Automatic Recognition Methods Supporting Pain Assessment: A Survey

As Published: 10.1109/TAFFC.2019.2946774

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: https://hdl.handle.net/1721.1/136497

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Automatic Recognition Methods Supporting Pain Assessment: A Survey

Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind W. Picard

Abstract—Pain is a complex phenomenon, involving sensory and emotional experience, that is often poorly understood, especially in infants, anesthetized patients, and others who cannot speak. Technology supporting pain assessment has the potential to help reduce suffering; however, advances are needed before it can be adopted clinically. This survey paper assesses the state of the art and provides guidance for researchers to help make such advances. First, we overview pain's biological mechanisms, physiological and behavioral responses, emotional components, as well as assessment methods commonly used in the clinic. Next, we discuss the challenges hampering the development and validation of pain recognition technology, and we survey existing datasets together with evaluation methods. We then present an overview of all automated pain recognition publications indexed in the Web of Science as well as from the proceedings of the major conferences on biomedical informatics and artificial intelligence, to provide understanding of the current advances that have been made. We highlight progress in both non-contact and contact-based approaches, tools using face, voice, physiology, and multi-modal information, the importance of context, and discuss challenges that exist, including identification of ground truth. Finally, we identify underexplored areas such as chronic pain and connections to treatments, and describe promising opportunities for continued advances.

Index Terms—Pain assessment, recognition, survey, review.

1 Introduction

Pain is a complex phenomenon and not fully understood yet. According to the most widely accepted definition pain is "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" [106]. However, basic research still advances the understanding of pain and there is an ongoing discussion about updating the definition [107], [108]. Pain is a personal experience, a mental episode [108], and always subjective. The meaning of the word is learned by each individual in early life through experiences related to injury [106]. Such pain, which is called acute pain, helps to identify harmful situations, to avoid tissue damage, and promotes healing by inhibiting activities that might cause further tissue damage [109]. Acute pain usually disappears with healing. For most conditions, if pain lasts longer than 3 months, it is called chronic or persistent pain. Pain may be also categorized as nociceptive (due to stimulation of sensory nerve fibers), neuropathic (due to impaired somatosensory nervous system), or psychogenic pain (caused, increased, or prolonged by mental, emotional, or behavioral factors).

Pain is a grave issue for many individuals and society as a whole. It is the primary reason that prompts people to seek medical attention [110]. *E.g.* according to Cordell *et al.*, pain was a chief complaint for 52.2% of all patient visits at an emergency department, whereas only 34.1% were not related to pain [111]. According to Gregory *et al.*, acute pain is a significant symptom in hospitalized patients, with up to 35% of patients reporting severe pain and approximately 50% of patients reporting pain [112]. In a study by Zoëga *et al.*, pain prevalence in hospitals was even 83% [113]. Medical progress has contributed to the increasing need for pain management: Many people now survive formerly lethal diseases such as cancer, HIV, and cardiovascular disease, but afterwards suffer from persistent pain caused by either the ongoing illness or by nerve damage due to the disease even after being cured [114]. Treatments such as surgery, chemotherapy, or radiotherapy also often cause pain [114]. Persistent pain has major implications for the individual suffering from pain and for her family and friends. The workplace suffers through loss of productive employees and the economic costs of caring for chronic pain patients are dramatic [114]. Chronic pain costs society more than cancer, heart disease and HIV combined [114]. *E.g.* in 2003, the financial impact on the U.S. economy was estimated to be $100 billion each year [115].

Valid and reliable pain assessment is necessary for differential diagnosis, choosing the adequate treatment, monitoring progress, and evaluating the need to continue or modify a treatment. Thus, pain assessment and management are not only important to provide comfort but also to prevent both immediate and long-lasting consequences that are harmful to the person's overall health [116], as uncontrolled pain not only causes suffering and reduces quality of life, but also compromises the nervous system [116], endocrine system [117] and immune function [114]. Untreated pain can lead to chronic pain syndrome which is often accompanied by decreased mobility, impaired immunity, decreased concentration, anorexia, and sleep disturbances. Moreover, wrong treatment may lead to problems and risks for the patients.

- *P. Werner and A. Al-Hamadi are with the Neuro-Information Technology Group, Otto von Guericke University, Magdeburg, Germany. E-mail: {Philipp.Werner, Ayoub.Al-hamadi}@ovgu.de*
- *D. Lopez-Martinez and R. W. Picard are with the Affective Computing Group, MIT Media Lab, Cambridge, MA 02139; USA.*
- *S. Walter and S. Gruss are with the Medical Psychology Group, University Clinic, Ulm, Germany.*

Manuscript received Dec 13, 2018; revised Jul 12, 2019; accepted Sep 24, 2019.

E.g. over-usage of opioids can depress breathing or lead to addiction [118]. Further, medication may have adverse effects like nausea, vomiting, or constipation [118].

Despite knowledge and technology, pain is still often poorly managed [114], [115], [119]–[121]. Although this is a general problem, it most severely affects patients with limited communication abilities, who cannot report their pain experience or whose report has low validity. Such vulnerable groups, who are often under- or overtreated, include: infants, toddlers, and children; adults with cognitive impairment (such as advanced dementia); persons with intellectual disability; critically ill or unconscious persons; and persons who are terminally ill [122]. In the future, automatic pain recognition systems based on pain behaviors (such as facial expressions, vocalizations, and body movements) and physiological responses, may complement current assessment methods for achieving better pain management. In contrast to traditional assessment tools, they could monitor pain continuously. This may improve clinical outcomes, *e.g.* by facilitating early intervention for patients that cannot call for help by themselves. Further, automatic systems may be more objective than a human observer, whose assessment is influenced by personal factors, such as the relationship to the sufferer [119] or the patient's attractiveness [123]. Recognition systems may also help to gain new knowledge about pain, *e.g.* about the dynamics of facial expressions [75], as they can be more sensitive to slight changes than common manual annotation by humans [104].

In the last decade, automatic pain recognition changed from an idea to a research topic of considerable interest. In order to review the published approaches, we conducted a systematic literature search. We searched the Web of Science as well as the proceedings of the major conferences on biomedical informatics and artificial intelligence for peer-reviewed papers on April 18th, 2018 using keywords related to pain, its measurement, and automatic / machine learning methods (details in supplemental material). We excluded animal studies, studies on only self-report of pain, and other studies that do not address automatic methods for pain assessment. This way, we found 61 papers about automatic methods for pain assessment, among which there was no review article. We also went through the reference lists of the identified papers, which yielded 65 additional papers for review. We identified some papers that are a subset of a later version; in those cases we only cite the more comprehensive version. Thus, dates of included papers do not indicate relative timing of the original work. Although not indexed in the Web of Science, there is a 2017 survey on automated pain assessment in infants [124]. To reduce overlap, we exclude most literature about pain in infants from our survey.

We start our discussion with an overview of the pain mechanisms and responses (Sec. 2), the clinically used pain assessment tools (Sec. 3), and the datasets and validation methods needed for advancing pain recognition (Sec. 4). In Sec. 5 we describe the automatic pain recognition approaches. In order to impact clinical practice, some challenges have to be overcome; Sec. 6 discusses these challenges and shows some promising directions. The paper is concluded in Sec. 7.

2 Pain Mechanisms and Responses

This section roughly overviews pain mechanisms and responses. For more details refer to the supplemental material.

Pain is a personal, private, subjective experience arising in the brain. Pain is not only a sensory phenomenon, but comprises sensory-descriminative, affective-motivational, and cognitive-evaluative components [125]: pain is characterized by severity, location, duration, and quality; it is unpleasant and motivates activity for pain relief; and it is influenced by cognitions such as evaluation of an injury's seriousness, distraction, or cultural values [115], [125]. The pain experience must be carefully distinguished from the pain cause (such as tissue damage with nociception), the pain response (verbal communication and non-verbal manifestations), and pain assessment (*e.g.* by a caregiver). The pain cause is often diagnosable (*e.g.* a fracture) and it may be controlled in deliberate pain stimulation (*e.g.* neurological assessments), but it may also be unknown or absent (especially in chronic pain). Typically, pain originates from noxious stimuli, *e.g.* due to tissue injury, that lead to a response of the sensory nervous system called nociception. Pain experience is modulated by personal and inter-personal factors, *e.g.* cognition, past experience, and situation [115], [125], [126]. As a result, the same stimulus may lead to different pain experiences. In rare cases, people do not experience any pain, and this brings them harm [127]; however, pain usually causes observable pain responses, which are modulated by personal and contextual factors. Pain responses may be categorized in physiological responses, behavioral responses, and self-report (which may be considered a controlled behavior). Currently, pain assessment is typically done by a caregiver by (1) obtaining self-report if available, (2) observing behavioral or physiological pain responses, and (3) using information about the pain cause. The assessment is influenced by personal, social, contextual factors and may be biased. Based on assessment, the caregiver initiates or adjusts pain management, which may comprise pharmacological intervention, physical therapy, and/or psychological therapy for easing pain (in trade off with potential side-effects).

2.1 Biological Mechanisms

Multiple areas of the nervous system, from peripheral nervous system to cerebral cortex, participate in the pain process. The process often starts with the activation of sensory neuronal pathways by noxious mechanical, heat, cold, chemical, or inflammatory stimuli. These stimuli activate nociceptors, primary sensory neurons with surface receptors specialized to detect noxious stimuli. The generated action potentials are conducted through nociceptive fibers and transduced to synapses in the spinal cord. Information about a noxious event in the periphery can activate both excitatory and inhibitory interneuronal circuits in the spinal cord leading to a protective reflexive withdrawal event. Further processing of nociceptive input occurs in numerous supraspinal structures, leading to the sensory discriminative perception of pain. While frequently a nociceptive stimulus will lead to pain, several conditions can change this perception. A person may also experience pain without activation of the nociceptive pathway, *e.g.* in psychogenic pain. For details about the biological mechanisms refer to other reviews [128]–[130].

2.2 Physiological Responses

Extensive interactions between the neural structures involved in pain sensation and autonomic control [131] lead to increased sympathetic outflow, resulting in measurable changes in different physiological signals [132].

Skin conductance [133] is an autonomically-modulated signal that changes in response to pain. Because sweat glands are exclusively innervated by sympathetic excitatory efferent neurons [131], the increased sympathetic outflow associated with pain causes sweat to be discharged into pores on the skin surface [134]. Sweat secretion alters the electrical properties of the skin (electrodermal activity, EDA), increasing the electrical conductance until the sweat is reabsorbed or evaporated.

Increased sympathetic activity also leads to major cardiovascular changes. It affects the **heart rate** [135], leading to tachycardia, and **heart rate variability** [136], an index of autonomic regulation of heart rate. Specifically, pain significantly increases low frequency power, as measured by power spectral analysis. Furthermore, pain also increases peripheral vascular resistance and stroke volume. Combined with the increased heart rate, this leads to an elevation in resting **blood pressure** [137]. Pain also affects **pupil diameter** due to the pupil dilation reflex, which is under dual sympathetic/parasympathetic control (dilating/constricting the pupil, respectively) [138], [139].

Finally, because pain processing involves a complex network of brain cortical regions [140], pain impacts the electrical and metabolic activity of these areas. **Electroencephalography (EEG)** can be used to detect changes in electrical activity in the brain cortex [141], whereas **functional magnetic resonance imaging (fMRI)** and **functional near-infrared spectroscopy (fNIRS)** can be used to detect brain hemodynamic changes in response to increased metabolic demand [142]. Both techniques have shown promise to detect patterns of response to pain [87], [101].

2.3 Behavioral Responses

Behavioral pain response fulfills two functions. (1) For protecting the own body, "pain grabs attention, interrupts associated behavior, and urges action towards mitigating it" [109], such as reflexive withdrawal of the hand from a hot surface. (2) We communicate pain: we show need for help to allies (potential help-givers) and hide weakness/vulnerability from antagonists – a behavior that probably developed since it increased chances of survival and reproduction [109]. Behavioral pain responses include facial expressions, body movements, and vocalizations. Chronic pain often also leads to permanent changes in everyday behavior and social interaction.

There are specific **facial expressions** associated with pain that occur relatively consistently across a range of clinical pain conditions and experimental pain modalities [143]–[146]. Moreover, the magnitude of facial movements increases with rising intensities of noxious stimulation [143], [146]. Facial expression research is usually conducted using the Facial Action Coding System (FACS) [147], which describes expressions by elementary Action Units (AUs) based on facial muscle activity. Fig. 1 shows an exemplary facial response during pain. There is good evidence that facial pain expression is not only sensitive but also specific to pain and can be distinguished from expressions of basic emotions [109], [143], [148]. However, pain also often co-occurs with emotions [109], often resulting in altered or blended expressions. Also, individuals differ in facial expressiveness [16], [58] and pain reaction threshold [126].

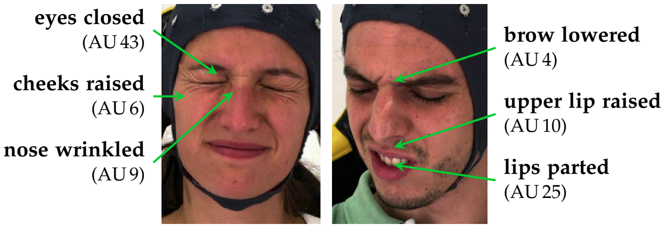

Fig. 1. Examples of facial expressions associated with pain.

Most **body movements** that have been identified to be pain-related serve to protect from further damage and to minimize pain. These include protective reflexes, rubbing, writhing, and guarding. The concrete body movement behaviors vary with the type of medical population. *E.g.* in chronic low back pain, movements may be characterized by guarding or stiffness, hesitation, bracing or support, abrupt action, or limping; or they may serve to rub or stimulate an affected body part [63]. With actors and naive observers, Walsh *et al.* [149] found that pain is communicated through averted head and trunk, hand touches to various sites, knee bending, and shoulder to front movements. Werner *et al.* [150] analyzed three pain datasets regarding head movements and postures during pain and found that they tend to be oriented downwards or towards the pain site and differ in the movement speed and range compared to other conditions.

Pain behavior also includes **vocalizations**, such as paralinguistic vocalizations (crying, moaning, groaning, gasping, and sighing), and voice quality aspects such as amplitude, timbre, and hesitancy observed during verbal self-report [143], [151]. In infants, cries are part of a typical pain response (but there is considerable variation) [152], [153]. Pain cries have been reported to have higher pitch and spectral energy, and to be harsher and less melodious than other infant cries [154]. Vocalizations have been researched mainly as part of the design and validation of observational pain scales (see Sec. 3.2). *E.g.* they are part of well-validated scales, such as CPOT [155], PACSLAC [156], and FLACC [122].

2.4 Pain and Emotion

Pain is defined not only as a sensory but also as an emotional experience. It includes an affective dimension incorporating many different emotions, which are primarily negative and related to the unpleasantness of the experience or to future implications. Anxiety, anger, and depression play an important role, especially but not exclusively in chronic pain [157]. Pain and emotions interact, *e.g.* emotional distress may be a modulating factor amplifying or inhibiting the severity of pain, or it may be a consequence of pain or a perpetuating factor [157]. Under high arousal levels, positive emotional states generally reduce pain and unpleasant emotional states exacerbate pain [158]. According to Gatchel *et al.*, neither acute nor chronic pain can be treated successfully without attending to the patient's emotional state [157]. In conclusion, emotion should be assessed along with pain. Whether pain can be disentangled from its accompanying emotions reliably and objectively, is an unanswered question for most physiological and behavioral response signals and their combinations. The challenging follow-up research question is whether blends of pain and emotion can be used successfully to simultaneously assess pain and the accompanying emotional state.

3 Clinically Used Pain Assessment Tools

In clinical practice, pain is usually diagnosed through the patient's self-report according to severity, sensory quality, location, temporal features, and factors that alleviate and intensify the pain. Self-report refers to conscious communication of pain-related information by the person in pain, typically with spoken or written language or using gestures, such as pointing to an image that best represents their feeling or nodding in response to a question. Clinical practice guidelines emphasize that self-report is the most valid way of assessing pain if the person is able to communicate [122], [159]. For the nonverbal patient, Herr *et al*. [122] recommend to (1) attempt obtaining self-report, (2) search for potential causes of pain, (3) observe patient behavior, (4) obtain surrogate reporting from a person who knows the patient well, and (5) attempt an analgesic trial. In the following, we introduce and discuss some clinically used assessment tools. We focus on pain intensity, since this is the current focus of automatic pain recognition research.

3.1 Self-Report

Self-report is often referred to as the gold standard in pain assessment [119], as pain is a subjective experience and therefore a person's report offers the best access to it. Self-reports provide retrospective accounts of events, experiences, and behaviors and are methodically convenient and economical; they are viewed as providing patient-centered care [119].

There exist different categories of intensity scales. Verbal Rating Scales (VRS) use discriminative verbal categories, such as: no pain, weak pain, severe pain, unbearable pain. Numerical rating scales (NRS) provide precise instruction [160], *e.g.* "Please describe your current pain on a scale of zero to ten, when zero means no pain and ten means the strongest conceivable pain". Visual Analogue Scales (VAS) consist of a line with the endpoints "no pain" and "worst pain imaginable". The patient reports his pain by marking a position on this line. Depending on the situation, different scales may be preferred [160]. *E.g.*, both VRS and NRS can be quickly and easily administered without paper and pencil, whereas VAS offers better pain intensity differentiation. NRS is used commonly in pain diaries. In addition to these, other scales exist for specific patient populations. *E.g.*, the smiley-based Wong Baker Scale (WBS) and Faces Pain Scales-Revised (FPS-R) are often used in pediatric and elderly patients respectively.

