UNIVERSITY OF TWENTE.

Faculty of Behavioural, Management and Social Sciences

Creation and Evaluation of a Database for Automatic Video-Based Pain Detection

Masterthesis HFE

Lisa Kiel

University of Twente – Enschede

February 2021

Supervisors:
Prof. Dr. Frank van der Velde
Dr. R.H.J. van der Lubbe

SPONSORED BY THE
Federal Ministry of Education and Research

In Cooperation with:

ixp institute of experimental psychophysiology

Abstract

Pain management can considerably influence the recovery of a patient. A system for automatic pain detection can therefore not only support health practitioners in their work but also improve the patient's well-being. Previous work on automatic painful facial expression recognition has been successful in discriminating painless from painful expressions. In this study, some of the limitations of these studies will be addressed by collecting a video database with continuous self-ratings of pain which enables to use self-ratings as ground truth at video frame level. Using machine learning algorithms, facial features that represent changes in the face due to expressions are ranked by their importance in discriminating non-painful from painful expressions. Different subsets of features are used to analyse which might be the most promising one in recognising painful facial expressions. The results show that the model performs well on new data if it was trained on data from the same participants, but on a chance level if a between-subjects design was used. Further, the model benefitted from excluding participants that indicated high levels of tiredness, which indicates that considering tiredness as a confounding variable could be important for future research about automatic painful facial expression. Overall the results show that a model built on the collected database can successfully be used to discriminate between no pain and pain classification when using subject-specific models.

Preface

This master thesis was written in cooperation with the Institute of Experimental Psychophysiology (IXP) in Düsseldorf, Germany. It is part of a bigger project that is a joint collaboration with the University of Lübeck, the Silesian University of Technology in Zabrze, Poland and the company APA and is partly funded by the German ministry of education and research (BMBF) within the research programme ICT 2020. The project aims to create a multimodal device for automatic pain detection that can be used in a clinical setting to support health care workers. The system will be based on on-line processed indicators such as facial expressions and different physiological features. IXP hereby is responsible for creating a machine learning model that is able to detect pain based on facial expressions. I was asked to create a database on which the model can be trained on. The following study therefore (1) describes the creation of the database including the method selection and (2) gives first insights into the performance of a machine learning model that is trained on this database.

1. Introduction

Pain control is a crucial part of the treatment for patients in the intensive care unit (ICU). It can influence the duration of infection, mortality and the length of stay in the ICU (Dale, Prendergast, Gélinas, & Rose, 2018). Yet, nearly half of the ICU patients can recall having experienced moderate or severe pain during and outside of treatment procedures and many physicians seem to underestimate the amount of pain common procedures can cause (Dale et al., 2018). When assessing pain, one of the most used and most simple method, that many of us who were unfortunate enough might be familiar with, is the oral numeric rating scale from 0 to 10. Self-reported pain assessments however quickly reach their limits as soon as the practitioners are working with patients who are unable to express their pain. Reasons for this can be that the patients are intubated, sedated or suffer from illnesses like dementia that impairs their cognitive ability (Ashraf et al., 2009), which are circumstances that are frequent in ICU. In these cases, there are alternatives to use assessments based on observation, e.g. the Critical-Care Pain Observation Tool (CPOT), a valid and reliable tool based on observing behavioural changes like facial expression, muscle tension and body movements (Dale et al., 2018). But even behavioural assessment tools deal with the disadvantage that they require an observer who takes their time to assess the patient's pain. Within the typically understaffed work environment that practitioners have to deal with, it is often impossible to detect potential pain early on. A system that can automatically detect pain can thus have a tremendously beneficial impact. The possibility to automatically detect pain at a very early stage would not only strongly reduce the patient's suffering and support their recovery but would also lighten the workload of the health care providers.

The current study aims to contribute to the research on developing a system for automatic pain detection. In this thesis, the focus will be on the facial expressions of people to detect pain. For this, a database consisting of video frames of facial expressions from people under varying degrees of pain will be created, in which the frames are annotated with the participant's pain ratings. The acquired facial expression database will then be analysed to answer the questions 1) if it can be used to discriminate painful from non-painful expressions and 2) which of the facial features in the frames might be the most suitable to detect pain.

In the following, an overview of the related work regarding automatic facial pain expression will be given to present the current state of research on this topic and to identify areas where the research on video-based pain detection might be improved on. Based on this, a second literature research will be conducted to select the appropriate methods for creating a

database consisting of video frames of painful facial expressions. This database will then be analysed to answer the research questions presented above.

2. Related Work

Facial expressions can provide relevant information about someone's pain and research has shown that it can serve as a reliable measuring instrument (Craig, Prkachin, & Grunau, 1992). One of the most widely used approaches for analysing facial expressions is the *Facial Action Coding System* (FACS), that was developed by Ekman and Friesen in 1978 (as cited in Chen, Ansari, & Wilkie, 2018; Pantic & Rothkrantz, 2004; Whitehill, Bartlett, & Movellan, 2013). The FACS uses so-called *Action Units* (AUs) which can define 46 movements of the face (Whitehill et al., 2013). It has been successfully utilised for the detection of a variety of cognitive states (see Ekman & Rosenberg, 2005 for an overview and review of studies using the FACS) and has achieved good results in the discrimination of "no-pain" vs. "pain" facial expressions (Ashraf et al., 2009).

A combination of twelve AUs has been repeatedly found to be related to pain expressions (see for example Craig, Hyde, & Patrick, 1991; Lucey et al., 2012; Prkachin, 1992). These AUs include tightening of the eyelids, pulling of the lip corners and closing of the eyes (see Table 1 for a full list). Out of these AUs, four combinations have been found to be relevant in particular. These are brow-lowering (AU 4), tightening the eyelids (AU 6) *or* cheek raising (AU 7), nose wrinkling (AU 9) *or* upper-lip raising (AU 10) and eye closure (AU 43, Kunz, Scharmann, Hemmeter, Schepelman, & Lautenbacher, 2008; Prkachin, 1992; Prkachin & Solomon, 2008). These findings led to the development of the Prkachin and Solomon Pain Intensity (PSPI) metric (Prkachin & Solomon, 2008), which combines mentioned pain-related AUs to calculate a pain intensity score. Werner et al. (2016) raised criticism regarding the use of the PSPI as ground truth in pain recognition. They state that the statistical analysis on which the PSPI is based on, was done on facial expressions on a video sequence level, so one has to be careful whether it is equally reliable on a frame by frame level. On a sequence level, a series of frames are annotated with one pain rating that was usually given after the pain stimulation instead of having one pain rating for each frame, which is the case for a frame level annotation. Analysing pain on a sequence level ground truth instead of a frame level ground truth could have the benefit of yielding information over the development of the pain but can have the disadvantage of overgeneralisation, as many different pain levels within this sequence all get categorised as only one pain rating. The PSPI further relies on only a very selective number of

AUs, although it is not unlikely that some people show atypical patterns in response to pain, which then would not be accounted for (Werner et al., 2016).

Table 1.
*Descriptions of facial Action Units (AUs) associated with pain. **

Action Units	Description
AU 4	Brow lower
AU 6	Cheek raise
AU 7	Lids tighten
AU 9	Nose wrinkle
AU 10	Upper lip raise
AU 12	Lip corner pull
AU 20	Lip stretch
AU 25	Lips part
AU 26	Jaw drop
AU 27	Mouth stretch
AU 43	Eyes closed
AU 45	Blink

* Information taken from Prkachin (1992) and Craig, Hyde and Patrick (1991).

Originally, a trained FACS-coder would code pictures or videos on a frame-by-frame basis according to the observed AUs. Since a minute of video material can take up to one hour to annotate, the benefit of its automation is obvious (Pantic & Rothkrantz, 2004). Over the last years, automatic facial expression detection has evolved rapidly. Looking at the development of automatic pain detection in particular, most of the recent work has been based on two publicly available databases: First (1) the UNBC McMaster Shoulder Pain Database (in the following referred to as UNBC database) by Lucey et al. (2012) and secondly (2) the BioVid Heat Pain Database by Walter et al. (2013). For the UNBC database, 129 people with shoulder pain were filmed while performing passive range-of-motion tests. The videos were subsequently coded on a frame-by-frame basis, using the FACS and the PSPI was used as the ground truth to define pain on a frame level. They achieved a performance of around 80% for pain detection at frame level. They further found significant variance in head movements between no-pain and pain conditions, especially for yaw (rotating the head left to right) and roll (tilting the head sideways). This is indicating that head movements could be relevant for

pain detection. Hammal and Cohn (2012) used the PSPI metric to classify four intensities of pain in 25 people from the UNBC database. They achieved a moderate to good interclass correlation, which suggests that facial expressions could reliably be used to measure pain intensity in this context. Another study related to this database was done by Sikka (2014), who found a Pearson correlation coefficient of .51 in their prediction of self-reported pain intensity, indicating that machine learning models based on FACS can reliably be used to predict pain intensity. Ashraf et al. (2009) compared the use of *rigid* face recognition (location, scale and rotation) and *non-rigid* face recognition, which are shape variations that cannot be expressed by its location, scale or rotation in 2D. This was a different approach since the detection of AU usually depends on rigid face detection, on which the work previously was focused on. Their model was trained using the Support Vector Machine, using a leave-one-subject-out approach, so that the test dataset did not include participants in the training set. They found that non-rigid shapes significantly improved the performance classifier. Additionally, they tested the performance of pain detection between frame-level and sequence-level. While the performance was better at frame-level, it still was above chance at sequence-level, which lead them to raise the question for future research to combine frame and sequence level. Another study on the UNBC database was done by Sikka et al. (2014) and aimed at predicting the self-ratings, rather than using a binary classification of *no-pain* and *pain* like other studies have done it previously. They found a positive correlation between the predicted ratings and the actual ones and a Pearson correlation of 0.51, which shows that the regression performed better than chance classification. In addition to the analysis on the UNBC-Database, Sikka et al. (2014) also conducted a study on facial expressions on children between 5 and 18 with post-operative and experimental pain. The pain recordings took place within 24 hours after the surgery and the no-pain recordings after clinical recovery. The participants rated their maximum perceived pain on a scale of 0-10 and the video segments were then analysed using AUs. AU 4, AU 7, AU 9, AU 25 and AU 45 were found to have a significant correlation with self-ratings.

The second database mentioned, the BioVid Heat Pain Database (Walter et al., 2013), contains video and physiological data (skin conductance (SC), electrocardiogram (ECG), electromyogram (EMG), and electroencephalography (EEG)) of 90 persons who received controlled heat pain in four intensities. Werner et al. (2013) used distance and gradient features to represent pain-related AUs. Among other features, they also included nasal wrinkles, nasolabial furrows and the head pose for their analysis. Generally, they found high variability in the facial expressiveness of their participants, i.e. some participants had a high variability of

facial expression between no pain and high pain intensities, while other's expressions barely varied between the two conditions. Nevertheless, for most participants, they have achieved a high performance when discriminating between no-pain and high-pain intensities when using person-specific models but had a performance not clearly above chance for lower pain levels. They further investigated the importance of head movements and found high variability mostly for pitch (moving the head up or down) but less so for yaw and roll as found by (Lucey et al., 2012). In a subsequent study, Werner et al. (2016) proposed new facial activity descriptors for pain assessment and estimation of pain intensity. They found that this new method outperformed their previous approaches. Another study on the BioVid database, by Lopez-Martinez and Picard (2017), focused on pain classification using only the physiological measures SC and ECG. For the classification, they used a multi-task learning approach, with which they were able to account for individual differences and achieved a good classification accuracy.

