and optimisation of an inference model, based on neural networks in an end-to-end manner. In [36–38], the authors propose a hybrid deep neural network pain detection architecture characterised by the combination of a feature embedding network consisting of a Convolutional Neural Network (CNN) [39] with a Long Short-Term Memory (LSTM) [40] Recurrent Neural Network (RNN), to take advantage of both spatial and temporal aspects of facial pain expressions in video sequences. Soar et al. [41] propose a similar approach by combining a CNN with a Bidirectional LSTM (BiLSTM), and using a Variable-State Latent Conditional Random Field (VRS-CRF) [42] instead of a conventional Multi-Layer Perceptron (MLP) to perform the classification. In [43], the authors also use a similar hybrid approach as in [36,37]; however, in this case, the feature embedding CNN is coupled to two distinct LSTM networks. The outputs of the LSTM networks are further concatenated and a MLP is used to perform the classification of the pain intensities in video sequences. Furthermore, Zhou et al. [44] propose a Recurrent Convolutional Neural Network (RCNN) [45] architecture for the continuous estimation of pain intensity in video sequences at the frame-level, whereas Wang et al. [46] propose a transfer learning approach, consisting of the regularisation of a face verification network, which is subsequently applied to a pain intensity regression task.

The current work focuses on the analysis of facial expressions for the discrimination of neutral, low and high levels of nociceptive pain in video sequences. Thereby, an end-to-end hybrid neural network characterised by the integration of spatial and temporal information at both the representational level of the input data (OFI and MHI) and the structural level of the proposed architecture (hybrid CNN-BiLSTM) is proposed. Furthermore, frame attention parameters [47] are integrated into the proposed architecture to generate an aggregated representation of the input data based on an estimation of the representativeness of each single input frame, in relation with the corresponding level of nociceptive pain. An extensive assessment of the proposed architecture is performed on both *BioVid Heat Pain Database (Part A)* [4] and *SenseEmotion Database* [6].

3. Proposed Approach

A video sequence can be characterised by both its spatial and temporal components. The spatial component represents the appearance (i.e., texture, shape and form) of each frame's content, whereas the temporal component represents the perceived motion across consecutive frames due to dynamic changes of the content's appearance through time. Most of the deep neural network approaches designed for the assessment of pain-related facial expressions generate spatio-temporal descriptors of the input data in two distinct and conjoint stages: a specific feature embedding neural network (which is commonly a pre-trained CNN) first extracts appearance based descriptors from the individual input frames (which are greyscale or colour images), and a recurrent neural network is subsequently used for the integration of the input's temporal aspect based on sequences of previously extracted appearance features, thus generating spatio-temporal representations of video sequences that are used for classification or regression tasks. Therefore, both the temporal and spatial aspects of video sequences are integrated uniquely at the structural level (e.g., the architecture of the neural network) of such approaches. The current approach extends this specific technique by additionally integrating motion information at the representational level (e.g., input data) of the architecture throughout sequences of motion history images [1] and optical flow images [2].

3.1. Motion History Image (MHI)

Introduced by Bobick and Davis [48], a MHI consists of a scalar-valued image depicting both the location and direction of motion in a sequence of consecutive images, based on the changes of pixel intensities of each image through time. The intensity of a pixel in a MHI is a function of the temporal motion history at that specific point. A MHI H_τ is computed using an update function $\Psi\left(x,y,t\right)$, and is defined as follows,

$$H_\tau\left(x,y,t\right) = \begin{cases} \tau & \text{if } \Psi\left(x,y,t\right) = 1 \\ max\left(0, H_\tau\left(x,y,t-1\right) - \delta\right) & \text{otherwise} \end{cases} \tag{1}$$

where (x,y) represents the pixel's location, t the time and τ the temporal extent of the observed motion (e.g., the length of a sequence of images); $\Psi\left(x,y,t\right) = 1$ is synonym of motion at the location (x,y) and at the time t; and δ represents a decay parameter. The update function $\Psi\left(x,y,t\right)$ is defined as follows,

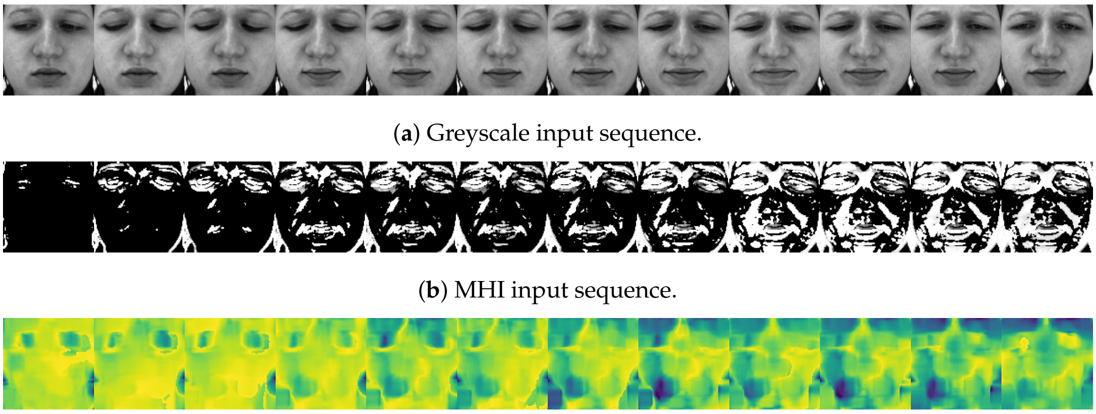
$$\Psi\left(x,y,t\right) = \begin{cases} 1 & \text{if } D\left(x,y,t\right) \geq \xi \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where ξ is a threshold; $D\left(x,y,t\right)$ represents the absolute value of the difference of pixel intensity values of consecutive frames and is defined as follows,

$$D\left(x,y,t\right) = |I(x,y,t) - I(x,y,t \pm \Delta t)| \tag{3}$$

where $I\left(x,y,t\right)$ represents the pixel intensity at the location (x,y) and at the time t; Δt represents the temporal distance between the frames.

Therefore, the computation of a MHI consists in first performing image differencing [49] between a specific, preceding frame and the current tth frame, and detecting the pixel locations where a substantial amount of movement has occurred (depending on the value assigned to the threshold ξ) based on Equation (2). Subsequently, Equation (1) is used to assign pixel values to the MHI. If a motion has been detected at the location (x,y) of the tth frame, a pixel value of τ is assigned at that location. Otherwise, the previous pixel value of that location is reduced by δ, thereby indicating the temporal occurrence of the motion at that specific location, according to the actual time t. This whole process is conducted iteratively until the entire sequence of images has been processed. The temporal history of motion is thereby encoded into the resulting MHI. Therefore, a whole sequence of images can be encoded into a single MHI. However, in the current work, a sequence of MHIs is generated from each single sequence of images by saving each single MHI generated at each single step of the iterative process described earlier. The resulting sequence of images is used as input for the designed deep neural networks. A depiction of such a sequence of MHIs can be seen in Figure 1b, with the corresponding sequence of greyscale images depicted in Figure 1a.



(**a**) Greyscale input sequence.

(**b**) MHI input sequence.

(**c**) OFI input sequence.

Figure 1. Data preprocessing. Following the detection, alignment, normalisation and extraction of the facial area in each frame of a video sequence, the images are converted into greyscale. MHI and OFI sequences are subsequently generated.

3.2. Optical Flow Image (OFI)

Optical flow refers to the apparent motion of visual features (e.g., corners, edges, textures and pixels) in a sequence of consecutive images. It is characterised by a set of vectors (optical flow vectors) defined either at each location (x, y) of an entire image (dense optical flow [50,51]), or at specific locations of a predefined set of visual features (sparse optical flow [2,52]). The orientation of an optical flow vector depicts the direction of the apparent motion, whereas the magnitude of an optical flow vector depicts the velocity of the apparent motion of the corresponding visual feature in consecutive frames. Thus, an OFI provides a compact description of the location, direction and velocity of a specific motion occurring in consecutive frames. The estimation of the optical flow is based on the brightness constancy assumption, which stipulates that pixel intensities are constant between consecutive frames. If $I\,(x, y, t)$ is the pixel intensity at the location (x, y) and at the time t, the brightness constancy assumption can be formulated as follows,

$$I\,(x, y, t) = I\,(x + \delta x, y + \delta y, t + \delta t) \tag{4}$$

where $(\delta x, \delta y, \delta t)$ represents a small motion. By applying a first-order Taylor expansion, $I\,(x + \delta x, y + \delta y, t + \delta t)$ can be written as follows,

$$I\,(x + \delta x, y + \delta y, t + \delta t) \approx I\,(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t. \tag{5}$$

Thus,

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \approx 0 \tag{6}$$

and by dividing each term by δt, the optical flow constraint equation can be written as follows,

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{dI}{dt} \approx 0. \tag{7}$$

Resolving the optical flow constraint equation (Equation (7)) consists of the estimation of both parameters $u = \frac{dx}{dt}$ and $v = \frac{dy}{dt}$. Several methods have been proposed to perform this specific task. The authors in [53,54] propose an overview of such approaches. In the current work, dense optical flow is applied, using the method of Farnebäck [50], to compute OFIs from consecutive greyscale images. A depiction of such a sequence of images can be seen in Figure 1c (both motion direction and motion velocity are color encoded).