While all these scales may be used to assess pain intensity, they require a level of cognitive processing and functioning that is not always present. Further, self-report is a controlled and goal-oriented response [143], which can be affected by reporting bias and variances in memory and verbal ability [123].

3.2 Observational Scales

Observational scales are commonly used when self-report fails. Numerous scales have been designed and validated for specific medical populations, such as infants and preverbal toddlers, (*e.g*. NIPS, CRIES, FLACC) [124]; elderly people with severe dementia, (*e.g*. PACSLAC, DOLOPLUS2, PAINAD) [156]; and critically ill and/or unconscious persons (*e.g*. BPS, CPOT, NVPS) [122]. Most scales consider facial expressions, vocalizations, and body language, while some include vital parameters. Details are available in Herr *et al*. [122] and the appendices of practice recommendations [161].

It is hard to assess and compare the validity of the various scales, because studies differ a lot in design, methodology, subjects, and conceptualization of the pain phenomenon. A considerable amount of prior training and experience is required to apply a scale reliably. Even if trained medical staff could record pain intensity by observation several times a day, such frequent measurement is likely to decline with economic pressures unless it is shown to provide cost savings. Therefore, relevant pain episodes may be missed or changes may be detected late by a human observer, whereas an automatic system could monitor pain continuously.

4 Datasets and Validation

Representative data are essential for developing a pain recognition system and proving its usefulness. In the following we address issues of recording and using data for pain recognition: (1) pain stimulation, (2) ways to label data and measures for evaluation, and (3) publicly available datasets. We strongly recommend to share datasets among researchers to accelerate progress in pain recognition research and facilitate comparison of competing approaches.

4.1 Pain Stimulation

In order to develop and validate pain recognition systems it is necessary to record data of people experiencing pain. This can be done with patients in clinical contexts or with healthy volunteers as common in basic and pharmaceutical research. Pain experience in patients is influenced by several factors, such as anxiety, disability, distraction, uncertainty, expectations, depression, and medications. Some of the bias can be avoided in experimental studies with healthy unmedicated volunteers. Further, the intensity, duration, frequency, and localization of experimental stimuli are controllable. Methods of **experimental pain stimulation** include thermal, electrical, mechanical, and chemical modalities [162]. Heat and electrical stimuli applied on the skin are the most common types in pain recognition research. Some of the advantages are very precise control of stimulus intensity and timing as well as repeatability. Other used stimulation methods include (1) the cold pressor task, in which the hand or forearm is immersed into cold water [163], (2) mechanical stimulation of the skin with an electronic hand-held pressure algometer [82], and (3) ischemic stimulation of muscle pain with a tourniquet applied on the arm [90]. For details on experimental pain stimulation refer to Olesen *et al*. [162].

In **clinical contexts** many patients suffer from ongoing pain due to disease or injury without external stimulation. But generally, pain experience and response is intensified by external stimuli or many necessary procedures and activities. *E.g.*, there are two forms of postoperative pain [95]: (1) pain at rest, *i.e.* the endogenous pain associated with disease and injury, including the preceding surgery, and (2) movement-evoked pain or exogenous pain brought on or aggravated by pain-evoking maneuvers (such as movement, clinical examination, or physiotherapy). In addition to measuring pain at rest, Sikka *et al.* [95] analyzed a transient exogenous pain condition stimulated by manual pressure on the surgical site (a typical clinical examination). In infants, heel lancing (drawing blood samples) and immunization are common clinical procedures that serve as pain stimuli for research [124]. Painful clinical procedures in critical care include turning, central venous catheter insertion, and wound drain removal; those and other procedures were used to identify pain behaviors by Puntillo *et al.* [164]. Movements that are likely to exacerbate pain in chronic low back pain patients are commonly used in research for classifying those patients versus healthy controls and assessing pain intensity.

4.2 Ground Truth and Performance Measures

Ground truth can be based on self-report, observer assessment, or study design. Due to the personal and subjective character of the pain experience, **self-report scales** (see Sec. 3.1) are considered the gold standard of measuring pain intensity. However, they cannot be applied with some vulnerable patient groups and are known to suffer from biases. Gathering self-report requires action from the person in pain, which may be perceived as an additional burden and may influence experiments. In experimental research, self-report may be used to calibrate stimulation in advance reducing the frequency of required action.

Another option is **pain rating by an observer**. Rating may be done with subjective Likert-type scales [58], [165], but validated systematic observation scales (see Sec. 3.2) should be preferred. Specialized scales are recommended for specific medical populations, such as people with advanced dementia or infants [122]. A problem with observer rating is that not all people show pain responses to the same extent; *e.g.*, several studies reported that about 20% of subjects did not display any facial response to pain [38], [58], [126], [166]. Although expressive variation is a general problem, this high number might be an experimental artifact due to ethical restrictions and low pain intensities [58]. Fortunately, severe pain intensities, which are practically more important, yield more responses and can be recognized more reliably.

A specific observer scale for facial expression, the **Prkachin and Solomon Pain Intensity (PSPI)** [144], [165] is widely used in pain recognition due to its association with the commonly used UNBC-McMaster database. It can be calculated for each individual image or video frame, after coding the intensity of certain action units (AU) according to the Facial Action Coding System (FACS). Several authors argued that PSPI offers a high temporal resolution, as it provides an independent score for each frame. But PSPI of a single frame should not be confounded with the *feeling* of pain at this particular moment, as it only measures the facial expression of pain. PSPI can go up and down with tension and relaxation of facial muscles although the felt pain is steadily increasing [38]. So the temporal resolution of PSPI might be misleading, especially if the pain persists for a longer time. Although a person actually experiences pain, the PSPI may be zero. There may be no facial reaction at all due to low pain intensity or expressiveness [126], [166], [167]. Further, the feeling of pain may induce a facial response that is not part of the prototypic pattern underlying PSPI (AU 4/6/7/9/10/43). Recently, Kunz and Lautenbacher [166] suggested that there are several "faces of pain". They found activity patterns that include raising of eyebrows (AU 1/2) or opening of the mouth (AU 25/26/27), which are not considered by PSPI. The PSPI may be also non-zero although the observed subject does not feel pain. Most obvious, AU43 (closed eyes) is not specific to pain, *e.g.* it also occurs during sleep and relaxation. Further, several facial expressions of emotions share AUs with PSPI [163], *e.g.* disgust (AU 9/10), fear and sadness (AU 4), and happiness (AU 6). Thus, if PSPI is used in a wider context, many frames are labeled as painful by mistake, which could be easily avoided by using alternative ground truth. As illustrated above and in more detail by Werner *et al.* [38], there are several shortcomings with PSPI. We recommend to either avoid its use, use it with more caution, or complement it with other ground truth. Also our work showing personalized differences suggests we should give the same cautionary warning for any measures derived from a small group of patients.

In experimental pain studies, ground truth with high temporal resolution can be obtained from the applied stimulus's time series. This belongs to the third category of ground truth, which originates from **study design and prior knowledge**. Many studies use well-established knowledge that a person feels more or less pain under certain circumstances compared to a reference. For instance, in heat pain stimulation a higher temperature usually induces more pain than a lower temperature; low back pain patients feel more pain during back-straining exercises than healthy controls; postoperative pain reduces with time after surgery due to healing; established analgesic drugs reduce pain; needle injections cause pain; pressure on wounds causes pain *etc*. Such knowledge about procedures and effects naturally defines categories that can be used as recognition targets. At the same time, it is important to hold context steady – especially because subtle ways it can change stress or emotion may affect pain.

Since every type of ground truth has its strengths and weaknesses, the best option is to create datasets with multiple types of ground truth and to evaluate and compare recognition systems with all the available ground truth types. For **measuring performance** we distinguish if outputs are categories or numbers. For categorical output, confusion matrices and Receiver Operating Characteristic (ROC) curves are very informative. If the dataset has a balanced class distribution, *i.e.* about the same number of samples in all classes, results can be condensed in one number using the accuracy measure, *i.e.* the total percentage of correctly classified samples. It is an intuitive measure, which however may be misleading if the class distribution is imbalanced. Imbalanced distributions are common in frame-based facial pain recognition, *e.g.* more than 80% of the UNBC-McMaster database has a PSPI score of zero ("no pain") [165]. Other measures such as the F1 score

Database	Subjects	Stimuli	Data Modalities (D) / Annotation (A)
UNBC-McMaster Shoulder Pain [165]	25 adult shoulder pain patients	200 range of motion tests with affected and unaffected limbs	**D:** video of face (low resolution, includes social interaction / talking) **A:** self-report (VAS, sensory & affective verbal scales), observer-assessed pain intensity (OPI), affected/unaffected limb, FACS coding
BioVid Heat Pain [56], [168], [169]	90 healthy adults (age 20-65)	14k heat pain (4 intensities × 20 repetitions × 2 parts × 90 participants); emotion elicitation, posed expression	**D:** video of face, EDA, ECG, sEMG (trapezius muscle; corrugator and zygomaticus for part B) **A:** stimulus (calibrated per person)
BP4D-Spontaneous [163]	41 healthy adults (age 18-29)	41 cold pressor task; emotion elicitation	**D:** video of face (color & 3D) **A:** stimulus, FACS coding
BP4D+ [170]	140 healthy adults (age 18-66)	140 cold pressor task; emotion elicitation	**D:** video of face (color, 3D, thermal), heart rate, respiration rate, blood pressure, EDA **A:** stimulus, FACS coding
MIntPAIN [80]	20 healthy adults (age 22-42)	2k electrical pain (40 stimuli in 4 intensities × 2 trials × 20 participants)	**D:** video of face (color, depth, thermal) **A:** stimulus (calibrated per person), self-report (VAS)
COPE [171]	26 neonates (age 18-36 hours)	60 heel lancing for blood collection; non-painful stimuli	**D:** 204 photographs of face **A:** category (pain, rest, cry, air puff, or friction)
YouTube [172]	142 infants (age 0-12 months)	immunizations (injection)	**D:** 142 videos with audio **A:** FLACC observer pain assessment
IIIT-S ICSD [173]	33 infants (age 3-24 months)	immunizations (injection) and other pain causes; non-painful cry causes	**D:** 693 audio cry samples **A:** category annotated by doctors and parents (pain, discomfort, hunger/thirst, and three others)
EmoPain^A [63]	22 chronic lower back pain patients (age $\mu = 50$) + 28 healthy controls (age $\mu = 37$)	physical exercises (therapy scenarios)	**D:** video, audio, motion capture, sEMG (trapezius, lumbar paraspinal muscles) **A:** self report, pain intensity assessed by naive observers from face, presence of pain behaviors assessed by experts from body movement
SenseEmotion^A [174]	45 healthy adults (age $\mu = 26$)	8k heat pain (3 intensities × 30 repetitions × 2 stimulus sites × 45 participants); emotion elicitation	**D:** video of face, audio, EDA, ECG, sEMG (trapezius muscle), RSP **A:** pain and emotion stimulus (pain calibrated per person)
X-ITE pain^A [175]	134 healthy adults (age 18-50)	24k phasic pain, 804 tonic pain (both by heat and electical stimulation, each with 3 intensities)	**D:** video of face (color, thermal), video of body (color, depth), audio, EDA, ECG, sEMG (trapezius, corrugator, zygomaticus) **A:** pain stimulus (calibrated per person)

^{A} Announced to be published, but not yet available. Check website in table caption for updates.
ECG: electrocardiogram EDA: electrodermal activity sEMG: surface electromyography FACS: Facial Action Coding System RSP: Respiration

TABLE 1
Pain recognition databases that are publicly available for research. For URLs and updates refer to https://github.com/philippwerner/pain-database-list.

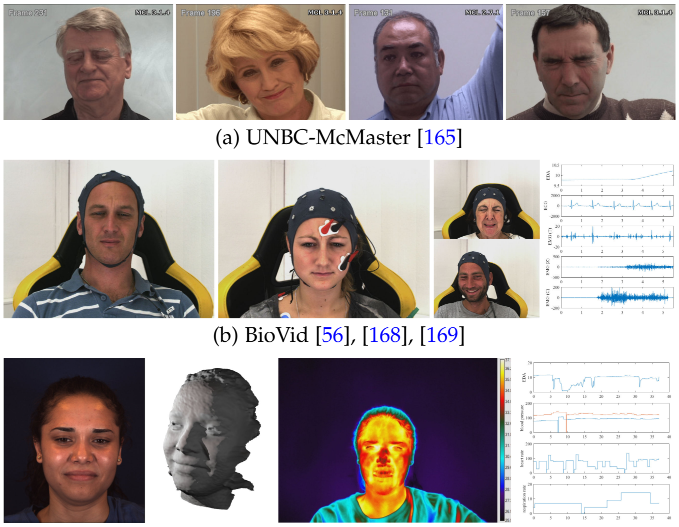
(a) UNBC-McMaster [165]
(b) BioVid [56], [168], [169]
(c) BP4D+ [170]

Fig. 2. Examples from some publicly available pain recognition databases.

(harmonic mean of precision and recall) tend to be more appropriate than accuracy with imbalanced distributions. Jeni *et al*. [176] recommended compensating skew bias by repeatedly applying random under-sampling on the test set and averaging the performance values obtained on the balanced test subsets. In the case of numeric output, *e.g*. intensities, it is common to report correlation values and errors. Werner *et al*. [177] compared characteristics of several performance metrics in an imbalanced scenario and recommended to focus on Intraclass Correlation Coefficients (ICC) for numeric output. Also informative is showing the distributions of predictions *vs*. ground truth, *e.g*. in a scatter plot, Bland-Altman plot, or confusion matrix (of discrete or discretized output). Generally, good practice is to report performances of the random, trivial, and perfect prediction models as reference points, since these numbers help to interpret the results independently of the used measure. Random refers to equally distributed predictions, trivial to predicting always the majority class or mean value, and perfect to always correctly predicting the ground truth label.

4.3 Available Datasets

Table 1 summarizes properties of databases for pain recognition research that are publicly available or have been announced for being available soon. The most commonly used database is the UNBC-McMaster Shoulder Pain Expression Archive Database [165]. It consists of 200 videos showing facial expressions of 25 participants suffering from shoulder pain, who underwent a series of range-of-motion tests with their affected and unaffected limbs. The dataset features rich annotation with self-report and observer measures of the pain intensity at video level and FACS coding at frame level. The BioVid heat pain database [56], [168], [169] was the first database including both, video and biopotentials. It comprises data of about 90 healthy adults, who were stimulated with heat of 4 intensities more than 14,000 times in total. BioVid consists of parts A, B, and C (pain stimulation) as well as part D (posed expression) and E (emotion elicitation). BP4D-Spontaneous [163] and BP4D+ [170] were recorded with 41 and 140 healthy volunteers respectively, but pain was only stimulated once per person by the cold pressor task. Next to color video of the face, BP4D+ also includes 3D and thermal video as well as some physiological signals. Fig. 2 illustrates a few examples from the UNBC-McMaster, BioVid, and BP4D+ databases. The MIntPAIN database [80] comprises color, depth, and thermal video of the faces of 20 healthy adults, who in total underwent about 1,600 electrical pain stimuli (in 4 intensities). In the infant pain domain, the COPE [66], the YouTube [172], and the IIIT-S ICSD database [173] are publicly available. In total, COPE consists of 204 static photographs taken of 26 neonates during several procedures. In the YouTube dataset there are 142 videos with sound of various infants' immunizations. The IIIT-S ICSD database comprises of 693 audio cry samples of 33 infants recorded in a doctor's room during immunization, routine check-up, and therapy sessions.

Not yet available, but announced to be published are the EmoPain [63], the SenseEmotion [174], and the X-ITE pain databases [175]. EmoPain addresses lower back pain and comprises video, audio, motion capture, and sEMG as well as several types of annotation. SenseEmotion and X-ITE were both recorded with healthy adults that underwent experimental pain stimulation. Both include audio and physiological modalities, but X-ITE further includes thermal video of the face, facial EMG, and video of body movement. Additionally, in X-ITE, pain occurs in four different qualities: phasic (short) and tonic (long) variants of each, heat and electrical stimuli.

5 Automatic Pain Recognition Approaches

In order to review the published automatic pain recognition approaches, we conducted a systematic literature search as detailed in the introduction. In the following subsections we address (1) the input of the recognition systems, (2) the system processing methods, (3) the output of the systems, and (4) the validation of the systems' usefulness. Tables 2-7 give an overview on the reviewed works that met all inclusion criteria. We organized the tables primarily by the used dataset. Works using the same database are grouped to simplify comparison, improve clarity, and reduce table sizes.

5.1 Input: Modalities and Sensors

Automatic pain recognition requires at least one channel of sensory input to provide the computer with information. Such a channel is often called a **modality**. The most important modalities for pain recognition may be categorized in behavior and physiology. Behavioral modalities are: facial expression; body movements (such as guarding, rubbing, restlessness, and head movements); paralinguistic vocalizations (such as crying or moaning), and spoken words (which may be transcribed by speech recognition and may contain self-report information). Modalities of interest in the physiology domain include the brain activity, cardiovascular activity, and electro-dermal activity. Additionally, classic direct human-computer interfaces such as keyboard or touch-display can be used to collect self- or observer-report of pain, possibly related activities, or context information. These may complement the other gathered information. A recognition system that processes information from one modality is called a unimodal system; if multiple modalities are used, it is called a multimodal system.