3. Theoretical foundations of pain

A widely used and accepted definition of pain is the one from the International Association for the Study of Pain (IASP), who defines pain as *"an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage.*" (IASP, 2017). The IASP further acknowledges that pain is always subjective and, even though it does not have to be tied to physical damage but can be of psychological cause, it still has to be accepted as pain. Hansen and Streltzer (2005) further emphasise the importance of psychological aspects when it comes to defining pain, as its perception is influenced by affective or evaluative factors. It thus strongly depends on the context it is perceived. For example, the amount of pain someone feels can be influenced by the amount of attention, anxiety, fear, or even past experiences a person has made (Hansen & Streltzer, 2005).

Pain is usually categorised into two types of pain, namely acute and chronic. An overview of the differences between acute and chronic pain can be found in Table 2. Acute pain occurs suddenly and lasts for only a short time and can usually be linked to a specific cause, like an injury (Institute of Medicine, 2011; Świeboda, Filip, Prystupa, & Drozd, 2013). Usually, pain fulfils a protective role by hindering people from activities that could worsen their condition and under ordinary circumstances the pain ends when its cause is resolved, e.g. if the pain stimulus is gone or the injury has healed (Institute of Medicine, 2011). Chronic pain however is defined as a pain that lasts for longer than three months and considered to be a

disease in itself (Institute of Medicine, 2011; Świeboda et al., 2013). It can occur without a detectable physical cause or stay prevalent after healing its related illness or injury (Świeboda et al., 2013).

Table 2. *Overview of the difference between acute and chronic pain*

	Acute pain	Chronic Pain
Duration	< 3 months	> 3 months
Cause	Known (illness or injury), treatable	Unknown, multifaceted
Therapy	Treatment / cure of underlying cause	Multi-therapeutic approach, cure unlikely

4. Method selection for creating the facial expression database

The aim of the current study, as outlined in the introduction, is to contribute to a project for the development of a device that is capable to automatically detect pain. The focus of this study hereby is solely on the detection of pain based on facial expressions. Given the experimental nature of this study, only acute pain will be studied as chronic pain cannot be induced in a controlled way and would require a longitudinal study. Reviewing above literature research, it can be concluded that the research on pain expression has been primarily focused on AUs over the last years and has only recently with the rise of machine learning been extended to also include additional descriptors like non-rigid features. Using AUs seems to be a viable strategy if human raters are involved in the process of categorising the facial expressions as it simplifies the complexity of them down to specific movements. However, this is not applicable for real-time applications and with now existing methods for automatic feature detection, there is the possibility to miss out on other, more effective features than only relying on action units to describe the high variability of facial expressions people can have. The research further was mostly done on a frame level, while the ground truth of pain was given through self-ratings at sequence level. I.e, the participant indicated their perceived pain level after the pain induction for the whole duration of the pain stimulus instead of having a continuous monitoring of their perceived pain throughout the stimulation.

The first goal of the current study is hence to create a database of video frames of facial expressions of people under varying levels of pain that uses continuous self-ratings for the pain annotation for each frame. The aim is to collect video recordings of facial expressions of a

sample of 50 participants. They will be recorded during experimental pain stimulation, while the subjects are continuously indicating their perceived pain level. In the following section, a literature research is described that aimed to select the instruments for the planned pain induction and pain measurements that will be used in the experiment to create the database.

4.1 Measuring pain

When speaking about measuring pain, a distinction can be made between one-dimensional scales, that measure only the pain *intensity* and multi-dimensional scales that might also include the pain *quality.* Pain intensity described the severity of the pain, while pain quality refers to the way the pain feels like, for example, 'hot', 'stinging' or 'dump', or where the pain is located. For this study, only the intensity is of interest, which is why only one-dimensional pain measuring instruments are explained in more detail. Further, there are a number of physiological measurements, like skin conductance or heart rate that have been proven to be successful in detecting pain (see e.g. Loggia, Juneau, & Bushnell, 2011; Lopez-Martinez & Picard, 2017). As this study focuses on facial expression only, these methods will not be explained in more detail.

As explained above, pain is a very complex, multifaceted and subjective experience. It is therefore not surprising that self-reports are considered as the "gold standard" in pain assessments (Hadjistavropoulos, Hunter, & Dever Fitzgerald, 2009). Probably the most used self-rating scale is the 0-10 *Numeric Rating Scale* (NRS), which is considered to be the scale with the most benefits and fewest weaknesses (Jensen, 2011). It has also shown the best results in adults because of its easy implementation and low rate of errors (Falch et al., 2014). To utilise this scale, patients are asked to indicate their perceived pain level between 0 (no pain) and 10 (pain as intense as you can imagine). According to Jensen (2011), answers in between two integers are also fine, as people can distinguish between more than ten pain levels.

Another scale is the *FACES-Pain Scale Revised* (FPS-R, Hicks, Baeyer, Spafford, van Korlaar, & Goodenough, 2001) which consists of six drawings of a face that shows expressions of increasing levels of pain (from no pain to very much pain) and is applied by pointing on a face that best suits the current paint level. It is assumed to be better suited for children as the FPS-R is more concrete and easier to understand for children (Hicks et al., 2001).

The *Verbal Rating Scale* (VRS) usually consists of four (Jensen, 2011) or six (Falch et al., 2014) verbal descriptions of pain, ranging from 'no pain', increasing to 'moderate pain' and up to 'severe pain' in the 4-point VRS, and up to 'worst pain imaginable' in the 6-point

VRS. This scale is more restricted in the amount of pain that can be expressed but can be easier to use for some patients (Jensen, 2011).

In the end, it was decided to use the NRS. Speaking could interfere with the facial expressions, so verbal ratings do not seem suitable. The NRS is further the easiest scale to operate and also allows for more accurate pain measuring as it offers a wider scale than the FPS-R.

4.2 Methods of pain induction

Pain stimuli can be separated into four categories of chemical, thermal, mechanical and electrical pain stimuli (Fillingim, Loeser, Baron, & Edwards, 2016). Chemical stimuli take some time to take effect and further last over a longer period of time. (Fraser & Grady, 2008). In contrast, the other stimuli usually are brief and phasic ones (Wong, Vierck, Riley, King, & Mauderli, 2010). Therefore, only electrical, thermal and mechanical stimuli will be considered for this study.

In Table 3, an overview of each pain stimuli is given, together with a rating of four different aspects that have been considered (feasibility, pain variation, data quality, and safety). For pain variation, a good rating was given if it is likely to obtain pain levels varying from very little pain up to the maximum pain imaginable. 'Data quality' concerned the quality of the video recordings, as this is the main goal of the study. Good ratings were given if the participant can sit straight and look straight forward. In a sitting position, little body movements would be ensured and would allow for an easy installation and adjustment of the camera positions. The category 'Safety' received a high rating if there is no risk for long-term harms for the participants.

For thermal stimulation, the cold pressor task and use of a heat thermode are considered. The cold pressor task has been used frequently in the pain literature (von Baeyer, Piira, Chambers, Trapanotto, & Zeltzer, 2005; Wehe, 2013). However, for good results, it requires special equipment because it is crucial to have a constant temperature and circulation of the water (von Baeyer et al., 2005). The cold pressor task has also been used to relieve pain (von Baeyer et al., 2005), which could conflict with the goal of achieving a good pain variation. Heat stimulation through a thermode has been used in several studies for pain induction (see e.g. Angst, Tingle, Phillips, & Carvalho, 2009; Hammal, Kunz, Arguin, & Gosselin, 2008; Walter et al., 2013). A thermode makes a highly controlled pain induction possible without the risk of causing any damage to the skin (Walter et al., 2013) as it can heat up precisely to a

defined temperature for a predefined period of time. This high control of the stimulus further allows to tailor the pain to each person individually by predefining the person's *pain threshold*, which is the point where a stimulus is perceived as painful for the first time (IASP, 2017) – and *pain tolerance level*, which is the maximum pain intensity a subject is able or willing to endure (IASP, 2017).

For mechanical pain stimulations, the pinprick test, using a pressure algometer and the Submaximum-Effort Tourniquet test have been found. For the pinprick test, a sharp object is used on the participant and gently applied to the skin ("pinprick test," 2009) which has the potential to penetrate the skin and thereby causing damage to the skin. The pressure algometer uses for example a rubber disk which presses against the participant's skin to induce pain (Park, Kim, Park, Kim, & Jang, 2011). As the pressure can potentially harm the participant too, by resulting in bruises, it is also omitted for further evaluation. For the Submaximum-Effort Tourniquet test, a tourniquet is inflated which causes ischemic pain (Carli, 2007). If certain safety regulations are met, it is a safe method to use and can also lead to high pain ratings. The drawback, however, is that the pain is not highly controllable as it takes a certain amount of time to build up.

The last stimuli considered are electrical. In combination with facial expressions, this method yielded in the study of Kunz, Mylius, Schepelmann and Lautenbacher (2004) very weak correlation with self-reported pain, suggesting that other mechanisms like shock reactions confound the painful expressions. Another mean for electrical pain stimulation are electrocutaneous stimuli that only activate a-delta fibres (see e.g. Blom & Lubbe, 2017; Mouraux, Iannetti, & Plaghki, 2010). The pain is delivered via a bipolar needle electrode that is inserted into the outermost layer of the skin and the applied pain is described as "pricking" (Mouraux et al., 2010) and could therefore circumvent the problem of a shock reaction and be a viable method for this study.

To conclude, it was decided that the thermal stimulation through a heat thermode seems to be the most suitable one for the present study. It makes it possible to control the pain stimulus with high accuracy and to tailor the pain to each person individually and has been used successfully in several studies about painful facial expressions.

Table 3. *Overview of Pain stimuli.*

Type	Method		Pain variation		Data quality		Safety
Thermal	Cold pressor	+	Good results in literature	o	Takes time to become painful	+	Good, participant is in control
	Heat thermode	+	Good results in literature	+	Good	+	Good, Safety regulations prevent any harm
Mechanical	Submaximum-Effort Tourniquet	+	Good results in literature	o	Takes time to become painful	+	Good if protocol is followed
	Pressure algometer	-	Used in literature for pain threshold, not tolerance	+	Good	-	Can leave bruising
	Pinprick	-	Used in literature for pain threshold, not tolerance	o	Fear of needles common, can affect data	-	Penetrates skin
Electrical	Electrical stimulation	+	Good results in literature	o	Startled expressions as possible confounder	+	Good

Note: Meaning of the ratings are + = positive, o = neutral, - = negative aspect, regarding the use the method for the current study

4.3 Feature detection through OpenFace

The above determined methods will be used to create a database of video frames of painful facial expressions. Before the video data can be analysed to answer the research questions if it is possible to distinguish between painful and painless facial expressions and which features are relevant to achieve this, they first have to be transformed into quantitative data. To do this, facial features for each frame will be extracted using the toolkit *'OpenFace'*. OpenFace was developed by Baltrušaitis, Zadeh, Chong Lim, and Morency (2018) and is capable to detect facial landmarks, estimate the head pose and eye-gaze and detect 18 facial action units from video frames or photos.