3.3. Network Architecture

As opposed to still images, the motion component of a video sequence is integrated into both MHIs and OFIs, therefore providing more valuable information for facial expressions analysis. Therefore, the proposed architecture consists of a multi-view learning [55] neural network with both OFIs and MHIs as input channels. An overall illustration of the proposed two-stream neural network can be seen in Figure 2. In a nutshell, an attention network specific to each input channel (OFIs and MHIs) first generates a weighted representation from the jth input sequence (h_j^{ofi} and h_j^{mhi}). The generated representation is subsequently fed into a channel specific classification model (which in this case is a MLP). The resulting class probabilities of each channel ($score_j^{ofi}$ and $score_j^{mhi}$) are further fed into an aggregation layer with a linear output function, where a weighted aggregation of the provided scores is performed as follows,

$$score_j = \alpha_{ofi} \cdot score_j^{ofi} + \alpha_{mhi} \cdot score_j^{mhi} \tag{8}$$

with the constraint

$$\alpha_{ofi} + \alpha_{mhi} = 1. \tag{9}$$

The entire architecture is trained in an end-to-end manner by using the following loss function,

$$\mathcal{L} = \lambda_{ofi} \cdot \mathcal{L}_{ofi} + \lambda_{mhi} \cdot \mathcal{L}_{mhi} + \lambda_{agg} \cdot \mathcal{L}_{agg} \tag{10}$$

where the loss functions of each input channel and of the aggregation layer are respectively depicted with $\mathcal{L}_{ofi}$, $\mathcal{L}_{mhi}$ and $\mathcal{L}_{agg}$. The parameters λ_{ofi}, λ_{mhi} and λ_{agg} correspond to the regularisation parameters of each respective loss function. Once the network has been trained, unseen samples are classified based on the output of the aggregation layer.

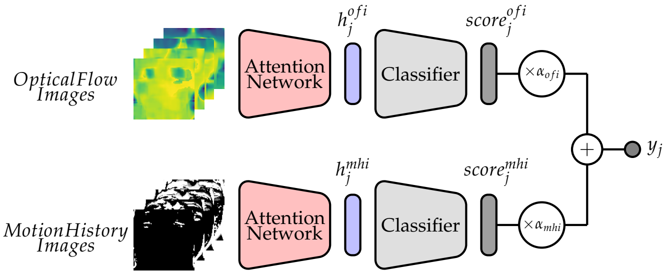


Figure 2. Two-Stream Attention Network with Weighted Score Aggregation.

The attention network (see Figure 3) consists of a CNN coupled to a BiLSTM with a frame attention module [47]. The CNN consists of a time distributed feature embedding network which takes a single facial image $im_{k,j}$ as input and generates a fixed-dimension feature representation $X_{k,j}$ specific to that image. Therefore, the output of the jth input sequence of facial images $\{im_{k,j}\}_{k=1}^{l}$ consists of a set of facial features $\{X_{k,j}\}_{k=1}^{l}$. The temporal component of the sequence of images is further integrated by using a BiLSTM layer. A BiLSTM [56] RNN is an extension of a regular LSTM [40] RNN, to enable the use of context representations in both forward and backward directions.

It consists of two LSTM layers, one processing the input sequence $\left\{X_{1,j}, X_{2,j}, \ldots, X_{l,j}\right\}$ sequentially forward in time (from $X_{1,j}$ to $X_{l,j}$) and the second processing the input sequence sequentially backward in time (from $X_{l,j}$ to $X_{1,j}$). A LSTM RNN is capable of learning long-term dependencies in sequential data, while avoiding the vanishing gradient problem of standard RNNs [57]. This is achieved throughout the use of cell states (see Figure 4), which regulate the amount of information flowing through a LSTM network throughout the use of three principal gates: forget gate (f_t), input gate (i_t) and output gate (o_t). The cell's output h_t (at each time step t) is computed, given a specific input x_t, the previous hidden state h_{t-1}, and the previous cell state C_{t-1}, as follows,

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{11}$$

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{12}$$

$$\tilde{C}_t = tanh\left(W_c s_t + U_c h_{t-1} + b_c\right) \tag{13}$$

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \tilde{C}_t \tag{14}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{15}$$

$$h_t = o_t \otimes tanh(C_t) \tag{16}$$

where σ represents the sigmoid activation function $\sigma(x) = (1 + exp(-x))^{-1}$ and $tanh$ represents the hyperbolic tangent activation function. The element-wise multiplication operator is represented by the symbol $\otimes$. The weight matrices for the input x_t are represented by W_i, W_f, W_o and W_c for the input gate, forget gate, output gate and cell state, respectively. Analogously, the weight matrices for the previous hidden state h_{t-1} for each gate are represented by U_i, U_f, U_o and U_c. The amount of information to be further propagated into the network is controlled by the forget gate (Equation (11)), the input gate (Equation (12)) and the computed cell state candidate $\tilde{C}_t$ (Equation (13)). These parameters are subsequently used to update the cell state C_t based on the previous cell state C_{t-1} (Equation (14)). The output of the cell can subsequently be computed using both Equation (15) and Equation (16).

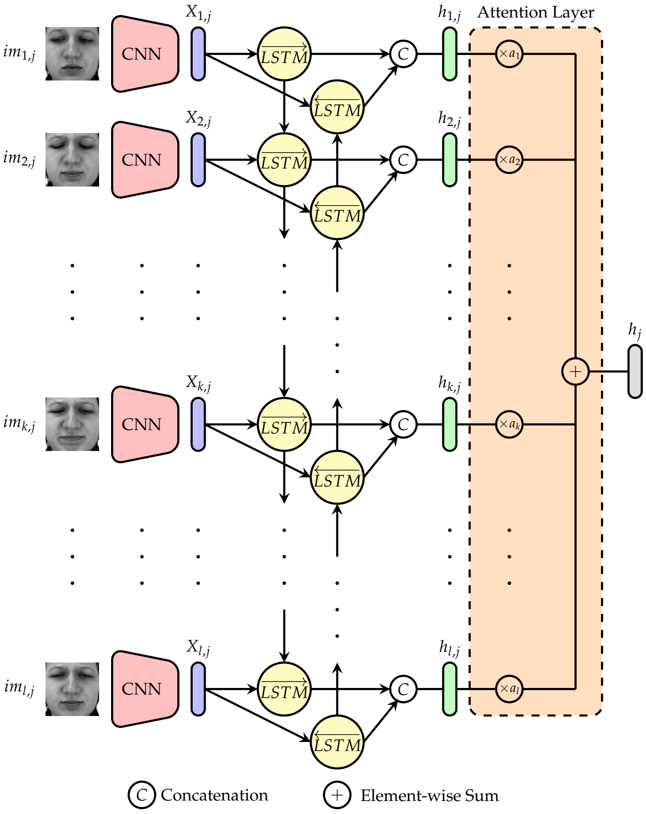


Figure 3. Attention Network.

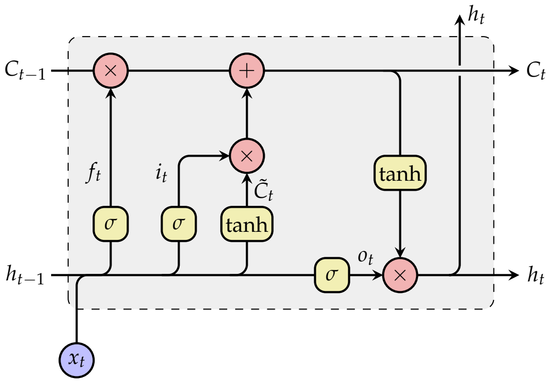


Figure 4. Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN).

