

FACS-Based Automated Pain Detection From Spontaneous Facial Expressions

BY

ZHANLI CHEN

B.S., Nanjing University of Posts and Telecommunications, 2006

M.S., Hong Kong University of Science and Technology, 2008

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:

Rashid Ansari, Chair and Advisor

Diana J. Wilkie, Co-advisor, University of Florida

Milos Zefran

Enis A. Cetin

Mojtaba Soltanalian

Copyright by

Zhanli Chen

2020

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisors Dr. Rashid Ansari and Dr. Diana J. Wilkie for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Cetin, Dr. Zefran, and Dr. Soltanalian, for their insightful comments and encouragement, but also for the hard question which incited me to widen my research from various perspectives.

I also appreciate my co-author Guglielmo Menchetti for his contribution on the two-stage deep learning implementation and insightful suggestion to this project.

I thank my fellow labmates, Ouday and Lubna, you are like my brother and sister. I shall memorize all funs we had in the past years.

Last but not the least, I would like to thank my family: my parents, my wife and our two puppy Yorkies for supporting me spiritually throughout writing this thesis and my life in general.

ZC

PREFACE

This dissertation is an original intellectual product of the author, Zhanli Chen. All of the work presented here was conducted in the Multimedia Communications Lab at the University of Illinois at Chicago.

The results of our research have been previously published (or will be published) as an article in the IEEE Transactions on Affective Computing (Chen et al., 2019) and several conference and symposium publications: IEEE Global Conference on Signal and Information Processing (GlobalSIP'19) (Chen et al., 2019), Medical Imaging 2012: Image Processing (SPIE 2012) (Chen et al., 2012). The copyright permissions for reusing the published materials are given in Appendix A. This dissertation also includes contents from a preprint uploaded to arXiv (Chen et al., 2018) as well as a manuscript to be submitted to an IEEE conference.

Zhanli Chen
Thursday 27th February, 2020

CONTRIBUTION OF AUTHORS

A version of Chapter 2 has been submitted to arXiv (Chen et al., 2018). I was responsible to conduct the literature survey and completed the manuscript. My advisors, Dr. Rashid Ansari and Dr. Diana J. Wilkie, helped proofread the manuscript and suggest modifications where necessary.

A version of Chapters 3 and 4 has been published in IEEE Transactions on Affective Computing (Chen et al., 2019). I was responsible for building the ideas, developing algorithms and writing the manuscript. Dr. Ansari provided advising on this research and proofread the entire manuscripts. One of the datasets involved in this research is developed by Dr. Wilkie and the pain evaluation metric is also from one of her early publications (Wilkie, 1995).

A version of Chapters 3 and 4 has been published in 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP 2019) (Chen et al., 2019). I was responsible for building and developing the majority of the ideas, and completed the most part of the manuscript. Guglielmo Menchetti developed the source code for the proposed system and help composite the implementation section of this paper. Dr. Enis A Cetin and Dr. Ansari was involved in manuscript composition. Dr. Wilkie proofread the manuscript and provided general advising.

TABLE OF CONTENTS

<u>CHAPTER</u>		<u>PAGE</u>
1	INTRODUCTION	1
1.1	Facial Expression and Pain	1
1.2	Pain Metrics based on FACS	3
1.2.1	Facial Action Coding System	3
1.2.2	PSPI Pain Metric	4
1.2.3	Wilkie’s Pain Metric	5
1.3	Motivations and Challenges for Automated Pain Detection	6
1.4	Main Contributions of This Research	8
2	REVIEW OF PROGRESS IN AUTOMATED PAIN DETECTION	11
2.1	Facial Expression Datasets	11
2.1.1	Standard Facial Expression Datasets for Conventional AFER	12
2.1.2	Facial Expression Dataset for Automated Pain Detection	13
2.1.3	Facial Expression Datasets for Deep Learning Based AFER	14
2.1.4	Selected Private Facial Expression Dataset for Pain Study	15
2.2	FACS-Based General AFER Framework	16
2.2.1	Face Detection	16
2.2.2	Face Alignment in 2D	17
2.2.3	Feature extraction	19
2.2.4	AU classification and regression	22
2.2.5	Performance Metrics	28
2.3	Automated Pain Detection from Facial Expressions	30
2.3.1	Genuine Pain Vs. Posed Pain	31
2.3.2	Binary Pain Classification	32
2.3.3	Pain Intensity Estimation	33
2.3.4	Pain Detection in Clinical Settings	35
2.3.5	Detecting Pain Event with ‘Weakly Labelled’ Data	37
2.3.6	Multi-Modality in Pain Detection	39
2.4	Evolution with Deep Learning	42
2.4.1	AFER with Deep Learning	42
2.4.1.1	AU Recognition Model with a Deep Learning Architecture	43
2.4.1.2	Multi-label AU Detection	46
2.4.1.3	AU Detection with Partially Labeled Data	49
2.4.2	Applications of Deep Learning in Pain Detection	51
2.4.3	Problems in Previous Research	55
3	DECOUPLED AUTOMATED PAIN DETECTION FRAMEWORK	57

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
3.1	Research Motivation	57
3.2	The Decoupled Framework and An End-to-Front Research Strategy .	60
3.3	Automated Facial Expression Recognition	62
3.3.1	Conventional CVML Based Emotient	62
3.3.2	Deep Learning Based Hybrid AFER	64
3.3.2.1	Face Tracking and Alignment	66
3.3.2.2	Preprocessing	66
3.3.2.3	Convolutional Neural Network Architecture	67
3.3.2.4	Multi-label AU output	68
3.3.3	Deep Learning Based End-to-end AFER	69
3.3.3.1	Deep Neural Network For Face Detection and Alignment Network .	69
3.3.3.2	Deep Neural Network For Multi-Label AU Recognition	70
3.4	Action Unit Combination Encoding	72
3.4.1	Compact Structure Vs. Clustered Structure	72
3.4.2	Bag of Words representation	74
3.5	Weakly Supervised Automated Pain Detection	77
3.5.1	Multiple Instance Learning	77
3.5.2	Multiple Clustered Instance Learning	81
4	EXPERIMENTAL RESULTS	84
4.1	Frame-level Pain-Related Action Units Prediction on UNBC-McMaster Dataset	84
4.1.1	Dataset Skew	84
4.1.2	AFER Performance Evaluation	85
4.1.3	AU Relations	86
4.2	Sequence-level Pain Detection In UNBC-McMaster Datasets	89
4.2.1	Correlations Between Pain Expression And Self-Report Pain	89
4.2.2	Evaluation Of The MIL Based Pain Detection	89
4.2.3	Multi-label AU Predictions And PSPI	93
4.3	Trans-dataset Validation For Human Coders With Wilkie's Dataset .	94
4.3.1	Pain Expression Evaluation Via AU Combinations	94
4.3.2	Consistency Evaluation Between The Automated System And FACS Coders	96
5	DISCUSSION AND FUTURE WORK	100
6	CONCLUSION	104
	APPENDIX	105
	CITED LITERATURE	108

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>	<u>PAGE</u>
VITA	122

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	ACTION UNIT DEFINITION AND PAIN-RELATED AU COMBINATIONS	6
II	KEY ANNOTATIONS IN THE MIL DERIVATION	78
III	A FER PERFORMANCE ON AUGMENTED UNBC-MCMMASTER . . .	87
IV	COMPARISON OF THE DECOUPLED FRAMEWORK WITH MS-MIL (Sikka et al., 2014)	91
V	VALIDATION WITH SELECTED PATIENTS FROM WILKIE DATASET	95
VI	COMPARISON OF MACHINE PREDICTION WITH HUMAN CODER DECISION ON SUBSEQUENCES IN WILKIE DATASET (Wilkie, 1995)	99

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Correlation between the frameworks used in addressing the AFER and APD problems	31
2	(a) Unified Pain Detection Framework.(b) Proposed Decoupled Pain Detection Framework.	59
3	The Decoupled Pain Detection Framework	61
4	The End-to-Front research strategy	62
5	Sample output from Emotient about (a) a patient from UNBC-McMaster Dataset, (b) a patient from Wilkie’s dataset.	65
6	(a) Original image from CK+ dataset. (e) Original image from UNBC-McMaster shoulder pain dataset. (b), (f) Image with landmarks. (c), (g) Aligned image. (d), (h) Aligned and masked image	67
7	CNN based Automated Facial Expression Recognition Framework	68
8	The Inception-Resnet-v1 architecture schema	71
9	AU combination structure: (a)Compact Structure, (b)Clustered Structure . .	73
10	Multi Label AU Predictions on Sample Frames	86
11	AU relations in UNBC McMaster (a)Ground Truth, (b)Predictions	88
12	Relations between frame-level and sequence-level ground truth on pain . . .	90
13	Frame Level Pain Prediction on Two Sample Video Sequences	93
14	A sample score sheet filled by human coders for pain analysis	97
15	Selected patients for test (a) P21, (b) P24, (c) P27, (d) P41 (e) P44	98
16	OptiTrack IR camera arrays setup	102

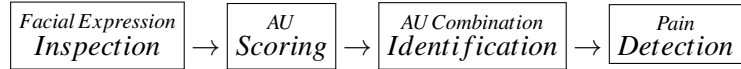
LIST OF FIGURES (Continued)

<u>FIGURE</u>		<u>PAGE</u>
17	(a) SoftImage compatible 33 control points layout in Optitrack. (b) Facial expression replication by Avatar.	103

SUMMARY

Patient pain can be detected highly reliably from facial expressions using a set of facial muscle-based action units (AUs) defined by the Facial Action Coding System (FACS). A key characteristic of facial expression of pain is the simultaneous occurrence of pain-related AU combinations, whose automated detection would be highly beneficial for efficient and practical pain monitoring. Existing general Automated Facial Expression Recognition (AFER) systems prove inadequate when applied specifically for detecting pain as they either focus on detecting individual pain-related AUs but not on combinations or they seek to bypass AU detection by training a binary pain classifier directly on pain intensity data but are limited by lack of enough labeled data for satisfactory training.

Inspired by the clinical demand for automated pain evaluation in end-of-life patient care, our research focuses on seeking an end-to-end automated pain detection approach that mimics the decision-making strategy of FACS-certified human coders by following the procedures below:



As our major contribution of this research, we proposed a new FACS-based automated pain detection framework comprised of two independent machine learning networks to infer the presence of Action Unit (AU) combinations that signify pain. The two networks are trained independently and are linked by novel AU-based data structures that are created by mapping single AU measurements per frame into pain-related low-dimensional feature vectors representing AU combinations. The decoupled architecture of two independent machine learning networks utilizes AU codings as the intermediate re-

SUMMARY (Continued)

sults and performs pain analysis in a low-dimensional AU score space, which would be more justifiable and efficient than the direct method.

The architecture of two independent machine learning networks not only improves data utilization from existing different pain-oriented video datasets, but also improves data fusion incorporating future acquired data, which addresses the most challenging problem arising from the data insufficiency. The decoupled architecture lends itself to flexible customization using different implementation techniques in the modules. The AFER system is realized with three types of configurations, including one conventional CVML system and two types of deep learning architectures. The automated pain detection is modeled as a weakly supervised problem, and the APD system is realized by two multiple instance learning frameworks (MIL and MCIL). AU combination are encoded from single AU scores by two novel data structures (Compact and Clustered), and the multiple instance learning frameworks are trained with low-dimensional features based on the pain-related AU combinations. In particular, we followed an **end-to-front** research strategy to develop the decoupled pain detection system in three research phases, with the ultimate goal of establishing a robust and generic automated pain analysis system for clinical applications.

Experimental results on the UNBC-McMaster Shoulder Pain Expression dataset show that the deep-learning based multi-label AFER system outperforms state-of-the-art AFER system that are based on classical machine learning (ML) techniques. Further tests with the Wilkie video dataset of lung cancer patients suggest the proposed decoupled framework has strong promise for effective pain monitoring in clinical settings, where segment-level patients' self-report pain is the only available ground truth.

CHAPTER 1

INTRODUCTION

Patient pain can be detected highly reliably from facial expressions using a set of facial muscle-based action units (AUs) defined by the Facial Action Coding System (FACS). In this chapter we describe FACS, its use in pain detection, and the relevant cues for pain detection. We also describe the motivation of our work and outline the main contributions of this dissertation.

1.1 Facial Expression and Pain

Pain assessment is essential to providing proper patient care and assessing its efficacy under clinical settings. A patients' self-report is commonly used as the 'gold standard' to report pain through manual pain measurement tools, including the verbal numerical scale (VNS) and visual analogue scale (VAS). However, human sensing and judgement of pain is subjective and the VAS or VNS report may vary significantly among individuals. Behavioral observation of a patient, in particular the use of facial expression, as a key behavioral indicator of pain, has been identified as an important modality for assessing pain (Williams, 2002) , especially when the patient's ability to communicate pain is impaired (HH, 2000). Patients who are dying, intellectually disabled (McGuire et al., 2010), critically ill and sedated (Payen et al., 2001)(Arif-Rahu and Grap, 2010), or have dementia (Manfredi et al., 2003), head and neck cancer, or brain metastasis (Herr et al., 2006)(Hadjistavropoulos et al., 2007) (Puntillo et al., 2004) are particularly vulnerable and in need of technology that could provide reliable and valid alerts about their pain to busy clinicians. The American Society for Pain Management Nursing (ASPMN), in its position

statement on pain assessment in the nonverbal patient (Herr et al., 2006), describes a hierarchy of pain assessment in which the observation of behavior including facial expressions is noted to be a valid approach to pain assessment. McGuire *et al* (McGuire et al., 2010) concluded that “pain in the intellectual disability population may be under-recognized and under-treated, especially in those with impaired capacity to communicate about their pain”. A study of patients undergoing procedural pain (Puntillo et al., 2004) showed a strong relationship between procedural pain and behavioral responses and it identified specific procedural pain behaviors that included facial expressions of grimacing, wincing, and shutting of eyes. It was relatively rare for facial expressions to be absent during a painful procedure. Findings by Payen *et al* (Payen et al., 2001) strongly support the use of facial expression as a pain indicator in critically ill sedated patients. A study of pain assessment (Manfredi et al., 2003) in elderly patients with severe dementia provides evidence that patients with pain show more pronounced behavioral response compared with patients without pain and notes that clinician observations of facial expressions are accurate means for assessing the presence of pain in patients unable to communicate verbally because of advanced dementia. Research has shown that facial expressions can provide reliable measures of pain across human lifespan and culture varieties (Williams, 2002) and there is also good consistency of facial expressions corresponding to painful stimuli. Assessment of facial expression of pain not only brings added value when verbal report is available but also serves as a key behavior indicator for pain in the scenario of non-communicating patients.

1.2 Pain Metrics based on FACS

1.2.1 Facial Action Coding System

There are two major approaches to study facial behaviors, namely message-based judgements and sign-based judgements. (Ekman and Friesen, 1969). The message-based judgements make inferences about emotions that underlie facial expressions including basic emotions and pain, while sign-based judgments focus on facial expressions itself by describing the surface of behavior created by muscular movement. The Facial Action Coding System (FACS) provides an objective description of facial expressions with 30 main action units(AU) based on the unique changes produced by muscular movements, where 12 action units are in the upper face and 18 are in the lower face. The action units are defined on the anatomical basis of muscular movement that are common to all people. Depending on the research question, observers commonly use two schemes to score AUs. The comprehensive scheme codes all AUs in a chosen video segment, and the selective scheme only codes predetermined AUs. The reliability of AU scoring, also referred to as inter-observer agreement, is assessed by considering four criteria: the occurrence of an AU, the intensity of an AU, the AU temporal precision defined as onset, apex, offset and neutral stage, and the aggregates that code certain AUs in combination. While the occurrence of AU is most widely used criterion to evaluate an AFER system, the latter three are evaluated depending on specific research problems. For example, a pain facial video event can be decomposed into a set of action units that overlap in time, which can be identified via AU aggregates coding. In this way, a message-based judgement can be inferred using predictions from the sign-based FACS system.

1.2.2 PSPI Pain Metric

Detailed coding of facial activity provides a mechanism for understanding of different parameters of pain that are unavailable in self-report measures (Craig et al., 1992). The Facial Action Coding System (FACS) (Ekman and Friesen, 1978)(Ekman, 2002)(Wilkie, 1995) has been found to be useful in assessing pain among seniors with and without cognitive impairment in research studies, in which pain behaviors are observed while a patient undergoes a series of structured activities (Keefe and Block, 1982). Research findings converge to support the validity of AUs 4, 6, 7, 9, 10, 20, 26, 27 and 43 defined in Table I as elements of the facial pain profile in adults with nonmalignant pain (Craig et al., 1991)(LeResche, 1982)(LeResche and Dworkin, 1984)(LeResche and Dworkin, 1988)(Prkachin and Mercer, 1989). The list of pain-related AUs has been further expanded in more extensive research (Williams, 2002) to include lip corner puller (AU12) and lips part (AU25). Prkachin *et al* (Prkachin and Mercer, 1989) proposed that AUs are displayed as a sequence that depends on pain intensity and durations. First eyebrows are lowered (AU 4), eyelids are narrowed (AU 7 then AU 6 with cheek raising and deepened nasolabial furrow), and then eyelids are closed (AU 43). Next the upper lip is raised (AU 10) and the nose wrinkled (AU 9). Finally, the mouth is opened (AU 26 proceeding to AU 27) and then the lips are stretched horizontally (AU 20). This finding leads to the development of an intensity-based metric known as Prkachin and Solomon Pain Intensity (PSPI) (Deyo et al., 2004) (Prkachin, 2009). The PSPI has been widely accepted in automated pain detection research in the engineering field, which is

a 16-level scale based on the contribution of individual intensity of pain related AUs and is defined as (Prkachin, 2009):

$$Pain = AU4 + \max(AU6, AU7) + \max(AU9, AU10) + AU43 \quad (1.1)$$

1.2.3 Wilkie's Pain Metric

Inspired by the work in (LeResche and Dworkin, 1984)(LeResche and Dworkin, 1988)(Prkachin and Mercer, 1989), Wilkie first adopted the FACS system to study patients with cancer pain (Wilkie, 1995), and proposed the collection of core pain-specific Action Units that occur singly or in combination as candidates for a cancer pain facial expression profile. Wilkie's pain metric measures the presence of AU 6 or 7, 20, or 27 alone, or combinations of AUs 4 and 6 or 7 or 43, 4 and 9 or 10, 4 and 26 or 27, 9 or 10 and 26, or 9 or 10 and 27, as summarized in Table I, which serves as the conceptual basis of our research. Note that although AU 27 (mouth stretch) is included in the pain expression recognition, it is excluded for the pain detection task from this study as it was found to occur highly infrequently. Facial expression annotation of videos using FACS is generally performed offline by trained experts who closely examine the video of a patient's face. Pain is assessed across the entire sequence based on the occurrence and frequency of pain-related AUs, where AUs are coded at each time step (i.e. each video frame) within the video subsequence. However, full FACS coding procedures are laborious, and they require 100 units of time coding for every unit of real-time, which makes its real-time clinical use prohibitive (Lucey et al., 2009),(Ashraf et al., 2009). Therefore, the development of an efficient real-

time automated FACS-based pain detection would be a significant innovation for enhanced patient care and clinical practice efficiency.