Facial landmarks refer to positions of (in this case) 68 points of the face, as it can be seen in the landmark index of Figure 1, and are detected in both 2D and 3D. These facial landmarks are also used to detect the facial AUs, which represent the change of a combination of facial landmarks and describe certain movements of the face, rather than just specific points (as discussed in section 2). An example of how these landmarks are mapped to a face can be seen in Figure 2. As OpenFace only detects 18 of the usually used 27 AUs, it also only detects

8 of the 12 AUs that have been found in the literature to be relevant to pain, which are AU 4, 6, 7, 9, 12, 20, 25 and 45.

OpenFace is further able to detect 6 rigid shape parameters and 34 non-rigid parameters, which are reduced components of the facials landmarks (Baltrušaitis, Robinson, & Morency, 2016). Rigid shapes can hereby be understood as shapes that can be transformed in terms of its location, scale and rotation, while non-rigid shapes represent deformations due to, for example, the person's expressions or their identity (Baltrušaitis, 2019). Overall, OpenFace tracks 709 features which will be used in the analysis of the created database.

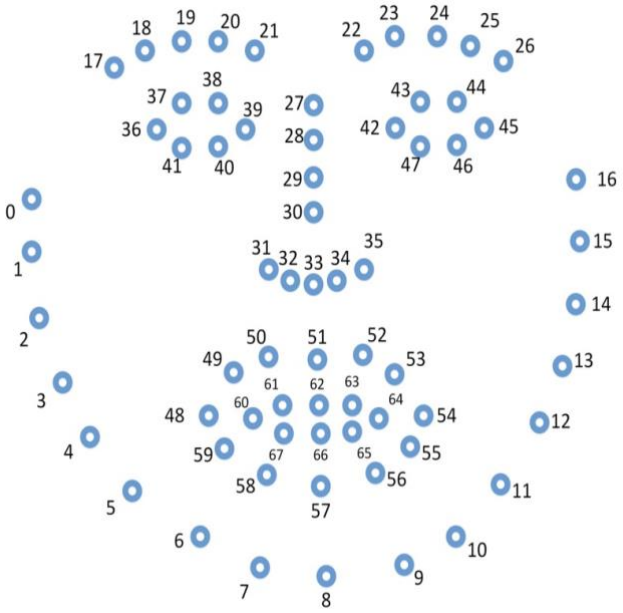

Figure 1. Facial Landmark index (taken from (Baltrušaitis, 2019).

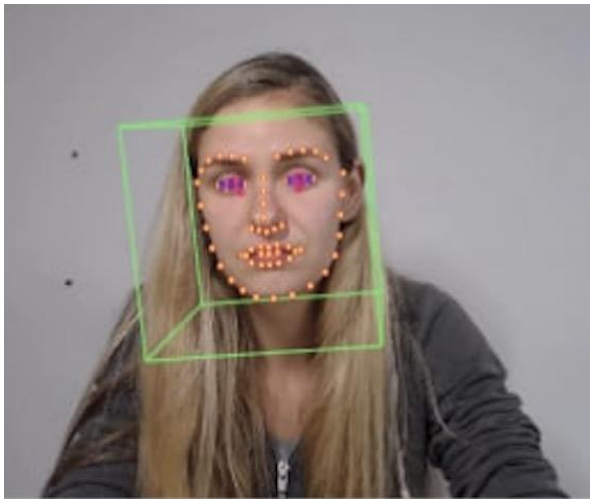

Figure 2. Example of OpenFace tracking facial landmarks (orange dots), head pose (green box) and eye gaze (pink dots).

5. Method

5.1 Participants

A total of 54 subjects were recruited and took part in the study in a laboratory at IXP. Twenty-nine were female, 25 male with an age range of 18-56 and a mean age of 27.7 (SD = 7.63). Excluded from participation in the survey were persons under 18 or over 70 years of age and persons who were pregnant or suffering from chronic pain, cardiovascular disease, neurological or psychiatric disorders. The participants also received monetary compensation for their travel of 50€. After investigating the data, 5 participants were excluded from the study due to missing data or poor video quality, leaving 49 participants remaining. The study design was approved by the ethical commission of the University of Lübeck, with whom we work together on the project for automatic pain detection.

5.2 Instruments

For the measurement of pain intensity, the Numeric Rating Scale (NRS) was used. The respondent selects their perceived pain on a scale of 0 (no pain) and 10 (highest conceivable pain). The evaluation of the pain intensity is done with the help of a digital scale on a screen, which is operated with a mouse (see Figure 3). Using a digital scale enables a continuous pain rating, so each frame in the video can be annotated with a self-rating. The ratings are recorded during the entire pain induction, with a sampling rate of 10 Hz and were later upsampled to 30 Hz to match the video frame rate of 30 frames per seconds. The test person's assessment of the pain intensity (self-rating) and an external assessment (observer-rating) are recorded by a

second test supervisor, who was able to monitor the test person on a screen during the test. The observer-rating will not be used in the current study but was recorded for the potential use in future analyses.

In addition to pain intensity and demographic data, the participants had to fill in the Pittsburgh Sleep Quality Index (PSQI, Buysse, Reynolds, Monk, Berman, & Kupfer, 1989), the German version by Riemann (1996) which checks the quality of sleep the participant had and the current state of tiredness, as being tired could potentially influence facial expressions due to for example an increase in eye blinking or slow reactions.

Figure 3. Evaluation of pain intensity by the test person on the NPRS.

6. Procedure

6.1 Technical Setup

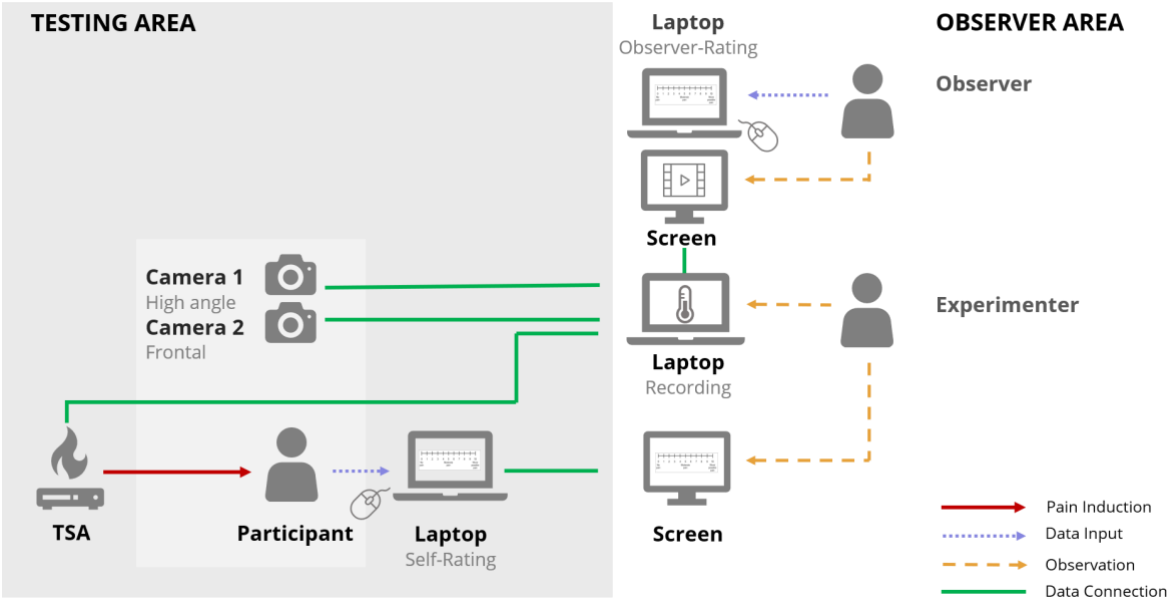

Figure 4. Technical Setup

An overview of the used technical setup during the procedure can be seen in Figure 4. The participant sat in front of a laptop, which was utilised to operate the NRS rating scale via a mouse. Two video cameras (Logitech C920) recorded the face of the subject frontally and from a higher angle (see Figure 5) with 30 frames per seconds. The test supervisor and observer were positioned orthogonally to the subject, as depicted on the right side of Figure 4. The observer also used a laptop to rate his observations, and a third laptop was used by the experimenter to operate the TSA II, whose thermode was attached to the participant's forearm. To minimise an observer effect, the view between the participant and the observer area was blocked by a curtain. The observer could see the participant on a screen in front of them to give observer ratings. In addition, the experimenter could see the self-ratings of the participant in real-time on another monitor. This allowed checking the distribution of the pain ratings to make any necessary temperature adjustments between the trials, for example setting the temperature lower if too many high pain ratings were reached.

Figure 5. Two camera angles used during pain stimulation. The left picture shows the recording from a higher angle, the left picture the frontal recording.

6.2 Pain induction and measurement of pain intensity

The pain induction protocol was based on methods found in the literature who have also used heat stimulation through a thermode (Angst et al., 2009; Hammal et al., 2008; Walter et al., 2013). In our experiment, the Thermal Sensory Analyser II (TSA II, Medoc, http://www.medoc-web.com) was used for the pain induction. After filling out the demographic questionnaire and informed consent (see Appendix A), the participant was asked to sit down in front of the laptop in a comfortable position. The thermode was then attached to the volar aspect of the dominant forearm. Prior to the start of the experiment, the individual pain limits of the volunteers were determined using the 'method of limits', whereby the

maximum of 50°C was not exceeded. With this method, the temperature increased at 1°C/s until the subject pressed the button of a mouse, which immediately brought the temperature down to the baseline of 32°C and then increased again. First, the participant was asked to press the button once they first felt a painful sensation (the *pain threshold*) and when they reached the highest pain imaginable (the *pain limit*). These pain levels were measured ten times in total in alternating order, five times for each pain level. Immediately afterwards the medium pain intensity was measured by asking the participant to press the button once a pain sensation of 5 out of 10 is reached, which was repeated 5 times as well. To determine the individual pain levels, the average value of the last 3 of the 5 measurements were used. Based on the three obtained components, 5 different levels of temperatures were calculated by taking the average values between two pain levels (see Figure 6).