Another criterion for classifying systems is the set of used **sensor** hardware. In pain measurement literature we may distinguish contact-based sensors (such as adhesive electrodes, wrist-bands, caps) and contact-free sensors (such as cameras and microphones). They differ regarding data and noise characteristics, tolerance of movement, ease of use, comfort for the observed person, privacy concerns, costs, and other factors. Widely used contact sensors record electrocardiogram (ECG, heart activity), electrodermal activity (EDA, sweat-gland activation, often measured using skin conductance level (SCL), and sometimes the old term "galvanic skin response" (GSR)), surface electromyogram (sEMG, muscle activity), photoplethysmogram (PPG, blood perfusion of the skin for pulse and other measures, also called blood volume pulse or BVP), respiration (RSP), electroencephalogram (EEG, electrical activity of the brain), or acceleration (ACM, movement). Although sensors and modalities are usually closely related, there are several modalities that can be measured via multiple sensors. *E.g.* heart rate can be measured using ECG, PPG, ACM, cameras, and several other sensors [178]–[180]. Facial expression is measurable through cameras and sEMG [181].

5.1.1 *Camera-Based Approaches*

The vast majority of pain recognition approaches to date (70 % of all reviewed works) have analyzed camera images containing facial expression (see Tables 2-7). As discussed in Sec. 2.3, facial expression plays an important role in communicating pain to others; thus, most early works in automatic pain recognition focused on this modality, *e.g.* [66], [86], [90]. This tendency was reinforced by the release of the first public database for pain recognition, the UNBC-McMaster database, which provides facial images with rich annotation, but no other modalities: 41 % of the reviewed works evaluated their methods on this database (see Table 2). Facial expression is often combined with head pose, which can be estimated from the same images (see Table 3, 4, and [95]). Irani *et al.* [82] and Haque *et al.* [80] not only use RGB images, but also thermal and depth images for analyzing facial expression. Adibuzzaman *et al.* [73] use photographs taken with smartphone cameras, which are heterogeneous regarding quality, resolution, sharpness, lighting *etc*. Generally, cameras suffer from a limited field of view (which also may be occluded or inadequately lighted) and interpreting images is more complex than other sensor signals. However, cameras are non-contact and may be more comfortable for the patient and convenient for the medical staff than contact-based sensors. Cameras are also a rich source of information with potential to measure not only facial expression, but also heart rate, respiration, or body movement. *E.g.* Zamzmi *et al.* [72] measured body movement (and facial expression) of neonates with a camera. Facial pain expression is widely reported as being not only sensitive, but also specific to pain (see Sec. 2.3). For other modalities, specificity to pain is either known to be low (*e.g.* other stressors can elicit the same response patterns as pain) or unknown. On the other hand, facial expression may be easier to fake than other physiological pain responses.

UNBC-McMaster	Processing					Output						Validation		
Paper	Features		Model / Fusion	Context		Objective				GT		# Subj.	# Vid.	Method
Ashraf '07 [1]	27k F	shape, SAPP, CAPP	SVM					VP			O	21	?	✓
Ashraf '09 [2]	27k F	shape, SAPP, CAPP	SVM			FP		VP		P	O	21	69	✓
Chen '13 [3]^{T}	4k F	LBP	AdaBoost		P	FP				P		25	200	✓
Chen '17 [4]	3k F	HOG, HOG-TOP	SVM / FF & DF	T		FP		VP		P	O	25	139	✓
Egede '17 [5]^{T}	8k F	pretrained CNN, HOG, shape	RVR / DF	T	P		FIC			P		25	200	✓
Florea '14 [6]^{T}	? F	learned projection of HoT	SVR (RBF) ensemble / DF				FIC			P		25	200	✓
Florea '16 [7]^{T}	56 F	learned projection of HoT	SVR (RBF) ensemble / DF	T			FIC			P		24	?	✓
Ghasemi '14 [8]	10 T	histogram of AUs	HCRF	T					VI3		O	25	200	✓
Hammal '12 [9]	4k F	log-normal filters	SVM				FI4			P		25	53	?/✓
Hong '16 [10]	4k F^{1}	2Standmap of texture	SVR				FIC			P		25	200	?
Irani '15 [11]	6 F	filter response histograms	rule-based	T			FI3			P		12	50	?
Kaltwang '12 [12]	5k F	shape, LBP, DCT	RVR (RBF) / DF				FIC			P		25	200	✓
Kaltwang '16 [13]	? F	LBP, LBP-TOP	Doubly sparse RVR (RBF)	T			FIC			P		25	200	(✓)
Khan '13 [14]	3k F	pyramid HOG & LBP	k-NN			FP				P		25	200	?
Kharghanian '16 [15]^{T}	? F	learned w/ un-sv. CDBN	SVM			FP				P		25	200	✓
Lopez-M. '17 [16]	40 F	PCA of shape	LSTM & HCRF	T	P		FIC		VI10	P	**V**	25	200	(✓)
Lo Presti '15 [17]	? T	Hankel matrix of shape	NN			FP^{2}				P		25	200	✓
Lo Presti '17 [18]	? T	Hankel of Haar & Gabor	AdaBoost			FP^{2}		VP		P	O	25	200	✓
Lucey '08 [19]	500 F	DCT	SVM			FP				P		20	142	✓
Lucey '11 [20]	16k F^{1}	shape, SAPP, CAPP	SVM / DF			FP				P		25	203	✓
Lucey '12 [21]	8k F^{1}	shape, CAPP	SVM			FP			VI3	P	O	25	200	✓
Meng '14 [22]	132 F	shape	k-NN, Hidden Markov Models	T		FP				P		25	200	✓
Neshov '15 [23]	270 F	SIFT around landmarks	SVM, reg. lin. regression			FP	FIC			P		25	200	?/✓
Pedersen '15 [24]	128 F	learned by semi-sv. autoencoder	SVM			FP				P		25	200	✓
Rathee '15 [25]	132 F	sv. DML from shape	SVM (RBF)				FI16			P		25	200	?
Rathee '16 [26]	? F	sv. DML from Gabor, HOG, LBP	SVM			FP	FI4			P		25	200	?
Rodriguez '17 [27]^{T}	4k F	finetuned CNN	CNN + LSTM	T		FP	FIC^{3}			P		25	200	✓
Romera-P. '13 [28]^{T}	? F	facial distances	reg. multi-task learning (lin. model)		P		FIC			P		24	?	(✓)
Roy '15 [29]	? F	PCA of Gabor	SVM			FP	FI4			P		?	?	?
Rudovic '13 [30]^{O}	2k F^{1}	LBP	Heteroscedastic CORF	T			FI6			P		22	147	(✓)
Rudovic '15 [31]^{O}	18 F	PCA of shape	context sensitive CORF	T	P		FI6			P		25	200	(✓)
Ruiz '14 [32]^{W}	3k F	spatiotemporal SIFT around landmarks	Regularized Multi-Concept MIL					VP			O	23	147	✓
Ruiz '16 [33]^{WO}	98 F	shape	Multi-Instance Dynamic Ordinal Random Fields	T			FI6		VI6	P^{4}	O	25	157	✓
Rupenga '16 [34]	132 F	shape	SVM, Extreme Learning Machine			FP				P		8	8	?
Sangineto '14 [35]^{T}	334 F	PCA of LBP	SVM, Multi-Output SVR (EMD kernel)		P	FP				P		24	?	✓
Sikka '14 [36]^{W}	200 F	Bag of Words of multi-scale dense SIFT	Multi-Segment MIL w/ Gradient Boosting	T		FP	FIC^{4}	VP		P^{4}	O	23	147	✓
Wang '17 [37]^{T}	N/A F	learned	CNN				FIC^{3}			P		25	200	✓
Werner '17 [38]	2k V	Time series descriptor (TSD) of shape & CAPP	SVM					VP	VI3		O	25	200	✓
Yang '16 [39]	? F	texture descriptors	SVM			FP		VP		P	O	21	147	✓
Zafar '14 [40]	1k F	shape	k-NN			FP	FI17^{4}			P		?	?	?
Zen '14 [41]^{T}	334 F	PCA of LBP	SVM, SVR (RBF)		P	FP				P		24	?	✓
Zhao '16 [42]^{WO}	2k F	PCA of shape, LBP, Gabor	Ordinal SVR				FIC^{3}			P		24	191	✓
Zhou '16 [43]	N/A	learned	Recurrent CNN	T			FIC			P		25	200	✓

^{W} Learning from weakly labeled data ^{O} Uses ordinal information ^{T} Involves transfer learning
^{1} Feature number is not given explicitly in paper, but approximated from feature description. ^{2} Time window of 10 frames is classified with thresholded PSPI sum as ground truth. ^{3} Ground truth intensity quantized to 6 levels. ^{4} Testing only
CAPP: Canonical APPearance CDBN: Convolutional Deep Belief Network CNN: Convolutional Neural Network
CORF: Conditional Ordinal Random Field DCT: Discrete Cosine Transform DML: Distance Metric Learning HCRF: Hidden Conditional Random Field
HOG: Histogram of Oriented Gradients HoT: Histogram of Topographical features LBP: Local Binary Pattern lin.: linear
LSTM: Long Short-Time Memory Recurrent Neural Network NN: Nearest Neighbor PCA: Principal Component Analysis SAPP: Similarity normalized APPearance SIFT: Scale Invariant Feature Transform sv.: supervised SVM: Support Vector Machine (classification) SVR: Support Vector Regression
reg.: regularized RBF: Radial Basis Function kernel RVR: Relevance Vector Regression

TABLE 2
Camera/facial expression-based pain recognition systems evaluated on **UNBC-McMaster Shoulder Pain database** [165]. We list **features** (number; F=frame-, T=time-window-, V=video-level; type), **model / fusion** (FF=feature fusion, DF=decision fusion, more abbreviations see above), used **context information** (T=temporal, P=person), **objective** of prediction (FP=frame-level presence of pain [binary], VP=video-level presence [binary], FIx=frame-level intensity of pain, VIx=video-level intensity, with x denoting the number of intensity levels or x=C for continuous output), used **Ground Truth (GT)** (P=Prkachin and Solomon Pain Intensity score [PSPI] [144], [165], O=Observer-assessed Pain Intensity, V=Visual Analog Scale [VAS] self-report), and number of subjects and videos used for validation, as well as the validation **method** (✓=recommended leave-one-subject-out cross validation; (✓)=other subject independent validation; ?=uncontrolled overlap of subjects between train and test set and/or lack of essential experimental details in paper, results may be biased, comparability is limited).

BioVid database	Input	Processing			Output	Validation	
Paper	Mod./S.	Features	Model / Fusion	Context	Objective	Part	# Subj.
Amirian '16 [44]	3E	? TSD, ptEDA	RBF Neural Network / mid-level, FF, DF	(PFS)	P I5 IC	A, C	86
Gruss '15 [45]	fE 3E	159 TSD, similarity	SVM (RBF) with FS	PFS	P I5	B	85
Kächele '15 [46] (Bio-Visual ...)	C fE 3E	425 TSD, 3dwFace	SVM with FS, Random Forest / FF, DF		P I2	A, B	87
Kächele '15 [47] (Multimodal ...)	C 3E	3k TSD, ptEDA, ldFace	Random Forest / FF, DF		P I2 IC	A, C	86
Kächele '16 [48]	3E	? TSD, ptEDA	Random Forest & k-NN / FF, DF	PMU	P I5 IC	A	87
Kächele '17 [49]	C 3E	3k TSD, ptEDA, ldFace	Random Forest / FF, DF	PMU	P I2 IC	A, C	87
Lopez-M. '17 [50] (Multi-task ...)	2E	17 TSD	Multi-task neural network	PMS	P	A	87
Lopez-M. '17 [51] (Physiological ...)	C 2E	290 TSD	Multi-task neural network	PMS	IC	A	85
Lopez-M. '18 [52] (Skin ...)	1E	6 TSD	Logistic Regression, SVM (linear & RBF)		P	A	87
Lopez-M. '18 [53] (Continuous ...)	2E	13 TSD	LSTM neural network		P IC	A	87
Walter '14 [54]	fE 3E	135 TSD	SVM (RBF) with FS	PFS	P I2	B	86
Walter '15 [55]	C fE 3E	? TSD, 3dwFace	Random Forest / FF, DF		P I2	B	86
Werner '13 [56]	C	299 TSD, 3dwFace	SVM (RBF)	(PMS)	P		90
Werner '14 [57]	C 3E	507 TSD, 3dwFace	Random Forest / FF	(PMS)	P	A	87
Werner '17 [38] (Automatic ...)	C	1k TSD, 3dwFace	Random Forest / FF	(PFS)	P I5	A	87
Werner '17 [58] (Analysis ...)	C	1k TSD, 3dwFace	Random Forest	PFS	P I3	A	7-87
Yang '16 [39]	F	34k texture descriptors	SVM		P	A	87

FS: Feature Selection k-NN: k-Nearest Neighbor SVM: Support Vector Machine (classification) RBF: Radial Basis Function

TABLE 3
Pain recognition systems evaluated on **BioVid database** [56], [168]. We list **modalities / sensors** (C=Camera/facial expression & head pose; F=Camera/facial expression; fE=facial sEMG [corrugator & zygomaticus]; 1E = EDA; 2E=EDA & ECG; 3E=EDA, ECG & trapezius sEMG) **features** (number; type: all papers use Heart Rate Variability features for ECG; TSD=time series statics descriptor with various features of amplitude, variability, entropy, similarity, frequency, and/or linearity; ptEDA=phasic & tonic decomposition of EDA; 3dwFace=hand-engineered facial 3D distances & wrinkle measures; ldFace=LBP-TOP & generic facial distances), **model / fusion** (FF=feature fusion [applies if no fusion mentioned], DF=decision fusion, more abbreviations see above), used **context information** (PFS = Person-specific Feature Standarization [sample distribution of test person used], PMS=Personalized Model by Supervised method [labeled samples of test person used], PMU=Personalized Model by Unsupervised method [no labeled samples of test person used]), **objective** of prediction (P=presence of pain [binary, with different pain intensities], I2=intensity classified in binary pairs, I5=intensity (5 classes), IC=intensity in continuous scale), and dataset part (see Sec. 4.3) and number of subjects used for validation. All papers classify time windows regarding their pain stimulus ground truth; pain intensities were individually calibrated by self-report.

SenseEmotion	Input	Processing		
Paper	Mod./S.	Features	Fusion	Context
Kessler '17 [59][P]	C RE	? TSD, dFace	HDF	
Thiam '16 [60]	A C	? various	DF	(PMS)
Thiam '17 [61] (Multi-modal...)	A C R3E	2k sv. dim. reduction of various features	FF & DF	(PMS)
Thiam '17 [62] (Hierarchical...)	C	? TSD, dFace, generic texture descriptors	HDF	PMU

[P] Involves remote photoplethysmography (rPPG)

TABLE 4
Pain recognition systems evaluated on **SenseEmotion database** [174] (40 subjects, classification of pain vs no pain, ground truth is the applied pain stimulus). All approaches classify with **Random Forest**. We list **modalities / sensors** (A=Audio, C=Camera/facial expression & head pose; RE=RSP & ECG; R3E=RSP, EDA, ECG & trapezius sEMG), **features** (number; type: TSD=time series statics descriptor, dFace=facial distances), **fusion** (FF=feature fusion, DF=decision fusion, HDF=hierarchical DF), used **context information** (PMS=Personalized Model Supervised, PMU=Personalized Model Unsupervised).

EmoPain	Input	Processing		Validation	
Paper	M./S.	Features	Model	#Su.	#Sa.
Aung '16 [63]	F	4k shape, LBP, DCT	SVM	17	317k[1]
Aung '16 [63]	B	30 hand-engineered	RF	?	152
Olugbade '14 [64]	B	13 hand-engineered	SVM w/ FS	31	49
Olugbade '15 [65]	B	15 hand-engineered	SVM, RF w/ FS	22	98

[1] Video frames FS: Feature Selection RF: Random Forest
SVM: Support Vector Machine

TABLE 5
Pain recognition systems evaluated on **EmoPain database (chronic low back pain)** [63]. We list **modalities / sensors** (F=Facial expression [camera], B=Body movement [motion-capture system & sEMG]), features (number, type), model, and number of **subjects** and **samples** used for validation. Olugbade et al. [64], [65] classify 3 pain intensities (high vs low self-report of patients vs healthy controls). Aung et al. [63] detect facial pain expression as labeled by naive observers and predict extent of occurrence of pain behaviors as labeled by expert observers.