In the current work, the hidden representation stemming from the forward pass $\left\{\overrightarrow{h_{1,j}}, \overrightarrow{h_{2,j}}, \ldots, \overrightarrow{h_{l,j}}\right\}$ and the one stemming from the backward pass $\left\{\overleftarrow{h_{1,j}}, \overleftarrow{h_{2,j}}, \ldots, \overleftarrow{h_{l,j}}\right\}$ are subsequently concatenated $\left\{\left[\overrightarrow{h_{1,j}}, \overleftarrow{h_{1,j}}\right], \left[\overrightarrow{h_{2,j}}, \overleftarrow{h_{2,j}}\right], \ldots, \left[\overrightarrow{h_{l,j}}, \overleftarrow{h_{l,j}}\right]\right\}$ and fed into the next layer. For the sake of simplicity, the output of the BiLSTM layer will be depicted as follows, $\left\{h_{1,j}, h_{2,j}, \ldots, h_{l,j}\right\}$ (with $h_{k,j} = \left[\overrightarrow{h_{k,j}}, \overleftarrow{h_{k,j}}\right]$). The next layer consists of an attention layer, where self-attention weights $\{a_k\}_{k=1}^{l}$ are computed and subsequently used to generate a single weighted representation of the input sequence. The self-attention weights are computed as follows,

$$\alpha_k = elu\left(W_k h_{k,j} + b_k\right) \tag{17}$$

$$a_k = \frac{exp(\alpha_k)}{\sum\limits_{i=1}^{l} exp(\alpha_i)} \tag{18}$$

where W_k are the weights specific to the input feature representation $h_{k,j} = \left[\overrightarrow{h_{k,j}}, \overleftarrow{h_{k,j}}\right]$ and elu represents the exponential linear unit activation function [58], which is defined as

$$elu_\alpha(x) = \begin{cases} \alpha \cdot (\exp(x) - 1) & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{19}$$

with $\alpha = 1$. Each self-attention weight expresses the relevance of a specific image for the corresponding emotional state expressed within the video sequence. Thereby, relevant images should be assigned significantly higher weight values as irrelevant images. The final representation of the input sequence is subsequently computed by performing a weighted aggregation of the BiLSTM output $\left\{h_{1,j}, h_{2,j}, \ldots, h_{l,j}\right\}$ based on the computed self-attention weights $\{a_k\}_{k=1}^{l}$ as follows,

$$h_j = \sum_{k=1}^{l} a_k \cdot h_{k,j} \tag{20}$$

and is further used to perform the classification task.

4. Experiments

In the following section, a description of the experiments performed for the evaluation of the proposed approach is provided. First, the datasets used for the evaluation are briefly described, followed by a depiction of the conducted data preprocessing steps. The experimental settings as well as the performed experiments are described subsequently. This section is finally concluded with a description and discussion of the experimental results.

4.1. Datasets Description

The presented approach is evaluated on both the *BioVid Heat Pain Database (Part A)* (BVDB) [4] and the *SenseEmotion Database* (SEDB) [6]. Both datasets were recorded with the principal goal of fostering research in the domain of pain recognition. In both cases, several healthy participants were submitted to a series of individually calibrated heat-induced painful stimuli, using the exact same procedure. Whereas the BVDB consists of 87 individuals submitted to four individually calibrated and gradually increasing levels of heat-induced painful stimuli (T_1, T_2, T_3 and T_4), the SEDB consists of 40 individuals submitted to three individually calibrated and gradually increasing levels of heat-induced stimuli (T_1, T_2 and T_3). Each single level of heat-induced pain stimulation was randomly elicited a total of 20 times for the BVDB and 30 times for the SEDB. Each elicitation lasted 4 s, followed by a recovery phase of a random length of 8 to 12 s during which a baseline temperature T_0 (32°C) was applied (see Figure 5). Whereas the elicitations were performed uniquely on one specific hand for the BVDB, the experiments were conducted twice for the SEDB, with the elicitation performed each time on one specific arm (left forearm and right forearm). Therefore, with the inclusion of the baseline temperature T_0, the dataset specific to the BVDB consists of a total of $87 \times 20 \times 5 = 8700$ samples, whereas the SEDB consists of a total of $40 \times 30 \times 4 \times 2 = 9600$ samples. During the experiments, the demeanour of each participant was recorded using several modalities consisting of video and bio-physiological channels for the BVDB, while the SEDB included audio, video and bio-physiological channels. The current work focuses uniquely on the video modality, and the reader should refer to the work in [10,14–16,33,59–64] for more experiments including the other recorded modalities.

4.2. Data Preprocessing

The evaluation performed in the current work is undertaken in both cases (BVDB and SEDB) on video recordings performed by a frontal camera. The recordings were performed at a frame rate of 25 frames per second (fps) for the BVDB and 30 fps for the SEDB. Furthermore, the evaluation is performed uniquely on windows of length 4.5 s with a shift of 4 s from the elicitation's onset, as proposed in [16] (see Figure 5). Once these specific windows are extracted, the facial behaviour analysis toolkit OpenFace [65] is used for the automatic detection, alignment and normalisation of the facial area (with a fixed size of 100 × 100 pixels) in each video frame. Subsequently, MHI sequences and OFI sequences are extracted using the OpenCV library [66]. Both MHIs and OFIs are generated relatively to a reference frame, which in this case is the very first frame of each video sequence. Concerning MHIs, the temporal extent parameter τ (see Equation (1)) was set to the length of the sequence of images ($25 \times 4.5 \cong 113$ frames for the BVDB and $30 \times 4.5 = 135$ frames for the SEDB). Furthermore, the threshold parameter ξ (see Equation (2)) was set to 1 to capture any single motion from two consecutive frames (in this case, the fluctuation of pixel intensities between the reference frame and the tth frame). Finally, to reduce the computational requirements, the number of samples in each sequence is reduced by sequentially selecting each second frame of an entire sequence for the BVDB (resulting in sequences with a total length of 57 frames), and each third frame of an entire sequence for the SEDB (resulting in sequences of length 45 frames). The dimensionality of the tensor input specific to the BVDB is, respectively, $(bs, 57, 100, 100, 3)$ for OFI sequences and $(bs, 57, 100, 100, 1)$ for MHI sequences (bs representing the batch size). The dimensionality of the tensor input specific to the SEDB is, respectively, $(bs, 45, 100, 100, 3)$ for OFI sequences and $(bs, 45, 100, 100, 1)$ for MHI sequences.

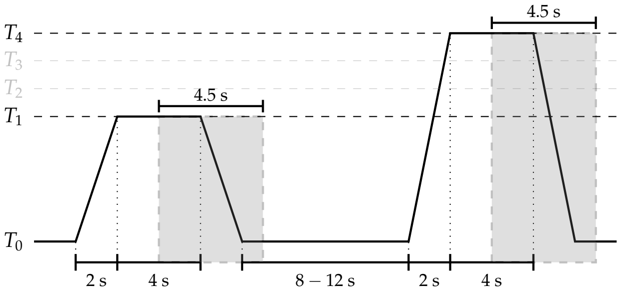


Figure 5. Video Signal Segmentation (BioVid Heat Pain Database (Part A)). Experiments are carried out on windows of length 4.5 s with a temporal shift of 4 s from the elicitations' onsets.

4.3. Experimental Settings

The evaluation performed in the current work consists of the discrimination between high and low stimuli levels. Therefore, two binary classification tasks are performed for each database: $T_0 vs. T_4$ and $T_1 vs. T_4$ for the BVDB, and $T_0 vs. T_3$ and $T_1 vs. T_3$ for the SEDB. Furthermore, the assessment of the proposed approach is conducted by applying a *Leave-One-Subject-Out* (LOSO) cross-validation evaluation, which means that a total of 87 experiments were conducted for the BVDB (40 experiments for the SEDB), during which the data specific to each participant is used once to evaluate the performance of the classification architecture optimised on the data specific to the remaining 86 participants (the data specific to 39 participants is used to optimise the architecture for the SEDB, and the data specific to the remaining participant is used to evaluate the performance of the architecture).

The feature embedding CNN used for the evaluation is adapted from the one proposed by the Visual Geometry Group of the University of Oxford *VGG16* [67]. The depth of the CNN model is substantially reduced to a total of 10 convolutional layers (instead of 13 as in the *VGG16* model), and the number of convolutional filters is gradually increased from one convolutional block to the next starting from 8 filters until a maximum of 64 filters. The activation function in each convolutional layer consists of the *elu* activation function (instead of the rectified linear unit (*relu*) activation function as in the *VGG16* model). Max-pooling and Batch Normalisation [68] are performed after each convolutional block. A detailed description of the feature embedding CNN architecture can be seen in Table 1. The coupled BiLSTM layer consists of two LSTM RNNs with 64 units each. The resulting sequence of spatio-temporal features is further fed into the attention layer in order to generate a single aggregated representation of the input sequence. The classification is further performed based on this representation and the architecture of the classification model is described in Table 2. The exact same architecture is used for the two input sequences (MHIs and OFIs). The outputs of the classifiers are further aggregated based on both Equation (8) and Equation (9). The whole architecture is subsequently trained in an end-to-end manner, using the Adaptive Moment Estimation (Adam) [69] optimisation algorithm with a fixed learning rate set empirically to 10^{-5}. The categorical cross entropy loss function is used for each network ($\mathcal{L}_{mhi} = \mathcal{L}_{ofi} = \mathcal{L}_{agg} = \mathcal{L}$), and is defined as follows,

$$\mathcal{L} = -\sum_{j=1}^{c} y_j log(\hat{y}_j) \tag{21}$$

where $\hat{y}_j$ represents the classifier's output, y_j is the ground-truth label value and $c \in \mathbb{N}$ is the number of classes for a specific classification task.