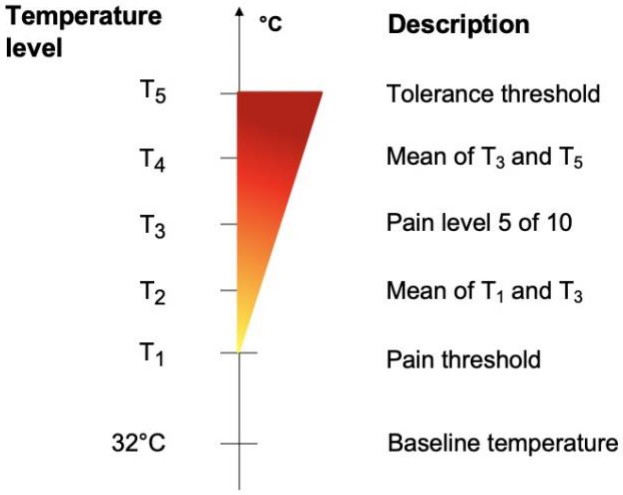

Figure 6. Used pain levels and their calculation (adapted from Werner et al., 2016)

For the following main examination, the thermode was attached to the non-dominant forearm. This had the purpose to exclude temperature adaptation of the skin or causing damage to it and further allowed the participant to operate the mouse with their dominant hand. The pain induction took place over 2x15 minutes, with a short break in between, in which the previously determined temperatures were applied in random order. Each temperature was applied twice with an increase of 0.5°C/s and 3 times with an increase of 2°C/s. The inter-stimulus interval was 5-15 seconds (randomised) to prevent temperature habituations.

The aim of using these five different temperatures was to achieve an even distribution of subjective pain levels from 0-10 while the different temperature increases were intended to

achieve more variability within the stimuli. During the break between the two pain stimulation sequences, the temperatures were adjusted if no satisfactorily even distribution of pain ratings was observed in the previous run. In addition, the position of the thermode was changed to avoid temperature habituations or damage to the skin.

7. Data Analysis

7.1 Data preparation

The obtained video data were analysed by using the facial analysis toolkit 'OpenFace' (Baltrušaitis et al., 2018) with which facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation were obtained for each frame in the video. The data were then merged with the self-rating, observer rating and demographic data of each participant. Due to the size of the obtained dataset, the dataset was manually reduced into two subsets before further analysis was possible with the given hardware. Beforehand, the data were filtered to contain only observations with the highest confidence that OpenFace indicated to have for the feature detection. Because the participants continuously operated a horizontal slider to indicate their pain levels, it was also decided to remove the features that represent eye movements and head movements on the yaw axis, which likely could be influenced by following the position of the slider. This left 419 of the initial 709 features.

The first subset consisted out of 100 'no pain' (0) self-ratings and 100 'high pain' (7-10) self-ratings for each participant and was used to do a feature selection, which was not possible with the whole dataset given the available hardware. The feature ranking was performed in the programme 'R' by building a Learning Vector Quantization (LVQ) model (adapted from Brownlee, 2019) using 10-fold cross-validation (see Appendix B for the used R code). In a 10-fold cross-validation, the data set is randomly divided into 10 folds. One of these folds is held back as test set while the rest is being used for training. This is repeated until each of the 10 folds has been used as a test set once. Using the 10-fold cross-validation, instead of a simple split of a test and training set, generally leads to a less biased estimate of the model's capability (Brownlee, 2018). The features were then ranked by their importance for discriminating painful from non-painful expressions using the varImp function of the caret package in R (Kuhn, 2019). The varImp function computes the ROC curve for each variable by computing the false positive rate against the true positive rate for a series of cutoffs. The area under the ROC curve is then used as measurement for the variable importance (Kuhn, 2019). Features with a high predictive power, i.e. which have a high impact on the performance

accuracy, therefore receive a higher importance than redundant features. From this feature ranking the most important 30 facial features that were obtained from OpenFace were then used for further analysis. The reduction to 30 features made it more feasible regarding computing power to do subsequent analysis on the dataset.

The second subset was reduced to contain all the 'high pain' ratings for each participant, matched with an equal amount of 'no pain' ratings and resulted in a dataset of 347,113 observations and was further used for the predictive analysis. Before each analysis, the self-ratings were converted into a categorical variable to represent the binary classification of 'no pain' and 'high pain'.

7.2 Predictive analysis

The overall project aims to develop a system that is capable to automatically detect pain through a multi-modal device, of which facial recognition will be one part of. It is therefore intended to implement machine learning algorithms that can classify the perceived pain of the patient in real-time. For this reason, a machine learning approach was also chosen for the analysis of the acquired dataset. In particular, the *k*-nearest neighbour (kNN) algorithm was used to analyse it (see Appendix C for an example code). The kNN algorithm is a lazy learning algorithm, which means that it learns by memorizing previous observations (Gama & de Carvalho, 2012). It then tries to map the new observation to the number of *k*-nearest neighbours to categorise it. Regarding the given dataset, the advantage of the kNN algorithm is that it is possible to give *k* a high number. As the observations consist out of 30 frames per second, the observations that are temporally close to each other will likely be quite similar to each other. Choosing a high *k* can therefore help for a more global pattern recognition. The chosen k for the following analysis is the square root of observations, which equals $k = 519$. The model therefore looks for the nearest 519 neighbours in the training set to classify a new observation in the test set. Approximately 80% of the datasets were used to train the model and the other 20% was set aside to test it.

The kNN algorithm was first performed on the subset consisting out of the resulting 30 features that were found during the feature selection, as explained in 7.1. The analysis was further done for both between and within-subjects. For the 'between-subjects' condition, the training and test set for the analysis were divided by participant, so that the test set does not contain any observations from a participant that is in the training set. For the 'within-subjects' condition, the test and training sets were randomly divided with an 80/20 split, so that the test

set consist of observations from participants that are also in the training set. It was checked that both datasets have data from every subject, however it was not controlled for an even distribution of data.

The same analysis was done with the dataset which features consist out of the AU's that were found in the literature to be relevant for pain expression (see p. 5). Out of these 12 identified AUs, OpenFace only computes 8 (AU 4, 6, 7, 9, 12, 20, 25 and 45), so the analysis has to be limited to these.

8. Results

8.1 Feature reduction

The goal of the feature reduction was to reduce the dataset to as few features as possible to improve the calculation time and also to make it more feasible for e.g. mobile applications. It was decided to initially reduce the features to 30 and do subsequent tests to see if even smaller feature subsets can yield promising results. Depicted in Figure 7 are the results of the top 30 features, ranked by their importance in distinguishing painful from non-painful states. The importance rating here is measured by the calculated area under the ROC curve. In the following, the subset consisting out of these top 30 features will be referred to as '*Subset A'.* It is visible that AU01 and AU02 seem to have the most importance in distinguishing pain from no pain, being the only features reaching an importance above .65. Note that all of the depicted features have an importance over .5 as the graph does not start at zero. The meaning of the features is explained in more detail in Table 4, however the authors did not provide further information about the non-rigid face parameters (the features starting with p_).

The goal of the feature analysis is to reduce the features to make the database more feasible to work with, not only during the analysis but also for potential real-time applications. To see if the number of features can be even further reduced, three additional subsets were selected to test the model's performance, as indicated by the lines in Figure 7. The first subset, Subset A.1, includes the two most important features which also clearly have a higher significance than any other feature in Subset A. Subset A.2 includes the following 4 features. The dividing line for Subset A.3 was chosen because the importance of the following features is relatively close to each other, which suggests that they might not add much more to the model's performance compared to the features ranked above.

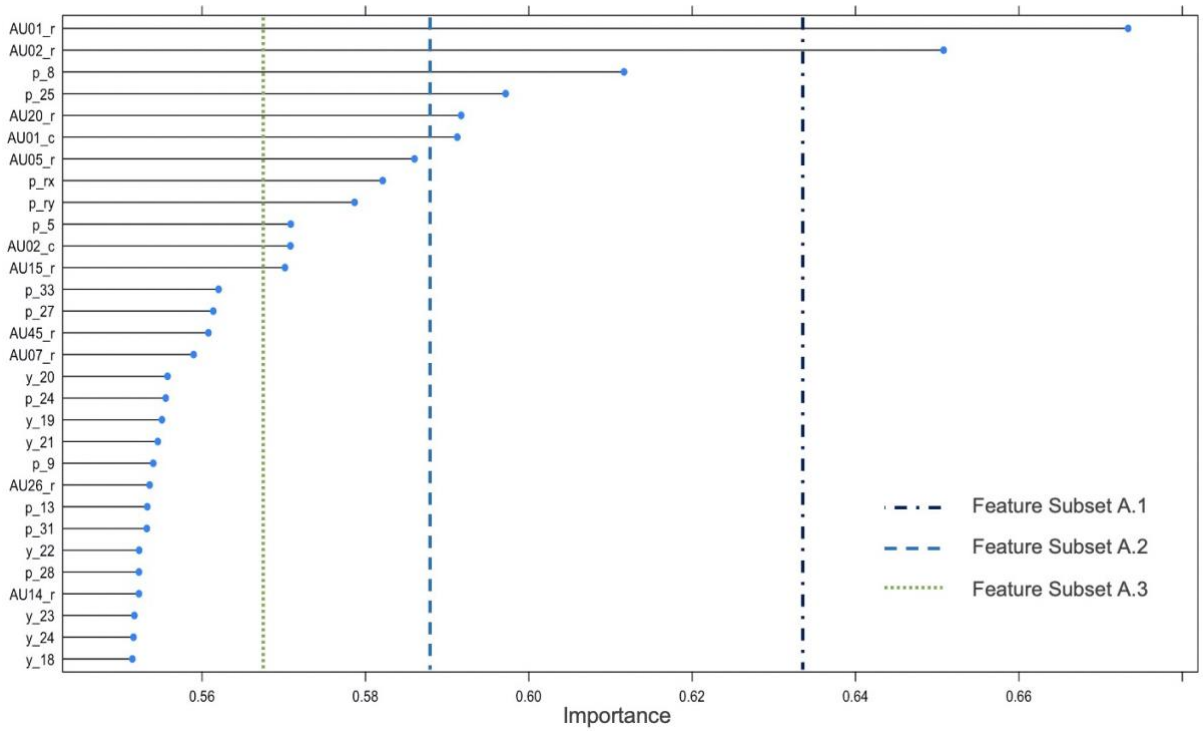

Figure 7. Top 30 features (Subset A) ranked by importance for discriminating painful from non-painful expressions. The lines indicate the different subsets of features used for further analysis.

Table 4.
*Description of the found features in the top 30 important features.**

Features	Description	Intensity (r)	Presence (c)
AU01	Inner brow raising	✓	✓
AU02	Outer brow raising (unilateral)	✓	✓
AU05	Upper lip raising	✓	
AU14	Dimpler	✓	
AU15	Lip corner depressor	✓	
AU20	Lip stretcher	✓	
AU26	Jaw drop	✓	
AU45	Blink	✓	
p_rx	Rigid face parameter (pitch rotation)	-	-
p_*n*	Non-rigid face parameters**	-	-
x_*n*, y_*n*	location of 2D landmarks in pixels**	-	-
X_*n*, Y_*n*	location of 3D landmarks in millimetres**	-	-

Note. '✓' = present in the top 30 features; '-' = not applicable
* Information for the AUs taken from Baltrušaitis et al. (2016)
** $n \in \{1,\ldots,34\}$, information taken from Baltrušaitis (2019), see Figure 1 for the Landmark Index

8.2 Results of the predictive analysis

The first two prediction calculations were done with Subset A, which includes all features presented in Figure 7. The trained model resulted in an accuracy of 51.01% when predicting the pain ratings on the dataset split between the subjects (see Table 5 for the confusion matrix). When predicting the accuracy on the randomly split dataset within the participants, an accuracy of 80.33% (see Table 6) was achieved on the test dataset.