5.1.2 Contact-Sensor Approaches

Since the release of the BioVid database, the contact-based sensors EDA and ECG have become the second-most widely used (13 % of reviewed papers, see Table 3, 4, and 7), followed by sEMG of the trapezius muscle (back of the neck, 10 % of papers, Table 3 and 4). Facial expression was also measured with sEMG at the corrugator supercilii (brow lowerer) and zygomaticus (mouth corner raiser) muscle (Table 3). Research in this direction has currently intensified due to the upcoming X-ITE database and other current studies, *e.g.* by Jiang *et al.* [83] who used 5 facial muscle sEMGs. In the context of low back pain assessment, sEMG was used to measure muscle activity during specific exercises [63]–[65], [81]; movement was measured via a motion capturing (MoCap) system [63]–[65] or inertial sensors [74]. Several researchers have analyzed brain activity to recognize pain from EEG, fMRI and fNIRS (see Table 7). Kessler *et al.* [59] and Thiam *et al.* [61] used respiration signals, Chu *et al.* [77], [78] used BVP. Wang *et al.* [102] and Yang *et al.* [105] recognized pain from various physiological parameters obtained from hospitals' electronic flow sheets. Hand movement and finger pressure were exploited by Rivas *et al.* [94] to detect pain and other states of stroke patients doing rehabilitation exercises.

Infant pain	Input	Processing			Validation	
Paper	Mod./S.	Features		Model/F.	#Su.	#Sa.
Brahnam '07 [66]	F	70	PCA	N. Netw.	26	204
Chang '16 [67]	A	?	Spectogram	CNN	?	2k
Pal '06 [68]	A F	6	pitch, formants; distances	k-NN; rule-based / DF	?	2k
Petroni '95 [69]	A	10	MFCC	N. Netw.	16	230
Rosales-P. '15 [70]	A	304	MFCC	GSFM	?	542
Sailor '18 [71]	A	39	ConvRBM	GMM	?	192
Zamzmi '17 [72]	A F VS	?	various	k-NN, RF, SVM / DF	18	?

CNN: Convolutional Neural Network
ConvRBM: Convolutional Restricted Boltzmann Machine GSFM: Genetic Selection of a Fuzzy Model GMM: Gaussian Mixture Model
k-NN: k-Nearest Neighbor N. Netw.: Neural Network
RF: Random Forest SVM: Support Vector Machine

TABLE 6
Pain recognition systems for **infants**. See [124] for more works. We list **modalities** / **sensors** (A=Audio [cry], F=Facial expression [camera], VS=Vital Signs [heart rate, respiration, oxagen saturation] and Body movement); features (number, type); **model** / **fusion** (DF=decision fusion); and number of **subjects** and **samples** used for validation. All papers classify pain vs other conditions. Pain was stimulated by heel lancing [66], [72] and immunization [69], [72].

Among the single modalities that have been compared so far, EDA quite consistently performs best [44], [49], [61], [78]; only facial expression (sEMG and camera-based) outperforms EDA in some experiments [46], [57]. EDA is less person-specific than other modalities [57], [61], which is beneficial for generalizing to unseen persons. However, EDA responses, like most physiological measures, are not specific to pain. EDA can increase with psychological or physiological arousal, as well as with certain neurological events [182]. The specificity of pain recognition has not been addressed adequately so far and future work should include distinguishing pain and other affective states, such as anxiety or anticipation. Downsides of most contact-based sensors are that (1) motion often leads to artifacts and (2) that signals are influenced by variation in the sensor's connectivity with skin, *e.g*. adhesive electrodes often come loose over time and have to be reattached.

5.1.3 Audio Approaches

In the audio domain, most efforts to recognize pain have focused on infant cries. Early work dates back to the 1990's [69] and there is still active research in this domain [67], [70]–[72], see Table 6. Analyzing cries is very valuable for recognizing pain among infants, who are a vulnerable group with limited communication abilities that would probably benefit a lot from the clinical adoption of pain recognition technology. Infant pain recognition has been recently covered in detail by the review paper of Zamzmi *et al*. [124]; thus, we do not repeat their coverage here. Aside from the infant cry domain, only a few papers have assessed pain with audio. Thiam *et al*. [60], [61] analyzed the audio signals of the SenseEmotion database, which do not contain verbal interaction, but mostly breathing noises and sporadic moaning sounds. In contrast, Tsai *et al*. [97], [98] and Li *et al*. [84] analyzed audio signals recorded during clinical interviews in an emergency triage situation. Whereas audio outperformed video-based facial expression recognition in Tsai *et al*. [97], the opposite results were found by Thiam *et al*. [60]. The verbal communication during the interview in Tsai's work leads to (1) more audio material with potentially discriminative information and (2) facial movements due to speaking that may interfere with facial expression recognition, which together may explain the superiority they found with audio. In Thiam's work, (1) all facial movements are expression-related and (2) moaning and pain-related breathing patterns may occur less consistently than facial responses. Oshrat *et al*. [91] analyzed prosody of patients with spinal cord and/or brain injuries for classifying significant *vs*. non-significant pain. A challenge in the analysis of audio data is to separate the sounds of interest from background noises, which may originate from medical devices, other people, or events.

5.1.4 Multimodal Approaches

A promising direction is to combine modalities in a multimodal system. Heterogeneous information sources may complement each other and lead to improved specificity and sensitivity. Generally, if the predictive performances of the single modalities are sufficiently good, their fusion tends to improve the results. This has been shown for combining facial expression and head pose [38], [56], [57]; EDA, ECG, and sEMG [44], [57]; video, EDA, ECG, and sEMG [46], [47], [49], [57]; video, RSP, ECG, and remote PPG [59]; video and audio [60], [97]; video, audio, EDA, ECG, EMG, and RSP [61], and MoCap and sEMG [64]. Similarly, unimodal systems for infant pain recognition have been outperformed by integrating facial expression, body movement, vital signs, and crying sound modalities [72]. Next to better performance, a multimodal system also facilitates improved flexibility and availability. In clinical environments, a modality may be unavailable due to various factors, *e.g*. the face may be injured or occluded by an oxygen mask hindering the measurement of facial expression. A multimodal system may be able to compensate for the lack of one or even multiple modalities and still provide a useful assessment.

5.2 Processing: Features, Models, and Use of Context

The input data are processed in order to find and use patterns for predicting a latent pain state that the observed person is in. For this purpose, **features**, which are a more discriminative and usually lower dimensional representation, are extracted from the raw input data. Features may be categorized as (1) generic features, (2) hand-designed features, and (3) learned features. Generic features are based on ideas that proved successful in other domains, but are not specifically adapted for pain recognition. Examples are local binary pattern (LBP) features for image data and frequency spectrum coefficients for one-dimensional signals. Hand-designed features are developed for the specific task taking advantage of expert knowledge; they are usually easy to interpret and lower dimensional. Examples are facial distances in image-based expression analysis or heart rate variability features extracted from ECG. With learned features, the feature extraction is optimized for the specific task during the training procedure. Most deep learning approaches fall into this category. The learned features are usually high dimensional and not easy to interpret, but facilitate highest recognition performance if trained with enough suitable data (which however may be not available). Generally, higher-dimensional feature vectors may contain more information, but also require more training data in order to identify the patterns that are relevant for the prediction task.

5.2.1 Frame-Level Facial Expression Features

In camera-based facial pain expression recognition, feature extraction is part of a processing pipeline. This may include (1) localizing facial landmarks (points along mouth, eyes, and eyebrows or more) and (2) registering landmarks and/or facial texture to gain invariance to translation, scale, and rotation. Those steps are generally applied in facial analysis (and extensively discussed in other reviews, such as [183]), where they localize relevant parts of the 2- to 4-dimensional signals (grayscale images to multispectral video), which are subsequently used to extract features. In this domain features can further be distinguished as (1) frame-level features *vs*. features that integrate information over time (time-window or video level), (2) geometric *vs*. appearance features, and (3) local *vs*. global features. A variety of frame-level features have been used for recognizing facial pain expression: (1) generic shape features (most often plain landmark coordinates) [1], [2], [5], [8], [12], [16], [17], [20]–[22], [25], [31], [33], [34], [38], [40], [42], [63], [89]; (2) generic appearance features, which include plain pixel representations ("SAPP", "CAPP", and similar) [1], [2], [8], [20], [21], [38], [43], [66], [73], Local Binary Pattern (LBP) [3], [12]–[14], [26], [30], [35], [39], [41], [42], [63], Histogram of Oriented Gradients (HOG) [4], [5], [14], [26], [60], [62], Gabor [18], [26], [29], [42], [75], [86], [95], other filters [9], [11], [82], Scale Invariant Feature Transform (SIFT) [23], [32], [36], Discrete Cosine Transform (DCT) [12], [19], [63], and others [6], [7], [10], [18], [39], [72]; (3) hand-engineered features, namely facial distances in 2D [28], [47], [49], [59], [60], [62], [68], [76], [89], [97] and 3D [90], which are often combined with measures of wrinkles, bulges, or/and furrows to capture some additional changes in appearance [79] (with 2D distances); [38], [46], [56]–[58], [103], [104] (with 3D distances); (4) features learned with neural networks [5], [15], [24], [27], [37], [43], [80]; and (5) features learned with other methods [6], [7], [25], [26].

Many researchers reduce the dimensionality and decorrelate features with principal component analysis (PCA); the resulting feature space may be considered to be learned in an unsupervised way. Other dimension reduction / feature learning methods appear to be more promising: Egede *et al*. [5] trained Convolutional Neural Networks (CNN) to recognize facial action units on the BP4D database and later applied the networks on the UNBC-McMaster database to extract features for predicting pain intensity. The idea to improve results by using knowledge from related domains is called transfer learning. It is a widely and successfully used strategy to cope with a lack of data, which is one of the problems in pain recognition. Florea *et al*. [6], [7] learned a feature transform on the CK+ emotion recognition dataset and transfered the data representation to the UNBC-McMaster database; they argued that this increases robustness, because CK+ has more subjects than UNBC-McMaster. Kharghanian *et al*. [15] trained a Convolutional Deep Belief Network (CDBN) to extract features; the unsupervised training involved parameterizing the first layer on natural images and the second layer on UNBC-McMaster. CNNs trained for face recognition were adopted and fine-tuned for facial pain recognition by Rodriguez *et al*. [27] and Haque *et al*. [80]. Similarly, Wang *et al*. [37] fine-tuned a face verification CNN for pain intensity estimation. Others did not transfer knowledge, but trained neural networks from scratch: Pederson *et al*. [24] trained a semi-supervised autoencoder to extract features for detecting pain and compared it with a fully unsupervised autoencoder. Zhou *et al*. [43] proposed a recurrent CNN for predicting pain intensity. In contrast to the other authors, both, Zhou *et al*. [43] and Wang *et al*. [37] did not explicitly extract features that are fed into a separate recognition model, but trained an all-in-one deep neural network combining the implicit feature extraction and the recognition model in a hardly separable and jointly optimized unit. Several works predicted action units (AUs) with a first set of models [8], [75], [85], [86], [95] (this can be considered a high-level supervised dimension reduction); the AUs were then used as features for a subsequent pain recognition model. In this context, Bartlett, Littlewort, and Sikka [75], [86], [95] also transfer knowledge, because the AU models were trained on other datasets.

5.2.2 Facial Expression Features beyond Frame-Level

Beyond the temporal scope of a single frame, facial expression features are often extracted through spatio-temporal descriptors. Every texture descriptor, such as LBP, can be extended to include the temporal domain by applying the Three Orthogonal Planes (TOP) principle. Instead of only extracting features from the spatial x-y plane, TOP also applies the same method on the spatio-temporal x-t and y-t planes, and concatenates the three feature vectors. Several works in pain recognition use LBP-TOP [39], [47], [49], [60]–[62], HOG-TOP [4], and LGBP-TOP (Local Gabor Binary Pattern TOP) [62]. Alternatively, in order to obtain time-window- or video-level features, each arbitrary frame-level feature can be considered a time series. Then, the time window or video can be represented by a Time series Statistics Descriptor (TSD). TSDs consist of statistical measures of the time series, such as mean and quartiles [95]. Works vary regarding the selection of measures, ranging from three [95], over five [86], ten [47], [49], up to 15 measures [97] per frame-level feature. Following Werner *et al*. [56] statistics have also been extracted from the first and second derivatives of the time series [46], [56], [57], [59]–[62]. Additionally, the descriptor has been extended by variables measuring durations, count of segments, and area under time series curve [38], [58]. Aside from TSDs, Bartlett *et al*. [75] proposed a Bag of Temporal Features (BoTF) descriptor, which condenses filter responses of frame-level features' time series into histograms. Other ways of pooling frame-level features over time are max-pooling [60], [62], [85], bag of words [97], and plain histograms [8]. A few comparisons between time-window-level features can be found in [39], [62], [97]. Werner *et al*. [38] argued that in the context of facial expression videos, temporal integration of frame-level features (*e.g*. with time window descriptors) is superior to integration of frame-level decisions, because many aspects of dynamics, such as speed, tendency, or overall variation are easier to represent.

5.2.3 Physiological Time-Series Features

Except for camera images and brain imaging, all other sensor signals are processed as time series. In the analysis of EDA and sEMG signals we also find several variants and

	Input		Processing				Output		Validation				
Paper	Modality (Sensor)		Features		Model / Fusion	Context	Objective	GT	Stimuli	Clinical context	Age	# Subjects	# Samples
Adibuzzaman '15 [73]	F	(smartphone)	35 F	PCA (Eigenfaces)	SVM, k-Nearest Neighbor	(PMS)	I3	VAS	-	breast cancer	35-48	6+513	454+513
Ashouri '17 [74]		inertial sensor	30 T	PCA of velocities & accelerations	SVM (RBF)	-	P	P vs HC	trunk motion	chronic low back pain	20-50	52	52
Bartlett '14 [75]	F		480 T	Bag of Temporal Features of AU scores (with FS)	SVM (RBF)	-	real vs fake	stim.	cold pressor, instruction	-	adults	25	50
Chen '12 [76]	F		11 F	distances	rule-based	PFN	P	FACS	?	lung cancer	adult	4	?
Chu '14 [77]		EDA, ECG, BVP	34 T	TSD, PCA, sv. dim. reduction (Fisher projection)	LDA	-	I6	stim.	electrical	-	22-25	6	?
Chu '17 [78]		EDA, ECG, BVP	36 T	TSD, Feature selection with genetic algorithm	LDA & SVM	(PDFS)	I4	stim. (PS)	electrical	-	22-25	6	1k
Hammal '12 [79]	F		7 F	dwFace	Transferable Belief Model (temporal fusion)	Place, T, PFN	pain vs emotion	stim.	instruction / heat	-	adults	15 / 20	85 / 40
Haque '18 [80][T]	F	(RGB, depth, thermal)	N/A	learned	CNN + LSTM / FF, DF	-	I5	stim. (PS)	electrical	-	22-42	20	2k
Hung '14 [81]		sEMG	32 T	energy & frequency	PCA-Neural Network	-	P	P vs HC	lifting weights	chronic low back pain	33±?	52	52
Irani '15 [82]	F	(RGB, depth, thermal)	15 F	filter response histograms	rule-based	T	I3	?	pressure	-	66-90	12	2k VF
Jiang '18 [83]		sEMG, EDA, ECG, RSP	13 T	TSD, heart/resp. rate	Neural Network	T	I3	VAS	heat, electrical	-	adults	30	348
Li '18 [84]		audio	1k T	TDS & autoencoder	SVM	PD, PFS	P, I3	NRS	-	triage (Tsai '16)	adults	141	335
Liu '18 [85]	F		1 F	modified PSPI	GMM	-	I3	self-report	?	emergency d.	18-82	83	83
Littlewort '09 [86]	F		100 T	TDS of AU scores	SVM (RBF)	-	real vs fake	stim.	cold pressor, instruction	-	adults	26	312 OT
Lopez-M. '18 [87]		fNIRS	? T	B-spline coeff.	MT-MKL	PMS	P	stim. (PS)	heat	-	19-38	20	228
Misra '17 [88]	E		3 T	spectral perturbation	SVM (RBF)	T	(I2)[1]	stim. (PS)	heat	-	20±2	30	60
Monwar '09 [89]	F		21	distances & shape	SVM	-	P	?	?	?	adults	34	68
Niese '09 [90]	F		10 F	3D distances & angles	SVM	PFN	pain vs emotion	?	tourniquet	-	adults	21	? VF
Oshrat '16 [91]		audio	2k T	various	SVM	-	P	self-report	-	CNS injury	23-65	27	400
Panavaranan '13 [92]	E		2 T	power spectral density of P3 channel	SVM (polynomial)	-	P	stim.	heat	-	adults	9	?
Pourshoghi '16 [93]		fNIRS	? T	B-spline coeff.	SVM, clustering		l2	stim. (PS)	cold pressor	-	adults	19	61
Rivas '17 [94]		hand movement & finger pressure	8 T	average state, speed, acceleration	Semi-Naive Bayesian classifier / DF	PMS	P	observer	rehabilitation exercises	stroke patients	adult	1	6k OT
Sikka '15 [95]	F	head pose	42 T	TSD of AU scores & head pose	logistic & linear regression	(PD)	P, IC	various[2]	pressure (examination)	appendectomy postoperative	5-17	50	150+297
Susam '18 [96]		EDA	? T	TSD of EDA, PCA	SVM	T	P	stim. & NRS	pressure	(see Sikka '15)	5-17	21	?
Tsai '16 [97]	F	audio	1k T	TSD & BoW of MFCCs & facial geometry features	SVM / FF, DF	PFS	P, I3	NRS	-	triage (emergency dep.)	adults	117	205
Tsai '17 [98][T]		audio	1k T	learned with LSTM autoencoder + sv. fine-tuning	SVM / DF	-	P, I3	NRS	-	triage (emergency dep.)	adults	63	126
Vatankhah '13 [99]	E		? T	various hand-engineered	SVM (RBF) fine-tuned by custom method	-	P, I2	?	cold pressor	-	28±?	13	?
Vatankhah '16 [100]	E		? T	wavelet decomposition	SVM (RBF)	-	P, I2	?	cold pressor	-	28±?	13	?
Wager '13 [101]		fMRI	? T	activation maps	LASSO-PCR		(I4)[1], I2	stim. (PS)	heat	-	24±5	114	?
Wang '13 [102]		physiological data	2 T	PCA of time series	outlier detection (un-sv.)	PMS	P	?	?	cancer / ICU	adult	53	?
Werner '12 [103][O]	F		11 F	3D distances & wrinkle measures	SVM (RBF), linear intensity model	PFM	P, IC	observer	instruction	-	adult	35	5k VF
Werner '14 [104][O]	F		11 F	3D dist. & wrinkle m.	SVM (RBF), ellipsoidal m.	PFM	P, IC	observer	instruction	-	adult	35	5k VF
Yang '16 [105]		physiological data	25 T	binarized states	Restr. Bolzmann Machine	PMS	P	?	?	ICU	adult	4	1k

[T] Involves transfer learning. [O] Uses ordinal information [1] Data points of same subject & pain intensity were averaged; with N intensities, N samples per subjects were classified.