TABLE I: ACTION UNIT DEFINITION AND PAIN-RELATED AU COMBINATIONS

AU	Description	Pain-Related Combinations
4	eye brow lower	6/7
6	cheek raiser	20
7	eye lid tightener	4+6/7/43
9	nose wrinkler	4+9/10
10	upper lip raiser	4+26
20	lip stretcher	9/10+26
26	jaw drop	
27	mouth stretch	
43	eyes closed	

1.3 Motivations and Challenges for Automated Pain Detection

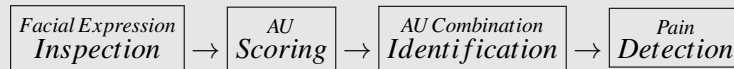
Over the past two decades, significant progress has been achieved in automated facial expression recognition. Advances in computer vision (CV) techniques have established robust facial image percep-

tion and modeling under naturalistic environments, and state-of-the-art machine learning (ML) methods have been adopted in the research on spontaneous facial expression recognition. The end-to-end solutions for facial expression recognitions propelled by recent advances in deep learning techniques have greatly motivated researchers seeking practical solutions to various affect-related problems, where the term “affect” in psychology is associated with experience of feeling or emotion. Identifying pain from facial expressions can be viewed as an extension to the spontaneous facial expression recognition problem, which is a more general research area for affect-related learning. An APD framework can be built naturally atop an AFER system as they can share most modules of system design. In addition, pain facial action units are also associated with other affects, including disgust, sadness, fear or anger (Williams, 2002). However, only a handful of studies have focused on automated pain detection from facial expressions. A literature survey conducted from 1982 to June 2014 with keywords “facial action coding system pain” by Rojo *et al* (Rojo et al., 2015) found only 26 relevant references. Prkachian notes that “information technology-based systems are currently relatively primitive, but there is little doubt that their use in assessing pain expression is feasible and their capacities will become more powerful with further research.” (Prkachin, 2009). One major challenge is the difficulty in collecting sufficient examples of annotated pain-related expressions, especially in the case of facial pain data collection from critically ill patients. Most publicly available video datasets are not targeted for pain, and only a small portion of labeled action units are pain-related. Existing pain expression datasets are mostly developed for targeted studies, which are small in size, highly imbalanced and lack sufficient diversity (Sikka et al., 2015)(Hamm et al., 2011) to train a robust automated system in general. As a result, current APD research has access to a very limited selection of pain datasets and has mainly focused on acute pain

elicited by a single type of stimuli, which is typically in a laboratory environment (Hammal and Cohn, 2014). On the other hand, human ability to perceive pain in others is significantly developed by the age of five to six and the sensitivity to more subtle facial signs of pain generally increases with age (Deyo et al., 2004), which suggests that an APD system mimicking human decision procedures of accessing pain is within the scope of present CVML techniques. Before we actually design our pain detection system, we first conduct a systematic review of APD development, with focus on how it can benefit from the advances in AFER research as well as from the progress in deep learning technology. In particular, we refer to AFER/APD methods that use only classic CVML techniques and are not based on deep learning as conventional approaches.

1.4 Main Contributions of This Research

Our research is inspired by the clinical demand for automated pain evaluation in end-of-life patient care. As our major contribution of this research, we proposed a decoupled pain detection framework that mimics the decision-making strategy of FACS-certified human coders by following the procedures below:



The proposed system decoupled the entire task into two sub-problems, pain expression (AU) detection and pain prediction from the detected pain expressions (AU). The two sub-problems are handled by an AFER system and an APD system, which are trained independently. The architecture of two indepen-

dent machine learning networks not only improves data utilization from existing different pain-oriented video datasets, but also improves data fusion incorporating future acquired data, which addresses the most challenging problem arising from the data insufficiency. The decoupled architecture lends itself to flexible customization using different implementation techniques in the modules. The AFER system is realized with three types of configurations, including one conventional CVML system and two types of deep learning architectures. The APD system is realized by two types of weakly supervised learning tools (MIL and MCIL) that are trained with low-dimensional features derived from predictions of AU combinations. We followed an end-to-front research strategy to develop the decoupled pain detection system in three research phases, and our ultimate goal is to establish a robust and generic automated pain analysis system for clinical applications.

The rest of the thesis is organized as follows, Chapter 2 reviews the progress in research that has contributed to automated pain detection, with focus on a) the framework-level similarity between spontaneous AFER and APD problems; b) the evolution of system design including the recent development of deep learning methods; c) the strategies and considerations in developing a FACS-based pain detection framework from existing research; and d) introduction of the most relevant databases that are available for FACS based AFER and APD studies. Chapter 3 presents the decoupled structure of the proposed automated pain detection framework comprised of an AFER and an APD system. Chapter 4 provides a demonstration of the advantages of the proposed framework based on the results of testing our method on the UNBC dataset and a unique dataset of lung cancer pain in terms of frame-level AU encoding and sequence-level pain detection. Chapter 5 discusses future work that should lead to a com-

mercialized product for automated pain detection. Finally, Chapter 6 presents the conclusions of this research.

CHAPTER 2

REVIEW OF PROGRESS IN AUTOMATED PAIN DETECTION

Parts of this chapter have been presented in a preprint (Chen et al., 2018) that has been uploaded to arXiv.

In this chapter, the relevant research that has contributed to automated pain detection is revealed. The focus is on a) the framework-level similarity between spontaneous AFER and APD problems; b) the evolution of system design including the recent development of deep learning methods; c) the strategies and considerations in developing a FACS-based pain detection framework from existing research; and d) the most relevant databases that are available for FACS based AFER and APD studies. We begin by describing the relevant databases.

2.1 Facial Expression Datasets

Availability of representative datasets is the fundamental basis for training and testing an AFER system. The performance evaluation of and comparison among different AFER approaches would not be fair or meaningful without context to the datasets used in the evaluation. Before proceeding to an overview of specific AFER and APD approaches, we will present an introduction to selected publicly available facial expression datasets in this section. The datasets are grouped depending on the context of the problem and presented in a chronological order of availability. These datasets were all FACS-coded and used in the studies to be covered in this thesis.

2.1.1 Standard Facial Expression Datasets for Conventional AFER

MMI (2005): The MMI dataset (Pantic et al., 2005) contains 740 static images and 780 video sequences captured in frontal and profile views from 19 subjects who were students and research staff. Two thirds of image samples and frames of 169 video sequences have been annotated by two certified FACS coders for 30 AUs. The temporal phases of an AU event (onset, apex, offset) are also coded for the 169 sequences. The MMI is a dataset of posed (not spontaneous) facial expressions under controlled settings.

RU-FACS (2005): The RU-FACS (or **M3**) dataset (Bartlett et al., 2006) captures spontaneous facial expressions from video-recorded interviews of 100 young adults of varying ethnicity. Each video segment lasts about 2 mins on average, where the head pose is close to frontal view with small to a moderate out-of-plane rotation. 33 subjects have been coded by two certified FACS coder for 19 AUs.

CK+ (2010): The extended Cohn-Kanade dataset (Lucey et al., 2010) was constructed by building upon the Cohn and Kanade's DFAT-504 dataset [**CK (2000)**] (Kanade et al., 2000). The CK+ dataset contains 593 videos captured under controlled conditions from 123 subjects with various ethnic backgrounds. Each brief video segment (approximately 20 frames on average) begins with a neutral expression and finishes at the apex phase. The entire dataset is annotated for 30 AUs at frame-level. The CK+ dataset has been made available since 2010 and is widely used for prototype and benchmark AFER systems.

GEMEP-FERA (2011): The GEMEP-FERA dataset (Valstar et al., 2012) is a fraction of the GENEva Multimodal Emotion Portrayal (GEMEP) corpus (Bänziger and Scherer, 2010) employed as a benchmark dataset for the first facial expression recognition challenge (FERA2011). The original GEMEP

corpus consists of over 7000 portrayals with 18 posed facial expressions portrayed by 10 actors. The GEMEP-FERA dataset employs 158 portrayals for AU detection sub-challenge and 289 portrayals for emotion sub-challenge. Ground truth is provided as frame-by-frame AU coding for 12 AUs and event coding of 5 discrete emotions.

GFT (2012): The Sayette Group Formation Task (GFT) dataset (Sayette et al., 2012) captures spontaneous facial expressions from multiple interacting participants in a psychological study to evaluate socioemotional effects of alcohol. The dataset include 172,800 video frames from 96 participants in 32 three-person groups (Girard et al., 2017). Multiple FACS coders are recruited for binary occurrence coding for 19 AUs on frame-level, and 5 AUs are selected for intensity coding (level: A-E). The baseline results are presented in a recent publication (Girard et al., 2017) for 10 AUs using maximal-margin framework (linear SVM) and deep learning framework (AlexNet).

2.1.2 Facial Expression Dataset for Automated Pain Detection

Establishing a pain-oriented facial expression dataset is much more challenging than establishing a general AFER dataset. **UNBC-McMaster (2011):** The UNBC-McMaster Shoulder Pain Expression Archive Dataset (Lucey et al., 2011) is the only publicly available spontaneous facial expression dataset targeting pain. It contains 48,398 FACS coded frames in 200 video sequences. The video segments are captured from patients suffering from shoulder pain and facial expressions are triggered by moving their affected and unaffected limbs. All frames are coded by certified FACS coders for 10 single pain-related AUs and the frame-level pain score is rated by the Prkachin and Solomon Pain Intensity (PSPI). A sequence-level pain label is assigned by self-reported VAS and Observer-rated Pain Intensity (OPI). In addition, 66-point facial landmarks from Active Appearance Model (AAM) are also provided for each

frame to facilitate the development of a user-customized AFER system. It is the only publicly available pain-oriented facial expression video dataset in which the spontaneous facial expressions are evoked solely by acute pain.

2.1.3 Facial Expression Datasets for Deep Learning Based AFER

SEMAINE (2012): As part of a project, Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression, the SEMAINE dataset (McKeown et al., 2012) provides high-quality, multimodal recording to study social signals occurring during interaction between human beings and virtual human avatars. The video frames were recorded from 150 participants of varying age and partitioned into subsets for training (48,000), development (45,000), and testing (37,695). Spontaneous facial expressions were annotated by three certified FACS coder for 6 AUs at frame-level. SEMAINE was selected as a benchmark dataset in the FERA 2015 challenge (Valstar et al., 2015).

DISFA (2013): The Denver Intensity of Spontaneous Facial Action (DISFA) Database (Mavadati et al., 2013) captures spontaneous facial expressions elicited by viewing a 4-mins video clip from 27 young adults via stereo video recording under uniform illumination condition. 4845 frames were recorded for each participant and coded by a single FACS coder for 12 AUs (5 upper face AUs and 7 lower face AUs) with 6 (0-5) level of intensities. The number of events (from onset to offset) and the number of frames for each intensity level were also recorded. A second FACS coder was recruited for annotating the videos from 10 randomly selected participants and resulted in a high interobserver reliability (ICC) ranging from 0.80 to 0.94.

BP4D (2014): The Binghamton-Pittsburgh 4D Spontaneous Expression Database (BP4D) (Zhang et al., 2014) contains spontaneous facial expression elicited by well-validated emotion instructions as-

sociated with eight tasks, which are captured from a diverse group of young adults. A total of 639,224 frames are produced from forty-one participants and annotated for 27 AUs by FACS certified coders in a 20-second segment interval. The video is available in both 3D and 2D, where geometric features are represented by an 83-point 3D temporal deformable shape model (3D-TDSM) and a 49-point 2D shape model tracked by a constrained local model (CLM). BP4D was also selected as a benchmark dataset in FERA 2015 (Valstar et al., 2015).

2.1.4 Selected Private Facial Expression Dataset for Pain Study

Wilkie (1995): One major distinguishing feature of our study is that we also conduct research on a unique dataset created by D.J.Wilkie (Wilkie, 1995), containing videos of 43 patients suffering from chronic pain caused by lung cancer. The videos are captured in natural settings in the subjects' homes. The patients were required to repeat a standard set of randomly ordered actions as instructed such as sit, stand up, walk, and recline, in a 10-minute video with a camera focused on the face area to record their facial expressions. Each video was partitioned into 30 equal-duration, 20-second subsequences. The subsequences were reviewed and scored for 9 AUs, possibly occurring in combination, by three trained human FACS coders independently, and the results were entered in a scoresheet that served as ground truth. Pain was declared to occur in a video subsequence if at least two coders agree with each other on the occurrence of a set of specific AU combinations listed on the scoresheet. The intensity of pain for the entire video was measured on the total number of subsequences that were associated with pain labels. Due to the issue in illumination and video quality degradation, only 1164 out of the total 1290 video subsequences are amenable for processing by Emotient, an AFER software described later and of these about 600 subsequences are suitable for pain analysis.

There are also other efforts seeking to provide publicly available pain-oriented datasets. Wener *et al* (Werner et al., 2013) addressed multi-modality pain detection by introducing the **BioVid Heat Pain** dataset, which includes 90 participants suffering from induced thermode pain in 4 levels. A recent Emotion and Pain project extends the option for chronic pain analysis by introducing a fully labeled multimodal dataset (named **EmoPain**) (Aung et al.,) targeting chronic lower back pain (CLBP). The details of both datasets and the baseline implementations are presented in section 2.3.6. However, the pain data in the BioVid dataset was collected under highly controlled settings, and the patients' self-report pain was not included, while the EmoPain dataset is still under construction without public access available at this time. Therefore we do not include these two datasets in this study.

2.2 FACS-Based General AFER Framework

Over the past two decades, research in general Automated Facial Expression Recognition systems, not necessarily pain-related, has generally adopted a framework that is comprised of four core modules: face detection, face alignment, feature extraction and facial expression classification.

2.2.1 Face Detection

The face detection module locates face regions with a rectangular bounding box in a video frame so that only localized image patches are extracted for further processing. One example of a state-of-the-art face detector is the real-time Viola-Jones detection framework (Viola et al., 2001) that employs the popular Adaboost learning algorithm in a cascade structure for classification. The ready-to-use Viola-Jones detector is capable of robustly detecting multiple faces over a wide range of poses and illumination conditions, as a result of which it is widely employed in most AFER research in practice. An extensive review for available face detection approaches is presented in (Zhang and Zhang, 2010).

2.2.2 Face Alignment in 2D

The face alignment module tracks a set of fiducial points on the detected face area in every video frame. The fiducial points are typically defined along the cues that best describe features of a face, including jawline, brows, eyes, nose, and lips. The set of fiducial points is also referred to as the shape model when a parametric model is applied to human face. While facial expressions produce non-rigid facial surface change, head pose rotations account for rigid motions. Although rigid motions are negligible in controlled experimental settings for posed facial expression datasets, occurrence of spontaneous facial expressions in natural environment is frequently accompanied by obvious rigid motions, where challenges arise for accurate feature extraction in the next step. Rigid motion normalization can be embedded within an alignment framework via a similarity or affine transform. Alternatively, an aligned face model can be rotated and frontalized with the estimated head pose parameters and registered to a front view mean face template with neutral facial expression.

Faces can be modelled as deformable objects which vary in terms of shape and appearance. The active appearance model (AAM) (Cootes et al., 2001)(Baker and Matthews, 2004) is one of the most commonly used face alignment method, where the optimization process relies on jointly fitting a shape model and an appearance model via the steepest decent algorithm over the holistic face texture. The original AAM is subject-dependent where the model is tuned specifically to the environmental conditions of the training dataset. This, however, may cause performance degradation on face images of an ‘unseen’ subject. This problem can be addressed by using the constrained local model (CLM) (Cristinacce and Cootes, 2006)(Cristinacce and Cootes, 2008), which is comprised of a shape model and an appearance model similar to AAM. It only utilizes rectangular local texture patches around the shape feature points

to optimize the shape model through a non-linear constrained local search. The appearance templates in CLM are represented by a small set of parameters that can be effectively used to fit unseen face images. Therefore a CLM framework can be conveniently modified as a subject-independent model (Asthana et al., 2013) for generic face alignment applications. The result in (Chew et al., 2012) indicates that CLM is generally more computationally efficient than AAM but it is slightly outperformed by AAM when rigid motions are presented.

Follow-up research led to continuous improvement to the family of AAM and CLM methods in terms of fitting accuracy and computational efficiency, which involves adaptation to the fitting methodology (Tzimiropoulos and Pantic, 2017), feeding rich annotated training data in unconstrained conditions (Zhu and Ramanan, 2012), combination with new appearance descriptors (Xiong and De la Torre, 2013), and fine tuning with cascaded regression techniques (Asthana et al., 2014). Such evolutions have boosted the performance of face alignment to handle illumination variations, out-of-plane rigid motions, as well as partial occlusions. It is worth noting that typical AU coding scenario is at close-to-front ($\pm 30^\circ$) view, which can be adequately handled by recent 2D alignment methods. Although 3D alignments provide much denser geometric modeling as well as depth information, computational inefficiency and difficulty in data acquisition makes it currently infeasible in many practical AFER and APD applications. Therefore we shall focus on automated pain detection from 2D facial expression recognition in this study. For an adequate survey on facial expression modeling in 3D, readers are referred to the survey (Sandbach et al., 2012).

2.2.3 Feature extraction

Facial expressions are revealed by non-rigid facial muscular movement. A feature extraction module is added between raw input pixels and classifiers and serves as an important interpreter to extract relevant non-rigid information for manifesting Action Units from computer vision perceptions, which is highly customizable in the framework structure with a rich option of feature descriptors. Features can be categorized into geometric-based and appearance-based. Geometric features extract geometric measurements related to coordinates of fiducial points, while appearance features are derived from pixel intensities. Predefined distance, curvatures and angles from displacement of the fiducial points were an intuitive way to extract geometric features in early AFER research (Tian et al., 2001)(Pantic and Patras, 2006). Static geometric measurements can be used as midlevel representations to extract dynamic features by encoding with temporal information for a more detailed description of neuromuscular facial activities (Valstar and Pantic, 2012). Geometric features are typically low in dimension and simple to compute and focus on describing the deformation of the conspicuous feature cues. However, geometric features are insufficient to model muscular actions off the feature cues and they are sensitive to registration errors from shape alignment. Therefore geometric features are typically applied in conjunction with other feature descriptors in the spontaneous AFER scenario.

On the other hand, appearance features could be extracted directly from the alignment modality (Ashraf et al., 2009), either by similarity normalized appearance (S-APP) that normalizes the rigid motion or by canonical appearance (C-APP) that warps the texture to the mean shape template. The resulting appearance descriptor is then concatenated in a high-dimensional feature vector based on raw pixel intensity for classification. In this case, facial-expression-related muscular activities are encoded-

ed together with a large portion of person specific features without discrimination, which potentially attenuates the discriminating power of the feature vector towards target AUs. Subtle texture changes like deepening the nasolabial furrow, raising the nostril wing, cheek raising, pushing the infraorbital triangle up, provide important reference for coding of certain AUs (Ekman and Friesen, 1978). These non-rigid motions are nonstructural, which are infeasible to be effectively described by raw geometric or pixel-based features. Descriptors developed for object detection are introduced in AFER research in order to exploit the features more in depth. These descriptors detect local intensity variations as feature points, edges, phase or magnitudes and assemble the low level appearance features into high-level feature representations, from which a meaningful classification can be made. A series of studies (Bartlett et al., 2006)(Wu et al., 2010)(Littlewort et al., 2011)(Wu et al., 2012) have applied a Gabor filter bank to extract features in 8 directions and 9 spatial frequencies and the output magnitudes are concatenated into a single feature vector for AU classification. The local binary pattern (LBP) encodes local intensity variation in the neighbourhood and as an extension to LBP, the local phase quantization (LPQ) captures local phase information from Fourier coefficients, and a histogram is assembled from the local descriptors as a high-level feature representation. If the facial video volume is sliced by three orthogonal planes (TOP), both LBP and LPQ static feature descriptors are then expanded to encode temporal information from $x - t$ and $y - t$ planes, which are known as LBP-TOP and LPQ-TOP (Jiang et al., 2011). A detailed survey of the application of the family of LBP descriptors to facial image analysis can be found in (Huang et al., 2011). More geometric-invariant feature descriptors have been successfully applied to AFER including the histogram of oriented gradient (HOG) (Chew et al., 2012) and scale-invariant feature transform (SIFT) (Girard et al., 2015). The appearance descriptors can also be associated with

an image pyramid representation to extract feature with multiple scales and resolutions (Dhall et al., 2011)(Sun et al., 2014).

In addition to extracting static features from a single frame, it is also beneficial to extract dynamic features comprised of spatially repetitive, time-varying patterns (Chetverikov and Péteri, 2005) to model motion-based information from an image sequence. Koelstra *et al* (Koelstra et al., 2010) employ two representation methods, motion history image (MHI) and free-form deformation (FFD), to derive non-rigid motions from consecutive frames. A MHI descriptor computes a motion vector field from a set of weighted binary difference images within a predefined time window, while a FFD descriptor captures non-rigid motions via a B-spline interpolation that computes local variations of a control point lattice. A quad-tree decomposition method is then applied to decompose the face region in a non-uniform grid representation, where most grids are placed on areas with high motion response. The quad-tree decomposition is performed in three orthogonal planes to extract features in terms of magnitude, horizontal motion, and vertical motions. This work (Koelstra et al., 2010) provides an example on feature customization in terms of geometric and appearance feature combination, dynamic feature extraction by utilizing temporal information, as well as a balance between global and local feature extraction. Each AU is activated by a combination of facial muscles, which can be geometrically located in a specific region on the face. Appearance features extracted from corresponding region of interests (ROI) bounded by geometric features potentially help to reduce the feature dimensionality and boost computation efficiency. Hamm *et al* (Hamm et al., 2011) divided a face image into 14 rectangular regions based on geometric landmarks' layout, from which a 72-dimensional Gabor response is pooled to generate a histogram-based feature representation. If multiple feature descriptors are employed in an AFER

framework, a feature fusion problem should also be considered. Wu *et al* (Wu et al., 2012) evaluate the performance multi-layer architectures for feature extraction involving a combination of two texture descriptors, the Gabor Energy Filter (GEF) and LBP. However, a more common way for feature fusion is to perform it in the classification stage, which will be addressed in the next section.