Table 5.
Confusion Matrix. Top 30 feature subset between-subjects.

	No Pain	Pain
No Pain	10948	10381
Pain	22741	23538
Accuracy	51.01	

Table 6.
Confusion Matrix. Feature Subset A, within-subjects.

	No Pain	Pain
No Pain	28847	8347
Pain	5311	26918
Accuracy	80.33	

In the following, the additional feature subsets of Subset A will be analysed, to see if even a smaller number of features can be used to achieve a comparable performance accuracy. The first subset, Subset A.1, with the two most important features fails due to too many ties which means that for the selected method, too many neighbours were equidistant to the target for the algorithm to come to a conclusion. It could be solved by increasing the allowed number of ties in the code, but this resulted in an increase of computing power that was not feasible with the given hardware. The second feature subset, Subset A.2, included 6 features (AU01_r, AU02_r, AU05_r, p_8, p_25, AU20_r and AU01_c). For the dataset that was split between-subjects, the accuracy was at chance level with 48.15% (see Table 7). Within the participants, the model performed slightly above chance with an accuracy of 66.26% (see Table 8).

Table 7.
Confusion Matrix of Feature Subset A.2, between-subjects

	No Pain	Pain
No Pain	6371	7738
Pain	27318	26181
Accuracy	48.15	

Table 8.
Confusion Matrix of Feature Subset A.2, within-subjects

	No Pain	Pain
No Pain	22499	11762
Pain	11659	23503
Accuracy	66.26	

Subset A.3 consists of 11 features. The results between the participants are slightly higher than A.2, but still at chance level with an accuracy of 53.42% (see Table 9). The model's performance within the participants resulted in a significantly higher accuracy compared to A.2, with an accuracy of 74.75% (see Table 10).

Table 9.
Confusion Matrix of Feature Subset A.3, between-subjects.

	No Pain	Pain
No Pain	7421	5224
Pain	26268	29695
Accuracy	53.42	

Table 10.
Confusion Matrix of Feature Subset A.3, within-subjects.

	No Pain	Pain
No Pain	26574	9946
Pain	7584	25319
Accuracy	74.75	

Before and during the experiment the participants were checked for their perceived tiredness. As some showed observable signs of fatigue, like closing their eyes for a long period of time, an increase of eye blinking and/or slow movements, it was decided to run an analysis where the participants that indicated a high level of fatigue (7 or above) were removed, to control for these factors. This left a database of 45 participants. The calculated model resulted in an accuracy of 58.62% between-subjects (see Table 11), which is higher than the models analysed above for between-subjects, which have been on chance level. For within-subjects, the accuracy was also higher compared to the models that were not controlled for tiredness, with an accuracy of 85.14% (see Table 12) compared to the previous highest score of 80.33%.

Table 11.
Confusion Matrix. Feature Subset A, between-subjects, controlled for tiredness.

	No Pain	Pain
No Pain	13283	8517
Pain	21236	28861
Accuracy	58.62	

Table 12.
Confusion Matrix. Feature subset A, within-subjects, controlled for tiredness.

	No Pain	Pain
No Pain	25741	5135
Pain	3722	25009
Accuracy	85.14	

In order to compare the performance of the found feature sets, a feature subset with only the pain associated AUs (AU 4, 6, 7, 9, 12, 20, 25 and 45) was created. An accuracy of 55.88% was achieved when predicting the pain ratings on the dataset split between-subjects (see Table 13) and an accuracy of 74.32% was achieved for the dataset split within-subjects (see Table 14). The performance is comparable to the Subset A.3, which had 3 more features included.

Table 13.
Confusion Matrix. AU subset between-subjects.

	No Pain	Pain
No Pain	16644	12783
Pain	17045	21136
Accuracy	55.88	

Table 14.
Confusion Matrix. AU subset within-subjects.

	No Pain	Pain
No Pain	26622	10296
Pain	7534	24971
Accuracy	74.32	

8.3 Feature subset excluding non-rigid face parameters

The analysed subset also included non-rigid face parameters. As these features not only describe distortions due to facial expressions but are also used to identify different faces, it could potentially have an influence on the model's performance for the between subject designs. Therefore, it was decided to run a second feature analysis, that excluded the non-rigid face parameters. The results of the top 30 features can be seen in Figure 8. In the following, it will be referred to this feature subset as '*Subset B*'. Subset B was also divided further into a smaller subset of the top 7 features, to compare the performance with the smaller subsets of Subset A.

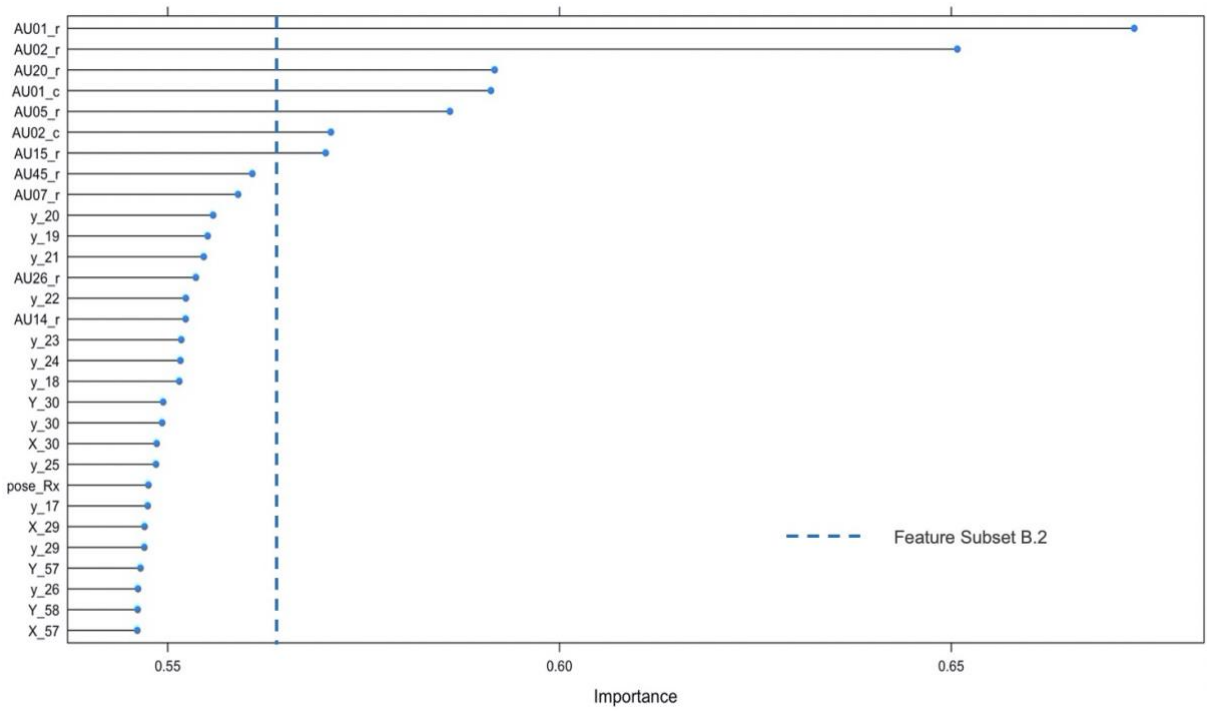

Figure 8. Top 30 features (Subset B), ranked by importance for discriminating painful from non-painful expressions. The line indicates the subset of features used for further analysis.

Comparing Subset A (Figure 7) with Subset B (Figure 8), it can be seen that the order of the features that are present in both datasets remain the same. The only difference is that by removing some of the previously top-scoring features, more features of lower importance were added in Subset B. Analysing Subset B, an accuracy of 54.93% was achieved between the subjects, and an accuracy of 79.17% within-subjects (see Table 15 and Table 16). The results are overall similar to Subset A, both for between and within-subjects.

Table 15.
Confusion Matrix. Feature Subset B, between-subjects.

	No Pain	Pain
No Pain	9016	5801
Pain	24673	28118
Accuracy	54.93	

Table 16.
Confusion Matrix. Feature Subset B, within-subjects.

	No Pain	Pain
No Pain	278510	8814
Pain	5643	26451
Accuracy	79.17	

Subset B.2 results in a feature set with 7 features. Between the participants, the performance accuracy was at 53.16% and within-subjects at 66.74% (see Table 17 & Table 18) which is comparable to the results of Subset A.2, which included 6 features (see Table 7 & Table 8).

Table 17.
Confusion Matrix. Subset B.2, between-subjects.

	No Pain	Pain
No Pain	14424	12404
Pain	19265	21515
Accuracy	53.16	

Table 18.
Confusion Matrix. Subset B.2, within-subjects.

	No Pain	Pain
No Pain	21973	10907
Pain	12185	24358
Accuracy	66.74	

In a previous analysis, it has been established that tiredness can potentially influence the accuracy of the model. To ensure a fair comparison between the subsets, another analysis was run on Subset B, which was controlled for tiredness by removing the participants who indicated a high level of fatigue. Similar to Subset A, the accuracy improved for the within-subject design (see Table 20). However, the accuracy for the between-subject design even

slightly decreased with an accuracy of 48.97% at chance level (see Table 19), while an accuracy slightly above chance was achieved for Subset A.

Table 19.
Confusion Matrix. Top 30 of adjusted Feature set controlled for tiredness, between-subjects.

	No Pain	Pain
No Pain	12936	15109
Pain	21583	22269
Accuracy	48.97	

Table 20.
Confusion Matrix. Top 30 of adjusted Feature set controlled for tiredness, within-subjects.

	No Pain	Pain
No Pain	25226	5873
Pain	4137	24371
Accuracy	83.21	

9. Discussion

This project aimed to develop a dataset for video-based pain detection and evaluating it to answer the questions whether or not it is possible to discriminate between painful and non-painful expressions with it and which features might be the most relevant to achieve this. Consequently, it was not surprising that discriminating pain was successful for person-specific models, but not for general models. The dataset overall performed well when discriminating new data from participants that the model was already trained on but performed merely better than guessing on the dataset split between-subjects. These findings are in line with Werner et al. (2013) who also found good results when predicting the pain levels within the participants and performance at chance level for a general model, and Lucey at al. (2012) who also achieved a performance at around 80% for automatic pain detection. This underlines that person-specific models seem to be the way to go for automatic pain recognition. They have the drawback of requiring a personalised trained model for every person, which is not feasible in many real-world applications. However, especially in intensive care people stay for a longer time and

setting aside a few minutes to train the model would still benefit the patient in the long term through a quicker detection of a painful state.