Table 1. Feature embedding CNN architecture.

Layer	No. Filters
2× Conv2D	8
MaxPooling2D	–
Batch Normalisation	–
2× Conv2D	16
MaxPooling2D	–
Batch Normalisation	–
3× Conv2D	32
MaxPooling2D	–
Batch Normalisation	–
3× Conv2D	64
MaxPooling2D	–
Batch Normalisation	–
Flatten	–

The size of the kernels is identical for all convolutional layers and is set to 3×3, with the convolutional stride set to 1×1. Max-pooling is performed after each block of convolutional layers over a 2×2 window, with a 2×2 stride.

The regularisation parameters of the loss function in Equation (10) are set as follows: $\lambda_{mhi} = \lambda_{ofi} = 0.2$ and $\lambda_{agg} = 0.6$. The value of the regularisation parameter specific to the aggregation layer's loss is set higher than the others in order to enable the architecture to compute robust aggregation weights. The whole architecture is trained for a total of 20 epoches with the batch size set to 40 for the BVDB and 60 for the SEDB. The implementation and evaluation of the whole architecture is done with the Python libraries Keras [70], Tensorflow [71] and Scikit-learn [72].

Table 2. Classifier Architecture.

Layer	No. Units
Dropout	–
Fully Connected	64
Dropout	–
Fully Connected	c

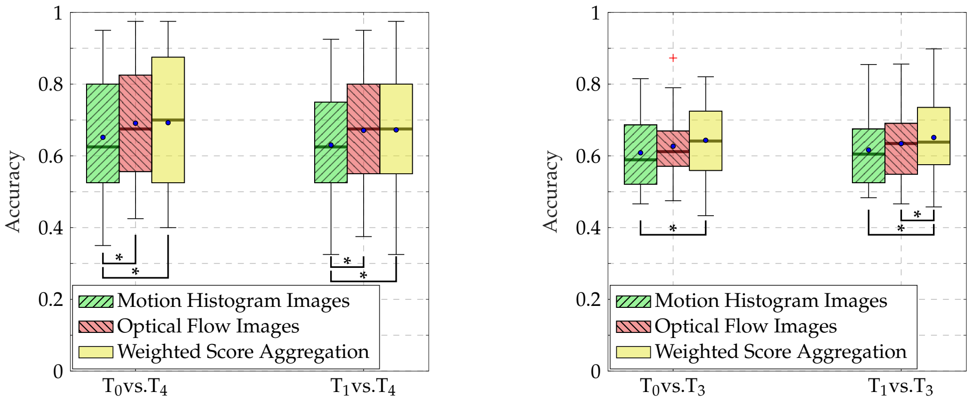
The dropout rate is empirically set to 0.25. The first fully connected layer uses the *elu* activation function, while the last fully connected layer consists of a *softmax* layer (whereby c depicts the number of classes of the classification task).

4.4. Results

The performance of the classification architectures specific to each input channel (MHIs and OFIs), as well as the performance of the weighted score aggregation approach are depicted in Figure 6. The performance metric in this case is the accuracy, which is defined as

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{22}$$

where tp refers to true positives, tn refers to true negatives, fp refers to false positives and fn refers to false negatives (since we are dealing with a binary classification task with two balanced datasets). For both datasets and both classification tasks, the aggregation approach significantly outperforms the classification architecture based uniquely on MHIs. Furthermore, the classification architecture based uniquely on OFIs outperforms the one based on MHIs for both databases and both classification tasks, with significant performance improvement in the case of the BVDB. The aggregation approach also performs slightly better than the architecture based uniquely on OFIs for both databases, although not significantly in most cases. The only significant performance improvement is achieved for the classification task T_1 vs. T_4 for the SEDB. However, the performance of both channel specific architectures and the performance of the score aggregation approach are significantly higher than chance level (which is 50% in the case of binary classification tasks) pointing at the relevance of the designed approach. Furthermore, the performance of the classification architecture is improved by using both channels and performing a weighted aggregation of the scores of both channel specific deep attention models.

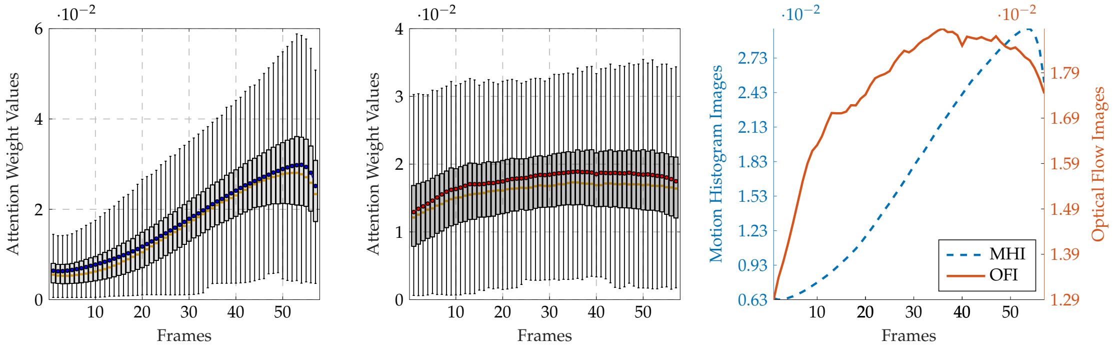


(**a**) BioVid Heat Pain Database (Part A).

(**b**) SenseEmotion Database.

Figure 6. Classification performance (Accuracy). An asterisk (*) indicates a significant performance improvement. The test has been conducted using a Wilcoxon signed rank test with a significance level of 5%. Within each boxplot, the mean and the median classification accuracy are depicted respectively with a dot and a horizontal line.

Moreover, to provide more insights into the self attention mechanism, the frame attention weight values computed at each evaluation step during the LOSO cross-validation evaluation process are depicted in Figure 7 for the BVDB and in Figure 8 for the SEDB (uniquely for the classification task T_0 vs. T_4, as the results for the classification task T_1 vs. T_4 are similar). The distribution of the weight values specific to the MHI deep attention models for both databases (Figure 7a,c for the BVDB, Figure 8a,c for the SEDB) is skewed left. It depicts a steady growth of weight values along the temporal axis of each sequence, with the MHIs located at the end of a sequence weighted significantly higher as the others. This is in accordance with the sequential extraction process of MHIs, as each extracted image contains more motion information as the previous one, with the last images accumulating almost the totality of motion information of an entire sequence. Therefore, concerning the actual classification task, the last MHIs are more interesting and relevant than the early images. Thus, such images should be weighted accordingly higher. The designed network is therefore capable of conducting this specific task by using self attention mechanisms.



(**a**) Motion History Images. (**b**) Optical Flow Images. (**c**) Mean Attention Weight Values.

Figure 7. BioVid Heat Pain Database (Part A): Attention network weight values for the classification task $T_0 vs. T_4$. Within each boxplot in (**a**,**b**), the mean and the median weight values are depicted, respectively, with a dot and a horizontal line. In (**c**), the average weight values are normalised between the maximum average value and the minimum average value to allow a better visualisation of the values distributions.

A similar observation can be made concerning the distribution of the weight values of OFIs (see Figure 7b,c for the BVDB, Figure 8b,c for the SEDB). Both depicted distributions are also skewed left, with gradually increasing weight values relative to the temporal axis. This shows that the recorded pain-related facial expressions for both BVDB and SEDB consist of gradually evolving facial movements, starting from a neutral facial depiction (not relevant for the actual classification task) to the apex of the facial movement (which is the most relevant frame for the depicted facial emotion) before gradually turning back to the neutral facial depiction. Therefore, the network assigns weight values according to this specific characterisation of pain-related facial movements using attention mechanisms, thus the relevance of such approaches for facial expression analysis.