2.2.4 AU classification and regression

FACS-based AFER studies typically focus on two core problems in Action Unit recognition: 1) which AUs are displayed in the input images or videos, and 2) how to measure the intensity of the observed AUs. Conventional AFER research frequently treats the detection of AU occurrence as a binary classification problem, where a binary classifier is trained for each of the target AU independently, and a separate set of classifiers or regressors are trained to discriminate AU intensity of different levels. Static classification approaches serve as the basis for AU recognition, which rely on features extracted from one static image frame. However, psychological studies (Bassili, 1979) suggest that facial behavior can be more reliably recognized from an image sequence than from a still image. Hence dynamic approaches include additional analysis steps on static predictions by utilizing temporal information from adjacent frames to improve reliability of frame-level AU estimations or perform high-level AU event coding across the video sequences.

Static Approach

Support vector machine (SVM) and boosting-based classifiers are commonly employed for the AU classification task. A binary support vector machine seeks a subset from training data known as support vectors, which defines a hyperplane that maximizes the margin between the hyperplane and closest points of both classes. In practice, the feature spaces can be lifted to a higher-dimensional feature

space using *kernels* (Szeliski, 2010), where linear and radial basis function (RBF) kernels are frequently used in SVM settings for AU classification. In fact, SVM was used as part of the baseline system in most facial expression datasets (Bartlett et al., 2006)(Lucey et al., 2010)(Lucey et al., 2011)(Mavadati et al., 2013)(Girard et al., 2017). A boosting-based approach learns a set of weak binary classifiers in a sequential order, where mis-classified samples are assigned higher weights and are considered as ‘hard’ problems to be handled with the next weak classifier. The final decision is made by a strong classifier that is generated from a linear combination of the set of weak classifiers.

Bartlett *et al* (Bartlett et al., 2006) assessed both SVM (linear and RBF kernels) and Adaboost (up to 200 features selected per AU) classifiers for facial action classification on 20 AUs, where Gabor wavelet features were extracted from the RU-FACS (spontaneous) and CK (posed) datasets. Chew *et al* (Chew et al., 2012) employ a linear SVM to test pixel-based features and more complex appearance features (HOG, Gabor Magnitudes) extracted from multiple datasets (CK+, M3, UNBC-McMaster and GEMEP-FERA) under different face alignment accuracy. The experiment suggests that the more complex appearance descriptors are robust to alignment errors on AU detection, but their advantages are limited under close-to-perfect-alignment. Jiang *et al* (Jiang et al., 2011) performed recognition of 23 upper and lower face AUs on MMI dataset with four different SVM kernel settings, including Linear, Polynomial, RBF and Histogram intersection. The SVM classifiers were trained with a set of features from LBP family with optional *Block* and *Pyramid* extensions, including LBP, LPQ, B-PLBP and B-PLPQ. Experimental results indicated that a SVM with histogram intersection provided slightly better performance than other kernel settings, and the block-pyramid based B-PLPQ feature outperformed other features but at the cost of increased computational complexity. Jeni *et al* (Jeni et al., 2013b) proposed a continuous AU in-

tensity estimation framework using support vector regression (SVR). Sparse features were learned from local patches via personal mean texture normalization followed by non-negative maxtrix factorization. A L_2 -loss regularized least-squares SVM (LS-SVM) model was trained for AU intensity regression in 6 ordinal levels (0-5). The proposed AU intensity estimation system was tested on 14 AUs on posed data from CK+ as well as on spontaneous data from BP4D for AU12 and AU14. Chu *et al* (Chu et al., 2017) investigated the problem of personalizing a classifier trained on a generic facial expression dataset with respect to unlabeled person-specific data through a transductive learning method, which is referred to as Selective Transfer Machine (STM). The idea behind STM is to re-weight the samples in the generic training set by minimizing the training and person-specific test distribution mismatch, such that a person-specific SVM classifier can be generated from the reweighted training data. The proposed system was evaluated for cross-subject AU detection on CK+ and RU-FACS, and cross-dataset AU detection in terms of RU-FACS→GEMEP-FERA and GFT→RU-FACS. The experimental results demonstrated that STM consistently outperformed generic SVM on both tasks.

Dynamic Approach

A dynamic approach for AU classification takes input from static AU predictions of multiple consecutive frames to perform temporal analysis that targets three problems in general: 1) improving prediction reliability on the current frame using information from past frames, 2) generating segment-level AU detection from frame-level prediction, and 3) locating AU events in a video sequence by modeling temporal phase transition (i.e. onset, apex, offset, and neutral) of an Action Unit.

Koelstra *et al* (Koelstra et al., 2010) trained a one-versus-all GentleBoost classifier for each AU in onset or offset phase independently with spatiotemporal features derived from Free-Form Deformations

(FFD) or Motion History Image (MHI). The outputs of each pair of onset and offset GentleBoost classifier were combined into one AU recognizer through a continuous Hidden Markov Model (HMM). The four-state HMM performed as a temporal filter as well as AU event encoder by modeling temporal phase transition for each AU, where the prior and transition matrices were estimated from the training set and the emission matrix was estimated from the outputs of the GentleBoost classifiers. The proposed system was tested for 27 AUs in MMI and 18 AUs in CK+. A follow-up study in (Valstar and Pantic, 2012) proposed a hybrid SVM-HMM framework to estimate the temporal phases of an AU event. A one-versus-one multi-class SVM was trained for each of the four temporal phases using 2520-dimensional geometric features. The outputs of SVMs were converted to probability measures via a sigmoid function, which were used to estimate the emission matrix of a HMM. Rudovic *et al* noticed the temporal phase of AUs are correlated with their intensities and proposed an approach based on Conditional Ordinal Random Field (CORF) to utilize the ordinal relations embedded in intensity levels. The proposed LAP-KCORF framework (Rudovic et al., 2012) extends the CORF model by introducing a kernel-induced Laplacian regularizer to the optimization process. The Composite Histogram Intersection (CHI) kernel was selected in the framework settings, which takes the input of LBP features from aligned training images. The LAP-KCORF model was trained for each AU separately. While neutral and apex phases can be discriminated solely from the ordinal score, the onset and offset phases have to be discriminated from the dynamic features. Experimental results on 9 upper face AUs drawn from MMI datasets demonstrated the advantages of the ordinal model over the SVM-HMM model on both AU detection and temporal phase estimation. Hamm *et al* (Hamm et al., 2011) proposed an AFER system to perform dynamic analysis of facial expressions on a private video dataset featuring patients with

neuropsychiatric disorders and healthy controls. A one-versus-all GentleBoost classifier was trained for each of the 15 selected AUs using Histogram of Gabor Features extracted from multiple pre-defined local regions. The dynamic analysis was measured by single and combined AU frequencies and affective flatness and inappropriateness, which was derived from the temporal profiles of AU predictions on frame level.

A complete AU event refers to activation of an AU in terms of its duration and start/stop boundaries. Realizing the event-based detection is more desirable than frame-based detection in many application scenarios in practice, Ding *et al* (Ding et al., 2016) carried out AU detection on frame-level, segment-level and transition encoding in a sequential order, and integrated the results from three detectors to Cascade of Tasks (COT) architecture to perform AU event coding. The frame-level detector was trained with SVM on SIFT features, where the frame-level detection was used to augment segment-level training data in both feature representation and sample weight. The segment-level data were represented by a Bag of Words (BOW) structure with geometric features. A weighted-margin SVM was employed for segment-level detection, where samples (segments) containing many positive frames (scored by frame-level detector) were associated with higher weights. Two transition detectors were then trained to refine the onset and offset boundaries from the detected AU segments using the segment-level features. The scores for segment-level prediction and transition predictions were linearly combined to predict the occurrence of AU events. The start and end frames of an AU event were determined based on the highest event score. Multiple events could be scored in a given video sequence using dynamic programming (DP) via a branching-and-bounding strategy. The proposed system was trained and tested for different set of AUs on CK+, RU-FACS, FERA and GFT datasets respectively. Sikka *et al* modeled an affective

event as a sequence of discriminative sub-events, and solved the event detection as a weakly supervised problem with ordinal constraints by the proposed Latent Ordinal Model (LOMo) (Sikka et al., 2016). A LOMo model seeks a prototypical appearance template to represent each sub-event, and the order of all sub-event permutations is associated with a cost value, where permutations far from the ground truth order were penalized with higher cost. A score function originating from a linear SVM was defined to measure the correlations between a weakly labeled frame (with only sequence-level label) and the templates with consideration of the ordinal cost. The entire model was learned using an algorithm based on stochastic gradient descent (SGD) with respect to a margin hinge loss minimization. The proposed method selected multiple frame-level feature descriptors (SIFT, LBP, geometric and deep-learned features) and was trained and evaluated on four facial expression datasets including CK+ and UNBC-McMaster.

Discovering AU Relations

The aforementioned methods consider single AU detection as one-versus-all classification problem. There are also efforts that attempt to make better use of FACS by exploiting dependencies and relations among AUs, which are potentially helpful in improving detection on highly skewed AUs in spontaneous facial video datasets. Tong *et al* (Tong et al., 2007) proposed a system based on Dynamic Bayesian Network (DBN) to model relations among different AUs in a probabilistic manner taking into consideration temporal changes in an AU event. An Adaboost classifier trained on Gabor features was employed as a baseline static AU detector. A Bayesian Network (BN) was used to model the relations of AUs in terms of their co-occurrence and mutual exclusivity. A DBN network is comprised of interconnected time slices of static BNs, where the state transitions between the consecutive time slices is modeled by a

HMM. The structure and parameters of the DBN were obtained by training it offline and AU recognition under temporal evolution was performed online with the trained DBN network. The DBN-based system was learned for 14 AUs from the CK+ dataset and the validation was further generalized using selected samples from the MMI dataset. Zhao *et al* (Zhao et al., 2015) formulated patch learning and multi-label learning problems in AU detection as a joint learning framework known as Joint Patch and Multi-label Learning (JPML). Patch learning sought local feature dependencies for each AU, where adaptive patch descriptors were placed around 49 facial fiducial points to extract 128D SIFT features from each patch. Multi-learning exploited strong correlations among AUs from a AU relation matrix that was learned on over 350,000 FACS annotated frames from multiple datasets. *Positive correlation* and *negative competition* were defined for relations of likely and rarely co-occurring AU pairs based on the correlation coefficients in the AU relation matrix. The JPML framework integrated the patch learning and multi-label learning into a logistic loss in the form of a patch regularizer and a relational regularizer, which were then jointly solved as an unconstrained problem. The JPML framework was trained and tested on CK+, GPT, BP4D datasets, from which specific active patch regions were identified for 11 AUs. Altogether 8 positive correlation and 14 negative competition pairs were discovered among 11 AUs.

2.2.5 Performance Metrics

After training a proposed AFER system on selected datasets, the next step is to evaluate the system performance with proper metrics. The criteria to select measure metrics depends on the type of classification problem, e.g. binary AU classification or multi-level AU intensity estimation.

Accuracy, *F1 score* and *Area Under the receiver operating Characteristic (AUC)* curve are three most frequently used metrics to evaluate a binary classifier. *Accuracy* is the percentage of correctly

classified positive and negative samples. *F1* score is the harmonic mean of between precision (*PR*) and recall (*RC*), so that the two measures are reflected through one score. The metric “*AUC* is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one” (Fawcett, 2006).

Evaluation of multi-level intensity classification and regression depends on two types of performance metrics in general. The first type includes the root mean square error (*RMSE*) and mean absolute error (*MAE*) metrics that effectively capture the difference between predicted value and the ground truth. While *MAE* is more intuitive for interpretation, *RMSE* possesses the benefit of penalizing large errors more, and is easier to be manipulated in many mathematical calculations. The second type of metric is correlation-based including Pearson’s correlation coefficients (*PCC*) and Intraclass correlation (*ICC*), which measure the trend of prediction following the ground truth without considering a potential absolute bias (Egede et al., 2017). The *ICC* is similar to *PCC* except the data are centered and scaled using a pooled mean and standard deviation (Girard et al., 2015), which require raters (e.g. a human coder and an AFER) to provide the same rating without multiplicative difference to achieve agreement on *ICC*. *RMSE/MAE* and *ICC* are often used simultaneously as complimentary metrics in AFER studies, as a low *ICC* is possible with deceptively low *RMSE/MAE* scores and vice-versa (Egede et al., 2017).

Most spontaneous facial expression datasets contain a significantly larger fraction of negative samples than the fraction of positive samples in practice, where the data imbalance can be defined by the skew ratio as $Skew = \frac{\text{negative examples}}{\text{positive examples}}$. Jeni *et al* (Jeni et al., 2013a) studied the influence of skew on multiple metrics and reported *Accuracy*, *F1*, *Cohen’s κ* and *Krippendorff’s α* were attenuated by skewed distribution. Although *AUC* was unaffected by skew ratio, it may mask poor performance due to data im-

balance. The author suggested reporting skew-normalized scores along with original ones to minimize skew-biased performance evaluation. This finding was also taken into account by several studies (Tóssér et al., 2016)(Zhao et al., 2015)(Rodriguez et al., 2017). It is worth mentioning that all metrics discussed above are also applicable to pain detection and deep learning based AFER approaches. We shall use the same abbreviations (as emphasized in *italic*) when referring to these metrics in the following sections.

2.3 Automated Pain Detection from Facial Expressions

Present APD studies generally follow two modalities to detect pain from facial expressions. In the indirect way video frames are first encoded with FACS-based Action Units and pain is then identified from pain-related AU predictions. Alternatively, a mapping can be learned directly from high-dimensional features to assign pain labels without going through the AU coding procedure. The progress in APD has benefited vastly from advances in AFER techniques which are applicable in both the direct and the indirect modalities, as AFER is the core module in the indirect method and most functional blocks in an AFER framework can be applied within a direct pain detection architecture with minor modifications. In fact, with few exceptions, all APD studies reviewed in this section are linked to some preliminary AFER research. The correlation between the frameworks used in addressing the AFER and APD problems is illustrated in figure 1.

Aside from the similarity to the AFER problem, there are additional considerations that should be factored in the design of an APD system. These stem from the nature of the pain data, methods used in labeling pain data, the system performance metrics used, and availability of supplemental pain data from non-facial sensing. For example, pain data may be collected under controlled, spontaneous, or clinical conditions; the ground truth may be obtained from patients' self-report, or certified FACS coder; a pain

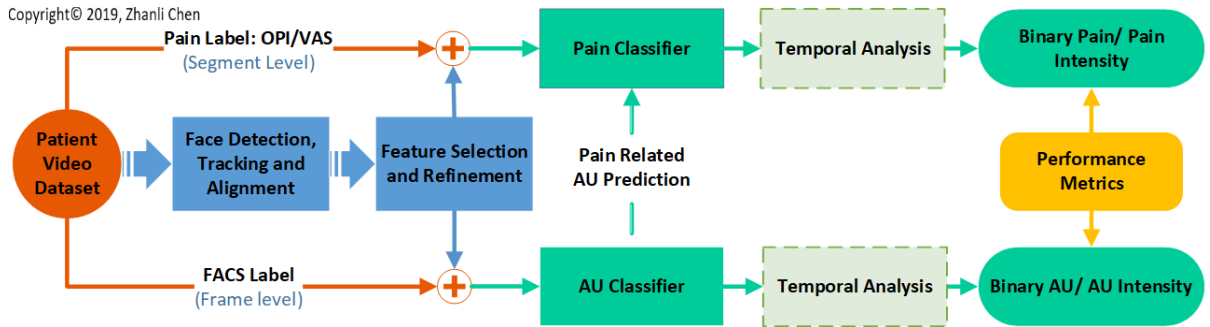


Figure 1: Correlation between the frameworks used in addressing the AFER and APD problems

event may be scored per frame or per segment for the binary pain decision or pain intensity estimation; acute and chronic pain may be distinguished from onset phase duration (Hammal and Cohn, 2014); suitable metrics could be used for performance evaluation and for comparison with human coders' decision. Last but not the least, signals from channels other than facial expressions may also help in pain detection. In this section, we will review existing APD research in a case study manner by addressing these considerations.

2.3.1 Genuine Pain Vs. Posed Pain

A pioneering APD study examined the problem of distinguishing posed pain from genuine pain through AU coding (Littlewort et al., 2007)(Littlewort et al., 2009). Facial expressions induced by posed pain and genuine pain are mediated by two neural pathway and thus have subtle differences in terms of muscular movements and dynamics. The system implementation was comprised of a two-level machine learning module. The first stage solved a general AFER problem by training independent linear SVM binary classifiers for a list of 18 single AUs and 2 AU combinations about prototypical expression

of fear and distress. Features were extracted by Gabor filters and selected by Adaboost. Differences of Z-scores were computed for three pain conditions consisting of real pain, fake pain, and absence of pain (baseline). AU 1, 4, and the distress brow combination were considered statistically significant for real vs. fake pain discrimination. Window-based statistics and event-based statistics were obtained from the first stage of AU prediction, and a second layer non-linear SVM was trained to discriminate genuine pain from faked pain. The AFER system was trained on 3 datasets (2 posed and 1 spontaneous), while the pain data were collected from 26 subjects subjected to cold pressure pain. The system was evaluated by *AUC* and *2FAC*, and the reported performance found to be superior to naive human judges. Although accuracy of individual AU classifier is still below that of human coders, this is among the first efforts to extend AFER to APD study.

2.3.2 Binary Pain Classification

Following the first study based on UNBC-McMaster dataset, (Ashraf et al., 2009) proposed a framework to detect pain at both frame and sequence level. The face video was aligned by AAM and a combination of geometric features (similarity normalized shape points - S-PTS) and appearance features (canonical appearance - C-APP) was extracted. A linear SVM was used to perform binary classification of “pain” vs. “no-pain”, and prediction of pain score at frame level was defined as the distance of the test observation from the separating hyperplane of SVM. Pain scores for every frame are summed and normalized for the duration of the sequence to produce a cumulative score for sequence level pain measure, and the decision threshold was determined by Equal Error Rate (*EER*) derived from ROC. In the follow-up research Lucey *et al* (Lucey et al., 2011) reported on recovering 3D pose from 2D AAM (Xiao et al., 2004) and performing statistical analysis based on ground truth to reveal that patients suffering

from pain displayed larger variance in head positions. A set of binary linear SVM classifiers was trained for 10 pain related AUs to replace the single binary pain SVM classifier in the previous work (Ashraf et al., 2009). Frame-level pain measurement was then derived from the output scores of AU detectors that were fused using linear logistical regression (LLR). A combination of S-PTS, S-APP and C-APP yields the best average AU recognition performance of 0.78 measured by *AUC* (Ashraf et al., 2009), and the performance was further improved to 0.81 by compressing the spatial information with a discrete cosine transform (DCT) (Lucey et al., 2008). Continuous research based on this framework was conducted in (Lucey et al., 2012), where OPI label was employed for sequence level pain evaluation to mimic the decision making of human observers. The OPI labels are segmented into three classes based on intensity (0-1, 2-3, and 4-5), and a one-versus-all binary SVM classifier was trained for each OPI group to generate a confusion matrix. The frame-level pain label assigned by the classifier model produced the highest probability score and a sequence level pain label is generated from all its member frames via a majority vote scheme. Although reasonable classification rate was obtained for no pain class (OPI 0-1), the system did not report desired performance to discriminate low pain level (OPI 2-3) from high pain level (OPI 4-5). Further findings from the testing results suggests that rigid head motions do not necessarily contribute to prediction of OPI intensities. This series of studies is valuable as it paves the way for APD research to replicate the job of a care-giver when monitoring the pain of a patient.