When implementing automatic and real-time pain recognition, a system would benefit from a small feature subset, as it would take less time and less powerful hardware to compute. For this reason, it was decided to analyse a maximum of 30 features, which were specified through the importance ranking. To see if the feature set could be reduced even more than the initial 30 features, additional subsets were created with which the dataset was analysed. Increasing the feature set from the top 6 to the top 12 features resulted in a performance increase of roughly 11% with 77.24% (Subset A.3) compared to 66.26% (Subset A.2). Including all 30 (Subset A), thus 21 more features, gave an increase in accuracy of only around 3%, with 80.33%. The performance accuracy did not increase massively from 12 to 30 features, so it could be reasonable to reduce the features to these 12 to decrease the processing time, especially if limited resources are available. Including more features still can result in a higher accuracy and could also potentially compensate for the differences people have in expressing pain and thus perform better over a higher variability of people.

During the experiment, the participants rated their tiredness as it could potentially influence the dataset, for example through slow movements or general lack of expressions. Indeed, removing the participants with high ratings of tiredness significantly increased the model's performance. For the within-subject design, the accuracy went from 80.33% up to 85.14% and the model built on the dataset split between-subjects performed slightly above chance level with an accuracy of 58.62%. This indicates that considering tiredness as a confounding variable in the research of automatic painful expression recognition could be important for future studies.

To compare how well the ranked feature set performs, the same analysis was done with only the AU's that were found to be relevant in the literature. As OpenFace does not compute all AUs available, only 8 of the 12 AUs could be used, so the results may not be fully representable to how the full set of pain-related AUs would perform. The results were comparable to Subset A.3, which also had a similar number of features. This indicates that the found features are comparably well suited for predicting pain as the AUs.

A lot of the features that had the highest importance concern ones that describe non-rigid face parameters. These are used not only to detect distortions of the face due to expressions but also to identify faces, hence, they could potentially impact the performance of the general, between-subjects model. Therefore, it was decided to do a second feature ranking

where those features were omitted. The model performance of the resulting feature set (Subset B) was comparable with the between and within-subjects performance obtained with Subset A. For the top 30 features within-subjects, a slightly lower but comparable accuracy of 79.17% for Subset B, compared to 80.32% for Subset A was achieved. Excluding participants with a high level of tiredness slightly increased the performance accuracy for the within-subjects design but not for between the participants, as it was the case for the Subset A that included the non-rigid face parameters. The only notable difference between Subset A and Subset B was thus the performance increase of the general model that was controlled for tiredness, otherwise the models performed equally well in distinguishing painful from non-painful behaviour given a similar number of features was included. This does not correspond to the results of Ashraf (2009) who found a performance increase when using non-rigid parameters for person-specific models.

The results found in the current study suggest that non-rigid face parameters perform equally well in facial pain recognition as AUs for person-specific model and the only performance increase compared to the AUs was found between the participants. Given that person-specific models seem to be the best option for facial expression recognition, non-rigid face parameters seem to have no benefit compared to AUs. As additional remarks, it was observed during the experiment that not everyone showed visual changes in their facial expressions even when they experienced high levels of pain, which has also been the case in the study of Werner et al. (2013). This strengthens the assumption that the expressiveness of pain differs widely between people and automatic painful facial expression might not work for everyone.

9.1 Limitations

When interpreting the results, one has to keep in mind the limitations of this study. As a laboratory study, the results naturally cannot completely be transferred to real-life situations. People can be inhibited showing emotions when they feel observed or might feel the need to prove themselves to not show any signs of pain. It also should be noted that the two test operators were both female, which due to sociological concepts could potentially influence how male participants behave during the pain stimulation.

Another limitation comes with the reliance on the participant's self-rating of pain. Before the experiment, they had to indicate the temperature at which the highest pain level they could endure was reached. Some participants reported being afraid of feeling pain which made

them indicate a much lower temperature initially. This became obvious for some participants when they did not continue indicating high pain levels during the trial. This further strengthens the need for using person-specific models for facial pain recognition, because many variables like someone's experience, fear or their expectations can influence the way they judge their perceived pain.

The self-rating was further indicated on a visual slider on a laptop which made a continuous pain rating possible but came with the disadvantage of influencing features like the gaze or the head-turning from right to left. The features describing this yaw rotation of the head was omitted, but it cannot be confidently said that the slider did not affect the roll or pitch of the head as well. It therefore could not be accounted for every head pose during the analysis although it potentially could be relevant for the prediction of pain as suggested in previous studies (e.g. Lucey et al., 2012; Werner et al., 2013). An alternative for using the slider could be to use for example a device to measure grip force, which the participant has to press harder the more pain he feels. This would ensure that the gaze is not affected while still making continuous pain ratings possible. Another limitation is that the pain induction was just a few seconds long. It is common for people in a clinical setting to experience pain over a longer period of time or to have a very slow increase in pain and their reaction might be different to a brief and phasic pain stimulus.

9.2 Future Research

The current study presented just a first impression of how a model could perform on the new dataset. Further studies should implement more advanced machine learning techniques and for example look for combinations of features that indicate pain instead of just looking at each feature individually. The current study was to this date also the first who studied facial expression using continuous self-ratings of pain on a frame level instead of using a sequence level ground truth. It can be expected that facial expressions are temporally close to the feeling of pain, but given the complexity of pain expressions, future research should look into the relationship between frame level and sequence level analysis of pain prediction and if a combination of the two could yield better results. Additionally, a study that also uses continuous pain ratings but with devices that enable the participant to move more freely – such as the proposed method of grip force – could help in understanding the relation of pain and eye or head movement better.

Conclusion

In summary, the collected dataset has been successfully used to predict pain through facial features when using person-specific models. For these, no significant difference was found when omitting non-rigid face parameters from the calculated feature set or when using only pain-related Action Units. However, the model benefitted from omitting persons with a high level of tiredness, indicating its relevance as a control variable in facial expression research. As this is the first database that offers self-ratings as ground truth on frame level, this study is also only a starting point in this regard. More research should be done on the presented database by using more advanced analysis techniques, e.g. for finding combinations of features that represent pain rather than looking at each feature individually.

References

Angst, M. S., Tingle, M., Phillips, N. G., & Carvalho, B. (2009). Determining heat and mechanical pain threshold in inflamed skin of human subjects. *Journal of Visualized Experiments : JoVE*, (23). https://doi.org/10.3791/1092

Ashraf, A., Lucey, S., Cohn, J., Chen, T., Ambadar, Z., Prkachin, K., & Solomon, P. (2009). The Painful Face - Pain Expression Recognition Using Active Appearance Models. *Image and Vision Computing*, *27*, 1788–1796. https://doi.org/10.1016/j.imavis.2009.05.007

Baltrušaitis, T. (2019). OpenFace Output Format. Retrieved September 22, 2020, from https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). *OpenFace: An open source facial behavior analysis toolkit*. 1–10. https://doi.org/10.1109/WACV.2016.7477553

Baltrušaitis, T., Zadeh, A., Chong Lim, Y., & Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.

Blom, J., & Lubbe, R. (2017). Endogenous spatial attention directed to intracutaneous electrical stimuli on the forearms involves an external reference frame. *International Journal of Psychophysiology*, *121*. https://doi.org/10.1016/j.ijpsycho.2017.08.006

Brownlee, J. (2018). A Gentle Introduction to k-fold Cross-Validation. Retrieved January 22, 2021, from https://machinelearningmastery.com/k-fold-cross-validation/

Brownlee, J. (2019). Feature Selection with the Caret R Package. Retrieved May 3, 2020, from https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/

Buysse, D. J., Reynolds, C. F. 3rd, Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. https://doi.org/10.1016/0165-1781(89)90047-4

Carli, G. (2007). *Tourniquet Test BT - Encyclopedia of Pain* (R. F. Schmidt & W. D. Willis, Eds.). https://doi.org/10.1007/978-3-540-29805-2_4539

Chen, Z., Ansari, R., & Wilkie, D. J. (2018). Automated Pain Detection from Facial Expressions using {FACS:} {A} Review. *CoRR*, *abs/1811.0*. Retrieved from http://arxiv.org/abs/1811.07988

Craig, K. D., Hyde, S. A., & Patrick, C. J. (1991). Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain. *Pain*, *46*(2), 161–171. https://doi.org/10.1016/0304-3959(91)90071-5

Craig, K. D., Prkachin, K. M., & Grunau, R. V. E. (1992). The facial expression of pain. In

Handbook of pain assessment. (pp. 257–276). New York, NY, US: The Guilford Press.

Dale, C. M., Prendergast, V., Gélinas, C., & Rose, L. (2018). Validation of The Critical-care Pain Observation Tool (CPOT) for the detection of oral-pharyngeal pain in critically ill adults. *Journal of Critical Care*, *48*, 334–338. https://doi.org/https://doi.org/10.1016/j.jcrc.2018.09.024

Ekman, P., & Rosenberg, E. L. (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS) (2nd ed.)*. https://doi.org/https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

Falch, C., Vicente, D., Haeberle, H., Kirschniak, A., Müller, S., Nissan, A., & Brücher, B. (2014). Treatment of acute abdominal pain in the emergency room: A systematic review of the literature. *European Journal of Pain (London, England)*, *18*. https://doi.org/10.1002/j.1532-2149.2014.00456.x

Fillingim, R. B., Loeser, J. D., Baron, R., & Edwards, R. R. (2016). Assessment of Chronic Pain: Domains, Methods, and Mechanisms. *The Journal of Pain : Official Journal of the American Pain Society*, *17*(9 Suppl), T10-20. https://doi.org/10.1016/j.jpain.2015.08.010

Fraser, N., & Grady, K. (2008). Review of Principles of Pain Management for Anaesthetists. *European Journal of Anaesthesiology (EJA)*, *25*(1). Retrieved from https://journals.lww.com/ejanaesthesiology/Fulltext/2008/01000/Review_of_Principles_of_Pain_Management_for.20.aspx

Gama, J., & de Carvalho, A. C. P. L. F. (2012). Machine Learning. In *Machine Learning: Concepts, Methodologies, Tools and Applications* (pp. 13–22). https://doi.org/http://doi:10.4018/978-1-60960-818-7.ch102

Hadjistavropoulos, T., Hunter, P., & Dever Fitzgerald, T. (2009). Pain assessment and management in older adults: Conceptual issues and clinical challenges. *Canadian Psychology/Psychologie Canadienne*, *50*(4), 241–254. https://doi.org/10.1037/a0015341

Hammal, Z., & Cohn, J. (2012). Automatic detection of pain intensity. *ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction*, 47–52. https://doi.org/10.1145/2388676.2388688

Hammal, Z., Kunz, M., Arguin, M., & Gosselin, F. (2008). Spontaneous Pain Expression Recognition in Video Sequences. In *In BCS International Academic Conference 2008 - Visions of Computer Science*.