[2] Accuracy assessed for predicting self-reported pain rating (NRS) and time since surgery, and compared to observer rating.

TABLE 7

Various pain recognition approaches. We report **modality and sensor** (F=facial expression [camera], E=EEG), **features** (number; F=feature extraction per single image/frame, T=time window/series level; type: TSD=time series statics descriptor), **model** / **fusion** (FF=feature fusion, DF=decision fusion), used **context information** (T=temporal; PFN=person-specific features relative to neutral state; PFS=Person-specific Feature Standarization; PDFS=person and day-specific feature standardization; PD=person demographics [gender, age, ethnicity]; PMS=Personalized Model by Supervised method [labeled samples of test person used], PMU=Personalized Model by Unsupervised method [no labeled samples of test person used]), **objective** of prediction (P=presence of pain [binary], Ix=intensity of pain, with x denoting the number of intensity levels or C for continuous output I2=pairwise classification of intensities), used **Ground Truth (GT)** (VAS=Visual Analog Scale self-report, P vs HC= Patients vs Healthy Controls, stim.=stimulus [PS=person-specific calibration based on self-report], NRS=Numeric Rating Scale self-report), **stimuli** for inducing pain (instruction means posed pain), clinical context, age and number of subjects and samples (deault: time series/videos, VF=video frames, OT=overlapping time windows).

extensions of Time series Statistics Descriptors (TSD). Walter *et al*. [54] explored the usefulness of amplitude, variability, frequency, stationarity, entropy, and linearity features for recognizing pain from sEMG (trapezius, corrugator and zygomaticus muscle) and EDA. Gruss *et al*. [45] additionally considered features that measure similarity of the current signal with the respective person-specific mean baseline signal. Later works use subsets of Walter *et al*.'s features [46], [55], [57] or combine a subset with additional features [44], [47]–[50], [59], [61], [83]. *E.g.*, EDA has been decomposed into phasic and tonic components based on physiological motivations, before extracting features from both components separately [44], [47]–[49], [52], [53], [61]. All of the above-mentioned work analyzed ECG with heart rate variability features. In contrast, Chu *et al*. [77], [78] only extract generic statistical features from ECG (as for BVP and EDA), not considering any frequency information. Yang *et al*. [105] binarize physiological parameters into normal or abnormal states before feeding them into the classifier. In the low back pain domain, sEMG signals were analyzed with energy and frequency features [81] as well as with hand-designed features based on the upper envelope of the rectified signal [63]–[65]. In this context, body movement is also represented by selected parameters of the human skeleton obtained with motion capturing [63]–[65]. In a rehabilitation context, Rivas *et al*. [94] extracted mean position, speed, and acceleration from hand movement and finger pressure time series.

5.2.4 Audio Features

In the audio domain the most widely used features are Mel Frequency Cepstral Coefficients (MFCC) [60], [61], [69], [70], [91], [97], [98], a spectral representation of sound that approximates the human auditory system's response. Other features include pitch [68], [84], [91], [97], [98], intensity [84], [91], [97], [98], Relative Spectral Perceptual Linear Predictive (RASTA-PLP) coefficients [60], [61], [91], Linear Predictive Coding (LPC) coefficients [60], [70], [91], harmonic to noise ratio [98], and formants [68]. It is common to include the first and second order temporal derivatives of features [60], [61], [91], [97], [98]. The features are extracted from short time-windows called frames (*e.g*. 60 ms) and are called low-level descriptors (LLDs). The time-series of LLDs in longer time windows are typically summarized through statistical functions to form high-level descriptors (as the TSDs above). Tsai *et al*. [98] used deep transfer learning on top of the above mentioned LLDs; they pre-trained an LSTM autoencoder on a large collection of Chinese TV talk show recordings (unsupervised) before fine-tuning the bottleneck layer to optimize the feature extraction for the pain recognition task. Li *et al*. [84] used the standard basic acoustic parameter set eGeMAPS [184] and trained a Maximum-Mean Discrepancy Conditional Variational Autoencoder to encode the eGeMAPS LLDs into a lower-dimensional latent representation. Next to using common features, Oshrat *et al*. [91] manually designed new features for pain recognition. Chang *et al*. [67] fed spectogram images into a CNN for recognizing infant cries. Sailor *et al*. [71] trained a Convolutional Restricted Boltzmann Machine (ConvRBM), learning an auditory filter bank optimized for infant cry recognition. See [124] for more work on infant cry recognition.

5.2.5 Brain-Activity Features

In brain-activity based pain recognition, feature extraction is very diverse. Panavaranan and Wongsawat [92] classify pain from EEG by considering the alpha and beta bands of a single channel's power spectral density. Vatankhah *et al*. use Lyapunov exponents, fractal dimension, entropy, and energy ratios [99] as well as sub-band statistics of wavelet decomposition [100]. Misra *et al*. [88] apply Independent Component Analysis and use event-related spectral perturbation, which were manually selected based on the study data, as features for pain classification. In fMRI experiments, Wager *et al*. [101] consider activation maps of brain regions that are pain-related according to a meta analysis of prior literature. In fNIRS experiments, both Pourshoghi *et al*. [93] and Lopez-Martinez *et al*. [87] extracted functional data analysis features from HbO_2 responses.

5.2.6 Recognition Models

Following feature extraction, the second essential processing component is the model that maps the features to the latent pain state. All types of machine learning models can be applied here. The model may involve data fusion, especially for integrating data in a multimodal system, which can be done at feature, decision, or an intermediate level.

Most approaches classify pain with Support Vector Machines (SVMs), either linear (see "SVM" in Tables 2-7) with Radial Basis Function (RBF) kernel [25], [45], [54], [56], [74], [75], [86], [88], [90], [99], [103], [104] or polynomial kernel [92]. Continuous valued output is obtained with the related Support Vector Regression [6], [7], [10], [35], [41], [42] and Relevance Vector Regression [5], [12], [13] models. Other widely used models are Random Forests (RF) [38], [46]–[49], [55], [58]–[63], [65], [72], [80], [179], Nearest Neighbor (NN) classifiers [14], [17], [22], [40], [48], [68], [72], [73], variations of Conditional Random Fields (CRF) [8], [16], [30], [31], [33], and various neural networks. Neural networks models used for pain recognition include Convolutional Neural Networks (CNN) [27], [37], [43], [67], [80], Long Short-Term Memory (LSTM) networks [16], [27], [80], Radial Basis Function (RBF) networks [44], [49], Restricted Bolzmann Machine (RBM) [71], [105], multi-task network [50], PCA neural network [81], and traditional multi-layer perceptrons [49], [66], [69], [83]. Interestingly, Brahnam *et al*. [66] trained their network with a genetic algorithm instead of backpropagation. In addition to this, researchers have used a Gaussian Mixture Model (GMM) [71], [85], a variant of naive Bayes classifier [94], and Genetic Selection of a Fuzzy Model (GSFM) [70].

For predicting pain intensity, several authors took advantage of the **ordinal relationship** between intensities. Werner *et al*. [103], [104] proposed to simplify frame-level labeling by avoiding the challenge of committing to an absolute intensity value; instead they labeled each video with ordinal pain intensities yielding sample pairs, each with a less and a more intense expression, which are then used to learn a continuous pain intensity estimation model. Zhao *et al*. [42] automatically constructed such pairs for learning an ordinal support vector regression model; they assumed that there is one apex (expression peak) per video, which needs to be manually selected, and that intensity is monotonically increasing and then decreasing, around the apex. Further, their proposed method can learn from both ordinal relationships and absolute intensities at the same time. Rudovic *et al*. [30], [31] and Ruiz *et al*. [33] enriched classification of pain intensity by considering the ordinal relationships among the classes with variants of Conditional Ordinal Random Field (CORF) models.

An important topic for pain assessment is learning from **weak labels**. Ground truth with coarse temporal granularity, such as pain intensity at video level (self-reported or observer-assessed), can be considered a weak label if the goal is to identify patterns at higher temporal resolution, such as facial expression in a single frame. Video-level ground truth has several advantages (see Sec. 4.2), including that it is much easier to gather. With Multiple Instance Learning (MIL), Sikka *et al*. [36] showed that it is possible to learn frame-level pain detection from video-level ground truth. In one video their prediction even outperformed frame-level ground truth; the learned model correctly discovered facial expression of pain in frames that FACS coders wrongly labeled with no action unit. Later, Ruiz *et al*. [32] applied the MIL in another framework showing *qualitatively* that the classifier learns to distinguish frames with painful and non-painful expressions from video labels only. In a follow-up work they combined MIL with ordinal regression; the proposed model outperformed several other approaches in predicting frame-level pain expression intensity after being trained with video-level intensities labels only [33]. Zhao *et al*. [42] evaluated three levels of supervision, including a weakly supervised setting in which they use 9% of the frame labels and ordinal information for training a frame-level intensity regressor. Wang *et al*. [102] experimented with unsupervised pain recognition based on outlier detection.

Related to supervised dimension reduction and learning of features (which we addressed above), **automatic feature selection** is another way of finding an optimized feature space. Authors used forward selection [45], [54], backward elimination [54], [64], [65], sequential floating forward selection [46], [61], all with wrapped SVM or Random Forest classification. Oshrat *et al*. [91] used Correlation-based Feature Selection. Feature selection can not only reduce dimensionality, but can also help with measuring the relevance of features and understanding the learned models. However, careful evaluation is needed to get unbiased and comparable results, see Sec. 5.4.

Data **fusion** is used to combine different modalities, features, decision scores, or other information sources for getting a single final prediction. Feature fusion (FF, also called early fusion), *i.e*. concatenating feature vectors, is very common and straightforward and we only explicitly mention it in Table 2-7 if the experiments include comparisons with alternatives, such as the performance obtained with the individual inputs or another fusion method [4], [38], [44], [46]–[49], [57], [80]. Decision fusion (DF, also called late fusion) methods combine the outputs of multiple models, either by a fixed rule, such as calculating the mean of outputs, or another trained model. Several authors extract multiple types of features from the facial expression modality, train an individual model for each feature, and fuse them through a second-level Linear Logistic Regression [20], Relevance Vector Regression model [5], [12], or Support Vector Regression ensemble [6], [7], usually outperforming the individual feature models. Feature fusion and several decision fusion options were compared in [44], [46], [47], [49], [60], [61], [97], but there is no clear winner and results depend on the classifier and data. Uncommon approaches include hierarchical decision fusion [59], [62], probabilistic decision fusion [4], [68], and mid-level fusion [44].

5.2.7 Use of Context

Context information, such as knowledge about the person, situation, or temporal context may be used by the system, usually improving predictive performance. However, exploiting knowledge about the person or situation requires assumptions (which sometimes may not hold) or manual interaction with the system. The use of temporal context often boils down to temporal smoothing of the predicted time series; alternatively it may contribute information about changes in an otherwise static representation. Note that the meaning of *temporal context* depends on the granularity of prediction. In frame-level prediction it refers to other frames; in time-window prediction it refers to other time windows. *I.e*. time-window or video-level methods that model dynamics independent of earlier or later time windows were *not* considered to be using temporal *context*.

Temporal context has been exploited (1) on the feature level, (2) through special models, and (3) by post-processing the output. Feature extraction often involves frames preceding and/or following the frame considered for classification, *e.g*. by extracting spatio-temporal descriptors with the TOP principle [4], [13] or a spatio-temporal volume as input for a CNN [5], by max pooling [36], or other methods [11], [82]. In the EEG domain, Misra *et al*. [88] apply baseline normalization of time-window samples, *i.e*. subtract the spectrum of the signals before pain stimulation from the spectrum of the following pain response to normalize features. Jiang *et al*. [83] applied test specific standardization, *i.e*. the samples of each of their tests (which took 108 s on average) were standardized independently to suppress within- and between-subject differences in feature value range. Special models that use temporal context of frame-level pain intensities are probabilistic graphical models [8], [16], [22], [30], [31], [33], recurrent neural networks [16], [27], [43], and a custom dynamic fusion [79]. Another alternative that can be combined with any model is post-processing of the output pain intensity. Florea *et al*. [7] apply temporal filtering of the pain intensity output for smoothing and removing blink-induced artifacts. To address person-specific biases in the intensity outputs, Egede *et al*. [5] subtracted the modal frame prediction of a video; their underlying assumption, which is reasonable for many applications, is that the person's face is neutral most of the time.

Differences between **persons**, *e.g*. in facial shape, appearance, and behavior, are a big challenge for automatic pain recognition and thus a common reason for using context information. Without further consideration, the sample distributions in feature space generally vary between individuals and lead to suboptimal recognition results. Next to post-processing as mentioned above, inter-individual variability has been addressed at both feature and model levels. Several works apply a person-specific feature transformation, either by calculating features relative to a person-specific mean or neutral state (PFN in Table 7, not marked in Table 2 and 5) [31], [63], [76], [90], [103], [104] or by applying a person-specific feature standardization (PFS in Table 3 and 7, not marked in Table 5), *i.e.* subtracting the mean and dividing by the standard deviation (or similar) [38], [44], [45], [48], [54], [57], [58], [64], [65], [84], [97]. Chu *et al.* [78] used a person and day specific feature standardization (PDFS) in their 7-day pain recognition experiment, since the physiological baseline signals are day dependent. The above mentioned baseline normalization as applied by Misra *et al.* [88] has a similar effect and may also normalize for situational influences to some extent. Apart from feature transformations, Lopez-Martinez *et al.* [16] personalized the estimation of self-reported pain intensity (VAS) by including an individual facial expressiveness score as an additional feature. At least one VAS and one observer pain intensity rating (OPI) are needed to calculate the expressiveness score. Olugbade *et al.* [65] found that adding a depression score as a feature can improve low back pain recognition over only using kinematic and muscle activity features. Sikka *et al.* [95] considered person demographics (gender, age, ethnicity) as additional features, but the inclusion did not alter performance of the facial behavior based recognition. In contrast, incorporating gender and age information improved the recognition performance significantly on the audio dataset used by Li *et al.* [84].

At model level, the use of context information can be classified into supervised methods, which require labeled samples of the test person, and unsupervised methods, which do not require labels but usually rely on the distribution of the test person's samples in the feature space. The supervised and unsupervised methods are marked PMS and PMU respectively in Tables 3, 4, and 7 (not specifically marked in 2). Several authors [4], [56], [57], [60], [61], [73], [94], [102], [105] trained person-specific models only using data of the person of interest, some comparing with generic models and other methods. Lopez-Martinez and Picard [50] and Lopez-Martinez *et al.* [51] used multi-task learning developing a set of person-specific classifiers that share a common hidden neural network layer. Romera-Paredes *et al.* [28] proposed a regularized multi-task learning, first learning a set of person-specific models and a common reference model while enforcing commonalities by a special regularization term. In a calibration stage, a personalized model was learned using a few labeled instances of the test person, while knowledge from the source domain was transfered through the regularization, which enforced similarity to the reference model. Chen *et al.* [3] applied inductive transfer learning by first training a set of weak classifiers with AdaBoost. Subsequently they selected and combined a set of K weak classifiers from the previously trained set with AdaBoost on the target data (test subject with few labeled samples). They also experimented with unsupervised person-adaptation, but the transductive transfer method did not outperform the generic model. More successfully, Sangineto *et al.* [35] proposed Transductive Parameter Transfer for personalized classification without labeled samples of the test subject by three steps: (1) training a subject-specific classifier for each subject in the labeled source training set, (2) training a regression model that maps the sample distribution of each subject to the corresponding decision boundary learned previously, and finally (3) using this model to compute the parameter vector of the target classifier for an unseen subject to get a personalized prediction model. The method is faster than other person-adaptation approaches, as there is no need to store and compare all the source and target samples. Zen *et al.* [41] modified this work to Support Vector-based Transductive Parameter Transfer by using a support vector based representation of the source distribution resulting in faster training and improved predictive performance. Kächele *et al.* [48] trained an individual model for each test person using similar persons from the training set. Measurement of similarity was based on (1) meta-information such as age, gender, or questionnaire items, (2) on distances of samples in feature space, and (3) on machine learning, *e.g.* on ranking of confidence of person-specific models on the test person. Using a subset of similar training subjects improved the recognition rate over training with all data with several measures. Thiam *et al.* [62] conducted similar experiments on other data. Another work by Kächele *et al.* [49] investigated unsupervised iterative person-adaptation based on confidence estimation; during online processing, samples with highly confident predictions were added to the training followed by a retraining of the model. Lopez-Martinez *et al.* [87] used multi-view multi-task learning to personalize the inference process while providing a neuroanatomical interpretation of the learned classifier weights.