2.3.3 Pain Intensity Estimation

Hammel *et al* (Hammal and Cohn, 2012) built a pain intensity classification system for the UNBC-McMaster dataset using the AAM-CAPP-SVM framework. A Log-Normal filter bank tuned to 15 orientations and 7 central frequencies was applied to the C-APP features to enhance the modeling of

deepening and orientation change of appearance features that characterize pain expressions. The pain metric is measured with the PSPI scale and the intensities are categorized into 4 classes (0, 1, 2, ≥ 3) and four linear SVM are trained with the Log-Normal filtered C-APP feature map for the four levels of pain intensity. The pain intensity is estimated on a frame-by-frame basis and the performance is evaluated by *F1* for 5-fold (91% – 96%) and leave-one-subject-out (45% – 67%) cross-validation. Moderate to high consistency between manual and automatic PSPI pain intensity was measured by *ICC* for the two cross-validation schemes.

Irani *et al* (Irani et al., 2015) adopted spatiotemporal oriented energy features to model facial muscular motions in a pain intensity detection framework on the same pain dataset. The face sequence was segmented and aligned using AAM, and then a face patch was further divided into 3 regions based on the prior knowledge of location of pain expression muscular activities. Facial muscular activities were extracted at the frame level by steerable and separable energy filters that were implemented by a second derivative Gaussian followed by a Hilbert transform tuned to 4 directions. A histogram was generated from pixel-based energy features as per directions and regions to form region-based energy descriptors. Vertical motion and horizontal motion were computed from the histogram of spatial features in up-down and left-right directions respectively. Frame-level motion features were summed in time domain to generate a final spatiotemporal descriptor for each region. Pain intensity was estimated by a weighted linear combination of the vertical and horizontal motion scores. Experimental results reveal an improvement in no-pain and weak-pain recognition in the work of (Hammal and Cohn, 2012).

Lundtoft *et al* (Lundtoft et al., 2016) modified the pain intensity framework in (Irani et al., 2015) by introducing a super pixel approach for face region segmentation. They argued a better recognition rate

for no pain category can be achieved by only exploiting vertical motions of the spatiotemporal descriptor from the middle region of the face.

2.3.4 Pain Detection in Clinical Settings

An encouraging clinical application of automated pain detection was reported in (Sikka et al., 2015)(Sikka, 2014). This study applied CVML models to assess pediatric postoperative pain and demonstrated good-to-excellent classification accuracy and strong correlation with pain ratings from human observers (parents and nurses). In the data collection stage, a single camera recorded video of 50 youth, 5 to 18 years old, for transient and ongoing pain conditions during 3 study visits after laparoscopic appendectomy. Facial activities were recorded for 5 minutes as a measure of ongoing pain. Then the surgical site was manually pressed for two 10-second periods to stimulate transient pain, and video recorded. Pain ratings from patients' self-report, parents, and nurses were recorded independently in a 11-point Numerical Rating Scale (NRS), where self-report pain rates were considered as the subjective ground truth and time since surgery provided the objective ground truth alternatively. Video segments with NRS rating ≥ 4 are defined as trials with pain, while those with NRS ratings of 0 are defined as trials without pain. The computer expression recognition toolbox (CERT) was employed to independently score 10 pain-related single AUs, one smile-related AU combination, and rigid head motion in 3D (yaw, pitch and roll) per frame. Three statistics (mean, 75th percentile, and 25th percentile) were computed for each of the 14 frame-level raw features across the duration of a pain event (ongoing and transient), which were comprised of a 42-dimensional feature vector for the pain event. Two linear regression models with L_1 regularization are trained on the same set of feature vectors for binary pain classification and pain intensity estimation tasks.

Performance of a binary pain model was evaluated via a 10-fold cross validation with two metrics, where *AUC* (in the range of 0.84-0.94) demonstrates good-to-excellent detection rate and Cohen's κ provides psychological measurement on categorical agreement between raters with consideration of chance agreement. Categorical agreement evaluated for transient and ongoing pain was fair to substantial ($\kappa = 0.36 - 0.61$) for the model trained on self-report ratings and substantial ($\kappa = 0.61 - 0.72$) for the model trained on objective ground truth, where the prediction of automated system was more consistent than human observers on ongoing pain according to the κ measurement. The pain intensity estimation model was evaluated by the Pearson correlations metric (both within and across subjects) using a leave-one-subject-out cross validation scheme. The best performance was achieved when the PIE system was trained with objective ground truth for both pain conditions, and the within-subject correlation ($r = 0.80 - 0.86$) was consistently higher than the correlation for all subjects ($r = 0.55 - 0.59$). The CVML model performance was at least equivalent to or exceeded that of a nurse for both BPC and PIE tasks, but was only equivalent to that of parents in PIE task. The analysis of results from PIE also supports the tendency for nurses to underestimate pain severity (Williams, 2002) when compared with children's self-report, especially under the ongoing pain condition. Model performances were not affected by including demographic information in the training features. This research outlines general procedures for applying a CVML model in clinical settings, which involve experimental design and data collection, CVML tools, model evaluation from both engineering and psychology aspects, as well as performance comparison with human observers. Therefore the methodology in this research serves as an important modality for future APD research.

2.3.5 Detecting Pain Event with ‘Weakly Labelled’ Data

A patient’s self-report is the gold standard for pain evaluation in patient care. A pain label is commonly available for a video sequence but not for every single frame. Such a situation is encountered frequently in computer vision since it is easier to obtain group labels for the data rather than individual labels, and is known as ‘weakly supervised’ learning problem. Sikka *et al* (Sikka et al., 2014) modeled a video sequence with a bag of words (BOW) structure that is comprised of multiple segments to handle the ‘weakly labelled data’, and solved the ‘weakly supervised’ pain detection problem with a multiple instance learning framework (MIL). Two partitioning schemes, normalized cuts (Ncuts) and scanning window (scan-wind), were employed to generate multiple segments (referred to as instances) from a video sequence (referred to as bag). According to BOW definition, a positive (pain) bag contains at least one positive instance while a negative (no pain) bag contains only negative instances. The frame-level feature extraction learns a mapping from the pixel space to a d -dimensional feature vector. Choice for frame-level feature descriptor is highly flexible, the multi-scale dense SIFT (MSDF) with a 4-level spatial pyramid was selected in this work (Sikka et al., 2014). The segment-level feature representation was then generated from all frames in the segment via a max pooling strategy, such that an instance in a bag is also represented by a d -dimensional vector. A Milboost (Zhang et al., 2006) framework based on gradient boosting was used to learn the posterior probabilities for both bags and instances, which were directly related to pain event prediction on segment level or sequence level. The frame-level pain score was also derived from posterior probability of segments by assigning the maximum score among all the segments to which it belongs. The frame level pain score can be used for pain localization in time domain. The proposed APD framework was trained and validated on the UNBC-McMaster dataset,

where sequences with OPI rating ≥ 3 are selected as positive samples and those with OPI rating = 0 are selected as negative samples. The binary pain detection was evaluated with *EER* (83.7%) in the ROC. The pain localization task was evaluated as pain intensity estimation on frame level with ground truth of PSPI, and the performance was evaluated by *FI* score (.471) and Spearman's rank correlation (Kendall, 1938) (.432).

The MIL framework detects pain (pain vs. no pain) in a Multi-Instance-Classification (MIC) setting, where the sequence-level intensity labels (VAS, OPI) of pain are binarized. It is, therefore, not suitable for a multi-level classification/regression problem like the pain intensity detection task. Ruiz *et al* embedded an ordinal structure of labels to the bag representation, which naturally corresponds to various rating or intensity estimation tasks. In the proposed Multi-Instance Ordinal Regression (MIOR) weakly supervised settings, a video sequence was modeled as a bag where instances are consecutive video frames in the sequence. The bag label belongs to an ordinal set and instance labels were treated as latent ordinal states, which were inferred from a learned mapping from feature space to the structural output space. A multi-instance dynamic ordinal random fields (MI-DORF) framework built upon the hidden conditional ordinal random fields (HCORF) (Kim and Pavlovic, 2010) framework was employed to handle the multiple instance learning (MIL) problem incorporating both ordinal and temporal data structures, where the temporal dynamics within the instances was encoded by transitions between ordinal latent states. The proposed MI-DORF framework was then applied to pain intensity estimation task on UNBC-McMaster dataset, where a total of 157 sequences from 25 subjects were selected based on the low-discrepancy criteria between sequence- and frame-level pain label. The sequence-level pain label in an ordinal scale (OPI label) between 0 and 5 was used to train the system, and the frame-level

label in PSPI was also normalized to the same scale and used for results validation. The frame-level (instance) features are represented by 49 fiducial points tracked by AAM. For a fair comparison with MIL in (Sikka et al., 2014), this study followed a similar experimental setup as in (Sikka et al., 2014), while the output probability for binary pain classification in MIC methods was also normalized to 0 – 5 as a representation of pain intensity. The performance on sequence-level prediction was measured by *PCC* (0.67), *MAE* (0.80), *ICC* (0.66), *Accuracy* (0.52), and *F1* (0.34), which outperformed other MIL methods (Sikka et al., 2014) (Hsu et al., 2014) as well as the baseline model (Kim and Pavlovic, 2010). Furthermore, the proposed approach achieved a comparable performance on frame-level pain intensity estimation ($ICC = 0.40$) using only sequence-level ground truth towards the state-of-the-art fully supervised (using frame level ground truth) Context-sensitive Dynamic Ordinal Regression method (Rudovic et al., 2015) ($ICC = 0.59 - 0.67$). This result suggests a good trade-off between system performance and labor intensive frame-level annotation.

2.3.6 Multi-Modality in Pain Detection

Pain is a subjective experience, which is possibly caused by various stimuli (e.g., spinal stenosis, rotator cuff tear, cardiac conditions or sickle-cell disease) and manifested through signals from multiple modalities including voice, facial and body expressions (Hammal and Cohn, 2014). Present efforts in APD largely focused on acute pain that is elicited by a single stimulus in a controlled environment, which poses a gap between existing APD research and requirements of practical pain-related clinical applications. The UNBC-McMaster is the first publicly available dataset with well-annotated data for acute pain that has facilitated APD research since its first availability in 2011. Although it has been

widely employed to benchmark the performance of both AFER and APD systems, its contribution is limited by its inability to lend itself to study other types of pain such as chronic pain.

A recent Emotion and Pain project extends the option for chronic pain analysis by introducing a fully labeled multi-modal dataset (named EmoPain) (Aung et al.,) targeting chronic lower back pain (CLBP). The data in EmoPain contains multiple signals collected with and across different modalities, including high-resolution multi-view face videos, full-body 3D motion capture, head-mounted and room audio signals, and electromyography signals from back muscles. Participants are requested to perform both instructed and non-instructed exercise to simulate therapist-directed and home-based-self-directed therapy scenarios. Level of pain from facial expressions was labelled by 8 raters, while six pain-related body behaviors were annotated by 4 experts. 22 CLBP patients together with 28 healthy control subjects were recruited as final participants for the CLBP study. The baseline experiment on facial expressions of pain used the full length of 34 unsegmented trials video captured by frontal-view camera from 17 patients, which are comprised of a total of 317,352 frames (33.3% contained pain). The baseline system tracked 49 inner facial landmarks that were aligned by a similarity transform. Two local appearance descriptors, the local binary pattern (LBP) and discrete cosine transform (DCT), are employed to extract features from local patches with a radius of 10 pixels centered on 30 fiducial points. The two sets of appearance features together with geometric features from the landmark points are fed separately into a linear SVM for binary pain recognition task. The best performance was achieved by the geometric feature, which is measured by *AUC* (.658) and *F1* (.446). Given facial expression as a communicative modality and body behavior as both communicative and protective (Sullivan et al., 2004), the authors further performed a qualitative analysis of the relation between pain-related facial and body expressions.

The results supported the finding that facial expressions of pain (70%) occur in connection with a protective behavior, which may help the pain research community to better understand the relation among movement, exercise, and pain experience.

Werner *et al* (Werner et al., 2013) addressed multi-modality pain detection by introducing the BioVid Heat Pain dataset, which include 90 participants suffering from induced thermode pain in 4 levels. Multi-view face videos were recorded from three synchronized optical cameras in conjunction with depth map from a Kinect sensor. The physiological signals are captured from the skin conductance level (SCL), the electrocardiogram (ECG), the electromyogram (EMG), and the electroencephalogram (EEG). The pain detection experiment was based on only the video data in this study, where data analysis based on physiological data would be conducted in future research. A 6D (3D position and 3D orientation) head pose vector was computed from a 3D point cloud recovered from the Kinect depth map. The face expression was uncoupled from rigid motion by projecting the tracked fiducial points (from 2D video frames) via a pinhole camera model onto the surface of a generic face model in 3D according to the current head pose. Geometric features are defined by 8 distance measures from the fiducial points. Some texture changes, including nasal wrinkles, nasolabial furrows, and eye closure are synthesized with mean gradient magnitudes extracted from local rectangular regions centered at 5 selected fiducial points. The final feature descriptor per frame was in 13D, which is a combination of geometric and appearance features. The temporal features are extracted from the 6D head pose vector and 13D facial expression feature vector within a 6s temporal window. This resulted in a 21D descriptor per signal defined by 7 statistic measures (mean, median, range, standard and median absolute deviation, interquartile and interdecile range) for each of the smoothed signal and its first and second derivatives

(Werner et al., 2013). A binary SVM classifier with radial basis function (RBF) kernel to discriminate pain from no pain was trained for each of the 4 pain levels, and the performance was measured by the F1 metric. Further experiments on the highest pain level data suggested that including head pose features would improve system performance of both person-specific and general pain classifiers. However this finding may be dataset-dependent as a contradictory conclusion was reported by studies (Lucey et al., 2011) using UNBC-McMaster dataset. A literature survey on machine-based pain assessment for infants (Zamzmi et al., 2016) reviewed progress on multimodal tools including behavior-based approaches (e.g. facial expression, crying sound, body movements) and physiological-based approaches (e.g. vital signs, and brain dynamics). However this is outside the scope of our review that is focused on pain experienced primarily by adult patients.

2.4 Evolution with Deep Learning

2.4.1 AFER with Deep Learning

Recent advances in deep learning techniques, especially the convolutional neural network (CNN), have provided an ‘end-to-end’ solution to learn facial action unit labels directly from raw input images, which significantly reduces the dependence on designing functional modules in a conventional CVML framework. The high-level deep features are learned from mass-scale simple features through a pipeline architecture which facilitates developing pose-invariant applications for AU detection. In fact, spontaneous facial expression datasets with large rigid motions that are challenging to conventional AFER studies, for example BP4D and DISFA, have been widely employed in recent deep learning-based research. On the other hand, the structure of multi-class outputs in a deep neural network (DNN) can be conveniently modified at the output layer to handle a multi-label AU detection problem. In this case,

AUs are jointly predicted by taking account for their co-occurrence and dependencies that are insufficiently exploited in previous one-versus-all binary classification settings. However, training a DNN from scratch relies on large labeled training datasets (100K+ samples in (Parkhi et al., 2015)), which is generally not available in most facial expression datasets due to the labor-intensive FACS coding procedure. Therefore, deep learning-based AFER with partially labelled data has also received attention in recent studies. The literature review in this section will focus on the three problems presented above as these are important common problems in APD studies using deep learning.

2.4.1.1 AU Recognition Model with a Deep Learning Architecture

A deep learning-based system generally follows a different paradigm than a conventional CVM-L system in performing the AU detection task, which focuses on three problems consisting of pre-processing, the DNN configuration and post-processing. In pre-processing, the input images are cropped, normalized and aligned with image processing and computer vision tools so that they have a uniform format when applied to the input layer of a DNN. The DNN combines the feature extraction and classification module into a single network, where the architecture, training scheme, and optimization settings need to be specified. Finally the post-processing step determines how the intermediate data from the output layer are interpreted to determine the targeted AUs.

An example of the deep learning-based framework (Gudi et al., 2015) for FACS AU occurrence and intensity estimation was proposed in FG 2015 Facial Expression Recognition and Analysis challenge (FERA 2015). The core architecture of the 7-layer CNN is composed of 3 convolutional layers and a max-pooling layer as feature descriptor and a fully connected layer to provide classification output. In the pre-processing step, faces are detected, aligned and cropped into a fixed size from the input images

followed by a global contrast normalization before they are fed to the CNN. The raw output values from the activation of the output layer are used as confidence scores to determine AU occurrence, where the optimal decision threshold is learned from a validation set based on highest *F1* measure. The proposed deep learning AFER system is trained from scratch on both BP4D and SEMAINE datasets, where the performance for binary AU detection was measured by *F1* (0.522 for BP4D, 0.341 for SEMAINE) and the intensity estimation is evaluated on BP4D only with *MSE* (1.181), *PCC* (0.621) and *ICC* (0.613).

A more extensive combination of features and learning framework is carried out (Tóssér et al., 2016) for frame-level AU detection under large head pose variations. Three types of features are extracted in the pre-processing step, including histogram of oriented gradients (HOG), similarity normalized facial images and a mosaic representation (cut and put together from patches around landmark positions). The AU detection is formulated with either single- or multi-label settings. Four types of network settings based on SVM and a customized small-scale CNN are investigated: 1) a single-label SVM with HOG features, 2) a single-label CNN with the normalized images, 3) a multi-label CNN with the normalized images, and 4) a multi-label CNN with the mosaic features. All the systems are trained on the BP4D dataset from the FERA 2015 challenge and tested on an augmented dataset that is created using 3D information and renderings of faces from BP4D with large pose variations. The performance is evaluated for binary detection task of 11 AUs with *F1* score, skew-normalized *F1* score, and AUC. In general, the single-label CNN selects similarity normalized features yielded the best performance over all four framework settings. Further testing with large head pose variation (0° to -72° for yaw and -36° to 36° for pitch) indicated that the *F1* score is a weak function of pose, which suggests that the deep-learning technique is capable of handling facial data under large pose variations.

Zhou *et al* addressed the AU intensity estimation under large pose variations by developing a multi-task deep network (Zhou et al., 2017) for the sub-challenge of FERA 2017. They investigate a pose-invariant model and a pose-dependant model, which are both built upon the bottom five layers of a pre-trained VGG16 network (Simonyan and Zisserman, 2014) via transfer learning as the feature descriptor. A two-layer fully connected regressor was trained for each AU independently in the pose-invariant model using data from one AU across all poses as the AU intensity estimator. In the pose-dependent model, the data of each AU were categorized into three groups based on the pitch pose. Three pose-dependent regressors were trained jointly with an auxiliary pose estimator under the assumption that the output of pose estimator is equal to the pose ground truth, and the final AU estimates were made from the dot product between the estimated pose vector and the three pose-dependent AU intensity estimates. Both models were trained on the BP4D datasets, and evaluated with the metrics of *RMSE*, *PCC* and *ICC*. The pose-invariant models using appearance features significantly outperform the baseline model in (Valstar et al., 2017) that relies on geometric features. Two likely reasons that the performance is affected are (i) the high failure rate on tracking geometric features under extreme pose change, and (ii) distortion to the facial structure introduced by the warping and alignment operation to the tracked landmarks. Comparison between the pose-dependent model and the pose-invariant model in terms of *ICC* revealed improvement on AU 1, 4, 12, and 17, but no improvement on AU 6, 10, and 14. Hence a mixed model (pose-invariant or pose-dependent) can be selected for each specific AU to further improve the performance.

2.4.1.2 Multi-label AU Detection

In practice, AUs frequently occur in combinations so that a single image frame may carry positive labels from multiple AUs. The one-versus-all binary classification in conventional AFER studies either perform region-specific AU detection or ignore the dependencies among co-occurring AUs. As a result AUs are detected independently in most cases and certain AU combinations are also treated as a single variable for classification. On the other hand, the architecture of a DNN is naturally designed for multi-classification problem, which can be conveniently converted to handle multi-label learning with a proper loss function (e.g. multi-class sigmoid entropy), such that the sum of probabilities from all the output classes are not necessarily equal to one. The multi-label setting combined with deep learning provides improvement in handling imbalanced samples in many spontaneous facial expression datasets. Furthermore, the joint estimation of multiple AUs provides high flexibility in describing various AU combinations, which is advantageous in encoding high-level event like pain or emotions.