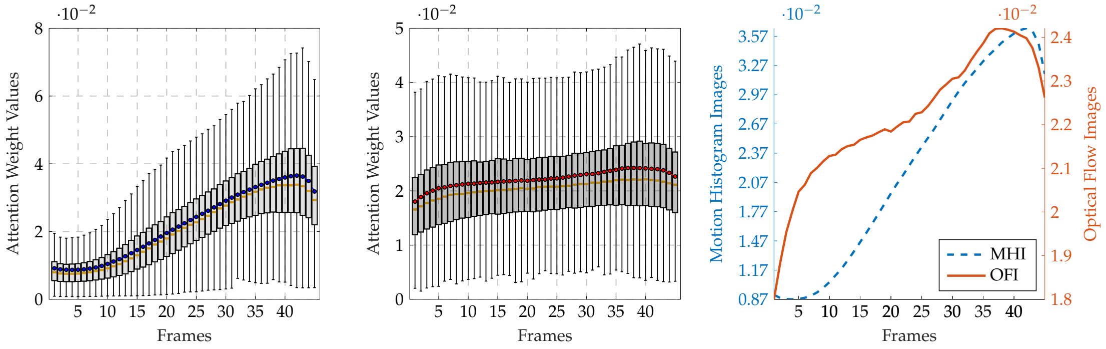
Furthermore, the performance of the weighted score aggregation approach is further assessed based on the following additional performance metrics,

$$Macro\ Precision = \frac{1}{c}\sum_{i=1}^{c}\frac{tp_i}{tp_i + fp_i} \tag{23}$$

$$Macro\ Recall = \frac{1}{c}\sum_{i=1}^{c}\frac{tp_i}{tp_i + fn_i} \tag{24}$$

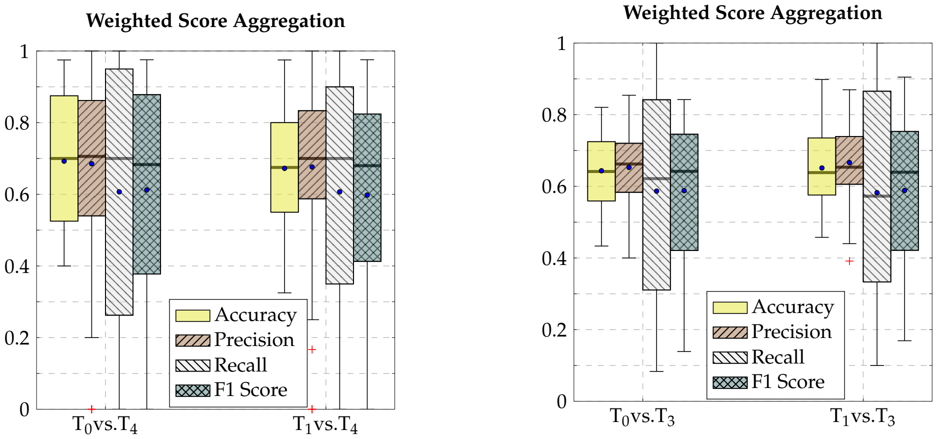
$$Macro\ F1\ Score = \frac{2 \times Macro\ Precision \times Macro\ Recall}{Macro\ Precision + Macro\ Recall} \tag{25}$$

where tp_i, fp_i and fn_i refer, respectively, to the true positives, false positives and false negatives of the ith class. The results of the evaluation are depicted in Figure 9, for both the BVDB (see Figure 9a) and the SEDB (see Figure 9b).



(**a**) Motion History Images. (**b**) Optical Flow Images. (**c**) Mean Attention Weight Values.

Figure 8. SenseEmotion Database: Attention network weight values for the classification task $T_0 vs. T_3$. Within each boxplot in (**a**,**b**), the mean and the median weight values are depicted respectively with a dot and a horizontal line. In (**c**), the average weight values are normalised between the maximum average value and the minimum average value to allow a better visualisation of the values distributions.



(**a**) BioVid Heat Pain Database (Part A). (**b**) SenseEmotion Database.

Figure 9. Weighted score aggregation classification performance. Within each box plot, the mean and median values of the respective performance evaluation metrics are depicted with a dot and a horizontal line, respectively.

These results depict a huge variance amongst all performance metrics, in particular the *Macro Recall*, which points at the fact that the classification tasks remain difficult. The evaluation on some participants yields a *Macro F1 Score* of null or nearly null, pointing at the fact that the architecture is unable to discriminate between low and high levels of pain elicitation for these specific participants. This is, however, similar and in accordance with previous works on these specific datasets. The authors of the BVDB in [73] were able to identify a set of participants who did not react to the levels of pain elicitation, therefore causing the huge variance in the classification experiments.

Finally, the performance of the weighted score aggregation approach is compared to other pain-related facial expressions classification approaches proposed in the literature. For the sake of fairness, we compare the results of the proposed approach with those results in related works which are based on the exact same dataset and were computed based on the exact same evaluation protocol (LOSO). The results are depicted in Table 3 for the BVDB and in Table 4 for the SEDB.

Table 3. Classification performance comparison to early works on the BioVid Heat Pain Database (Part A) in a LOSO cross-validation setting for the classification task $T_0 vs. T_4$.

Approach	Description	Performance
Yang et al. [27]	BSIF	65.17
Kächele et al. [31,62]	Geometric Features	65.55 ± 14.83
Werner et al. [8]	Standardised Facial Action Descriptors	**72.40**
Our Approach	Motion History Images	65.17 ± 15.49
Our Approach	Optical Flow Images	69.11 ± 14.73
Our Approach	Weighted Score Aggregation	69.25 ± 17.31

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold and the second best approach is underlined.

Table 4. Classification performance comparison to early works on the SenseEmotion Database in a LOSO cross-validation setting for the classification task $T_0 vs. T_3$.

Approach	Description	Performance
Kalischek et al. [38]	Transfer Learning	60.10 ± 00.06
Thiam et al. [15]	Standardised Geometric Features	**66.22 ± 14.48**
Our Approach	Motion Histogram Images	60.86 ± 09.81
Our Approach	Optical Flow Images	62.70 ± 09.24
Our Approach	Weighted Score Aggregation	64.35 ± 10.40

The performance metric consists of the average accuracy (in %) over the LOSO cross-validation evaluation. The best performing approach is depicted in bold and the second best approach is underlined.

In both cases, the performance of the weighted score aggregation approach is on par with the best performing approaches. However, it has to be mentioned that the authors of the best performing approaches for both the BVDB [8] and the SEDB [15] perform a subject-specific normalisation of the extracted feature representations in order to compensate for the differences in expressiveness amongst the participants. Although this specific preprocessing step has proven to significantly improve the classification performance of the architecture [61], it is not realistic as it requires that the whole testing set is already available beforehand. The normalisation parameters should be learned on the available training material and subsequently applied to the testing material during the inference phase. Nevertheless, the proposed approach based on the weighted aggregation of the scores of both MHI- and OFI-specific deep attention models generalises well and is capable of achieving state-of-the-art classification performances.

5. Conclusion

In the current work, an approach based on a weighted aggregation of the scores of two deep attention networks based, respectively, on MHIs and OFIs has been proposed and evaluated for the analysis of pain-related facial expressions. The assessment performed on both BVDB and SEDB shows that the proposed approach is capable of achieving state-of-the-art classification performances and is on par with the best performing approaches proposed in the literature. Moreover, the visualisation of the weight values stemming from the implemented attention mechanism shows that the network is capable of identifying relevant frames in relation with the current level of pain elicitation depicted by a sequence of images, by assigning significantly higher values to the most relevant images in comparison to the weight values of irrelevant images. Furthermore, as the proposed architecture was trained from scratch in an end-to-end manner, it is believed that transfer learning, in particular, for the feature embedding CNN used to generate the feature representation of each frame, could potentially improve the performance of the whole architecture. Such an analysis was not conducted in the current work, as the optimisation of the presented approach was not the goal of the conducted experiments, but rather the assessment of such an architecture for the analysis of pain-related facial expressions. Moreover, a multi-stage training strategy could also potentially improve the overall performance of the architecture, as the end-to-end trained approach is likely to suffer from overfitting, in particular, when considering the coupled aggregation layer. The representation of the input sequences should be further investigated as well. Both MHIs and OFIs have the temporal aspect of the sequences integrated into their properties. The performed evaluation has shown that a model based on OFIs significantly outperforms the one based on MHIs in most cases. However, it has also been shown that most of the interesting frames in MHI sequences are located at the very end of the temporal axis of each sequence. Therefore, single MHIs extracted from entire sequences could also be used as input for deep architectures. Overall, the performed experiments show that the discrimination between lower and higher pain elicitation levels remains a difficult endeavour. This is due to the variety of expressiveness amongst the participants. However, personalisation and transfer learning strategies could potentially help improve the performance of inference models applied in this specific area of research.