Generally, pain recognition systems are validated on specific datasets, which narrows down **situational context**, allowing the recognition to indirectly take advantage of knowledge about the situation. This is most obvious for the body movement based low back pain recognition systems [63]–[65], [74], [81], which are designed for specific exercises and (generally) will not work outside these contexts. In vocalization based pain recognition, the SenseEmotion database (without social interaction) differs significantly from Tsai *et al.*'s [97], [98] emergency triage interview situational context. An explicit situational context variable was used by Hammal and Kunz [79]; their "place" variable may indicate a medical or non-medical context, and it biases the classifier towards the most relevant expression.

5.3 Output: Objectives and Ground Truth

The reviewed automatic pain assessment systems pursue different objectives. The most common are detecting the presence of pain (a binary classification) or assessing the pain intensity. Suitable ground truth is required for developing and evaluating such systems. As detailed in Sec. 4.2 the ground truth may originate from self-report, observation, or knowledge about the pain inducing procedure, the healing progress *etc.* An important aspect is temporal granularity of the ground truth. Usually, ground truth is sparse and only available once per pain stimulus or procedure. Finer granularity may be available, but the temporal resolution may be misleading (see Sec. 4.2).

The FACS-based Prkachin and Solomon Pain Intensity (PSPI) is one of the oldest and most commonly used **ground truth on the UNBC-McMaster database** ("P" in "GT" column of Table 2). It assesses facial expression on a frame level as an integer in a range 0-16 and is either thresholded to classify pain *vs.* no pain expressions ("FP" in "objective" column), quantized to classify n intensity levels ("FIn"), or modeled on a continuous scale with regression ("FIC"). Additionally, Observer-assessed Pain Intensity (OPI, integer in range 0-5) is available as ground truth per video ("O" in "GT" column of Table 2). Similarly to PSPI, several works have binarized OPI to classify the presence of pain ("VP" in "objective" column) or have quantized it to n levels to classify intensity ("VIn"). The Visual Analog Scale self-reported pain intensity (VAS, integer in range 0-10, "V" in "GT" column) is also provided per video, but was used only once so far [16].

The **temporal granularity** of ground truth and predicted output are usually the same due to the need for comparison in quantitative validation, but there are a few exceptions: In the context of weak labels, frame-level pain recognition can be learned from video-level ground truth, such as OPI. Sikka *et al*. [36] learned a binary classifier with binarized OPI labels and compared its frame-level predictions with binarized PSPI; they also showed a correlation between the classifier's decision score and PSPI, indicating some capability to infer intensity. Ruiz *et al*. [33] combined the underlying multi-instance learning idea with ordinal regression for directly modeling pain intensity, validating frame-level intensity predictions with PSPI, also only using video-level OPI for training. Generally, every model can be applied with a sliding time window to provide temporally continuous pain intensity estimates at a desired repetition rate, even if it has been learned from ground truth with coarser temporal granularity. This has been demonstrated on the BioVid Heat Pain dataset [44], [47]–[49]. For instance, Amirian *et al*. [44] trained with time windows that were temporally aligned with the pain stimuli and predicted across the whole pain experiment with a sliding window.

The ground truth available and used with the **BioVid heat pain database** is the applied stimulus intensity, *i.e.* one of four pain intensities or baseline (no pain stimulus). The temperatures were calibrated per person in order to compensate for different pain sensitivities. All works predict the presence of pain ("P" in Table 3) by classifying baseline versus one pain intensity (usually comparing the results of different pain levels). Prediction of pain intensity has been evaluated with all 5 classes ("I5") [38], [44], [45], [48], baseline and the two highest intensities [58], and pairwise intensity classifications ("I2") [44]–[47], [49], [54], [55]. Kächele *et al*. [47]–[49] and Amirian *et al*. [44] also apply regression for continuous valued intensity estimation. In the **SenseEmotion database** (Table 4) the ground truth is the applied heat pain stimulus intensity (with person-specific temperatures) as well, but with three intensities plus baseline. So far, only baseline *vs*. pain (of the three levels) has been addressed on this dataset, but no pain intensity measurement.

The **EmoPain database** (chronic low back pain) provides different ground truths for facial expression and body movement [63]. Facial expressions were labeled for pain intensity on frame level by eight naive observers during real-time playback. The resulting label time series were post-processed to account for some problems such as varying reaction time. For their facial pain detection experiments, Aung *et al*. [63] combined the eight raters' labeling into one by considering a frame to be painful if at least 3 raters agreed on that. Body movements were labeled by a group of experts, who first agreed on a set of protective behavior categories and subsequently labeled the frame-level occurrence of the categories. In subsequent experiments, the objective was to predict the extent of occurrence of the protective pain behaviors for each specific exercise [63]. In other works with the same dataset, Olugbade *et al*. [64], [65] classified body movement during specific exercises in three categories: high and low pain intensity, and no pain. Samples of low back pain patients were labeled as more or less painful depending on their self-report; samples of healthy controls were labeled with no pain. Ashouri *et al*. [74] and Hung *et al*. [81] classified chronic low back pain patients *vs*. healthy controls as well, but with other datasets.

In other **clinical contexts**, self report is generally the preferred ground truth, see Table 7. Adibuzzaman *et al*. [73] classified three intensities of breast cancer pain based on VAS self-report labels. Tsai *et al*. [97], [98] and Li *et al*. [84] used NRS labels to predict presence and intensity of pain during emergency triage interviews. In a pediatric post-operative setting, Sikka *et al*. [95] predicted presence and continuous-valued intensity of pain according to self-reported NRS and time since surgery (an objective ground truth in this context). Studies by Liu *et al*. [85] (emergency room) and Oshrat *et al*. [91] (CNS injuries) used custom self-report scales. In a study with lung cancer patients, Chen *et al*. [76] tried to detect pain based on prior knowledge of FACS action units.

Most **other work** used the applied stimulus as ground truth, either with [78], [80], [88], [101] or without person-specific calibration [75], [77], [79], [86], [92]. Few used annotation by trained observers [94], [103], [104], or did not (or not clearly) report on the origin of the used ground truth labels [82], [89], [90], [99], [100], [102], [105]. Jiang *et al*. [83] segmented their tests of increasing pain intensity into different levels based on self-report, similar to the the calibration procedure used for the BioVid database [168]. Again, the objective of most of these works was to detect the presence of pain or to predict the intensity of pain in discrete categories [77], [78], [80], [82]–[84], [88], [99]–[101]. A few works classify pain *vs*. emotions [79], [90] or distinguish genuine *vs*. simulated pain [75], [86]. The latter studies showed that an automatic computer vision system can outperform trained human observers in distinguishing real from faked facial expression of pain.

In **infant pain recognition**, ground truth usually originates from procedures. Pain is stimulated by heel lancing (for obtaining a blood sample) or immunization and classified *vs*. other stressors (such as hunger at feeding time, diaper change, or fear of a jack-in-the-box) and rest [66]–[69]. In contrast, Zamzmi *et al*. [72] use the Neonatal Infant Pain Scale, a multimodal observer pain scale, as ground truth; they classify pain *vs*. rest and the level of cry. Some other works [70], [71] do not clearly describe the ground truth labeling method. See Zamzmi *et al*. [124] for more details on pain assessment in infants.

5.4 Validation: Datasets and Methodologies

A proposed system needs to be validated for showing its usefulness. The validation's significance heavily depends on the **dataset** used. Important dataset factors are the number of subjects and *independent* samples — the more, the better. However, consecutive video frames are highly correlated and thus should *not* be considered as independent samples, as discussed later. Datasets may be classified by pain stimulation method (if any), ground truth, and medical population / subject group. The latter may be characterized by age group (form newborns to the elderly) and health condition. Validation with healthy subjects is useful, but validation in a **clinical context** is even more valuable, since it examines the potential real use case of the technology.

Most studies were conducted with healthy adults and experimental pain stimulation. Only a few works explicitly include elderly people, which are a highly relevant group since prevalence of pain increases with age. About 30 of the BioVid database subjects are 50-65 years old (see Table 3 for related works). Irani *et al*. [82] studied pressure pain responses in 12 healthy volunteers aged 66-90 years, but their publication lacks important methodological details. Most pain experiments were done with younger adults or with adults of unspecified age. In clinical contexts, none of the work so far has addressed patients with dementia, which is an important population since the capacity to report pain is diminished in moderate and severe dementia [146]. The age of the 25 shoulder-pain patients in the UNBC-McMaster database has not been reported [165], but the complete study's 129 participants were 42 ± 14 years old. Nine of the 22 chronic low back pain patients in the EmoPain dataset were aged 60-67. Liu *et al*. [85] studied pain responses of 140 adult emergency room patients; 21% of them were 60-82 years old. The breast cancer study by Adibuzzaman *et al*. [73] is outstanding regarding the number of subjects. They used 454 photographs collected from 6 patients in a longitudinal study to train a model to predict VAS and applied it to photographs of 513 other subjects. Regarding the number of samples, the BioVid database is currently the most comprehensive available dataset (13k pain stimuli of about 90 subjects). An overall outstanding validation was demonstrated by Sikka *et al*. [95] in the context of pediatric post-operative pain following laparoscopic appendectomy. They enrolled 50 patients aged 5-17 years and collected 150 video samples of ongoing pain and 297 of transient pain (manual pressure at the surgical site for typical clinical examination). They evaluated models for predicting self-reported pain and time since surgery for both conditions and compared the results with observer ratings from nurses. The automatic method was as successful (in ongoing pain) or better (in examination-induced pain) in estimating children's self report than nurses. From the clinical perspective, this is the most comprehensive validation of all reviewed papers. Another strong evaluation was done by Tsai *et al*. [97] with 205 interview sessions of 117 patients in emergency triage and NRS ground truth. Furthermore, they also analyzed clinical outcomes and showed that the system can help to predict prescription of pain-killers and hospitalization. In follow-up work, Li *et al*. [84] evaluate using a superset of these data, comprising 335 clinical interviews of 141 patients.

An essential aspect for comparing recognition systems is **generalization performance**, *i.e*. an estimate of how well we can expect the system to fulfill its recognition task on unseen data. Within each good recognition study, there is at least one test of generalization performance on data that were not used to train or tune the model. However, generally the results of different papers are not comparable. Even when different authors use the the same database, comparability is limited due to differences in (1) prediction tasks, (2) evaluation methodology, (3) performance measures, (4) degree of automation and manual intervention, and (5) used subsets of the data. Therefore, better published numerical results do not necessarily indicate a superior system. Listing performance numbers can cause readers of a survey to think that a method reporting 90% is better than a method reporting 80% even when the latter might work better in a head-to-head comparison where the comparison conditions are properly controlled. Reproducing results is also hard because many papers lack essential details (and some authors do not perfectly respond to mail). We strongly recommend to establish precise evaluation protocols to improve comparability of results in the future. Guidelines of how to use the data should be published along with every new database as has become common practice in domains such as face recognition [185].

Werner *et al*. [58] analyzed facial expressiveness of subjects in the BioVid dataset, illustrating that system validation results heavily depend on the selection of the **subject subset** (performance varies between 49% and 93% with exactly the same algorithm). Consequently, papers should mention if parts of the recorded data were excluded from experiments and why they were excluded, because only using a subset may bias the results and reduce comparability. On the UNBC-McMaster dataset (see Table 2), several works used different subsets of the data limiting comparability. Moreover, the distributed UNBC-McMaster database itself is only a small subset of the study data by Prkachin and Solomon [144] (25 of 129 subjects), which were selected without clarifying the selection criteria. Similarly, Bartlett *et al*. [75] does not explain why only 25 of 45 recorded subjects were used in the experiments. In a preceding work, Littlewort *et al*. [86] probably used the same dataset, but without mentioning that there were 45 study participants in total.

Further, the evaluation **method** is a critical factor if video frames are considered as samples, as in UNBC-McMaster. This is evident in work by Hammal and Cohn [9], who reported results with two evaluation methods, (1) the leave-one-subject-out cross validation proposed by the database providers [165] and (2) classic cross validation, in which data of a subject occurs in both the training and test sets. By only changing the validation method, obtained performances dropped from 91-96% to 40-67%. The main problem is that the frames of a video (and more generally samples from the same person) are not statistically independent in the distribution of all people. However, independent training and test sets are required to get unbiased estimates of generalization capabilities. *E.g*. consider one frame is in the training and the following frame in the test set; then the frame in the test set is not really unseen, as it is nearly identical to the preceding frame in the training set. Another quite common methodological flaw is to use test set labels in an earlier step that seems to be unrelated to the classification/regression, *e.g*. during feature extraction, feature selection, or supervised dimension reduction. *E.g*. if cross validation is applied, such an earlier step has to be repeated for every fold without using any of the respective test set labels. Misra *et al*. [88] manually selected features using all data and cross validated afterwards. The method may be valuable to get basic insights about pain responses in EEG. However, the obtained classification performance may be overly optimistic, because it is not a good estimate of how well the recognition would work on unseen persons. Further, Misra *et al*. [88] and Wager *et al*. [101] averaged all trials of the same subject and condition before training and testing, *i.e*. they classified averaged pain responses. This procedure requires test set labels for grouping samples, is not applicable in the prospective applications that we can imagine, and similarly yields performances that may be overly optimistic. Several other papers lack clarity regarding such methodological aspects making it hard to interpret the results and assess generalization potential. Papers often lack parameters that are essential for reproducing results, *e.g*. of machine learning or feature extraction methods. As a whole, more clarity is needed regarding methodology and we suggest that reviewers demand more detail.

6 Challenges and Promising Directions

In the following we address open challenges in automatic pain recognition and propose directions that we believe are promising to overcome them. We would like to encourage the scientific community to work on these topics for accelerating progress towards better pain assessment. Some general goals are to: (1) Advance pain recognition systems to **meet requirements for clinical application**. This includes adapting and validating systems on clinical populations; improving the specificity, sensitivity, and robustness in non-laboratory conditions; and focusing on non-technical topics such as acceptance of the technology and cost-effectiveness analysis. (2) Work on **other types of pain or other objectives**. Most current work addresses acute nociceptive pain. Chronic pain is in general more challenging than acute pain, but also has a larger impact on society (see Sec. 1). We encourage researchers to address chronic pain and to collect real-world ambulatory data to assess existing models under long-term chronic pain scenarios and develop refined methods that are helpful for affected patients. Visceral and neuropathic pain, for example, are other relevant types of pain that are rarely addressed. Further, most work focuses on presence and intensity of pain; other interesting objectives include suppression or amplification of pain as well as pain quality and location. (3) Explore the **connections to treatment and rehabilitation**. A broader view on specific pain problems offers potential for better solutions.

In the following subsections we discuss challenges and promising directions regarding (1) knowledge, (2) data and validation, and (3) algorithms and hardware.

6.1 Knowledge

The development of pain recognition systems would benefit from more knowledge about: (1) the physiology of pain and its measurable responses. Automatic systems can help to gain knowledge, *e.g*. Bartlett *et al*. [75] revealed aspects of pain expression that had been unavailable to observers before. (2) More knowledge about factors influencing pain would help automatic recognition, as it may allow to better leverage context information. (3) Another relevant topic is the interaction of pain with other affective states (such as negative emotions [109], anticipation, and startle) and their impact on pain responses. (4) Dataset quality would benefit from finding better pain elicitation methods and experimental protocols for controlling confounding factors, *e.g*. for reducing the impact of differences in pain sensitivity and expressiveness. (5) Finding more reliable and valid ground truth would reduce label noise and improve learned models; combining multiple types of ground truth (similar to a consent of multiple experts) may be promising. It would be also useful to find more specific and sensitive objective observational measures than PSPI (see discussion in Sec. 4.2), *e.g*. a non-linear combination of AUs, which may also include AUs that do not occur in pain to differentiate it from other states. (6) Better understanding the requirements for clinical practice and adoption by busy medical personnel.

6.2 Data and Validation

A major challenge for advancing pain recognition is the availability of data, which are hard to collect. Hence, datasets should be shared to accelerate progress. An optimal dataset should be multimodal, include high quality annotations, comprise not only pain but also other relevant states to assess specificity and reduce false alarm rate, and should be published with strict evaluation protocols to improve comparability of results. Shared datasets of clinical pain are needed for validating recognition systems in real use cases, *e.g*. with patients in post-operative phases and with dementia. The datasets should also cover aspects that may be relevant in those use cases, *e.g*. social interaction or long-term consistency. Chu *et al*. [77], [78] evaluated pain recognition on seven consecutive days showing that long-term consistency is challenging due to variations in the physiological signals that are not related to pain. Another related issue is generalization and overfitting to datasets. Few works to date [38], [39] have validated methods on two datasets that are significantly distinct (BioVid and UNBC-McMaster). Future work should be validated on multiple datasets to show consistent performance across diverse data and how well a system generalizes to other conditions, medical populations, pain types *etc*. Another way to strengthen the significance of results is to evaluate a system with multiple types of ground truth [95]. Further, ready-to-use recognition systems need to be validated in independent clinical studies to convincingly demonstrate their clinical utility.