The deep region and multi-label (DRML) (Zhao et al., 2016) framework solved the problem of region-based feature extraction and joint estimation of multiple AUs in a unifying deep learning network. The core architecture of DRML is a CNN comprised of five convolutional layers and one max-pooling layer, and the multi-label option is enabled by configuring the output layer as a multi-label sigmoid cross-entropy loss. An additional region layer is attached after the first convolutional layer and the feature map is partitioned from the convolutional layer with a fixed 8 by 8 grid. Each sub-region is processed by the same sub-network architecture formed by a batch normalization (BN) layer with ReLU activation and a convolutional layer. The coefficients of each sub-network in the region layer are updated independently with respect to local features and AU correlations, and the output of the region

layer is a concatenation of all re-weighted patches. However, this settings requires good face alignment such that local features can be consistently captured in the same sub-regions. The entire network is trained from scratch with data from BP4D (12 AUs) and DISFA (8 AUs) respectively and the performance is evaluated with *F1* (0.483 on BP4D, 0.267 on DISFA) and *AUC* (0.560 on BP4D, 0.523 on DISFA). The experimental results indicated that the detection of AUs with higher skewness benefited from multi-label learning, and the region layer also improves the detection rate of most AUs by better exploiting structural information from local facial regions.

Region segmentation paradigm based on geometric features in conventional AFER can also be extended to deep learning-based approach. Li *et al* (Li et al., 2017) built a deep learning frame-level binary AU detection framework for region of interest (ROI) adaptation, multi-label, and LSTM based temporal fusing. Twenty regions of interest (14×14) related to local muscular motions that activate corresponding AUs are generated from and centered around 20 fiducial points derived from the shape model. A VGG16 network is used as the base feature descriptor, where the output from 12th convolutional layer (224×224) was chosen as the feature map and cropped by the pre-defined sub-regions. A CNN-based ROI cropping net (ROI Net) is trained for each individual sub-region as an AU-specific feature descriptor. The AU detection is formulated as a multi-label problem at the output layer to jointly predict all AUs per frame. The static predictions are then fed into a LSTM network to refine the AU estimates using temporal information from the past 24 frames. The proposed framework is trained for 12 AUs in the BP4D dataset and tested on both BP4D and DISFA datasets. Experimental results demonstrate the contributions from all three aspects of region cropping, multi-label, as well as LSTM. The settings with multi-label ROI nets followed by a single layer LSTM yields the best performance based on *F1* score

(0.661 on BP4D, 0.523 on DISFA). The performance improvement over existing DRML approach is attributed to the adaptive way of sub-region selection and the adoption of a deeper pre-trained network (VGG) as base feature descriptor.

Walecki *et al* (Walecki et al., 2017) integrated a conditional graph model to the output of a CNN to learn the co-occurrence structure of AU intensity levels from deep features in a statistical manner. The CNN is comprised of 3 convolutional layers each with RELU activation followed by max-pooling, which is used as a feature descriptor to extract deep features from input image frames. Each image frame in the dataset bears a label vector of all AUs of interest and each entry in a label vector assumes a value from a six-point ordinal scale. The ordinal representation of AU intensities and their pairwise co-occurrence relations are defined by unary (ordinal) and binary (copula) cliques in the output graph that is learned by a Conditional Random Fields (CRF) model. The CRF framework estimates the posterior probability of a label vector conditioned on the deep features in the form of a normalized energy function, which is defined by a set of unary and pairwise potential functions. The parameters of CNN, unary, and pairwise potential functions are then jointly optimized through a 3-step iterative balanced batch learning algorithm. A data augmentation learning approach is also applied to leverage information from multiple datasets, such that a single 'robust' model is trained across multiple datasets instead of training an individual 'weak' model on each dataset. Under the data-augmentation settings, the CNN is trained using data from multiple datasets simultaneously. However different CRF pairwise connections for AU dependencies are learned separately for each dataset, as these datasets may contain non-overlapping AUs with considerable variation in dynamics. The training process of the frame-level based CRF-CNN AU intensity framework involves data from BP4D in FERA 2015, DISFA and UNBC-

McMaster dataset, where the UNBC-McMaster dataset was only used for data augmentation purpose. The settings of CRF-CNN with iterative balanced batch approach and data-augmentation option achieve the best performance, which is measured using *ICC* (0.63 on BP4D, 0.45 on DISFA) and *MAE* (1.23 on BP4D, 0.61 on DISFA). The DISFA dataset benefits from the augmentation yielding a 7% improvement of ICC.

2.4.1.3 AU Detection with Partially Labeled Data

In general, FACS-based data annotation is far more labor-intensive than facial expression data acquisition, such that a complete label assignment to all the training images in a dataset is not always available. By assuming intra-class similarity on feature extraction or a distribution depicting the dependencies among AUs, labels of unannotated images may be inferred using the set of labeled data. Learning with partially labeled data could enrich the feature space by improving utilization of all available data, which is especially beneficial to deep learning-oriented approaches requiring training with extensive data.

In addition to the binary label set $\{-1, 1\}$ that indicates negative samples and positive samples, a 0 label is assigned to an unlabeled sample in the partial label learning problem. Wu *et al* proposed a multi-label learning with missing labels (MLML) framework (Wu et al., 2015) to solve this problem in a multi-label setting, which is applicable to AU recognition. The goal of this method is to learn a mapping that is specified by a set of coefficients between the extracted frame-level feature and the predicted label. The missing labels are handled based on two local-label assumptions. An instance-level smoothness assumes that two frames with similar features should have label vectors that are close, and a class-level smoothness assumed that two classes with close semantic meanings should have similar

instantiations across the whole dataset. The consistency between predicted labels and provided labels is enforced on the MLML objective functions, which is comprised of a Hinge loss and constraints from two smooth matrices generated upon the label smoothness assumptions. The AU classification task is tested on the CK+ dataset with the available portion of labels on four levels ranging from 20% to 100%. The performance is evaluated by average precision (PR) (0.81 – 0.92) and AUC (0.892 – 0.949).

The label smoothness assumption in (Wu et al., 2015) could be invalid as the closeness of two samples in feature space may arise because the subject is same rather than indication of the same AU. Wu *et al* argued that regular spatial and temporal patterns embedded in AU labels can be described probabilistically as a label distribution, and AU labels including the missing ones can be modeled as samples from such underlying AU distributions (Wu et al., 2017). In their method, a deep neural network is used as a feature descriptor to extract high-level features from a raw input image, where only the top hidden layer and the output layer are involved in learning AU label distributions and AU classifiers. A Restricted Boltzmann Machine (RBM) is employed to learn AU distributions with respect to the partially annotated ground truth labels from the bias of AU labels in the output layer, bias of the hidden layer, and the weights between the output layer and the hidden layer. More specifically, the weights between the two layers are assumed to capture the mutually exclusive and co-occurrence dependencies among multiple AUs. Multiple SVMs are used as classifiers for multiple AU classification based on the high-level features. The object loss function is a Hinge loss constrained by the log-likelihood of the learnt AU label distributions. The loss function is modified such that all data are utilized in the training process but only the annotated data contributed to the loss optimization. The two-step procedure first trained the DNN with the partially annotated data in an unsupervised manner and then fine-tuned the parameters of

both DNN and SVM with respect to the object function. The system is first trained and tested with fully annotated data from BP4D and DISFA respectively and minor improvement is reported on the proposed system toward other state-of-the-art methods (e.g. DRML) in term of $F1$ measure. Further testing with partially labeled data also demonstrates that the proposed system outperforms MLML with missing data ratio ranging from 0.1 to 0.5.

2.4.2 Applications of Deep Learning in Pain Detection

Deep learning techniques have also been adopted in recent APD research, with focus on improving feature representation and exploiting temporal dynamics by utilizing corresponding deep neural network architectures. To our best knowledge, the handful of studies on APD involving deep learning are all based on UNBC-McMaster dataset.

An end-to-end deep learning pain detector presented in (Rudovic et al., 2015) is an upgrade of the classic pain detection framework that is based on hand-crafted features. The proposed two-stage deep learning framework is comprised of a convolutional neural network (CNN) and long short-term memory network (LSTM). The CNN takes raw images as input and predicts pain at frame level, which is built upon a pre-trained VGG-16 CNN and served as a baseline system. The feature frames (4096 dimensions) from the fully connected layer (fc6) before the output layer of the CNN are grouped into sequences of length 17, and all the frames are assigned the label from the preceding frame in the sequence. The frame-level features are fed into the LSTM network to improve the pain prediction of a frame by taking into account the past 16 frames. The input images are preprocessed in several settings including: 1) alignment with generalized Procrustes Analysis followed by cropping to a fixed size, 2) removal of background with a binary mask, and 3) frontalization with down sampled shape model (C-APP). The

proposed system is trained on UNBC-McMaster dataset with frame-level pain labels, and the best performance on binary pain detection task ($Accuracy = 90.3\%$, $AUC = 0.913$) is reported by the Aligned Crop scheme combined with exploiting temporal information using LSTM. The author argues that removing background information causes worse performance as the zero background value is converted to nonzero in training CNN. Further test with skew-normalization on the dataset shows that the Accuracy measurement varies significantly with skew in testing data while AUC is less affected, which is in accordance with the findings in (Jeni et al., 2013a).

One obstacle in applying deep learning to the APD problem is that the size of a pain-oriented dataset is too limited to support training a deep neural network from scratch. Egede *et al* (Egede et al., 2017) addressed this issue by combining hand-crafted and deep-learned features to estimate pain intensity as a multi-modality fusing problem that is tailored to small sample settings on the UNBC-McMaster dataset. The deep-learned features are obtained from two pre-trained CNN models for AU detection task on BP4D dataset (Jaiswal and Valstar, 2016), which are associated with the eye and mouth region of the face respectively. A difference descriptor is applied to the local patch sequences with a temporal window of 5 frames to obtain frame-level features in the form of image intensity and binary mask, which are combined and fed as input of CNN. The output of the last fully connected convolution layer (3072D) of each CNN is retrieved and combined into a (6144D) deep-learned feature vector. In addition, a 218D geometric feature vector extracted from the 49-points shape model and 2376D HOG feature vectors extracted from local patches around facial points are employed as hand-crafted features at frame-level. A relevance vector regressor (RVR) is trained for each of the geometric, HOG, and deep features, and a second-level RVR is trained to fuse the RVR predictions from each modality in the first level. All RVR

learners are trained on the frame level ground truth annotated by PSPI. System performance is evaluated by *RMSE* (0.99) and *PCC* (0.67), with the most significant contributions from HOG, followed by geometric feature and least from deep learned feature. The authors attribute these findings to the likelihood that the deep features are fine-tuned to the original AU detection problem in (Jaiswal and Valstar, 2016), whereas the hand-crafted features are more generic in nature. The author further concludes that MSE effectively measures the difference between the predictions and the ground truth, while Pearson's correlation measures how well the prediction follows the ground truth trend. Although a good system performance should possess low RMSE and high Pearson's correlation at the same time, a good rating on either metric is not necessarily correlated with the same good rating of the other.

Liu *et al* (Liu et al., 2017) proposed a direct estimation framework based on the UNBC-McMaster dataset to predict the subjective self-report pain intensity score represented by VAS via hand-crafted personal features and multi-task learning. The proposed DeepFaceLIFT (LIFT = Learning Important Features) framework is a two-stage personalized model, which is comprised of a 4-Layer neural network (NN) with ReLU activation, and Gaussian process (GP) regression models. A fully connected NN is trained on frame-level AAM facial landmark features to predict frame-level VAS score in a weakly supervised manner, where all frames within a sequence carry the sequence VAS label. Personal features (complexion, age, and gender) are also involved in the training stage via three separate settings (Not used, Appended to the 3rd layer of NN, Appended to Input Features), and the multi-task learning is referred to the training data labeled with VAS, and labeled with both VAS and OPI. Instead using a temporal model as in typical APD settings, the authors solve the sequence-level pain estimation as a static problem based on sequence-level statistics. A set of ten statistics (mean, median, minimum,

maximum, variance, 3rd moment, 4th moment, 5th moment, sum of all values, interquartile range) is generated from frame-level predictions of pain score, as these *sufficient statistics* (Hogg and Craig, 1995) capture important information that help to infer unknown parameters in distribution models. The ten-dimensional statistical sequence-level feature vectors are fed into a GP regression model to obtain a personalized estimate of VAS on sequence-level. The system performance is evaluated by *MAE* (2.18) and *ICC* (0.35), where the first stage NN is trained with both VAS and OPI label in conjunction with personal features appended to the 3rd NN layer, and the second stage GP input uses only the predicted VAS feature. This result suggests that using OPI scores from an external observer contribute to estimation of the subjective VAS score by exploiting the benefits of multi-task learning.

In the past decades, significant progress in computer vision and machine learning techniques (CVM-L) has led to the development of increasingly sophisticated systems for Automated Facial Expression Recognition (AFER). There is growing demand for facial expression applications, as a result of which the emphasis of AFER research evolved from posed expressions acquired in a controlled setting to spontaneous expressions captured in a natural setting. The details of AFER progress can be found in survey papers (Bousmalis et al., 2013)(Corneanu et al., 2016)(Zeng et al., 2009). Meanwhile, studies that explore machine learning-based pain interpretation from facial expressions have seen only a limited effort. Ashraf et al. (Ashraf et al., 2009) studied the UNBC-McMaster dataset and proposed the use of Active Appearance Model (AAM) to extract three feature types to train SVM pain classifiers. In their method the sequence-level labels are generated using an average scheme. In their follow-up studies (Lucey et al., 2011)(Lucey et al., 2012), the same set of features is used to train a binary classifier for each single pain-related AU at the frame level while a pain intensity classifier is adopted at the sequence level using

the OPI labels. The temporal dynamics of AUs are investigated by Chen *et al* (Chen et al., 2012) in a simple rule-based method to detect pain of patients suffering from lung cancer in the Wilkie's dataset (Wilkie, 1995). Sikka *et al* (Sikka et al., 2015) proposed a CVML-based model to assess pediatric postoperative pain on a video dataset of neurotypical youth. In their method, a total of 14 single AUs are extracted under 3 statistics to form a 42-dimensional descriptor for each pain event.

There is other research about deep learning-based AFER application on the estimation of pain expression intensity. Wang *et al* extend a face verification network (Wang et al., 2017a) to the expression intensity regression task using a regularized regression loss, and a weighted evaluation metric is proposed to address the issue of imbalanced expression intensity data in the UNBC-McMaster Dataset (Wang et al., 2017b). Majumder *et al* (Majumder et al., 2016) addressed the same imbalanced data issue by introducing over-sampling techniques, and introduced a system based on Gaussian Mixture Regression (GMR) and trained on geometric features to recognize the fifteen different intensity levels in the PSPI scale. Irani *et al* (Irani et al., 2015) adopted spatiotemporal oriented energy features to model facial muscular motions in a pain intensity detection framework on the same dataset, where pain intensities are estimated by a weighted linear combination of the vertical and horizontal motion scores. Neshov *et al* (Neshov and Manolova, 2015) employed a classical CVML framework to detect binary pain with SVMs and solved the problem of continuous pain intensity estimation with linear regression.

2.4.3 Problems in Previous Research

Our research is also built on existing AFER methods and benefits from the progress in deep learning techniques. However, we noted four shortcomings in previous research : (1) AU recognition and pain recognition are treated as separate problems which are handled by an AU classifier and a pain classifier

respectively, without giving sufficient attention to their relationship; (2) pain analysis is more focused on single AU detection rather than pain-relevant AU combinations; (3) Existing research on pain intensity estimation is largely focused on frame-level PSPI, which tends to be benchmarking-driven. As noted by Werner *et al* (Werner et al., 2017): “The 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”. We move one step beyond by investigating the relationship between machine predicted FACS scores of pain expressions and patients’ self-report of pain that is more commonly available in clinical applications in practice.

CHAPTER 3

DECOUPLED AUTOMATED PAIN DETECTION FRAMEWORK

Parts of this chapter have been presented in (Chen et al., 2019) and (Chen et al., 2019). Copyright © 2019, IEEE.

3.1 Research Motivation

An important observation that motivated our research is that existing research on FACS-based automated AU recognition focuses on detection of single AUs. We note that pain-related AUs could occur in conjunction with other AUs to form combinations irrelevant to pain. Therefore inference based only on occurrence of individual AUs is not sufficient for pain identification. While the ground truth of facial expressions and action units in video databases is available at frame level, the ground truth about pain is typically available at sequence level and only via self-report, which is an example of ‘weakly labeled’ data. In early attempts of automated pain analysis, pain is declared to occur in a video if the average output of frame-level pain score exceeds a threshold (Lucey et al., 2009)(Ashraf et al., 2009). However, pain-related frames may constitute a small fraction of all frames in a long video, the averaging paradigm could therefore severely attenuate the signal of interest. Recent research (Sikka et al., 2014) suggests that video-based pain detection can be formulated as a weakly supervised learning problem, which can be effectively handled by a machine learning tool called multiple instance learning (MIL). In this study (Sikka et al., 2014), a binary pain classifier is trained directly via the high-dimensional features extracted from video frames without going through the AU coding procedure. Although encouraging results are reported from experiments on UNBC-McMaster dataset, this setting is vulnerable to perfor-

mance degradation for trans-dataset application, due to interference from person-specific features and demographic variations encoded in the high-dimensional features.

In a commonly used procedure for the manual detection of pain, FACS-certified coders first perform AU coding for every video frame and then infer the sequence-level pain label from the occurrence and frequency of pain-related AU combinations. On the other hand, most existing automated pain detection research involving segment-level pain label preferred to learn a direct mapping between raw features and self-reported pain with a unified machine learning framework, as shown in Figure 2(a). In general, AU occurrence is highly correlated with the appearance of pain in facial expressions, and the reliability of pain detection strongly depends on the accuracy of AU coding. AU coding relies on observable facial muscular movements or facial expressions, whereas pain is more like a latent variable and is not always manifested continuously in facial expressions, especially in the scenario of chronic pain. Facial expressions of pain are more likely to appear when the pain intensity level is high or when intensity of pain changes to a higher level. Due to the sparsity of pain expressions, video captured in clinical settings usually lacks sufficient positive samples to train a reliable pain classifier by learning a direct mapping between high-dimensional facial features and self-report pain labels. On the other hand, texture variations can be sufficiently synthesized by the deep learning technique to learn a reliable mapping between the high-dimensional facial features and the underlying muscular movement corresponding to Action Units.

Inspired by these observations, we proposed a new FACS-based automated pain detection framework comprised of two independent machine learning networks to infer the presence of Action Unit (AU) combinations that signify pain (Chen et al., 2019). The two networks are trained independently

and are linked by novel AU-based data structures that are created by mapping single AU measurements per frame into pain-related low-dimensional feature vectors representing AU combinations, as shown in 2(b). The decoupled architecture of two independent machine learning networks utilizes AU codings as the intermediate results and performs pain analysis in a low-dimensional AU score space, which would be more justifiable and efficient than the direct method. In addition, the decoupled framework alleviates the difficulty in training large scale pain-specific datasets by facilitating data fusion from different pain-labelled video datasets, which helps to develop a robust and generic automated pain analysis system with potentials for analyzing multiple types of pain.

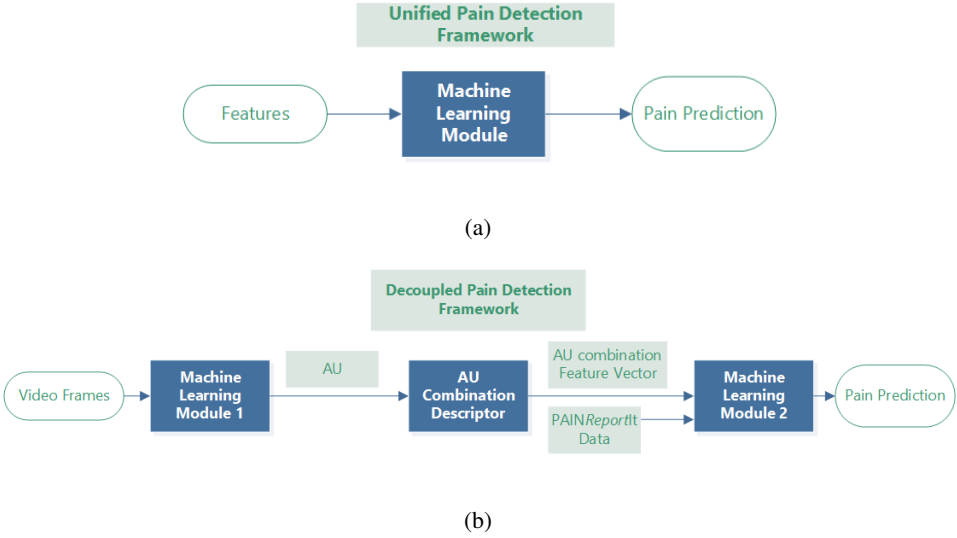


Figure 2: (a) Unified Pain Detection Framework.(b) Proposed Decoupled Pain Detection Framework.