Author Contributions: Conceptualisation, P.T. and F.S.; Methodology, P.T.; Software, P.T.; Validation, P.T.; Formal Analysis, P.T.; Investigation, P.T. and F.S.; Writing—Original Draft Preparation, P.T.; Writing—Review and Editing, P.T., H.A.K. and F.S.; Visualisation, P.T.; Supervision, H.A.K. and F.S.; Project Administration, H.A.K. and F.S.; Funding Acquisition, H.A.K. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results has received funding from the Federal Ministry of Education and Research (BMBF, SenseEmotion) to F.S., (BMBF, e:Med, CONFIRM, ID 01ZX1708C) to H.A.K., and the Ministry of Science and Education Baden-Württemberg (Project ZIV) to H.A.K.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ahad, M.A.R.; Tan, J.K.; Kim, H.; Ishikawa, S. Motion History Image: its variants and applications. *Mach. Vis. Appl.* **2012**, *23*, 255–281.
2. Horn, B.K.P.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203.
3. Lucey, P.; Cohn, J.F.; Prkachin, K.M.; Solomon, P.E.; Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In Proceedings of the Face and Gesture, Santa Barbara, CA, USA, 21–25 March 2011; pp. 57–64.
4. Walter, S.; Gruss, S.; Ehleiter, H.; Tan, J.; Traue, H.C.; Crawcour, S.; Werner, P.; Al-Hamadi, A.; Andrade, A. The BioVid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In Proceedings of the IEEE International Conference on Cybernetics, Lausanne, Switzerland, 13–15 June, 2013; pp. 128–131.
5. Aung, M.S.H.; Kaltwang, S.; Romera-Paredes, B.; Martinez, B.; Singh, A.; Cella, M.; Valstar, M.; Meng, H.; Kemp, A.; Shafizadeh, M.; et al. The automatic detection of chronic pain-related expression: requirements, challenges and multimodal dataset. *IEEE Trans. Affect. Comput.* **2016**, *7*, 435–451.
6. Velana, M.; Gruss, S.; Layher, G.; Thiam, P.; Zhang, Y.; Schork, D.; Kessler, V.; Gruss, S.; Neumann, H.; Kim, J.; et al. The SenseEmotion Database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In Proceedings of the Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Cancun, Mexico, 4 December 2016; pp. 127–139.
7. Thiam, P.; Kessler, V.; Schwenker, F. Hierarchical combination of video features for personalised pain level recognition. In Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 26–28 April, 2017; pp. 465–470.
8. Werner, P.; Al-Hamadi, A.; Limbrecht-Ecklundt, K.; Walter, S.; Gruss, S.; Traue, H.C. Automatic Pain Assessment with Facial Activity Descriptors. *IEEE Trans. Affect. Comput.* **2017**, *8*, 286–299.
9. Tsai, F.S.; Hsu, Y.L.; Chen, W.C.; Weng, Y.M.; Ng, C.J.; Lee, C.C. In Proceedings of the Toward Development and Evaluation of Pain Level-Rating Scale For Emergency Triage Based on Vocal Characteristics and Facial Expressions. Interspeech 2016, San-Francisco, CA, USA, 8–12 September 2016; pp. 92–96.

10. Thiam, P.; Schwenker, F. Combining deep and hand-crafted features for audio-based pain intensity classification. In Proceedings of the Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Beijing, China, 20 August, 2018; pp. 49–58.
11. Walter, S.; Gruss, S.; Limbrecht-Ecklundt, K.; Traue, H.C.; Werner, P.; Al-Hamadi, A.; Diniz, N.; Silva, G.M.; Andrade, A.O. Automatic pain quantification using autonomic parameters. *Psych. Neurosci.* **2014**, *7*, 363–380.
12. Chu, Y.; Zhao, X.; Han, J.; Su, Y. Physiological signal-based method for measurement of pain intensity. *Front. Neurosci.* **2017**, *11*, 279.
13. Lopez-Martinez, D.; Picard, R. Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medecine and Biology Society, Honolulu, HI, USA, 18–21 July 2018; pp. 5624–5627.
14. Thiam, P.; Schwenker, F. Multi-modal data fusion for pain intensity assessement and classification. In Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications, Montreal, QC, Canada, 28 November–1 December 2017; pp. 1–6.
15. Thiam, P.; Kessler, V.; Amirian, M.; Bellmann, P.; Layher, G.; Zhang, Y.; Velana, M.; Gruss, S.; Walter, S.; Traue, H.C.; et al. Multi-modal pain intensity recognition based on the SenseEmotion Database. *IEEE Trans. Affect. Comput.* **2019**, doi:10.1109/TAFFC.2019.2892090.
16. Thiam, P.; Bellmann, P.; Kestler, H.A.; Schwenker, F. Exploring deep physiological models for nociceptive pain recognition. *Sensors* **2019**, *19*, 4503.
17. Ekman, P.; Friesen, W.V. *The Facial Action Unit System: A Technique for the Measurement of Facial Movement*; Consulting Psychologist Press: Mountain View, CA, USA, 1978.
18. Senechal, T.; McDuff, D.; Kaliouby, R.E. Facial Action Unit detection using active learning and an efficient non-linear kernel approximation. In Proceedings of the IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 7–13 December 2015; pp. 10–18.
19. Lucey, P.; Cohn, J.; Lucey, S.; Matthews, I.; Sridharan, S.; Prkachin, K.M. Automatically detecting pain using Facial Actions. In Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–8.
20. Abe, S. *Support Vector Machines for Pattern Classification*; Springer: Berlin, Germany 2005.
21. Brümmer, N.; Preez, J.D. Application-independent evaluation of speaker detection. *Comput. Speech Lang.* **2006**, *20*, 230–275.
22. Zafar, Z.; Khan, N.A. Pain intensity evaluation through Facial Action Units. In Proceedings of the 22nd International Conference on Pattern Recognition,Stockholm, Sweden, 24–28 August 2014; pp. 4696–4701.
23. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
24. Prkachin, K.M.; Solomom, P.E. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain* **2008**, *139*, 267–274.
25. Xu, X.; Craig, K.D.; Diaz, D.; Goodwin, M.S.; Akcakaya, M.; Susam, B.T.; Huang, J.S.; de Sa, V.S. Automated pain detection in facial videos of children using human-assisted transfer learning. In Proceedings of the International Workshop on Artificial Intelligence in Health, Stockholm, Sweden, 13–14 July 2018; pp. 162–180.
26. Monwar, M.; Rezaei, S. Pain recognition using artificial neural network. In Proceedings of the IEEE International Symposium on Signal Processing and Information Theory, Vancouver, BC, Canada, 27–30 August 2006; pp. 8–33.
27. Yang, R.; Tong, S.; Bordallo, M.; Boutellaa, E.; Peng, J.; Feng, X.; Hadid, A. On pain assessment from facial videos using spatio-temporal local descriptors. In Proceedings of the 6th International Conference on Image Processing Theory, Tools and Applications, Oulu, Finland, 12–15 December 2016; pp. 1–6.
28. Zhao, G.; Pietikaeinen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928.
29. Ojansivu, V.; Heikkilä, J. Blur insensitive texture classification using local phase quantization. In Proceedings of the Image and Signal Processing, Cherbourg-Octeville, France, 1–3 July 2008; pp. 236–243.
30. Kannala, J.; Rahtu, E. BSIF: Binarized Statistical Image Features. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, 11–15 November 2012; pp. 1363–1366.