6.3 Algorithms and Hardware

Despite the recent advances in pain recognition, there is a lot of room for improving performance in non-laboratory, clinical settings. Combining multiple modalities, while adding complexity, may improve sensitivity and specificity. Improving recognition within each single modality is valuable as well, since the multimodal performance typically benefits from better input modalities. One of the modalities, the audio channel, may be underrated. Aside from infant cry analysis, few have investigated the use of audio for recognizing pain. However, recent work by Tsai *et al*. [97], [98] and Li *et al*. [84] achieved promising results that should encourage more work on this modality, which also plays an important role in related applications of affective computing [186]. In general, better algorithms and/or hardware are needed to cope with several challenges: (1) Inter-individual differences often account for more variation than the signal of interest. Person-specific calibration can help, but most approaches have used either labeled data or the full sample distribution of the test subject, which are unavailable in most real use cases. In a clinical context, a baseline-only calibration seems most realistic, while group-specific models may be an alternative; (2) Measuring low intensity pain, which only yields low amplitude responses, is underexplored; (3) Coping with interfering factors and artifacts. *E.g.*, in camera-based systems, these include lighting changes, occlusions and motion, as well as poor views of the face; Contact sensors suffer from movement artifacts and potential loss of contact; (4) Learning despite small datasets or lack of fine-grained labels. In the reviewed work we found promising approaches for transfer learning and for exploiting both ordinal relations and weak labels; (5) Developing online processing, while many of the published approaches assume that long time series and/or the test person's full sample distribution are available, which is generally not the case; (6) Pain often occurs with emotions and a technology is required to identify blends of facial expressions [109] and to assess pain and accompanying emotional states simultaneously.

7 Conclusions

Research on automatic methods supporting pain assessment has yielded many promising ideas and successful approaches, *e.g.* concerning (1) characteristics of modalities and multi-modal systems, (2) learning from weak and ordinal ground truth and from few data, and (3) personalizing models and using temporal context. Despite the significant progress, in order to impact clinical practice, more effort is needed in advancing knowledge and technology, in gathering the necessary data, and in improving and demonstrating the usefulness of recognition systems in real use cases.

Acknowledgments

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

References

[1] A. Ashraf *et al.*, "The painful face - Pain expression recognition using active appearance models," *ACM Int. Conf. Multimodal Interfaces*, 2007.

[2] A. B. Ashraf *et al.*, "The Painful Face - Pain Expression Recognition Using Active Appearance Models," *Image Vis. Comput.*, vol. 27, no. 12, 2009.

[3] J. Chen *et al.*, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognit. Letters*, vol. 34, no. 15, 2013.

[4] ——, "A new framework with multiple tasks for detecting and locating pain events in video," *Comput. Vis. Image Understanding*, vol. 155, 2017.

[5] J. Egede *et al.*, "Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation," in *IEEE Conf. Autom. Face Gesture Recognit.*, 2017.

[6] C. Florea *et al.*, "Learning Pain from Emotion: Transferred HoT Data Representation for Pain Intensity Estimation," in *Eur. Conf. Comput. Vis. Workshops*, 2014.

[7] ——, "Pain intensity estimation by a self-taught selection of histograms of topographical features," *Image Vis. Comput.*, vol. 56, 2016.

[8] A. Ghasemi *et al.*, "Social signal processing for pain monitoring using a hidden conditional random field," in *IEEE Workshop on Statistical Signal Processing*, 2014.

[9] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in *ACM Int. Conf. Multimodal interaction*, 2012.

[10] X. Hong *et al.*, "Capturing correlations of local features for image representation," *Neurocomputing*, vol. 184, 2016.

[11] R. Irani *et al.*, "Pain Recognit. using Spatiotemporal Oriented Energy of Facial Muscles," in *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015.

[12] S. Kaltwang *et al.*, "Continuous pain intensity estimation from facial expressions," in *Advances in Visual Computing*, ser. LNCS. Springer, 2012, vol. 7432.

[13] ——, "Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 38, no. 9, 2016.

[14] R. A. Khan *et al.*, "Pain detection through shape and appearance features," in *IEEE Int. Conf. Multimedia and Expo*, 2013.

[15] R. Kharghanian *et al.*, "Pain detection from facial images using unsupervised feature learning approach," *Int. Conf. IEEE Engineering in Medicine and Biology Society*, 2016.

[16] D. Lopez-Martinez *et al.*, "Personalized Automatic Estimation of Self-Reported Pain Intensity from Facial Expressions," in *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017.

[17] L. L. Presti *et al.*, "Using Hankel Matrices for Dynamics-based Facial Emotion Recognit. and Pain Detection," in *Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2015.

[18] L. Lo Presti and M. La Cascia, "Boosting Hankel matrices for face emotion recognition and pain detection," *Comput. Vis. Image Understanding*, vol. 156, 2017.

[19] P. Lucey *et al.*, "Improving Pain Recognit. Through Better Utilisation of Temporal Information." in *Int. Conf. Auditory-Visual Speech Processing*, 2008.

[20] ——, "Automatically detecting pain in video through facial action units," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 3, 2011.

[21] ——, "Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database," *Image Vis. Comput.*, vol. 30, no. 3, 2012.

[22] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Trans. Cybernetics*, vol. 44, no. 3, 2014.

[23] N. Neshov and A. Manolova, "Pain detection from facial characteristics using Supervised Descent Method," in *Int. Conf. Intell. Data Acquisition Advanced Comput. Systems*, 2015.

[24] H. Pedersen, "Learning Appearance Features for Pain Detection Using the UNBC-McMaster Shoulder Pain Expression Archive Database," in *Int. Conf. Comput. Vis. Systems*. Springer, 2015.

[25] N. Rathee and D. Ganotra, "A novel approach for pain intensity detection based on facial feature deformations," *Journal of Visual Communication and Image Representation*, vol. 33, 2015.

[26] ——, "Multiview Distance Metric Learning on facial feature descriptors for automatic pain intensity detection," *Comput. Vis. Image Understanding*, vol. 147, 2016.

[27] P. Rodriguez *et al.*, "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," *IEEE Trans. Cybernetics*, 2017.

[28] B. Romera-Paredes *et al.*, "Transfer learning to account for idiosyncrasy in face and body expressions," in *IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2013.

[29] S. D. Roy *et al.*, "An Approach for Automatic Pain Detection through Facial Expression," in *Int. Conf. Intell. Human Comput. Interaction*. Elsevier, 2015.

[30] O. Rudovic *et al.*, "Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields," in *Advances in Visual Computing*, ser. LNCS, vol. 8034. Springer, 2013.

[31] ——, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 37, no. 5, 2015.

[32] A. Ruiz *et al.*, "Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization," *British Machine Vis. Conf.*, 2014.

[33] ——, "Multi-Instance Dynamic Ordinal Random Fields for Weakly-Supervised Pain Intensity Estimation," in *Asian Conf. Comput. Vis.*, ser. LNCS. Springer, 2016.

[34] M. Rupenga and H. B. Vadapalli, "Automatic spontaneous pain recognition using supervised classification learning algorithms," in *Pattern Recognit. Assoc. South Africa Int. Conf.*, 2016.

[35] E. Sangineto *et al.*, "We are not All Equal: Personalizing Models for Facial Expression Analysis with Transductive Parameter Transfer," in *ACM Int. Conf. Multimedia*, 2014.

[36] K. Sikka *et al.*, "Classification and weakly supervised pain localization using multiple segment representation," *Image Vis. Comput.*, vol. 32, no. 10, 2014.

[37] F. Wang *et al.*, "Regularizing face verification nets for pain intensity regression," in *Int. Conf. Image Processing*. IEEE, 2017.

[38] P. Werner *et al.*, "Automatic Pain Assessment with Facial Activity Descriptors," *IEEE Trans. Affective Computing*, vol. 8, no. 3, 2017.

[39] R. Yang *et al.*, "On pain assessment from facial videos using spatio-temporal local descriptors," in *Int. Conf. Image Processing Theory, Tools and Applications*. IEEE, 2016.

[40] Z. Zafar and N. A. Khan, "Pain Intensity Evaluation through Facial Action Units," in *Int. Conf. Pattern Recognit.* IEEE, 2014.

[41] G. Zen *et al.*, "Unsupervised Domain Adaptation for Personalized Facial Emotion Recognition," in *ACM Int. Conf. Multimodal Interaction*, 2014.

[42] R. Zhao *et al.*, "Facial Expression Intensity Estimation Using Ordinal Information," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[43] J. Zhou *et al.*, "Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video," *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016.

[44] M. Amirian *et al.*, "Using Radial Basis Function Neural Networks for Continuous and Discrete Pain Estimation from Bio-physiological Signals," in *Artificial Neural Networks in Pattern Recognition*, 2016.

[45] S. Gruss *et al.*, "Pain intensity recognition rates via biopotential feature patterns with support vector machines," *PLoS ONE*, vol. 10, no. 10, 2015.

[46] M. Kächele *et al.*, "Bio-visual fusion for person-independent recognition of pain intensity," in *Multiple Classifier Systems*, 2015.

[47] ——, "Multimodal data fusion for person-independent, continuous estimation of pain intensity," in *Engineering Applications of Neural Networks*, 2015.

[48] ——, "Methods for Person-Centered Continuous Pain Intensity Assessment From Bio-Physiological Channels," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, 2016.

[49] ——, "Adaptive confidence learning for the personalization of pain intensity estimation systems," *Evolving Systems*, vol. 8, no. 1, 2017.

[50] D. Lopez-Martinez and R. Picard, "Multi-task Neural Networks for Personalized Pain Recognition from Physiological Signals," in *Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2017.

[51] D. Lopez-Martinez *et al.*, "Physiological and Behavioral Profiling for Nociceptive Pain Estimation Using Personalized Multitask Learning." in *NIPS Workshop on Machine Learning for Health*, 2017.

[52] D. Lopez-Martinez and R. Picard, "Skin conductance deconvolution for pain estimation," in *IEEE Conf. Biomedical and Health Informatics*, 2018.

[53] ——, "Continuous Pain Intensity Estimation from Autonomic Signals with Recurrent Neural Networks," in *Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018.

[54] S. Walter *et al.*, "Automatic pain quantification using autonomic parameters," *Psychology and Neuroscience*, vol. 7, no. 3, 2014.

[55] ——, "Data fusion for automated pain recognition," in *Int. Conf. Pervasive Computing Technologies for Healthcare*, 2015.

[56] P. Werner *et al.*, "Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges," in *British Machine Vis. Conf.*, 2013.

[57] ——, "Automatic pain recognition from video and biomedical signals," in *Int. Conf. Pattern Recognit.*, 2014.

[58] ——, "Analysis of facial expressiveness during experimentally induced heat pain," in *Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2017.

[59] V. Kessler *et al.*, "Multimodal fusion including camera photoplethysmography for pain recognition," *Int. Conf. Companion Technology*, 2017.

[60] P. Thiam *et al.*, "Audio-Visual Recognit. of Pain Intensity," in *Multimodal Pattern Recognit. of Social Signals in Human-Comput.-Interaction Workshop*, 2016.

[61] P. Thiam and F. Schwenker, "Multi-modal data fusion for pain intensity assessment and classification," in *Int. Conf. Image Processing Theory, Tools and Applications*. IEEE, 2017.

[62] P. Thiam *et al.*, "Hierarchical Combination of Video Features for Personalised Pain Level Recognit." in *Eur. Symp. Artif. Neural Networks, Computational Intell. and Machine Learning*, 2017.

[63] M. S. Aung *et al.*, "The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset," *IEEE Trans. Affective Computing*, vol. 7, no. 4, 2016.

[64] T. A. Olugbade *et al.*, "Bi-Modal Detection of Painful Reaching for Chronic Pain Rehabilitation Systems," in *ACM Int. Conf. Multimodal Interaction*, 2014.

[65] ——, "Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain," in *Int. Conf. Affective Comput. Intell. Interaction*. IEEE, 2015.

[66] S. Brahnam *et al.*, "Machine assessment of neonatal facial expressions of acute pain," *Decision Support Systems*, vol. 43, no. 4, 2007.

[67] C.-Y. Chang and J.-J. Li, "Application of deep learning for recognizing infant cries," in *IEEE Int. Conf. Consumer Electronics-Taiwan*, 2016.

[68] P. Pal *et al.*, "Emotion Detection From Infant Facial Expressions And Cries," in *IEEE Int. Conf. Acoustics Speed and Signal Processing*, vol. 2, 2006.

[69] M. Petroni *et al.*, "Identification of pain from infant cry vocalizations using artificial neural networks (ANNs)," in *Proc. SPIE 2492, Applications and Science of Artificial Neural Networks*, 1995.

[70] A. Rosales-Pérez *et al.*, "Classifying infant cry patterns by the Genetic Selection of a Fuzzy Model," *Biomedical Signal Processing and Control*, vol. 17, 2015.

[71] H. B. Sailor and H. Patil, "Auditory Filterbank Learning Using ConvRBM for Infant Cry Classification," in *Interspeech*. ISCA, 2018.

[72] G. Zamzmi *et al.*, "Automated Pain Assessment in Neonates," in *Scandinavian Conf. Image Analysis*. Springer, 2017.

[73] M. Adibuzzaman *et al.*, "Assessment of Pain Using Facial Pictures Taken with a Smartphone," in *IEEE Annual Comput. Software and Applications Conf.*, 2015.

[74] S. Ashouri *et al.*, "A novel approach to spinal 3-D kinematic assessment using inertial sensors: Towards effective quantitative evaluation of low back pain in clinical settings," *Computers in Biology and Medicine*, vol. 89, 2017.

[75] M. S. Bartlett *et al.*, "Automatic decoding of facial movements reveals deceptive pain expressions," *Current Biology*, vol. 24, no. 7, 2014.

[76] Z. Chen *et al.*, "Automated detection of pain from facial expressions: a rule-based approach using AAM," in *SPIE 8314, Medical Imaging 2012: Image Processing*, 2012.

[77] Y. Chu *et al.*, "Physiological Signals Based Quantitative Evaluation Method of the Pain," *IFAC Proc. Volumes*, vol. 47, no. 3, 2014.

[78] ——, "Physiological Signal-Based Method for Measurement of Pain Intensity," *Frontiers in Neuroscience*, vol. 11, no. 279, 2017.

[79] Z. Hammal and M. Kunz, "Pain monitoring: A dynamic and context-sensitive system," *Pattern Recognit.*, vol. 45, no. 4, 2012.

[80] M. A. Haque *et al.*, "Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities," in *IEEE Int. Conf. Automatic Face & Gesture Recognit.*, 2018.

[81] C.-C. Hung *et al.*, "Using surface electromyography (SEMG) to classify low back pain based on lifting capacity evaluation with principal component analysis neural network method," in *Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2014.

[82] R. Irani *et al.*, "Spatiotemporal Analysis of RGB-D-T Facial Images for Multi-Modal Pain Level Recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015.

[83] M. Jiang *et al.*, "Acute pain intensity monitoring with the classification of multiple physiological parameters," *Journal of Clinical Monitoring and Computing*, vol. 33, no. 3, 2019.

[84] J.-L. Li *et al.*, "Learning Conditional Acoustic Latent Representation with Gender and Age Attributes for Automatic Pain Level Recognition," in *Interspeech*, 2018.

[85] P. Liu *et al.*, "Clinical valid pain database with biomarker and visual information for pain level analysis," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2018.

[86] G. C. Littlewort *et al.*, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image Vis. Comput.*, vol. 27, no. 12, 2009.

[87] D. Lopez-Martinez *et al.*, "Multi-task multiple kernel machines for personalized pain recognition from functional near-infrared spectroscopy brain signals," in *Int. Conf. Pattern Recognit.*, 2018.

[88] G. Misra *et al.*, "Automated classification of pain perception using high-density electroencephalography data," *Journal of Neurophysiology*, vol. 117, no. 2, 2017.

[89] M. M. Monwar and S. Rezaei, "Support vector machine for automatic pain recognition," in *Computational Imaging VII*, 2009.

[90] R. Niese *et al.*, "Towards Pain Recognit. in Post-Operative Phases Using 3D-based Features From Video and Support Vector Machines," *Int. J. Digital Content Techn. Applic.*, vol. 3, no. 4, 2009.

[91] Y. Oshrat *et al.*, "Speech prosody as a biosignal for physical pain detection," in *Speech Prosody*, 2016.

[92] P. Panavaranan and Y. Wongsawat, "EEG-based pain estimation via fuzzy logic and polynomial kernel support vector machine," in *Biomedical Engineering Int. Conf.* IEEE, 2013.

[93] A. Pourshoghi *et al.*, "Application of functional data analysis in classification and clustering of functional near-infrared spectroscopy signal in response to noxious stimuli," *Journal of Biomedical Optics*, vol. 21, no. 10, 2016.

[94] J. J. Rivas *et al.*, "Automatic Recognition of Pain, Anxiety, Engagement and Tiredness for Virtual Rehabilitation from Stroke: A Marginalization Approach," in *Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2017.

[95] K. Sikka *et al.*, "Automated Assessment of Children's Postoperative Pain Using Computer Vision," *Pediatrics*, vol. 136, no. 1, 2015.