3.2 The Decoupled Framework and An End-to-Front Research Strategy

The proposed automated pain detection system is comprised of a cascade of two machine learning systems: an Automated Facial Expression Recognition (AFER) system that computes frame-level confidence scores for single AUs (shown on the left side of Fig.3) and a MIL system that performs sequence-level pain prediction based on contributions from a pain-relevant set of AU combinations (shown on the right side of Fig.3). MIL is well-suited to handle the ‘weakly labelled’ pain data that can be conveniently represented by a bag of word (BOW) structure, and details will be covered in the following sections. AU combinations are described by two distinct low-dimensional novel feature structures: (i) a compact structure that encodes all AU combinations of interest into a single vector, which is then analyzed in the MIL framework; (ii) a clustered structure that uses a sparse representation to encode each AU combination separately followed by analysis in the multiple clustered instance learning (MCIL) framework, which is an extension of MIL. To our knowledge, this is the first work that applies MCIL in facial expression-related research. Both machine learning systems are trained on the UNBC-McMaster dataset independently with frame-level and sequence-level labels, as illustrated by the three circular blocks in the center of Fig.3.

After an adequate literature survey presented in Chapter 2, we realized recent state-of-the-art generic AFER systems (iMotions A/S, 2016)(McDuff et al., 2013) have proved to be robust to context variations, especially illumination and rigid motions, and have shown good performance on new datasets with different demographics. Meanwhile, little attention has been paid to the problem of learning pain from pain expressions. In order to minimize risk while boosting progress of our research, we divided the development of the decoupled automated pain detection system into three phases, as shown in Fig. 4.

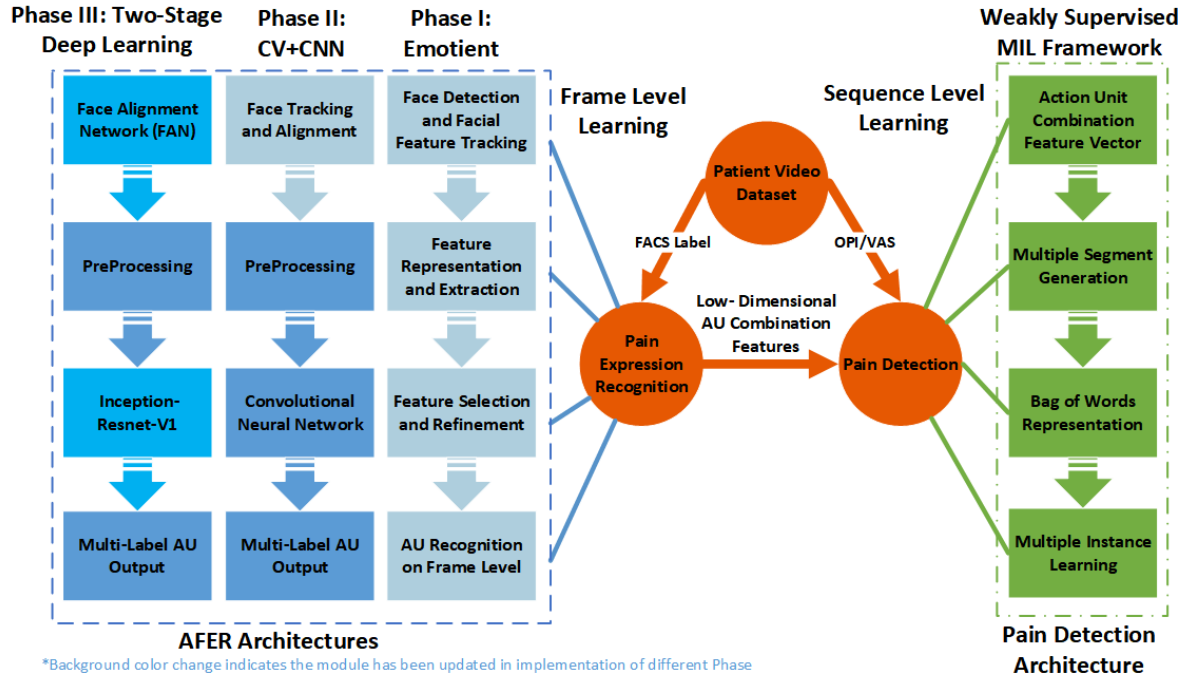


Figure 3: The Decoupled Pain Detection Framework

We start the work of Phase I by focusing on the APD system at the end of the diagram in Fig. 4, by assuming that the problem in preceding blocks has been perfectly solved. Next we move one step backward to improve the performance of AFER with deep learning techniques as the major goal of Phase II task, while we still assume performance of the face alignment is sufficiently reliable. Finally we move one more step backward for Phase III by upgrading the classic face alignment module with a state-of-the-art deep learning network for robust face alignment in both 2D and 3D, and establish an end-to-end multi-stage deep learning based APD system for clinical applications. In contrast to the typical front-

to-end learning process of a CVML framework, we are actually implementing an end-to-front research strategy.

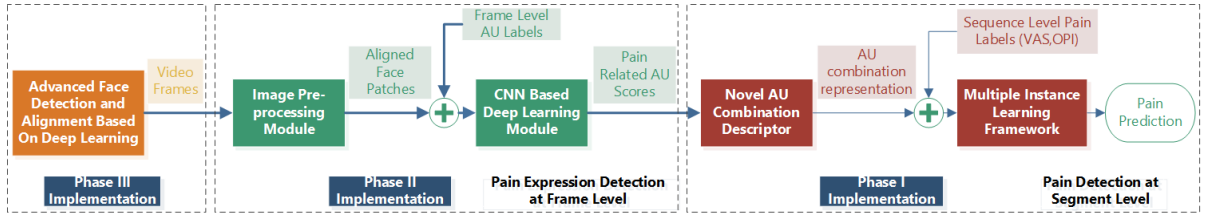


Figure 4: The End-to-Front research strategy

3.3 Automated Facial Expression Recognition

We employed three different AFER architectures corresponding to each phase of system implementation as shown on the left side of Fig.3. The details are discussed in the following subsections.

3.3.1 Conventional CVML Based Emotient

The AFER architecture in Phase I is based on conventional CVML techniques, which can be conveniently described with four key blocks consisting of face detection, feature representation, feature selection, and classification. The first block identifies the face area with a rectangular boundary box in every video frames. The second block aligns the detected face areas and employs various descriptors to extract features from the facial images. The third block selects features most relevant to the non-rigid motions caused by facial expressions, and applies dimension reduction techniques to compress the fea-

ture vector to a size that is tractable for the classifier. The fourth block contains a set of one-versus-all classifiers that are trained on the refined feature vectors for each AU of interest, and the output could be either binary decisions or soft scores that reflect the probability or confidence about the target AUs. Existing research (Bartlett et al., 2006)(Koelstra et al., 2010)(Lucey et al., 2011)(Jiang et al., 2011) on spontaneous facial expression recognition shows AFER systems are highly customizable, and more blocks could be added to this framework to boost performance depending on the application. In the proposed decoupled framework settings, the AFER system can be trained and updated independent of the pain detection system using FACS labels as the ground truth.

One example of such an AFER system is the computer expression recognition toolbox (CERT) (Littlewort et al., 2011), which is the core of the Emotient (iMotions A/S, 2016) system. In the system setup, face detection is based on an extension of the classic Viola-Jones approach. Ten facial feature points are tracked using GentleBoost and the detected face area is aligned to a canonical template patch through an affine warp estimated from the feature positions. A Gabor filter bank is then applied to extract features in 8 directions and 9 spatial frequencies and the filter outputs are concatenated into a single feature vector. The feature vector is then fed into separate linear support vector machines (SVM) for individual AU recognition.

We use Emotient to track and label a set of eight AUs $\{4, 6, 7, 9, 10, 20, 26, 43\}$, that are commonly used in most pain-oriented research. The processing results are represented by the flow of Evidence numbers, where an Evidence number, ranging between -2 to 2 (iMotions A/S, 2016), is assigned to each AU on every frame. The Evidence output for an expression channel represents the odds in logarithm (base 10) scale of a target AU being present. For example: Evidence = 2 ($10^2 = 100$) means the observed

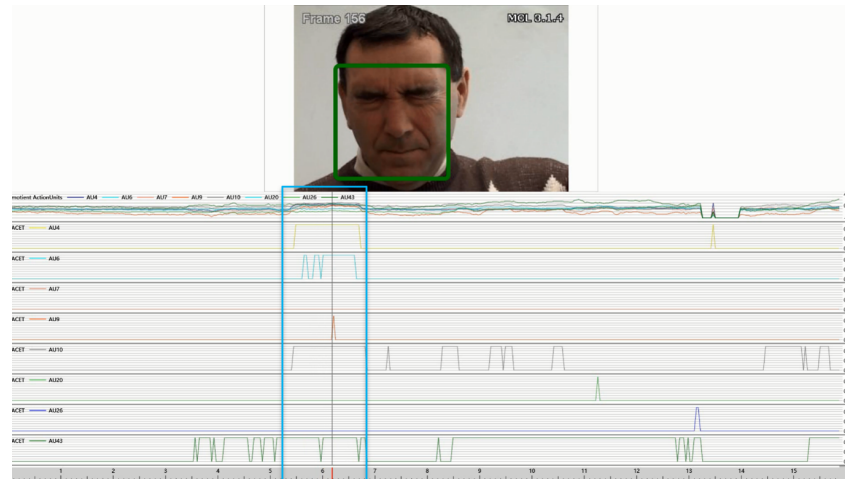
expression is 100 times more likely than not to be categorized by an expert human coder as target AU and Evidence = 0 means the chances that the expression to be categorized by an expert human coder as target AU or not are equal. The Evidence scores can be conveniently transformed to Probability measurements by the following equation:

$$Probability = \frac{1}{1 + 10^{-Evidence}} \quad (3.1)$$

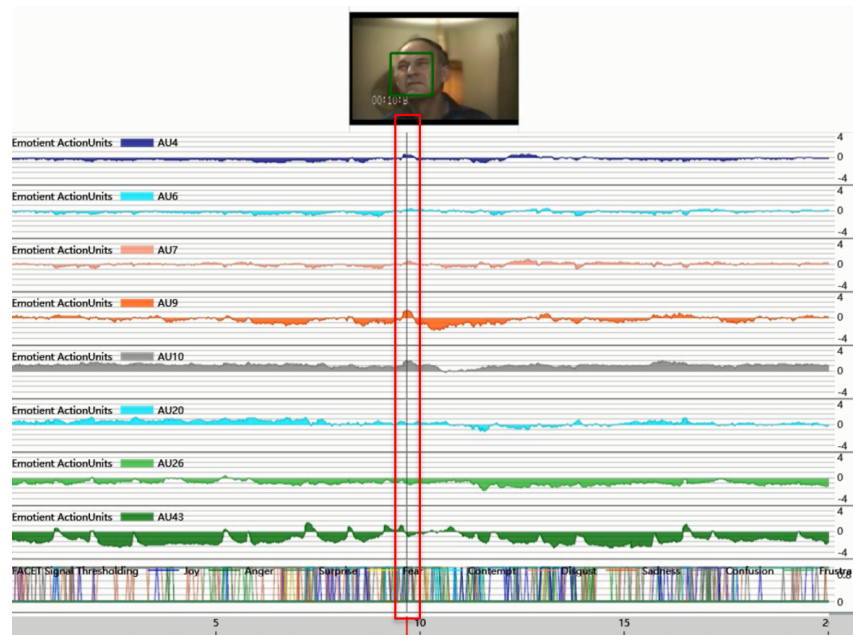
The probability measurements derived from the Evidence scores are used as the features of pain analysis framework. The evidence/probability score profile of every AU can be viewed as a 1D time domain signal. The sample output from Emotient about two patients selected from the UNBC-McMaster dataset and the Wilkie's dataset are presented in Fig. 5

3.3.2 Deep Learning Based Hybrid AFER

In the Phase II study, we employed a convolutional neural network (CNN) architecture (Zhao et al., 2016) to detect pain-related AUs at the frame level, and trained it with combined data from the UNBC-McMaster dataset and the extended Cohn-Kanade dataset (CK+). The CK+ dataset is annotated for a collection of 30 AUs, which not only add positive samples to reduce the data skewness for pain expressions, but also enrich the data diversity with non-pain-related AUs when it is used jointly with the UNBC-McMaster dataset. In observation of the fact that a single frame can be annotated with multiple AUs, the CNN is configured to solve a multi-label problem where multiple AUs are jointly learned as a single classification problem. This multi-label setting will also improve detection of highly skewed AUs in turn by taking AU correlations into account (Zhao et al., 2016). Meanwhile the classic face



(a)



(b)

Figure 5: Sample output from Emotient about (a) a patient from UNBC-McMaster Dataset, (b) a patient from Wilkie's dataset.

tracking and alignment modules are preserved in the Phase II AFER architecture, and we refer this type of implementation as Hybrid CNN AFER.

3.3.2.1 Face Tracking and Alignment

The first block is responsible for reliably detecting faces and aligning a set of facial fiducial points on the detected face area in every video frames. The fiducial points are typically defined along the cues that best describe features of a face, including jawline, brows, eyes, nose, and lips. While the Viola-Jones face detector (Viola et al., 2001) and Active Appearance Models (AAM) (Cootes et al., 2001) are classic tools in many AFER research, recent deep learning based research (Bulat and Tzimiropoulos, 2017) also provides an end-to-end solution to the face tracking and alignment task. It is worth noting that in recent AFER studies, the face detection and alignment modules were less emphasized, as state-of-the-art algorithms are available as off-the-self toolboxes.

3.3.2.2 Preprocessing

In the preprocessing procedure, a video frame is first cropped and centralized to the detected face patches. A 2D-similarity transform is then applied to normalize the in-plane rigid face motions. The background is removed using a convex hull generated from the face alignment result in the following step since it is unrelated to facial expressions. Finally, the resulting face image is resized to 170×170 and is passed to the input layer of the convolutional neural network, as shown in Fig.7.

In Fig. 6 (a) and (e) an example facial image from the CK+ and UNBC-McMaster dataset is shown. The aligned and masked version of this image is shown in Fig. 6 (d) and (h). Deep neural networks are not shift- and rotation-invariant. Therefore, aligning the image is necessary to increase the recognition rate of AUs.

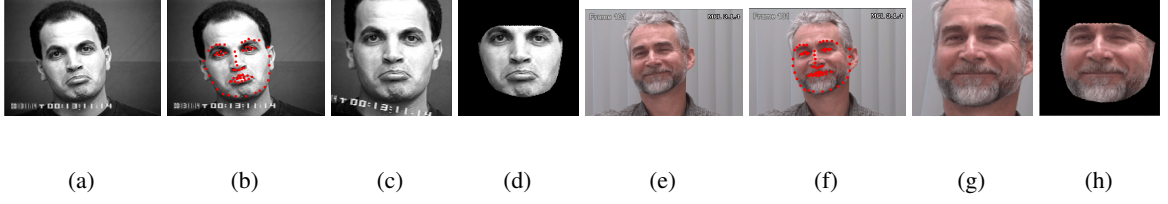


Figure 6: (a) Original image from CK+ dataset. (e) Original image from UNBC-McMaster shoulder pain dataset. (b), (f) Image with landmarks. (c), (g) Aligned image. (d), (h) Aligned and masked image

3.3.2.3 Convolutional Neural Network Architecture

We employed an 8-layer convolutional neural network architecture described in (Zhao et al., 2016) to synthesize deep features from the raw face images as shown in Fig.7. The RGB face images received at the input layer is passed to the first convolutional layer (conv1) comprised of 32 convolutional kernels of size 11×11 , and the output feature map has a size of 160×160 . The layer can be conveniently described by the notation of $32 \times 11 \times 11 @ 160 \times 160$. Note that the depth of the convolutional kernels should match the channel of input feature maps, and the number (channel) of feature maps at the output of a convolutional layer equals to the number of convolutional kernels. The pooling layer (pool2) following conv1 performs max operation over 2×2 spatial neighbourhood with a stride 2 on all the 32 feature maps, which is noted as $32 \times 2 \times 2 @ 80 \times 80$. Another 4 convolutional layers (conv3-conv6) are connected subsequently to the pool2 layer for advanced local feature abstraction. Finally two fully connected layers are employed to explore global correlations from features extracted by the convolutional layer (conv6), which results a 2048D vector as the deep feature representation of the entire face image.

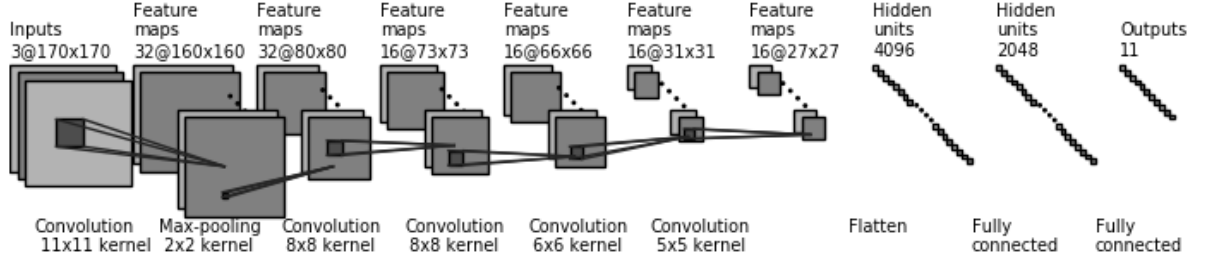


Figure 7: CNN based Automated Facial Expression Recognition Framework

3.3.2.4 Multi-label AU output

Classical deep learning networks (Krizhevsky et al., 2012)(Taigman et al., 2014) are capable of handling classification among thousands of categories, where each positive sample is only labeled for one category. On the other hand, the pain-related action units involved in the facial expression problem is small in number, and each positive sample frame could have multiple AUs activated and labeled. The label for a frame is represented by a $1 \times M$ vector $\{0, 1\}^{1 \times M}$, where M is the number of AUs of interest. Let N_F be the total number of frames, the ground truth matrix Y is given by $Y = \{0, 1\}^{N_F \times M}$, and Y_{nm} is the label of the m th AU in frame n . The loss function in a multi-class CNN is modified for the multi-label case with a multi-label sigmoid cross-entropy loss that is defined as,

$$L(Y, \hat{Y}) = \frac{1}{N_F} \sum_{n=1}^{N_F} \sum_{m=1}^M [Y_{nm} \log \hat{Y}_{nm} + (1 - Y_{nm}) \log (1 - \hat{Y}_{nm})] \quad (3.2)$$

where \hat{Y} is the prediction of Y , which has been normalized to the range $[0, 1]$. We trained the CNN to jointly track a set of 11 AUs $\{4, 6, 7, 9, 10, 12, 20, 25, 26, 27, 43\}$, where AU12 and AU25 are included to help improve dataset imbalance by exploring AU correlations under the multi-label settings.

3.3.3 Deep Learning Based End-to-end AFER

We attempt to further extend the application of deep learning techniques in the Phase III implementation by developing an end-to-end deep learning based AFER that jointly detects the complete set of pain-related AUs (Chen et al., 2019). The proposed system is a two-phase learning model that is comprised of two deep neural networks (DNN) connected sequentially. The first DNN is a face detection module that is used to extract facial landmarks robustly under challenging conditions. The second DNN, is a Convolutional Neural Network (CNN) that is used for AU detection. We refer to this end-to-end deep learning architecture as two-stage DNN AFER.