Hansen, G. R., & Streltzer, J. (2005). The psychology of pain. *Emergency Medicine Clinics of North America*, *23*(2), 339–348. https://doi.org/10.1016/j.emc.2004.12.005

Hicks, C., Baeyer, C., Spafford, P., van Korlaar, I., & Goodenough, B. (2001). The Faces Pain Scale - Revised: Toward a common metric in pediatric pain measurement. *Pain*, *93*, 173–183. https://doi.org/10.1016/S0304-3959(01)00314-1

Institute of Medicine. (2011). Relieving pain in America: A blueprint for transforming prevention, care, education, and research. In *Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research*. https://doi.org/10.17226/13172

International Association for the Study of Pain. (2017). IASP Terminology. Retrieved December 14, 2017, from https://www.iasp-pain.org/Education/Content.aspx?ItemNumber=1698#Pain

Jensen, M. P. (2011). *The pain stethoscope: A clinician's guide to measuring pain*. Tarporley: Springer Healthcare Ltd. : Imprint: Springer Healthcare.

Kuhn, M. (2019). The caret Package. Retrieved January 25, 2021, from https://topepo.github.io/caret/index.html

Kunz, M., Mylius, V., Schepelmann, K., & Lautenbacher, S. (2004). On the relationship between self-report and facial expression of pain. *The Journal of Pain : Official Journal of the American Pain Society*, *5*(7), 368–376. https://doi.org/10.1016/j.jpain.2004.06.002

Kunz, M., Scharmann, S., Hemmeter, U., Schepelman, K., & Lautenbacher, S. (2008). The facial expression of pain in patients with dementia. *Pain*, *133*, 221–228. https://doi.org/10.1016/j.pain.2007.09.007

Loggia, M. L., Juneau, M., & Bushnell, M. C. (2011). Autonomic responses to heat pain: Heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity. *Pain*, *152*(3), 592–598. https://doi.org/10.1016/j.pain.2010.11.032

Lopez-Martinez, D., & Picard, R. (2017). *Multi-task neural networks for personalized pain recognition from physiological signals*. https://doi.org/10.1109/ACIIW.2017.8272611

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., Chew, S., & Matthews, I. (2012). Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing*, *30*(3), 197–205. https://doi.org/https://doi.org/10.1016/j.imavis.2011.12.003

Mouraux, A., Iannetti, G. D., & Plaghki, L. (2010). Low intensity intra-epidermal electrical stimulation can activate Aδ-nociceptors selectively. *Pain*, *150*(1), 199–207. https://doi.org/10.1016/j.pain.2010.04.026

Pantic, M., & Rothkrantz, L. J. M. (2004). Facial action recognition for facial expression

analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(3), 1449–1461. https://doi.org/10.1109/TSMCB.2004.825931

Park, G., Kim, C. W., Park, S. B., Kim, M. J., & Jang, S. H. (2011). Reliability and usefulness of the pressure pain threshold measurement in patients with myofascial pain. *Annals of Rehabilitation Medicine*, *35*(3), 412–417. https://doi.org/10.5535/arm.2011.35.3.412

pinprick test. (2009). Retrieved August 19, 2019, from Mosby's Medical Dictionary, 8th edition website: https://medical-dictionary.thefreedictionary.com/pinprick+test

Prkachin, K. M. (1992). The consistency of facial expressions of pain: A comparison across modalities. *Pain*, Vol. 51, pp. 297–306. https://doi.org/10.1016/0304-3959(92)90213-U

Prkachin, K. M., & Solomon, P. E. (2008). The structure, reliability and validity of pain expression: evidence from patients with shoulder pain. *Pain*, *139*(2), 267–274. https://doi.org/10.1016/j.pain.2008.04.010

Riemann, D., & Backhaus, J. (1996). *Behandlung von Schlafstörungen: ein psychologisches Gruppenprogramm*. Retrieved from https://books.google.de/books?id=ZcfWAAAACAAJ

Sikka, K. (2014). *Facial Expression Analysis for Estimating Pain in Clinical Settings*. https://doi.org/10.1145/2663204.2666282

Świeboda, P., Filip, R., Prystupa, A., & Drozd, M. (2013). Assessment of pain: types, mechanism and treatment. *Annals of Agricultural and Environmental Medicine : AAEM*, . *1*(July 2014), 2–7.

von Baeyer, C. L., Piira, T., Chambers, C. T., Trapanotto, M., & Zeltzer, L. K. (2005). Guidelines for the cold pressor task as an experimental pain stimulus for use with children. *The Journal of Pain*, *6*(4), 218–227. https://doi.org/10.1016/j.jpain.2005.01.349

Walter, S., Gruss, S., Ehleiter, H., Junwen Tan, Traue, H. C., Werner, P., … Moreira da Silva, G. (2013). The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. *2013 IEEE International Conference on Cybernetics (CYBCO)*, 128–131. https://doi.org/10.1109/CYBConf.2013.6617456

Wehe, S. (2013). *Analyse der optimalen Stimulationstemperatur zur Messung der Schmerztoleranz*. Georg-August-Universität zu Göttingen.

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., & Traue, H. (2016). Automatic Pain Assessment with Facial Activity Descriptors. *IEEE Transactions on Affective Computing*, *PP*, 1. https://doi.org/10.1109/TAFFC.2016.2537327

Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., & Traue, H. (2013). *Towards Pain*

Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges. https://doi.org/10.5244/C.27.119

Whitehill, J., Bartlett, M. S., & Movellan, J. R. (2013). Automatic facial expression recognition. *Social Emotions in Nature and Artifact*, *88*.

Wong, F., Vierck, C. J., Riley, J. L. I., King, C., & Mauderli, A. P. (2010). A new thermal stimulation method for human psychophysical studies: Pain intensity clamping. *Journal of Neuroscience Methods*, *188*(1), 83–88. https://doi.org/10.1016/j.jneumeth.2010.02.001

Appendix A

Informed Consent

Beschreibung der Studie

Wir bedanken uns für Ihr Interesse an unserer Studie. Für Ihre vollständig abgeschlossene Teilnahme erhalten Sie unmittelbar nach Abschluss des Experiments eine monetäre Entschädigung über 50 Euro. Den Empfang der Aufwandsentschädigung quittieren Sie uns mit Ihrem Namen und Ihrer Unterschrift. Die Quittung wird getrennt von allen anderen Daten dieser Studie aufbewahrt und dient dem Auftraggeber als Nachweis darüber, dass die Gelder für die Aufwandsentschädigung bestimmungsgemäß verwendet wurden.

Während des Experiments wird Ihnen am Unterarm ein Sensor angelegt, der sich soweit erhitzt, dass es für Sie teilweise schmerzhaft werden kann. Es besteht allerdings nie die Gefahr durch die zugeführte Wärme Schäden davonzutragen und Sie können das Experiment jederzeit unterbrechen. Die Gesamtdauer der Studie wird etwa 120 Minuten in Anspruch nehmen, wovon das Experiment 90 Minuten dauern wird. Der Zweck der Studie ist es, mit Hilfe der ausgewerteten Daten ein System zur automatischen Schmerzerkennung zu entwickeln.

Zur Verfolgung des oben genannten Studienzwecks werden von Ihnen personenbezogen Daten erhoben, gespeichert und ausgewertet. Hierzu gehören neben demografischen Daten und Angaben zu Ihrer Gesundheit auch Video- und Tonaufnahmen, die während der experimentellen Durchführung aufgezeichnet werden. Video- und Tonaufnahmen sind personalisierte Daten, d.h. hierüber lassen sich Rückschlüsse auf Ihre Identität ziehen. Die Verwendung dieser Daten setzt vor der Teilnahme an der Studie folgende freiwillig abgegebene Einwilligungserklärung voraus, d.h. ohne die nachfolgende Einwilligung können Sie nicht an der Studie teilnehmen.

Einwilligungserklärung

Ich habe die Beschreibung der Studie gelesen und erkläre mich damit einverstanden, dass im Rahmen dieser Studie erhobene Daten nach der aktuellen Datenschutzverordnung (DSGVO) auf verschlüsselten elektronischen Datenträgern (mindestens 256-Bit-AES-Verschlüsselung) am Institut für experimentelle Psychophysiologie, aufgezeichnet, gespeichert und ausgewertet werden. Gegebenenfalls in Papierform erhobene Daten werden digitalisiert, ebenfalls elektronisch gespeichert und die Originale umgehend und unwiderruflich vernichtet. Ich bin

darüber aufgeklärt worden, dass die Speicherung meiner Daten in pseudonymisierter Form erfolgt – „pseudonymisiert" heißt in diesem Zusammenhang, dass die Daten nicht mit meinem Namen, sondern unter einem Probandencode gespeichert werden, der Ihnen vom Versuchsleiter zugewiesen wird. Mir ist bewusst, dass eine Pseudonymisierung für die Speicherung der aufgenommenen Video- und Tonaufnahmen nicht möglich ist, d.h. es findet eine Speicherung in „personalisierter" Form statt. Alle mit meinem Code versehenen Daten werden mit Erfüllung des Forschungszwecks spätestens aber am 01.05.2035 gelöscht.
Ich bin damit einverstanden, dass die Daten in anonymisierter Form – also ohne meinen Code sowie ohne Alters- und Geschlechtsangabe – an andere Wissenschaftlerinnen und Wissenschaftler und/oder Auftraggeber und Auftraggeberinnen weitergegeben, veröffentlicht und in Datenrepositorien gespeichert werden. Mir ist ebenfalls bewusst, dass die Daten in anonymisierter Form zur Nachnutzung zwecks wissenschaftlicher Veröffentlichungen genutzt werden. Ich weiß, dass ich wegen des fehlenden individuellen Codes eine Löschung dieser anonymisierten Daten nicht mehr veranlassen kann.
Ich weiß, dass die Personen, die mich während meiner Studienteilnahme betreuen, in Belangen des Datenschutzes unterwiesen sind und der Schweigepflicht unterliegen. Sie nehmen keinen Einblick in die erhobenen Rohdaten. Mir ist bekannt, dass die Studie ausschließlich zu Forschungszwecken dient und dass keine klinische Begutachtung oder individuelle Rückmeldung über die erhobenen Daten erfolgt.

Ich bin darüber aufgeklärt worden, dass ich jederzeit ohne Angabe von Gründen die weitere Teilnahme an der Studie ablehnen und die Einwilligungserklärung widerrufen kann, ohne dass mir daraus Nachteile entstehen. Im Fall eines solchen Widerrufs meiner Einwilligung, an der Studie teilzunehmen, werden keinerlei Daten von mir abgespeichert bzw. bereits gespeicherte Daten umgehend unwiederbringlich gelöscht.
Ich weiß, dass ich mich bei Anmerkungen oder Fragen zur Studie sowie zum Zweck des Widerrufs dieser Einwilligungserklärung wenden kann an:
Lisa Kiel
Institut für Experimentelle Psychophysiologie
Gustav-Poensgen-Straße 29
40215 Düsseldorf
+49 211 975 326 53
l.kiel@ixp-duesseldorf.de

Mir ist bewusst, dass die monetäre Entschädigung für diese Studie im Falle eines Widerrufs nur anteilig ausgezahlt werden kann.
Ich habe die Einwilligungserklärung verstanden und erkläre mich bereit, an der Studie teilzunehmen.