[96] B. T. Susam *et al.*, "Automated Pain Assessment using Electrodermal Activity Data and Machine Learning," in *Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018.

[97] F.-S. Tsai *et al.*, "Toward Development and Evaluation of Pain Level-Rating Scale for Emergency Triage based on Vocal Characteristics and Facial Expressions," in *Interspeech*, 2016.

[98] ——, "Embedding stacked bottleneck vocal features in a LSTM architecture for automatic pain level classification during emergency triage," in *Int. Conf. Affective Comput. Intell. Interaction*. IEEE, 2017.

[99] M. Vatankhah *et al.*, "Perceptual pain classification using ANFIS adapted RBF kernel support vector machine for therapeutic usage," *Applied Soft Computing*, vol. 13, no. 5, 2013.

[100] M. Vatankhah and A. Toliyat, "Pain Level Measurement Using Discrete Wavelet Transform," *Int. Journal of Engineering and Technology*, vol. 8, no. 5, 2016.

[101] T. D. Wager *et al.*, "An fMRI-Based Neurologic Signature of Physical Pain," *New England J. Medicine*, vol. 368, no. 15, 2013.

[102] S. Wang *et al.*, "When you can't tell when it hurts: a preliminary algorithm to assess pain in patients who can't communicate." in *AMIA Annual Symp. Proc.*, 2013.

[103] P. Werner *et al.*, "Pain recognition and intensity rating based on Comparative Learning," in *Int. Conf. Image Processing*, 2012.

[104] ——, "Comparative learning applied to intensity rating of facial expressions of pain," *Int. Journal of Pattern Recognit. & Artificial Intelligence*, vol. 28, no. 5, 2014.

[105] L. Yang *et al.*, "PATTERN: Pain Assessment for paTients who can't TEll using Restricted Boltzmann machiNe," *BMC Medical Informatics & Decision Making*, vol. 16, no. S. 3, 2016.

[106] H. Merskey *et al.*, "Pain terms: a list with definitions and notes on usage," *PAIN*, vol. 6, no. 3, 1979.

[107] A. C. Williams and K. D. Craig, "Updating the definition of pain," *PAIN*, vol. 157, no. 11, 2016.

[108] M. Aydede, "Defending the IASP Definition of Pain," *The Monist*, vol. 100, no. 4, 2017.

[109] A. C. Williams, "Facial expression of pain: an evolutionary account." *The Behavioral and brain sciences*, vol. 25, no. 4, 2002.

[110] P. Mäntyselkä *et al.*, "Pain as a reason to visit the doctor: a study in Finnish primary health care." *Pain*, vol. 89, no. 2-3, 2001.

[111] W. H. Cordell *et al.*, "The high prevalence of pain in emergency medical care," *The American Journal of Emergency Medicine*, vol. 20, no. 3, 2002.

[112] J. Gregory and L. Mcgowan, "An examination of the prevalence of acute pain for hospitalised adult patients: A systematic review," *Journal of Clinical Nursing*, vol. 25, no. 5-6, 2016.

[113] S. Zoëga *et al.*, "Quality Pain Management in the Hospital Setting from the Patient's Perspective," *Pain Practice*, vol. 15, no. 3, 2015.

[114] M. E. Lynch, "The need for a Canadian pain strategy." *Pain research & management*, vol. 16, no. 2, 2011.

[115] D. C. Turk and R. Melzack, "The Measurement of Pain and the Assessment of People Experiencing Pain," in *Handbook of Pain Assessment*. Guilford Press, 2011.

[116] A. Mitchell and B. J. Boss, "Adverse effects of pain on the nervous systems of newborns and young children: a review of the literature." *Journal of Neuroscience Nursing*, vol. 34, no. 5, 2002.

[117] F. Tennant, "The Physiologic Effects of Pain on the Endocrine System," *Pain and Therapy*, vol. 2, no. 2, 2013.

[118] H. McQuay *et al.*, "Treating acute pain in hospital." *BMJ: British Medical Journal*, vol. 314, no. 7093, 1997.

[119] K. D. Craig, "The social communication model of pain." *Canadian Psychology*, vol. 50, no. 1, 2009.

[120] S. L. Beck *et al.*, "Dissemination and Implementation of Patient-centered Indicators of Pain Care Quality and Outcomes," *Medical Care*, 2018.

[121] W. Meissner *et al.*, "Management of acute pain in the postoperative setting: the importance of quality indicators," *Current Medical Research and Opinion*, vol. 34, no. 1, 2018.

[122] K. Herr *et al.*, "Pain assessment in the patient unable to self-report: position statement with clinical practice recommendations." *Pain management nursing*, vol. 12, no. 4, 2011.

[123] K. D. Craig, "The facial expression of pain: Better than a thousand words?" *APS Journal*, vol. 1, no. 3, 1992.

[124] G. Zamzmi *et al.*, "A Review of Automated Pain Assessment in Infants: Features, Classification Tasks, and Databases," *IEEE Reviews in Biomedical Engineering*, 2017.

[125] R. Melzack and K. Casey, "Sensory, motivational and central control determinants of pain: a new conceptual model," in *The skin senses*, D. Kenshalo, Ed., Springfield, IL, 1968.

[126] K. M. Prkachin and K. D. Craig, "Expressing pain: The communication and interpretation of facial pain signals," *Journal of Nonverbal Behavior*, vol. 19, no. 4, 1995.

[127] P. Yancey and P. Brand, "The gift of pain," *Grand Rapids, MI: Zondervan*, 1997.

[128] W. D. Willis and K. N. Westlund, "Neuroanatomy of the Pain System and of the Pathways That Modulate Pain," *Journal of Clinical Neurophysiology*, vol. 14, no. 1, 1997.

[129] S. Khalid and R. S. Tubbs, "Neuroanatomy and Neuropsychology of Pain." *Cureus*, vol. 9, no. 10, 2017.

[130] A. Schnitzler and M. Ploner, "Neurophysiology and functional neuroanatomy of pain perception." *Journal of Clinical Neurophysiology*, vol. 17, no. 6, 2000.

[131] E. E. Benarroch, "Pain-autonomic interactions: A selective review," *Clinical Autonomic Research*, vol. 11, no. 6, 2001.

[132] G. Stewart and A. Panickar, "Role of the sympathetic nervous system in pain," *Anaesthesia and Intensive Care Medicine*, vol. 14, no. 12, pp. 524–527, 2013.

[133] W. Boucsein, *Electrodermal Activity*. Springer Science & Business Media, 2012.

[134] M. Benedek and C. Kaernbach, "Decomposition of skin conductance data by means of nonnegative deconvolution," *Psychophysiology*, vol. 47, pp. 647–658, 2010.

[135] A. J. Terkelsen *et al.*, "Acute pain increases heart rate: Differential mechanisms during rest and mental stress," *Autonomic Neuroscience: Basic and Clinical*, vol. 121, no. 1-2, 2005.

[136] B. M. Appelhans and L. J. Luecken, "Heart rate variability and pain: Associations of two interrelated homeostatic processes," *Biological Psychology*, vol. 77, no. 2, pp. 174–182, 2008.

[137] M. Saccò *et al.*, "The Relationship Between Blood Pressure and Pain," *The Journal of Clinical Hypertension*, vol. 15, no. 8, 2013.

[138] C. R. Chapman *et al.*, "Phasic pupil dilation response to noxious stimulation in normal volunteers: Relationship to brain evoked potentials and pain report," *Psychophysiology*, vol. 36, no. 1, 1999.

[139] E. Szabadi, "Modulation of physiological reflexes by pain: role of the locus coeruleus," *Frontiers in Integrative Neuroscience*, vol. 6, no. October, pp. 1–15, 2012.

[140] R. D. Treede, "Transduction and transmission properties of primary nociceptive afferents." *Rossiiskii fiziologicheskii zhurnal imeni I.M. Sechenova*, vol. 85, no. 1, pp. 205–11, 1999.

[141] E. S. d. S. Pinheiro *et al.*, "Electroencephalographic Patterns in Chronic Pain: A Systematic Review of the Literature," *PLOS ONE*, vol. 11, no. 2, p. e0149085, 20 16.

[142] B. D. Kussman *et al.*, "Capturing pain in the cortex during general anesthesia: Near infrared spectroscopy measures in patients

undergoing catheter ablation of arrhythmias," *PLoS ONE*, vol. 11, no. 7, pp. 1–13, 2016.

[143] K. D. Craig *et al.*, "The facial expression of pain," in *Handbook of Pain Assessment*. Guilford Press, 2011.

[144] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, 2008.

[145] K. M. Prkachin, "The consistency of facial expressions of pain: a comparison across modalities," *Pain*, vol. 51, no. 3, 1992.

[146] M. Kunz *et al.*, "The facial expression of pain in patients with dementia," *PAIN*, vol. 133, no. 1, 2007.

[147] P. Ekman *et al.*, *Facial Action Coding System: The Manual on CD ROM*. A Human Face, 2002.

[148] D. Simon *et al.*, "Recognition and discrimination of prototypical dynamic expressions of pain and emotions," *Pain*, vol. 135, no. 1, 2008.

[149] J. Walsh *et al.*, "Pain communication through body posture: The development and validation of a stimulus set," *PAIN*, vol. 155, no. 11, 2014.

[150] P. Werner *et al.*, "Head movements and postures as pain behavior," *PLOS ONE*, vol. 13, no. 2, 2018.

[151] V. Warden *et al.*, "Development and psychometric evaluation of the Pain Assessment in Advanced Dementia (PAINAD) scale," *J. American Medical Directors Association*, vol. 4, no. 1, 2003.

[152] R. V. Grunau and K. D. Craig, "Pain expression in neonates: facial action and cry," *Pain*, vol. 28, no. 3, 1987.

[153] C. Johnston and M. E. Strada, "Acute pain response in infants: a multidimensional description," *Pain*, vol. 24, no. 3, 1986.

[154] B. Stevens *et al.*, "Premature Infant Pain Profile: Development and Initial Validation," *The Clinical Journal of Pain*, vol. 12, no. 1, 1996.

[155] M. Stites, "Observational pain scales in critically ill adults," *Critical care nurse*, vol. 33, no. 3, 2013.

[156] S. M. Zwakhalen *et al.*, "Pain in elderly people with severe dementia: a systematic review of behavioural pain assessment tools," *BMC Geriatrics*, vol. 6, no. 3, 2006.

[157] R. J. Gatchel *et al.*, "The biopsychosocial approach to chronic pain: scientific advances and future directions." *Psychological bulletin*, vol. 133, no. 4, 2007.

[158] M. A. Lumley *et al.*, "Pain and emotion: a biopsychosocial review of recent research." *Journal of clinical psychology*, vol. 67, no. 9, 2011.

[159] Registered Nurses' Association of Ontario, "Practice Recommendations," in *Assessment and Management of Pain*, 3rd ed. Toronto: Registered Nurses' Association of Ontario, 2013, p. 21.

[160] P. Nilges, "Klinische Schmerzmessung," in *Praktische Schmerzmedizin*, R. Baron *et al.*, Eds. Springer, 2013.

[161] Registered Nurses' Association of Ontario, *Assessment and Management of Pain*, 3rd ed. Toronto: Registered Nurses' Association of Ontario, 2013.

[162] A. E. Olesen *et al.*, "Human Experimental Pain Models for Assessing the Therapeutic Efficacy of Analgesic Drugs," *Pharmacological Reviews*, vol. 64, no. 3, 2012.

[163] X. Zhang *et al.*, "BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, 2014.

[164] K. A. Puntillo *et al.*, "Pain behaviors observed during six common procedures: Results from Thunder Project II," *Critical Care Medicine*, vol. 32, no. 2, 2004.

[165] P. Lucey *et al.*, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *IEEE Int. Conf. Automatic Face Gesture Recognit. Workshops*, 2011.

[166] M. Kunz and S. Lautenbacher, "The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain," *European Journal of Pain*, vol. 18, no. 6, 2014.

[167] M. Kunz *et al.*, "On the relationship between self-report and facial expression of pain," *Journal of Pain*, vol. 5, no. 7, 2004.

[168] S. Walter *et al.*, "The biovid heat pain database: Data for the advancement and systematic validation of an automated pain recognition," in *IEEE Int. Conf. Cybernetics*, 2013.

[169] L. Zhang *et al.*, "BioVid Emo DB: A multimodal database for emotion analyses validated by subjective ratings," in *IEEE Symp. Series Comput. Intell.*, 2016.

[170] Z. Zhang *et al.*, "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[171] S. Brahnam *et al.*, "SVM Classification of Neonatal Facial Images of Pain," in *Int. Workshop on Fuzzy Logic and Applications*, I. Bloch *et al.*, Eds. Springer, 2006, vol. LNCS 3849.

[172] D. Harrison *et al.*, "Too many crying babies: a systematic review of pain management practices during immunizations on YouTube," *BMC Pediatrics*, vol. 14, no. 134, 2014.

[173] V. K. Mittal, "Discriminating the Infant Cry Sounds Due to Pain vs. Discomfort Towards Assisted Clinical Diagnosis," in *Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. ISCA, 2016.

[174] M. Velana *et al.*, "The SenseEmotion Database: A Multimodal Database for the Development and Systematic Validation of an Automatic Pain- and Emotion-Recognition System," in *IAPR Workshop on Multimodal Pattern Recognit. of Social Signals in Human-Comput. Interaction*. Springer, 2016.

[175] S. Gruss *et al.*, "Multi-Modal Signals for Analyzing Pain Responses to Thermal and Electrical Stimuli," *Journal of Visualized Experiments*, vol. 146, 2019.

[176] L. A. Jeni *et al.*, "Facing Imbalanced Data: Recommendations for the Use of Performance Metrics," in *Int. Conf. Affective Comput. Intell. Interaction*. IEEE, 2013.

[177] P. Werner *et al.*, "Handling Data Imbalance in Automatic Facial Action Intensity Estimation," in *British Machine Vis. Conf.*, 2015.

[178] J. Kranjec *et al.*, "Non-contact heart rate and heart rate variability measurements: A review," *Biomedical Signal Processing and Control*, vol. 13, 2014.

[179] P. Werner *et al.*, "Automatic heart rate estimation from painful faces," in *Int. Conf. Image Processing*, 2014.

[180] J. Hernandez *et al.*, "Biowatch: Estimation of heart and breathing rates from wrist motions," in *Int. Conf. Pervasive Computing Technologies for Healthcare*, 2015.

[181] K. Limbrecht-Ecklundt *et al.*, "Mimische Aktivität differenzierter Schmerzintensitäten: Korrelation der Merkmale von Facial Action Coding System und Elektromyographie," *Schmerz*, vol. 30, no. 3, 2016.

[182] R. W. Picard *et al.*, "Multiple arousal theory and daily-life electrodermal activity asymmetry," *Emotion Review*, vol. 8, no. 1, pp. 62–75, 2016.

[183] C. A. Corneanu *et al.*, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognit.: History, Trends, and Affect-Related Applications," *IEEE Trans. Pattern Analysis Machine Intell.*, vol. 38, no. 8, 2016.

[184] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affective Computing*, vol. 7, no. 2, 2016.

[185] B. F. Klare *et al.*, "Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[186] S. Poria *et al.*, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, 2017.

Philipp Werner is a research team leader and PhD candidate at the Neuro-Information Technology Group of the Otto-von-Guericke University Magdeburg, Germany. His research focuses on pain recognition, facial expression recognition, human behavior recognition, computer vision, pattern recognition, and deep learning. See http://philipp-werner.info for more details.

Daniel Lopez-Martinez is a PhD candidate in Medical Engineering and Medical Physics at the Harvard-MIT Program in Health Sciences and Technology. See www.daniellopezmartinez.com for more details.

Steffen Walter is a researcher in the Medical Psychology at the University of Ulm, Germany. He received his PhD in human biology from the University of Ulm in 2008. His research focuses on automated pain recognition, affective computing, and psychotherapy processes.

Ayoub Al-Hamadi is an adjunct professor and head of the Neuro-Information Technology Group in the Otto-von-Guericke-University Magdeburg, Gemany. He received his PhD in Technical Computer Science in 2001 and the Habilitation in Artificial Intelligence and the Venia Legendi in Pattern Recognition and Image Processing in 2010, both from Otto-von-Guericke-University Magdeburg. Al-Hamadi is the author of more than 300 articles in peer-reviewed international journals, conferences, and books. His research focuses on computer vision, pattern recognition, and image processing. See http://www.iikt.ovgu.de/al_hamadi.html for more details.

Sascha Gruss received his PhD from the University of Ulm, Germany in 2015. Currently he is a research team leader in automatic pain recognition projects. His main research interests include pattern recognition, bio signal processing, machine learning, and companion technology.

Rosalind W. Picard, ScD, FIEEE, is Professor of Media Arts and Sciences at MIT, founder and director of the Affective Computing Research Group at the MIT Media Lab, chief scientist, chairman, and co-founder of Empatica Inc, and co-founder of Affectiva, Inc. She has a Bachelor in Electrical Engineering from the Georgia Institute of Techology and an SM and ScD in Electrical Engineering and Computer Science from MIT. Picard is author of the 1997 book Affective Computing, and author or co-author of over three hundred scientific articles. She is a fellow of the IEEE and of the AAAC. In 2019, she became a member of the National Academy of Engineering, recognized for contributions to wearable computing and affective computing.