3.3.3.1 Deep Neural Network For Face Detection and Alignment Network

The first module of our system is responsible for reliably detecting a face in a frame, and accurately tracking a set of facial fiducial points in every video frame. This block is expected to work in close to real time and the tracked fiducial points are used to align the detected face with respect to a frontal-view face with neutral expressions. The aligned face patches are then fed into the deep learning network in the second stage after necessary pre-processing procedures, where we follow the same pre-processing settings as described in Phase II.

We take advantage of the progress made in a state-of-the-art deep learning-based landmark detection method called *Face Alignment Network (FAN)* (Bulat and Tzimiropoulos, 2017), which is capable of handling the face alignment problem in existing 2D and 3D datasets. The FAN network is able to

handle not only ‘traditional’ face alignment challenges including large rigid head motions, initialization and resolutions, but also features for the capability to convert 2D landmark annotations to 3D. The FAN network is pre-trained and ready to use, and the details may be found in (Bulat and Tzimiropoulos, 2017).

The FAN network is used in our study to track 68 fiducial points from an image containing a face, which is comprised of a face detection and a landmark detection parts. The first part is implemented using the *Single-Shot Scale-Invariant Face Detector (S^3FD)*(Zhang et al., 2017), a real-time face detector based on a CNN, which is capable of detecting faces at different scales. The landmark detection of the FAN is based on a stack of four Hourglass networks followed by a hierarchical, parallel and multi-scale convolution block.

3.3.3.2 Deep Neural Network For Multi-Label AU Recognition

In order to extract relevant features for the problem of AU recognition, we used the deep neural network model Inception-ResNet-v1 (Szegedy et al., 2017). The architecture is composed of 24 blocks as shown in Figure 8.

The main part of the model consists of the *inception modules*. These modules apply 1x1, 3x3 and 7x7 convolutions to the input image, and the results are concatenated and the output goes through an *ReLU* activation layer. These modules are interleaved with a *reduction block* whose purpose is to reduce the dimensions of the grid while keeping the number of channels the same. The Inception-ResNet-v1 is a very deep network and has been pre-trained for tasks not directly related with pain expression recognition. Therefore we use the same data from the combination of the UNBC-McMaster and CK+ datasets to fine-tune the Inception-ResNet-v1 network for AU recognition via transfer learning.

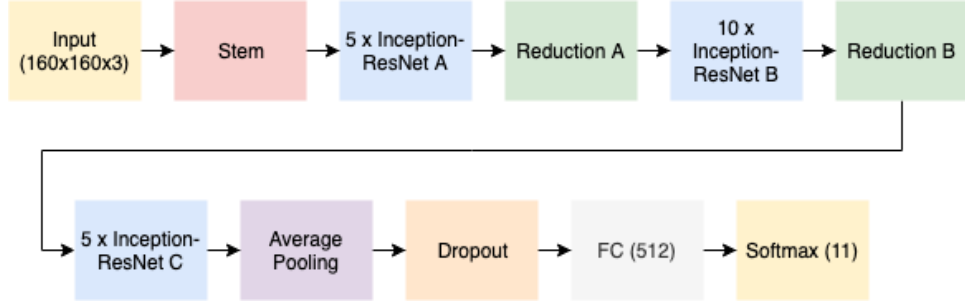


Figure 8: The Inception-Resnet-v1 architecture schema

The bottleneck of the architecture consists of an *average pooling* layer, followed by a *dropout* layer and two fully connected layers whose dimensions are 512 and 11, respectively. We have a total of 11 classes in this network. Our goal is to recognize the 9 AU's described in Fig.1. In addition we included two other classes corresponding to neutral expressions and another class representing all the other AUs of the FACS. Since some AUs co-occur frequently, we model the problem as that of multi-label prediction. In order to do this, we use the *softmax* activation function for the last dense layer and the model is trained using the multi-label cross entropy loss, defined as:

$$L_E(y, \hat{y}) = -\frac{1}{L} \sum_{l=1}^L 1[y > 0] \log \hat{y}_l + 1[y < 0] \log (1 - \hat{y}_l) \quad (3.3)$$

where L is the number of labels, and $1[x]$ is the indicator function returning 1 if x is true and 0 otherwise.

3.4 Action Unit Combination Encoding

3.4.1 Compact Structure Vs. Clustered Structure

In practice, Action Units 6/7 are the most frequently observed pain-related AUs in FACS coding (Prkachin, 2009). Multiple AU combinations could also be activated in a video segment for pain evaluation. When we define feature vectors based on AU combinations, it is important to have a comprehensive characterization including individual AU contribution, activation of individual AU combination, as well as correlation among activation of multiple AU combinations. The measurement for AU combinations can be conveniently derived from the single AU predictions. Two different feature vector structures based on AU combination scores, which are referred to as compact or clustered, are proposed as follows,

Compact Structure: Let $A = \{4, 6, 7, 9, 10, 20, 26, 43\}$ be the set of single pain-related AUs. The AU combination feature vector for frame i of a video sequence S is an 11-dimensional column vector

$$v_i = [P_{AU_6} P_{AU_7} P_{AU_{20}} P_{AU_{4 \oplus 6}} P_{AU_{4 \oplus 7}} P_{AU_{4 \oplus 43}} \\ P_{AU_{4 \oplus 9}} P_{AU_{4 \oplus 10}} P_{AU_{4 \oplus 26}} P_{AU_{9 \oplus 26}} P_{AU_{10 \oplus 26}}]^T$$

in which each entry of the feature vector is the probability estimate of corresponding AU or AU combination and the \oplus operator denotes the co-occurrence of AUs. In our method, we have chosen to assign the probability of combination $AU_{i \oplus j}$ as the smaller of P_{AU_i} and P_{AU_j} so that $P_{AU_{i \oplus j}} = \min(P_{AU_i}, P_{AU_j}), \forall i, j \in A$. As a result, the pain information about frame i is conveyed by the probability measurement of all pain-related AU combinations that are compressed in a single low-dimensional vector, as shown in Fig. 9(a).

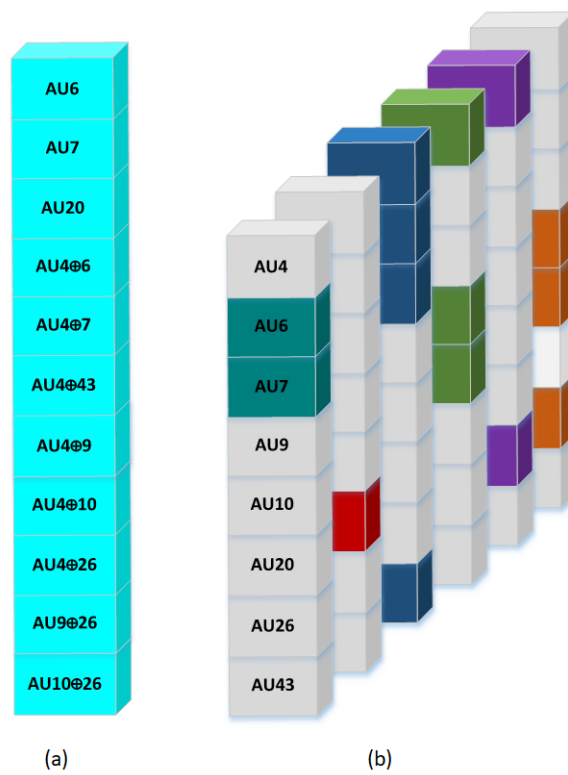


Figure 9: AU combination structure: (a)Compact Structure, (b)Clustered Structure

Clustered Structure: Here we follow Wilkie’s coding strategies (Wilkie, 1995) to group the 11 pain-related single AU and AU combinations into clusters according to two criteria, 1) there is a common AU shared by the combinations in the cluster and 2) the AU combinations within a cluster are actuated in adjacent area on the face. Six clusters are formed in this way, including $\{AU_{6/7}\}$, $\{AU_{20}\}$, $\{AU_{4\oplus 6/7/43}\}$, $\{AU_{4\oplus 9/10}\}$, $\{AU_{4\oplus 26}\}$ and $\{AU_{9/10\oplus 26}\}$. The AU combination feature representation of a video frame i under the clustered structure is composed of an 8×6 matrix, where the column j is highlighted in Fig. 9(b) by all the single AUs involved in the combinations belonging to cluster j , where $j \in \{1, 2, 3, 4, 5, 6\}$. The non-zero entries in column j are the probability measurements of the highlighted AUs in cluster j and all the remaining entries of the feature matrix are set to zero, which results in a sparse representation of features, as shown in Fig. 9(b).

3.4.2 Bag of Words representation

A patient’s self-report is still the golden rule for pain evaluation in patient care. A pain label is commonly available for video segment but not for every single frame. Such a situation is encountered frequently in computer vision since it is easier to obtain group labels for the data rather than individual labels, and is known as ‘weakly supervised’ learning problem. On the other hand, although temporal dynamics of spontaneous facial expressions have good intra-dataset consistency, it could vary significantly under clinical settings depending on the level and type of pain a patient is suffering. While AUs evoked by acute pain usually last less than a second, those evoked by chronic cancer pain could last for minutes. Hence conventional temporal modeling with a fixed moving window or a preset duration parameter (Lucey et al., 2009) (Koelstra et al., 2010) is inadequate to handle practical applications. To

address the challenges from weakly labeled data and complicated temporal dynamics, we employ the bag of word (BOW) representation as suggested in (Sikka et al., 2012)(Sikka et al., 2014).

A video sequence S_i in a dataset can be represented as a bag that contains a number of segments generated from S_i . The bag is defined as $\{s_{ij}\}_{j=1}^{N_i}$, where s_{ij} is the j th segment in the bag containing N_{ij} frames i.e. $s_{ij} = \{f_i^{m_j}, f_i^{m_j+1}, \dots, f_i^{N_{ij}-m_j+1}\}$. s_{ij} contains only contiguous frames and N_{ij} is the total number of frames in s_{ij} taken from S_i and $f_i^{m_j}$ is the m_j th frame in S_i and the 1st frame in s_{ij} . The bags are then associated with the label of sequence S_i as $B = \{S_i, y_i\}_{i=1}^N$, $y_i \in \{0, 1\}$, which defines two kinds of bags, positive and negative. A positive bag contains at least one positive instance, while a negative bag contains no positive instances (Sikka et al., 2014). Adopting this representation for the pain detection problem, a positive bag refers to a video sequence that contains pain-related facial expressions, and a negative bag refers to a video sequence that does not contain any pain-related facial expression. In practice, pain-related AU temporal segments occupy only a small portion of the entire video sequence. The sparsity of positive training samples fits well in the context of BOW structure, which is another motivation to adopt this type of data structure. It takes three steps, **S1-S3** as described below, to generate a BOW representation from the feature space.

(S1) Feature Extraction at Frame Level: Define a mapping $\phi_F : R^{m \times n} \rightarrow R^d$ as the feature extraction process on frame level that maps a frame of size $m \times n$ in image space to a d -dimensional feature vector, where $d = 11$ for the case of the compact structure. In the case of the clustered structure, the mapping is defined as $\phi_F : R^{m \times n} \rightarrow R^D$, where R^D refers to the space of 8×6 sparse feature matrices. Feature vectors are typically of very high dimension in existing unified framework. However, R^d or

R^D in our case is simply the low-dimensional AU combination feature vector space in the proposed decoupled pain detection framework.

(S2) Multiple Segment Generation: The instances in a bag are video segments containing consecutive frames belonging to the sequence. The boundary of each segment can be generated conveniently in two ways. A typical way is to run overlapping temporal scanning windows at multiple scales known as Sc-Wind. A parallel way is clustering the frames in a sequence using normalized cuts (Ncuts). Each element of the weight matrix of Ncut algorithm is obtained by a similarity measure between frames f_i^u and f_i^v in sequence i measured by

$$W_i(u, v) = \exp\left(-\left|\frac{\phi_F(f_i^u) - \phi_F(f_i^v)}{\sigma_f}\right|^2\right) + \exp\left(-\left|\frac{t_u - t_v}{\sigma_t}\right|^2\right) \quad (3.4)$$

where t_u refers to frame index of f_i^u , and σ_f and σ_t are constants selected for feature domain and time domain respectively. Details of Ncuts are provided in (Shi and Malik, 2000) (Sikka et al., 2014).

(S3) Feature Representation at Segment Level: The feature representation of a video segment is denoted by the mapping $\phi_S : S \rightarrow R^d$ that transform video segment in sequence space S to a d -dimensional feature vector for the compact case, and $\phi_S : S \rightarrow R^D$ for the clustered case accordingly. This mapping is specified by a max-pooling strategy from the feature representation of all the frames in the segment as:

$$\phi_S(s_{ij}) = \max_k (\phi_F(f_i^k) | f_i^k \in s_{ij}) \quad (3.5)$$

The instance in a bag is now represented by a single feature vector with the same dimension d as the frame-level feature vector. After associating the pain label to the bag, a multiple instance learning (MIL) framework can be trained on the BOW data for automated pain detection.

3.5 Weakly Supervised Automated Pain Detection

3.5.1 Multiple Instance Learning

The general idea for solving machine learning problems is to establish a classifier and optimize it with respect to a loss function. Viola *et al.* (Viola et al., 2005) first solved the MIL problem with a boosting framework, which is known as MILboost, and discussed its application in object detection from images. In this section, we give a brief overview of MILboost and how it can be customized for pain detection. The decision on the presence of pain is based on the probability of bags being positive. The posterior probabilities of bags and instances are defined as:

$$p_i = \mathcal{P}(y_i = 1 | S_i) \quad (3.6)$$

$$p_{ij} = \mathcal{P}(y_i = 1 | s_{ij}) \quad (3.7)$$

The only available ground truth is the label of the bag, and all the instances in a bag carry the same label as the bag.

To help readers follow the derivation, we summarize the key annotations in Table II. The posterior probability of a segment may be extended to all the frames in it by placing a Hamming window centered within the segment. While each frame in a sequence may be members of multiple segments in practice,

TABLE II: KEY ANNOTATIONS IN THE MIL DERIVATION

Annotation	Description	Level
S_i	Bag i	Sequence
p_i	Probability of Bag i	Sequence
s_{ij}	Instance j in Bag i	Segment
$\phi_S(s_{ij})$	Feature Representation of s_{ij}	Segment
p_{ij}	Probability of Segment s_{ij}	Segment
GM	General Mean Softmax	Segment
$h(s_{ij})$	Weak Classifier trained on Segments	Segment
$H(s_{ij})$	Strong Classifier constructed by combing Weak Classifiers via Boosting	Segment
f_i^k	Frame k in Bag i	Frame
$\phi_F(f_i^k)$	Feature Representation of f_i^k	Frame

we assign the maximum scores from all these segments as the posterior probability of the frame (Sikka et al., 2014), i.e.

$$p_{f_k^i} = \mathcal{P}(y_i = 1 | f_k^i) = \max_j (\omega(s_{ij}) * p_{ij} | f_k^i \in S_{ij}) \quad (3.8)$$

A classifier $H_T : R^d \rightarrow R$ is trained on the feature vectors of instances, and a posterior probability is assigned to each instance based on the classifier output, *s.t.*

$$p_{ij} = \sigma(H_T(\phi_s(s_{ij}))) \quad (3.9)$$

where $\sigma()$ is a sigmoid function *s.t.*

$$\sigma(x) = \frac{1}{1 + \exp(-x)}, \forall x \in R \quad (3.10)$$

The loss function is defined by negative log-likelihood, which is the same as that used in the logistic regression problem:

$$\mathcal{L} = - \sum_i^N (y_i * p_i + (1 - y_i)(1 - p_i)) \quad (3.11)$$

Since a positive bag contains at least one positive instance, the probability p_i of a bag being positive depends on the probability of the instance that is most likely to be classified as positive, i.e.

$$p_i = \max_j (p_{ij}) \quad (3.12)$$

The MILboost uses the boosting procedure to construct a strong classifier $H_T(s_{ij})$ by iteratively combining a set of weak classifiers $h_t(s_{ij})$ as,

$$H_T(s_{ij}) = \sum_{t=1}^T \alpha_t h_t(s_{ij}) \quad (3.13)$$

where H_T denotes the classifier constructed in the T^{th} iteration and all weak classifiers are the same type of learners that are generated from space \mathcal{H} . Note $H_T(s_{ij})$ and $h_t(s_{ij})$ are simplified notations of $H_T(\phi_s(s_{ij}))$ and $h_t(\phi_s(s_{ij}))$ respectively. The boosting algorithm updates the weight of instances at the end of each iteration by taking the gradient of the loss function \mathcal{L} w.r.t $H_T(s_{ij})$. A weak classifier h_t is added for each iteration that gives the lowest error on the weighted data, and the coefficient α_t is determined by the error of the classifier. Details about the MILboost algorithm can be found in (Viola et al., 2005).

$$\omega_{ij} = -\frac{\partial \mathcal{L}}{\partial H_T(s_{ij})} = -\frac{\partial \mathcal{L}}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial H_T(s_{ij})} \quad (3.14)$$

The instance weights are then normalized as $\omega'_{ij} = \frac{|\omega_{ij}|}{\sum_{ij} |\omega_{ij}|}$. Misclassified instances will be assigned higher weights and a weak classifier $h_{T+1}(s_{ij})$ is trained on the reweighted data and added to H_{T+1} in the $(T+1)^{th}$ iteration

$$h_{T+1} = \arg \max_{h \in \mathcal{H}} \sum_i^N \sum_j^{N_i} \omega'_{ij} h(s_{ij}) \quad (3.15)$$

where N is the total number of bags and N_i is the number of instances in the i th bag. However, since the max function is not differentiable, a soft-max function g is used as an approximation. Define a vector

$\mathbf{p}_i = [p_{i1} p_{i2} \dots p_{iN_i}]$, then the softmax function with respect to the subscript j , $j \in \{1, 2, \dots, N_i\}$ is given by:

$$p_i = g(\mathbf{p}_i) \approx \max_j(p_{ij}) \quad (3.16)$$

Among all the options for soft-max function, the generalized mean (GM) is a preferable model as suggested by past research (Sikka et al., 2014). For the instances $\{s_{ij}\}_{j=1}^{N_i}$ in the bag of S_i , the GM approximation is given by:

$$g_{GM}(\mathbf{p}_i) = \left(\frac{1}{N_i} \sum_j p_{ij}^u \right)^{\frac{1}{u}} \quad (3.17)$$

where u is the parameter that controls sharpness and accuracy in the GM model *s.t.* $g_{GM}(\mathbf{p}_i) \rightarrow \max_j(p_{ij})$ as $u \rightarrow \infty$. Now the gradient of the GM soft-max is given by:

$$\frac{\partial p_i}{\partial p_{ij}} = p_i \frac{p_{ij}^{(u-1)}}{\sum_{s=1}^{N_i} p_{is}^u} \quad (3.18)$$

3.5.2 Multiple Clustered Instance Learning

In the compact structure feature settings, scores of all AU combinations are encoded in one feature vector, which can be conveniently handled by the original MIL framework. However, it may be desirable to distinguish the contribution of individual AU combinations for more precise analysis of different types of pain. In practice, the clustered representation is a more natural way that is used by human coders and a positive decision on any of the clusters is sufficient to identify pain. Multiple Clustered Instance Learning (MCIL) proposed by Xu *et al.* (Xu et al., 2014) is an extension of MIL that was proposed to provide patch-level clustering of 4 subclasses of cancer tissues, and facilitates both image-level

classification and pixel level segmentation (cancer vs. non-cancer). Based on the structural similarity of the problems, we have adapted the MCIL framework to handle the clustered feature structure for pain recognition.

MCIL assumes there are K clusters in a positive bag and an associated hidden variable $y_{ij}^k \in \{0, 1\}$ that indicates whether the instance s_{ij} belongs to the k th cluster. An instance could be considered positive if it belongs to one of the K clusters and a bag is labeled as positive bag only if it contains at least one positive instance. The goal of MCIL is to learn one boosting classifier $H_T^k(s_{ij}^k)$ for each of the K clusters. In our clustered data representation settings, each video frame is encoded by a 8×6 matrix. If we treat each column vector as an independent instance, all the instances in one bag will form six clusters naturally. A MCIL learner can be trained and the overall decision is based on the cluster classifier that gives maximum output:

$$H_T(s_{ij}) = \max_k (H_T^k(s_{ij})) \quad (3.19)$$

Similar to the core of MIL, the posterior probability of bag i is given by,

$$p_i = \max_j \max_k (p_{ij}^k) \quad (3.20)$$

where $j \in \{1, 2, \dots, N_i\}$ and $k \in \{1, 2, \dots, K\}$.