_____ _____

Ort, Datum Name in Druckbuchstaben, Unterschrift

Name der aufklärenden Person:

_____

Appendix B

R Code used to rank the features

```
# load the data

data_subset<-read.csv("/Users/lisa/05_Big_Merge/subset200.csv")

data_subset <- data_subset[c(1,15,23:731)]


#Categorise pain ratings

data_subset$Selfrating <- ifelse(between(data_subset$Selfrating,7,10), 1, 0)

data_subset$Selfrating <- as.factor(data_subset$Selfrating)


#missing values

row.has.na <- apply(data_subset, 1, function(x){any(is.na(x))})

sum(row.has.na)

data_subset <- data_subset[!row.has.na,]


#Feature Selection - Rank by Importance - Learning Vector Quantization (LVQ) model

TrainClass <- data_subset[,1]

TrainData <- data_subset[,2:711]

control <- trainControl(method="repeatedcv", number=10, repeats=3)

model <- train(TrainData, TrainClass, method ="lvq", preProcess ="scale", trControl=control)


# estimate variable importance

importance <- varImp(model, scale=FALSE)

# summarize importance

print(importance)

plot(importance, cex.lab=0.1, top=30)
```

Appendix C

Example of the R Code used to calculate the models

```
# load the data
rawdata <-read.csv("/Users/lisa/05_Big_Merge/Big_merge.csv")
#unique(raw_data$id)
subset2 <- rawdata [c("id", "Selfrating","AU01_r", "AU02_r",  "AU01_r", "AU02
_r","AU20_r", "AU01_c", "AU05_r", "AU02_c", "AU15_r")]


head(subset2, n=5)

##      id Selfrating AU01_r AU02_r AU01_r.1 AU02_r.1 AU20_r AU01_c AU05_r AU
02_c
## 1 vp041          0   0.00      0     0.00        0   0.00      0      0
0
## 2 vp041          0   0.04      0     0.04        0   0.00      1      0
1
## 3 vp041          0   0.06      0     0.06        0   0.00      1      0
1
## 4 vp041          0   0.11      0     0.11        0   0.01      1      0
1
## 5 vp041          0   0.09      0     0.09        0   0.01      1      0
1
##   AU15_r
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0

# convert numerical values to categorical 0 and 1
subset2$Selfrating <- ifelse(between(subset2$Selfrating,7,10), 1, 0)
subset2$Selfrating <- as.factor(subset2$Selfrating)

head(subset2, n=5)

##      id Selfrating AU01_r AU02_r AU01_r.1 AU02_r.1 AU20_r AU01_c AU05_r AU
02_c
## 1 vp041          0   0.00      0     0.00        0   0.00      0      0
0
## 2 vp041          0   0.04      0     0.04        0   0.00      1      0
1
## 3 vp041          0   0.06      0     0.06        0   0.00      1      0
1
## 4 vp041          0   0.11      0     0.11        0   0.01      1      0
1
## 5 vp041          0   0.09      0     0.09        0   0.01      1      0
1
```

```
##   AU15_r
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
```

```
#check for N/A
row.has.na <- apply(subset2, 1, function(x){any(is.na(x))})
sum(row.has.na)
```

```
## [1] 3
```

```
subset2 <- subset2[!row.has.na,]
```

```
sum(subset2$Selfrating==1)
```

```
## [1] 175688
```

```
sum(subset2$Selfrating==0)
```

```
## [1] 171425
```

```
#divide df into train and test set by ID
# Randomly assign train/test groups to all values of ID
set.seed(7)
groups <-
  subset2 %>%
  select(id) %>%
  distinct(id) %>%
  rowwise() %>%
  mutate(group = sample(
    c("train", "test"),
    1,
    replace = TRUE,
    prob = c(0.8, 0.2) # Set weights for each group here
  ))

groups
```

```
## # A tibble: 50 x 2
## # Rowwise:
##    id    group
##    <chr> <chr>
##  1 vp041 test
##  2 vp012 train
##  3 vp013 train
##  4 vp015 train
##  5 vp016 train
##  6 vp017 train
##  7 vp018 train
```

```
##  8 vp019 test
##  9 vp014 train
## 10 vp020 train
## # … with 40 more rows

# Join group assignments to my_dat
subset2group <- subset2 %>%
  left_join(groups)

## Joining, by = "id"

#create training + test dataframes
train_subset2 <- filter(subset2group, group == "train")
test_subset2 <-  filter(subset2group, group == "test")

head(train_subset2, n=5)

##      id Selfrating AU01_r AU02_r AU01_r.1 AU02_r.1 AU20_r AU01_c AU05_r AU
02_c
## 1 vp012          0      0      0        0        0   0.39      0      0
0
## 2 vp012          0      0      0        0        0   0.66      0      0
0
## 3 vp012          0      0      0        0        0   0.76      0      0
0
## 4 vp012          0      0      0        0        0   0.76      0      0
0
## 5 vp012          0      0      0        0        0   0.78      0      0
0
##   AU15_r group
## 1      0 train
## 2      0 train
## 3      0 train
## 4      0 train
## 5      0 train

###https://towardsdatascience.com/k-nearest-neighbors-algorithm-with-examples
-in-r-simply-explained-knn-1f2c88da405c
# normalise function
nor <-function(x) { (x -min(x))/(max(x)-min(x)) }
##Run nomalisation on the predictor columns
train_sub2_norm <- as.data.frame(lapply(train_subset2[,c(3:11)], nor))
test_sub2_norm <- as.data.frame(lapply(test_subset2[,c(3:11)], nor))

summary(train_sub2_norm)

##      AU01_r            AU02_r          AU01_r.1          AU02_r.1
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00200   Median :0.0000   Median :0.00200   Median :0.0000
##  Mean   :0.06451   Mean   :0.0369   Mean   :0.06451   Mean   :0.0369
```

```
##  3rd Qu.:0.07000   3rd Qu.:0.0100   3rd Qu.:0.07000   3rd Qu.:0.0100
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      AU20_r            AU01_c           AU05_r            AU02_c
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.03363   Mean   :0.2032   Mean   :0.02111   Mean   :0.2055
##  3rd Qu.:0.02800   3rd Qu.:0.0000   3rd Qu.:0.00200   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      AU15_r
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.04202
##  3rd Qu.:0.03800
##  Max.   :1.00000
```

```
##extract 2nd column of train dataset because it will be used as 'cl' argumen
t in knn function. The "," serves to make a vector

target_cat_sub2 <- train_subset2[,2]
test_cat_sub2 <- test_subset2[,2]

##run knn function
set.seed(7)
pr_sub2 <- knn(train_sub2_norm,test_sub2_norm,cl=target_cat_sub2, k=519, prob
= TRUE)


##create confusion matrix
tab_sub2 <- table(pr_sub2,test_cat_sub2)

print(tab_sub2)
```

```
##        test_cat_sub2
## pr_sub2     0     1
##       0 14690 12528
##       1 18999 21391
```

```
##this function divides the correct predictions by total number of prediction
s that tell us how accurate the model is.
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(tab_sub2)
```

```
## [1] 53.36794
```

```
### KNN not separated by participant
##Generate a random number that is 8 Run for dataset within participants 0% of the tot
al number of rows in dataset.                                           set.seed(7)
ran <- sample(1:nrow(subset2), 0.8 * nrow(subset2))
subset2_norm <- as.data.frame(lapply(subset2[,c(3:11)], nor))
```

```
summary(subset2_norm)
```

```
##      AU01_r            AU02_r           AU01_r.1          AU02_r.1
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00200   Median :0.0000   Median :0.00200   Median :0.0000
##  Mean   :0.05949   Mean   :0.0333   Mean   :0.05949   Mean   :0.0333
##  3rd Qu.:0.06200   3rd Qu.:0.0080   3rd Qu.:0.06200   3rd Qu.:0.0080
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      AU20_r            AU01_c           AU05_r            AU02_c
##  Min.   :0.00000   Min.   :0.000   Min.   :0.00000   Min.   :0.00
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00
##  Median :0.00000   Median :0.000   Median :0.00000   Median :0.00
##  Mean   :0.03322   Mean   :0.191   Mean   :0.01946   Mean   :0.21
##  3rd Qu.:0.03000   3rd Qu.:0.000   3rd Qu.:0.00200   3rd Qu.:0.00
##  Max.   :1.00000   Max.   :1.000   Max.   :1.00000   Max.   :1.00
##      AU15_r
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.002
##  Mean   :0.043
##  3rd Qu.:0.040
##  Max.   :1.000
```

```
#Run nomalization on first 4 coulumns of dataset because they are the predict
ors
subset2_within_train <- subset2_norm[ran,]
##extract testing set
subset2_within_test <- subset2_norm[-ran,]
##extract 5th column of train dataset because it will be used as 'cl' argumen
t in knn function.
subset2_within_target <- subset2[ran,2]
subset2_within_testcat <- subset2[-ran,2]

subset_train <- subset2[ran,]
subset_test <- subset2[-ran,]
unique(subset_train$id)
```

```
##  [1] "vp029" "vp065" "vp024" "vp038" "vp056" "vp035" "vp034" "vp039" "vp05
0"
## [10] "vp053" "vp047" "vp060" "vp018" "vp013" "vp030" "vp032" "vp044" "vp05
7"
## [19] "vp062" "vp021" "vp016" "vp048" "vp012" "vp019" "vp042" "vp017" "vp04
5"
## [28] "vp036" "vp061" "vp033" "vp058" "vp054" "vp014" "vp041" "vp070" "vp01
5"
## [37] "vp055" "vp059" "vp037" "vp043" "vp051" "vp069" "vp049" "vp023" "vp02
2"
## [46] "vp046" "vp064" "vp052" "vp020" "vp040"
```

```r
unique(subset_test$id)
```

```
##  [1] "vp041" "vp012" "vp013" "vp015" "vp016" "vp017" "vp018" "vp019" "vp01
4"
## [10] "vp020" "vp021" "vp022" "vp023" "vp024" "vp029" "vp030" "vp032" "vp03
3"
## [19] "vp034" "vp035" "vp036" "vp037" "vp038" "vp039" "vp040" "vp042" "vp04
3"
## [28] "vp044" "vp045" "vp046" "vp047" "vp048" "vp049" "vp050" "vp051" "vp05
2"
## [37] "vp053" "vp054" "vp055" "vp056" "vp057" "vp058" "vp059" "vp060" "vp06
1"
## [46] "vp062" "vp064" "vp065" "vp069" "vp070"
```

```r
set.seed(7)
pr_sub2_within <- knn(subset2_within_train,subset2_within_test,cl=subset2_wit
hin_target,k=519, prob=TRUE)

## confusion matrix
tab_sub2_within <- table(pr_sub2_within,subset2_within_testcat)
print (tab_sub2_within)
```

```
##                subset2_within_testcat
## pr_sub2_within     0     1
##              0 17136  9885
##              1 17020 25382
```

```r
accuracy(tab_sub2_within)
```

```
## [1] 61.24483
```