31. Kächele, M.; Thiam, P.; Amirian, M.; Werner, P.; Walter, S.; Schwenker, F.; Palm, G. Engineering Applications of Neural Networks. Multimodal data fusion for person-independent, continuous estimation of pain Intensity, In Proceedings of the Engineering Applications of Neural Networks, Rhodes, Greece, 25–28 September 2015; pp. 275–285.
32. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
33. Thiam, P.; Kessler, V.; Walter, S.; Palm, G.; Scwenker, F. Audio-visual recognition of pain intensity. In Proceedings of the Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, Cancun, Mexico, 4 December 2016; pp. 110–126.
34. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 401–408.
35. Almaev, T.R.; Valstar, M.F. Local Gabor Binary Patterns from Three Orthogonal Planes for automatic facial expression recognition. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 356–361.
36. Bellantonio, M.; Haque, M.A.; Rodriguez, P.; Nasrollahi, K.; Telve, T.; Guerrero, S.E.; Gonzàlez, J.; Moeslund, T.B.; Rasti, P.; Anbarjafari, G. Spatio-temporal pain recognition in CNN-based super-resolved facial images. In Proceedings of the International Conference on Pattern Recognition: Workshop on Face and Facial Expression Recognition, Cancun, Mexico, 4 December 2016; pp. 151–162.
37. Rodriguez, P.; Cucurull, G.; Gonzàlez, J.; Gonfaus, J.M.; Nasrollahi, K.; Moeslund, T.B.; Roca, F.X. Deep Pain: Exploiting Long Short-Term Memory networks for facial expression classification. *IEEE Trans. Cybern.* **2018**, doi:10.1109/TCYB.2017.2662199.
38. Kalischek, N.; Thiam, P.; Bellmann, P.; Schwenker, F. Deep domain adaptation for facial expression analysis. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, Cambridge, UK, 3–6 September 2019; pp. 317–323.
39. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and application in vision. In Proceedings of the IEEE International Symposium on Circuits and Systems, 2010, Paris, France, 30 May–2 June 2010; pp. 253–256.
40. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
41. Soar, J.; Bargshady, G.; Zhou, X.; Whittaker, F. Deep learning model for detection of pain intensity from facial expression. In Proceedings of the International Conference on Smart Homes and Health Telematics, Singapore, 10–12 July 2018; pp. 249–254.
42. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, Williams College, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
43. Bargshady, G.; Soar, J.; Zhou, X.; Deo, R.C.; Whittaker, F.; Wang, H. A joint deep neural network model for pain recognition from face. In Proceedings of the IEEE 4th International Conference on Computer and Communication Systems, Singapore, 23–25 February 2019; pp. 52–56.
44. Zhou, J.; Hong, X.; Su, F.; Zhao, G. Recurrent convolutional neural network regression for continuous pain intensity estimation in Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1535–1543.
45. Liang, M.; Hi, X. Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3367–3375.
46. Wang, F.; Xiang, X.; Liu, C.; Tran, T.D.; Reiter, A.; Hager, G.D.; Quaon, H.; Cheng, J.; Yuille, A.L. Regularizing face verification nets for pain intensity regression. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp.1087–1091.
47. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame attention networks for facial expression recognition in videos. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 3866–3870.
48. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267.
49. Yin, Z.; Collins, R. Moving object localization in thermal imagery by forward-backward MHI. In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop, New York, NY, USA, 17–22 June 2006; pp. 133–140.

50. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 29 June–2 July 2003; pp. 363–370.
51. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; pp. 25–36.
52. Lucas, B.D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, University of British Columbia, Vancouver, BC, Canada, 24–28 August 1981; pp. 674–679.
53. Beauchemin, S.S.; Barron, J.L. The computation of optical flow. *ACM Comput. Surv.* **1995**, *27*, 433–466.
54. Akpinar, S.; Alpaslan, F.N. Chapter 21—Optical flow-based representation for video action detection. In *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*; Deligiannidis, L., Arabnia, H.R., Eds.; Morgan Kaufmann: Boston, MA, USA, 2015; pp. 331–351.
55. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **2013**, *23*, 2031–2038.
56. Schuster, M.; Paliwal, K.K. Bidirectional Recurrent Neural Network. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681.
57. Hochreiter, S.; Bengio, Y.; Frasconi, P. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *Field Guide to Dynamical Recurrent Networks*; IEEE Press: Piscataway, NJ, USA, 2001.
58. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2016**, arXiv:1511.07289. Available online: https://arxiv.org/abs/1511.07289 (accessed on 3 February 2020)
59. Werner, P.; Al-Hamadi, A.; Niese, R.; Walter, S.; Gruss, S.; Traue, H.C. Automatic pain recognition from video and biomedical signals. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4582–4587.
60. Walter, S.; Gruss, S.; Traue, H.; Werner, P.; Al-Hamadi, A.; Kächele, M.; Schwenker, F.; Andrade, A.; Moreira, G. Data fusion for automated pain recognition. In Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, Istanbul, Turkey, 20–23 May 2015; pp. 261–264.
61. Kächele, M.; Thiam, P.; Amirian, M.; Schwenker, F.; Palm, G. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE J. Sel. Top. Sign. Process.* **2016**, *10*, 854–864.
62. Kächele, M.; Amirian, M.; Thiam, P.; Werner, P.; Walter, S.; Palm, G.; Schwenker, F. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evol. Syst.* **2016**, *8*, 1–13.
63. Bellmann, P.; Thiam, P.; Schwenker, F., Computational Intelligence for Pattern Recognition. In *Computational Intelligence for Pattern Recognition*; Pedrycz, W., Chen, S.M., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 83–113.
64. Bellmann, P.; Thiam, P.; Schwenker, F. Using a quartile-based data transtransform for pain intensity classification based on the SenseEmotion Database. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, Cambridge, UK, 3–6 September 2019; pp. 310–316.
65. Baltrusaitis, T.; Robinson, P.; Morency, L.P. OpenFace: An open source facial behavior analysis toolkit. In Proceedinggs of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
66. Bradski, G. The OpenCV library. *Dr Dobb's J. Softw. Tools* **2000**, *25*, 120–125.
67. Simonyan, K.; Zisserman, A. Very deep convolution networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556. Available online: https://arxiv.org/abs/1409.1556 (accessed on 3 February 2020)
68. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167. Available online: https://arxiv.org/abs/1502.03167 (accessed on 3 February 2020)
69. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2015**, arXiv:1412.6980. Available online: https://arxiv.org/abs/1412.6980 (accessed on 3 February 2020)
70. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 21 January 2020).
71. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, C.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org/ (accessed on 21 January 2020).

72. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
73. Werner, P.; Al-HamadiAl-Hamadi, Ayoub, S. Analysis of facial expressiveness during experimentally induced heat pain. In Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, San Antonio, TX, USA, 23–26 October 2017; pp. 176–180.

Sample Availability: The BioVid Heat Pain Database (Part A) is publicly available for non-commercial research and can be acquired by contacting the authors of the database at the following web-page: http://www.iikt.ovgu.de/BioVid.print.

II

List of Publications

Journal Articles

Kächele, Markus, Mohammadreza Amirian, Patrick Thiam, Philipp Werner, et al. (2017). “Adaptive confidence learning for the personalization of pain intensity estimation systems”. In: *Evolving Systems* 8(1), pp. 71–83. DOI: 10.1007/s12530-016-9158-4.

Kächele, Markus, Patrick Thiam, Mohammadreza Amirian, Friedhelm Schwenker, and Günther Palm (2016). “Methods for person-centered continuous pain intensity assessment from bio-physiological channels”. In: *IEEE Journal of Selected Topics in Signal Processing* 10(5), pp. 854–864. DOI: 10.1109/JSTSP.2016.2535962.

Öner, Tuba, Patrick Thiam, Gregor Kos, Rudolf Krska, Friedhelm Schwenker, and Boris Mizaikoff (2019). “Machine learning algorithms for the automated classification of contaminated maize at regulatory limits via infrared attenuated total reflection spectroscopy”. In: *World Mycotoxyn Journal* 12(2), pp. 113–122. DOI: 10.3920/WMJ2018.2333.

Thiam, Patrick, Peter Bellmann, Hans A. Kestler, and Friedhelm Schwenker (2019). “Exploring Deep Physiological Models for Nociceptive Pain Recognition”. In: *Sensors* 4503(20). DOI: 10.3390/s19204503.

Thiam, Patrick, Viktor Kessler, Mohammadreza Amirian, Peter Bellmann, et al. (2019). “Multi-Modal Pain Intensity Recognition Based on the SenseEmotion Database”. In: *IEEE Transactions on Affective Computing*. © 2019 IEEE. DOI: 10.1109/TAFFC.2019.2892090.

Thiam, Patrick, Hans A. Kestler, and Friedhelm Schwenker (2020). “Two-Stream Attention Network for Pain Recognition from Video Sequences”. In: *Sensors* 20(839). DOI: 10.3390/s20030839.

Thiam, Patrick, Sascha Meudt, Günther Palm, and Friedhelm Schwenker (2018). "A Temporal Dependency Based Multi-modal Active Learning Approach for Audiovisual Event Detection". In: *Neural Processing Letters* 48(2), pp. 709–732. DOI: 10.1007/s11063-017-9719-y.

Book Chapters

Bellmann, Peter, Patrick Thiam, and Friedhelm Schwenker (2018). "Multi-Classifier-Systems: Architectures, Algorithms and Applications". In: *Computational Intelligence for Pattern Recognition.* Ed. by Witold Pedrycz and Shyi-Ming Chen. Vol. 777. Cham: Springer International Publishing, pp. 83–113. DOI: 10.1007/978-3-319-89629-8_4.

Schwenker, Friedhelm, Ronald Böck, Martin Schels, Sascha Meudt, et al. (2017). "Multimodal Affect Recognition in the Context of Human-Computer Interaction for Companion-Systems". In: *Companion Technology: A Paradigm Shift in Human-Technology Interaction.* Ed. by Susanne Biundo and Andreas Wendemuth. Cham: Springer International Publishing, pp. 387–408. DOI: 10.1007/978-3-319-43665-4_19.

Conference Papers

Bellmann, Peter, Patrick Thiam, and Friedhelm Schwenker (2020). "Person Identification based on Physiological Signals: Conditions and Risks". In: *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM).* Vol. 1. INSTICC. SciTePress, pp. 373–380. DOI: 10.5220/0008865503730380.

Kächele, Markus, Martin Schels, Patrick Thiam, and Friedhelm Schwenker (2015). "Fusion Mappings for Multimodal Affect Recognition". In: *2015 IEEE Symposium Series on Computational Intelligence*, pp. 207–313. DOI: 10.1109/SSCI.2015.53.

Kächele, Markus, Patrick Thiam, Mohammadreza Amirian, Philipp Werner, et al. (2015). "Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity". In: *Engineering Applications of Neural Networks, EANN 2015.* Ed. by Lazaros Iliadis and Chrisina Jayne. Vol. 517. Cham: Springer International Publishing, pp. 275–285. DOI: 10.1007/978-3-319-23983-5_26.

Kessler, Viktor, Patrick Thiam, Mohammadreza Amirian, and Friedhelm Schwenker (2017a). “Multimodal Fusion including Camera Photoplethysmography for Pain Recognition”. In: *2017 International Conference on Companion Technology (ICCT)*, pp. 1–4. DOI: 10.1109/COMPANION.2017.8287083.

Kessler, Viktor, Patrick Thiam, Mohammadreza Amirian, and Friedhelm Schwenker (2017b). “Pain Recognition with Camera Photoplethysmography”. In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–5. DOI: 10.1109/IPTA.2017.8310110.

Shoukry, Nadeen, Omar Elkilany, Patrick Thiam, Viktor Kessler, and Friedhelm Schwenker (2020). “Subject-independent Pain Recognition using Physiological Signals and Para-linguistic Vocalizations”. In: *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM).* Vol. 1. INSTICC. SciTePress, pp. 142–150. DOI: 10.5220/0008912201420150.

Thiam, Patrick, Markus Kächele, Friedhelm Schwenker, and Günther Palm (2015). “Ensembles of Support Vector Data Description for Active Learning Based Annotation of Affective Corpora”. In: *2015 IEEE Symposium Series on Computational Intelligence*, pp. 1801–1807. DOI: 10.1109/SSCI.2015.251.

Thiam, Patrick, Viktor Kessler, and Friedhelm Schwenker (2017). “Hierarchical Combination of Video Features for Personalised Pain Level Recognition”. In: *25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 465–470.

Thiam, Patrick, Hans A. Kestler, and Friedhelm Schwenker (2020). “Multimodal Deep Denoising Convolutional Autoencoders for Pain Intensity Classification based on Physiological Signals”. In: *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM).* Vol. 1. INSTICC. SciTePress, pp. 289–296. DOI: 10.5220/0008896102890296.

Thiam, Patrick and Friedhelm Schwenker (2017). “Multi-Modal Data Fusion for Pain Intensity Assessement and Classification”. In: *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6. DOI: 10.1109/IPTA.2017.8310115.

Workshop Papers

Amirian, Mohammadreza, Markus Kächele, Patrick Thiam, Viktor Kessler, and Friedhelm Schwenker (2016). “Continuous Multimodal Human Affect Estimation Using Echo State Networks”. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.* New York, NY, USA: Association for Computing Machinery, pp. 67–74. DOI: 10.1145/2988257.2988260.

Bellmann, Peter, Patrick Thiam, and Friedhelm Schwenker (2019). “Using a Quartile-based Data Transformation for Pain Intensity Classification based on the SenseEmotion Database”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 310–316. DOI: 10.1109/ACIIW.2019.8925244.

Kächele, Markus, Martin Schels, Sascha Meudt, Viktor Kessler, et al. (2015). “On Annotation and Evaluation of Multi-modal Corpora in Affective Human-Computer Interaction”. In: *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. Ed. by Ronald Böck, Francesca Bonin, Nick Campbell, and Ronald Poppe. Cham: Springer International Publishing, pp. 35–44. DOI: 10.1007/978-3-319-15557-9_4.

Kächele, Markus, Patrick Thiam, Günther Palm, and Friedhelm Schwenker (2014). “Majority-Class Aware Support Vector Domain Oversampling for Imbalanced Classification Problems”. In: *Artificial Neural Networks in Pattern Recognition*. Ed. by Neamat El Gayar, Friedhelm Schwenker, and Cheng Suen. Cham: Springer International Publishing, pp. 83–92. DOI: 10.1007/978-3-319-11656-3_8.

Kächele, Markus, Patrick Thiam, Günther Palm, Friedhelm Schwenker, and Martin Schels (2015). “Ensemble Methods for Continuous Affect Recognition: Multi-Modality, Temporality, and Challenges”. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: Association for Computing Machinery, pp. 9–16. DOI: 10.1145/2808196.2811637.

Kalischek, Nikolai, Patrick Thiam, Peter Bellmann, and Friedhelm Schwenker (2019). “Deep Domain Adaptation for Facial Expression Analysis”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 317–323. DOI: 10.1109/ACIIW.2019.8925055.

Ruiz, Adrian T., Patrick Thiam, Friedhelm Schwenker, and Günther Palm (2018). “A k-Nearest Neighbor Based Algorithm for Multi-Instance Multi-Label Active Learning”. In: *Artificial Neural Networks in Pattern Recognition*. Ed. by Luca Pancioni, Friedhelm Schwenker, and Edmondo Trentin. Cham: Springer International Publishing, pp. 139–151. DOI: 10.1007/978-3-319-99978-4_11.

Sellner, Jan, Patrick Thiam, and Friedhelm Schwenker (2019). “Visualizing Facial Expression Features of Pain and Emotion Data”. In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Ed. by Friedhelm Schwenker and Stefan Scherer. Cham: Springer International Publishing, pp. 101–115. DOI: 10.1007/978-3-030-20984-1_9.

Thiam, Patrick, Markus Kächele, Friedhelm Schwenker, and Günther Palm (2015). “Ensemble Methods and Active Learning in HCI”. In: *Proceedings of the Workshop on New Challenges in Neural Computation*, pp. 65–67.

Thiam, Patrick, Viktor Kessler, and Friedhelm Schwenker (2014). “A Reinforcement Learning Algorithm to Train a Tetris Playing Agent”. In: *Artificial Neu-*

ral Networks in Pattern Recognition. Ed. by Neamat El Gayar, Friedhelm Schwenker, and Cheng Suen. Cham: Springer International Publishing, pp. 165–170. DOI: 10.1007/978-3-319-11656-3_15.

Thiam, Patrick, Viktor Kessler, Steffen Walter, Günther Palm, and Friedhelm Scwenker (2017). "Audio-Visual Recognition of Pain Intensity". In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction.* Ed. by Friedhelm Schwenker and Stefan Scherer. Cham: Springer International Publishing, pp. 110–126. DOI: 10.1007/978-3-319-59259-6_10.

Thiam, Patrick, Sascha Meudt, Markus Kächele, Günther Palm, and Friedhelm Schwenker (2014). "Detection of Emotional Events Utilizing Support Vector Methods in an Active Learning HCI Scenario". In: *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems.* New York, NY, USA: Association for Computing Machinery, pp. 31–36. DOI: 10.1145/2668056.2668062.

Thiam, Patrick, Sascha Meudt, Friedhelm Schwenker, and Günther Palm (2016). "Active Learning for Speech Event Detection in HCI". In: *Artificial Neural Networks in Pattern Recognition.* Ed. by Friedhelm Schwenker, Hazem M. Abbas, Neamat El Gayar, and Edmondo Trentin. Cham: Springer International Publishing, pp. 285–297. DOI: 10.1007/978-3-319-46182-3_24.

Thiam, Patrick and Friedhelm Schwenker (2019). "Combining Deep and Handcrafted Features for Audio-based Pain Intensity Classification". In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction.* Ed. by Friedhelm Schwenker and Stefan Scherer. Cham: Springer International Publishing, pp. 49–58. DOI: 10.1007/978-3-030-20984-1_5.

Velana, Maria, Sascha Gruss, Georg Layher, Patrick Thiam, et al. (2017). "The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Pain- and Emotion-Recognition System". In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction.* Ed. by Friedhelm Schwenker and Stefan Scherer. Cham: Springer International Publishing, pp. 127–139. DOI: 10.1007/978-3-319-59259-6_11.