Define a vector $\mathbf{p}_i^K = [p_{i1}^1 \dots p_{iN_i}^1 p_{i1}^2 \dots p_{iN_i}^2 \dots p_{i1}^K \dots p_{iN_i}^K]$. The max function is approximated by a soft-max function,

$$p_i = g^K(\mathbf{p}_i^K) \quad (3.21)$$

Taking the GM model as an example, derivation of the softmax function for the care of K clusters goes as follows,

$$\begin{aligned}
 g_{GM}^K(\mathbf{p}_i^K) &= \left(\frac{1}{K} \sum_k (g_{GM}([p_{i1}^k \ p_{i2}^k \ \dots \ p_{iN_i}^k])) \right)^{\frac{1}{u}} \\
 &= \left(\frac{1}{K} \sum_k \left(\left(\frac{1}{N_i} \sum_j (p_{ij}^k)^u \right)^{\frac{1}{u}} \right)^u \right)^{\frac{1}{u}} \\
 &= \left(\frac{1}{KN_i} \sum_{k,j} (p_{ij}^k)^u \right)^{\frac{1}{u}}
 \end{aligned} \tag{3.22}$$

Finding the strong classifier of a cluster follows a standard boosting procedure and all the classifiers are trained on the same set of BOW instances. However, the weight of instances are updated per cluster,

$$\omega_{ij}^k = -\frac{\partial \mathcal{L}}{\partial H_T^k(s_{ij})} = -\frac{\partial \mathcal{L}}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}^k} \frac{\partial p_{ij}^k}{\partial H_T^k(s_{ij})} \tag{3.23}$$

The partial derivative of $\frac{\partial p_i}{\partial p_{ij}^k}$ for the GM model is given by,

$$\frac{\partial p_i}{\partial p_{ij}^k} = p_i \frac{(p_{ij}^k)^{(u-1)}}{\sum_{s=1}^{N_i} \sum_{t=1}^K (p_{is}^t)^u} \tag{3.24}$$

For the remaining two terms in the partial derivative of the weight update expression, $\frac{\partial \mathcal{L}}{\partial p_i}$ is the same as

in MILboost and $\frac{\partial p_{ij}^k}{\partial H_T^k(s_{ij})} = p_{ij}^k(1 - p_{ij}^k)$, which is the derivative *w.r.t* a sigmoid function.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 Frame-level Pain-Related Action Units Prediction on UNBC-McMaster Dataset

4.1.1 Dataset Skew

Data imbalance is a significant challenge for system training using spontaneous facial expression dataset, where the number of positive training examples is often much smaller than that of negative examples for action unit detection. The imbalance of data is measured by the skew ratio, as defined in (Jeni et al., 2013a):

$$Skew = \frac{\text{negative examples}}{\text{positive examples}} \quad (4.1)$$

Common procedures to address this issue include employing over-sampling techniques (Majumder et al., 2016) or weighted metrics (Wang et al., 2017b). The UNBC-McMaster dataset is only encoded for 11 pain-related AUs and the AU data is highly imbalanced with skew varying from 4.86 to 2243. In this study, we use the CK+ as an auxiliary dataset, which contains generic facial expressions and is encoded for a complete set of 30 AUs. Data consolidation from CK+ to UNBC-McMaster dataset not only augmented available positive examples for each AU classes, but also introduced non-pain-related AUs as negative examples to increase the system robustness. As a result, the AU skew ratio after consolidation varies from 5.58 to 68.86, and the dataset is more balanced across all pain-related AUs, as illustrated in Table III , column 2-3.

4.1.2 AFER Performance Evaluation

Accuracy and *F1* score are frequently used metrics to evaluate a binary classifier (Chen et al., 2018). *Accuracy* is the percentage of correctly classified positive and negative samples, where the decision threshold is set to 0.5 in general. *F1* score is the harmonic mean between precision (*PR*) and recall (*RC*), so that the two measures are reflected through one score. We evaluated the Hybrid CNN AFER and the Two-Stage DNN AFER on the UNBC-McMaster dataset with a 5-fold validation, and the result is reported by *Accuracy* and *F1*. The performance metric is normalized with dataset skew as per the methods in (Jeni et al., 2013a). The Multi-Label AU predictions on two sample frames is shown in 10. We also employ the commercialized Emotient (iMotions A/S, 2016) for performance comparison with the multi-label deep learning networks on the pain-related AU detection task. Although tracking results are not available for AU12, AU25, and AU27 using Emotient, this does not affect the performance of other AUs as Emotient treats the prediction of single AUs independently. The results are summarized in TableIII, column 4-9, which indicate the deep learning-based techniques outperform state-of-the-art AFER system built with conventional CVML techniques. In addition, the performance of the trained-from-scratch CNN in small scale does not differ significantly from state-of-the-art very deep Inception-Resnet fine-tuned with transfer learning. This finding is also supported by a survey on deep learning-based facial expression recognition (Li and Deng, 2018). Note that the two-stage DNN was tested with a setting (Chen et al., 2019) that is slightly different from those the Hybrid CNN, so the result is not a strict like-for-like comparison. Taking advantage of the block-based design of the entire system, it is preferable to use the FAN face alignment network as the front end of the 10-layer CNN in the Hybrid CNN schema at the current stage of research, because it allows a better control of the entire system.

Unless stated otherwise, the core AFER system in the following experiments is based on the 10-layer CNN architecture.



Action Units	Descriptions		
None		0.0005	0.0027
4	Eyebrow Lowerer	0.9989	0.9132
6	Cheek Raiser	0.9998	0.0056
7	Eye Lid Tightener	0.9999	0.0041
9	Nose Wrinkler	0.9956	0.0044
10	Upper Lip Raiser	0.0107	0.0078
20	Lip Stretcher	0.0038	0.0067
26	Jaw Drop	0.0022	0.0097
27	Mouth Stretch	0.0019	0.0120
43	Eyes Closed	0.9991	0.0060

Figure 10: Multi Label AU Predictions on Sample Frames

4.1.3 AU Relations

Concurrence rate is a common measure for evaluating the relations between two AUs. We evaluate this relation as the chance that AU_i is activated given AU_j is activated, *s.t.*

$$P(AU_i|AU_j) = \frac{\text{number of } AU_i \& AU_j \text{ activated examples}}{\text{number of } AU_j \text{ activated examples}} \quad (4.2)$$

TABLE III: AFER PERFORMANCE ON AUGMENTED UNBC-MCMaster

Action	Orig.	Comb.	<i>F1-Score</i>			<i>Accuracy</i>		
Unit	<i>Skew</i>	<i>Skew</i>	Emo- tient	Hybrid CNN	2-Stage DNN	Emo- tient	Hybrid CNN	2-Stage DNN
4	36.61	12.97	0.743	0.872	0.874	0.838	0.967	0.977
6	6.27	7.03	0.802	0.910	0.922	0.736	0.954	0.980
7	11.00	10.3	0.557	0.865	0.900	0.849	0.970	0.978
9	94.50	41.5	0.715	0.825	0.725	0.935	0.928	0.984
10	75.95	68.86	0.610	0.850	0.681	0.598	0.939	0.986
12	4.86	5.59	-	0.903	-	-	0.951	-
20	56.2	30.7	0.370	0.848	0.785	0.882	0.937	0.981
25	15.78	7.7	-	0.810	-	-	0.919	-
26	18.30	19.9	0.374	0.901	0.833	0.931	0.935	0.980
27	2243	54.6	-	0.839	0.710	-	0.861	0.984
43	15.60	21.68	0.670	0.901	0.851	0.798	0.957	0.985

(a) Ground Truth AU Relations											
AU4	1	0.58	0.3	0.18	0.16	0.41	0.026	0.24	0.073	0	0.7
AU6	0.11	1	0.38	0.068	0.084	0.83	0.054	0.15	0.087	0	0.18
AU7	0.097	0.62	1	0.071	0.056	0.61	0.059	0.15	0.1	0.0039	0.21
AU9	0.45	0.9	0.56	1	0.39	0.57	0.052	0.46	0.069	0	0.67
AU10	0.32	0.89	0.36	0.31	1	0.55	0.13	0.62	0.18	0	0.41
AU12	0.064	0.67	0.3	0.035	0.042	1	0.02	0.1	0.077	0	0.11
AU20	0.04	0.42	0.28	0.031	0.099	0.2	1	0.4	0.19	0	0.11
AU25	0.11	0.35	0.21	0.081	0.14	0.3	0.12	1	0.2	0	0.16
AU26	0.037	0.23	0.17	0.014	0.046	0.25	0.065	0.23	1	0	0.26
AU27	0	0	0.72	0	0	0	0	0	0	1	1
AU43	0.31	0.41	0.29	0.12	0.088	0.31	0.031	0.16	0.22	0.0074	1
	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU27	AU43
(b) Deep-learned AU Relations											
AU4	1	0.56	0.32	0.19	0.18	0.35	0.041	0.24	0.08	0	0.68
AU6	0.12	1	0.4	0.072	0.095	0.82	0.075	0.16	0.09	0	0.17
AU7	0.098	0.55	1	0.057	0.046	0.53	0.066	0.14	0.092	0.0032	0.19
AU9	0.52	0.9	0.51	1	0.41	0.62	0.018	0.42	0.042	0	0.72
AU10	0.4	0.96	0.34	0.33	1	0.48	0.14	0.65	0.22	0	0.44
AU12	0.06	0.65	0.3	0.04	0.037	1	0.025	0.11	0.069	0.00028	0.092
AU20	0.068	0.56	0.36	0.011	0.1	0.24	1	0.36	0.18	0	0.094
AU25	0.12	0.35	0.22	0.073	0.14	0.29	0.1	1	0.17	0	0.17
AU26	0.051	0.26	0.19	0.0097	0.064	0.25	0.069	0.23	1	0	0.29
AU27	0	0	1	0	0	0.15	0	0	0	1	1
AU43	0.33	0.39	0.31	0.13	0.097	0.26	0.028	0.17	0.22	0.0051	1
	AU4	AU6	AU7	AU9	AU10	AU12	AU20	AU25	AU26	AU27	AU43

Figure 11: AU relations in UNBC McMaster (a)Ground Truth, (b)Predictions

where AU_i and AU_j are in the pain-related AU sets. The ground truth of AU relations for UNBC-McMaster dataset is summarized in Fig.11(a), and the pairwise conditional probabilities computed from the CNN predictions is summarized in Fig.11(b). Note that the matrices are not symmetric due to the different skewness for each AU, and the relation of AU27 is not fully revealed as there is a very limited number of AU27 samples in UNBC-McMaster dataset. The deep-learned AU relations match the ground truth statistics precisely, with an average element-wise Euclidean distance between the two matrices in Fig.11 at 0.04. The information of AU relations is embedded in the CNN training through the multi-label

setting, which improves AU predictions at frame level. In addition, exploring such relations through AU combinations also helps to learn the correlations between pain expressions and self-reported pain of patients.

4.2 Sequence-level Pain Detection In UNBC-McMaster Datasets

4.2.1 Correlations Between Pain Expression And Self-Report Pain

The UNBC-McMaster dataset provides two types of labels for the study of pain: the pain expression label at frame level (PSPI), and the patients' self-report pain label at sequence level (VAS, OPI). In order to evaluate the correlation between these two types of ground truth on pain, we first divided sequences in this dataset into 3 categories, namely No Pain, Weak Pain and Strong Pain, based on the maximum frame-level PSPI score in each sequence similar to the settings in (Irani et al., 2015). Next, the PSPI-OPI relations were plotted in a stacked bar chart for each OPI level as shown in Fig. 12, where the portion of sequences in each PSPI category was marked in blue, orange, and yellow for the cases of No Pain, Weak Pain and Strong Pain respectively. Pain expression with higher intensity is observed with increasing level of pain feeling in general. However, we do notice that pain expressions are less likely to be evoked by low-level pain ($OPI = 1, 2$). Past research (Ashraf et al., 2009)(Lucey et al., 2008) uses the protocol that treats video sequences with OPI ratings ≥ 3 as positive samples and those with $OPI = 0$ are treated as negative samples. This yielded the same set of 147 sequences from 23 subjects as in (Sikka et al., 2014).

4.2.2 Evaluation Of The MIL Based Pain Detection

In this study, we also trained and validated the pain detection system on the UNBC-McMaster dataset. Video sequences are first fed frame-by-frame into the deep learning-based AFER to generate

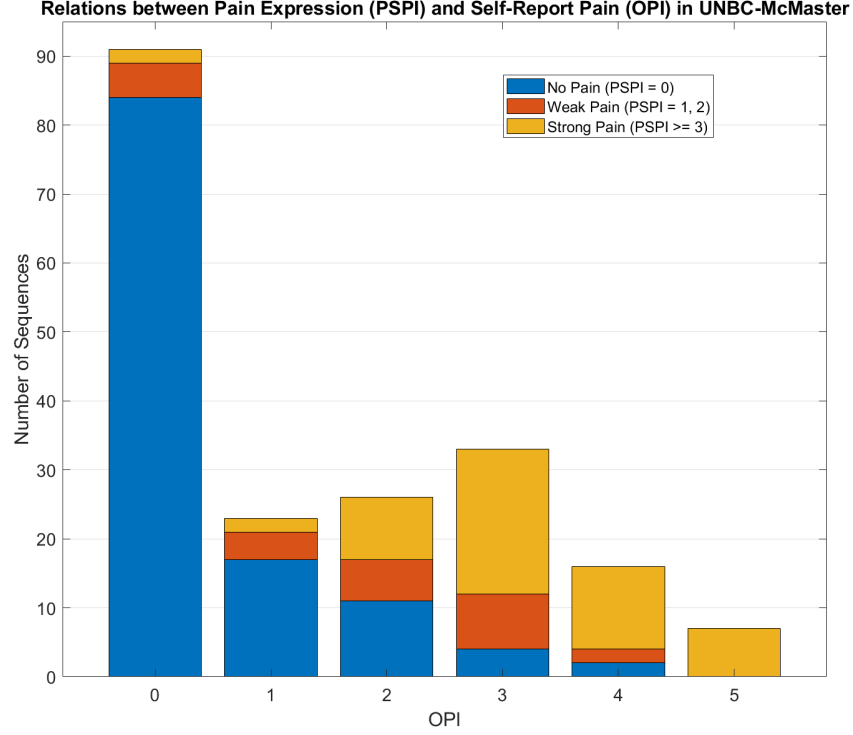


Figure 12: Relations between frame-level and sequence-level ground truth on pain

predictions on pain-related AUs. Dataflows of the jointly detected single AUs are then encoded with the compact and clustered AU combination structures separately. Finally, these two types of BOW features about AU combinations are used to train MIL (referred to as Compact-MIL) and MCIL (referred to as Clustered-MCIL) models. Instances in a bag are generated by two temporal aggregation methods, Sc-wind and Ncuts. We set the multiple scaling window size at 30,40,50 for Sc-wind. The size of segment in a cluster is limited between 21 and 81 for Ncuts, and $\sigma_f = 1$, $\sigma_t = 30$ in computing the frame correlation matrix, where the parameter values are selected via a grid search. In order to illustrate

the flexibility of the decoupled framework as well as its benefit by utilizing deep learning techniques, we also replaced the CNN-based AFER with Emotient and report the performance for comparison.

Four metrics are chosen to evaluate the performance of MIL and MCIL learners, including *Accuracy*, *Accuracy-EER*, *F1* score and Area Under the receiver operating Characteristic (*AUC*). Here *Accuracy-EER* refers to the *Accuracy* computed at Equal Error Rate (*EER*) in the Receiver Operation Curve (*ROC*) (Ashraf et al., 2009)(Lucey et al., 2008), and *AUC* metric equals to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

TABLE IV: COMPARISON OF THE DECOUPLED FRAMEWORK WITH MS-MIL (Sikka et al., 2014)

Pain Label	Framework Configuration		Sc-wind				Ncuts			
	AFER	APD	<i>Accuracy</i> (%)	<i>Accuracy</i> (%) at <i>EER</i>	<i>AUC</i>	<i>F1</i>	<i>Accuracy</i> (%)	<i>Accuracy</i> (%) at <i>EER</i>	<i>AUC</i>	<i>F1</i>
$OPI \geq 3$		MS-MIL(Sikka et al., 2014)		83.7				82.99		
	Emotient	Compact-MIL	83.96	83.93	0.875	0.776	85.04	85.7	0.9	0.783
	Emotient	Clustered-MCIL	85.18	85.6	0.92	0.824	86.84	88.3	0.936	0.836
	CNN	Compact-MIL	87.07	87.5	0.914	0.854	88.43	89.47	0.907	0.857
	CNN	Clustered-MCIL	89.11	89.47	0.924	0.862	90.47	91.23	0.948	0.876
$OPI \geq 2$	Emotient	Clustered-MCIL	81.18	82.50	0.888	0.809	80.39	81.67	0.876	0.796
	CNN	Clustered-MCIL	87.05	87.5	0.894	0.869	88.23	88.75	0.907	0.875

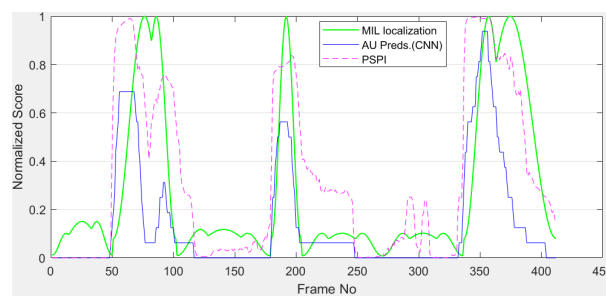
We first select video sequences with OPI ratings ≥ 3 as positive samples (i.e. examples with $OPI = 1, 2$ are excluded), and evaluate the performance of MIL and MCIL learners with two segmentation

scheme referred to as Sc-Wind and Ncuts. As part of the flexibility feature of the decoupled framework, we also report the performance using Emotient and the CNN network respectively as the frontend to the APD system, which makes a total of 8 different configurations. We consider the study in (Sikka et al., 2014), referred to as MS-MIL, for comparison and the results based on a 10-fold cross-validation are summarized in Table IV, Row 3-7. In general, Clustered-MCIL outperforms Compact-MIL as a pain detection system. This improvement may be attributed to the feature sparsity from the clustered representation, which not only increases the margin on features but also follows more naturally a human coder's decisions. On the other hand, the decoupled framework configurations using CNN as the frontend outperform their Emotient-based counterparts. The best performance is achieved by the MCIL framework using Ncuts in conjunction with the CNN based AFER. This observation is further demonstrated through a more challenging pain detection experiment by including video sequences with OPI=2 as the positive examples, as shown in Table IV, Row 8-9. The accuracy of pain-related AU encoding is increased by introducing deep learning based AFER, which leads to the overall improvement on pain detection at sequence level using the weakly supervised MIL framework.

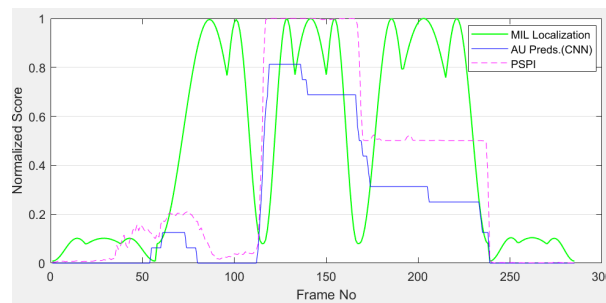
AU coding is a 'hard' problem involving learning a mapping from a very high-dimensional pixel space to the low-dimensional AU space in a complex environment. However, due to the fact that AUs are evoked by certain facial muscular movement, reliable AU coding can be achieved by a robust AFER system if trained on sufficient data. On the other hand, pain is a subjective measure and can be viewed as a latent variable. It will be easier to synthesize similarity from low-dimensional features that are highly correlated to pain under the impact of problem uncertainty. It is worth pointing out that the selected data

has $skew = 1.607$ for $OPI \geq 3$ and $skew = 1.05$ for $OPI \geq 2$ respectively with sequence-level binary pain label in contrast to the highly imbalanced FACS labels on frame-level data.

4.2.3 Multi-label AU Predictions And PSPI



(a)



(b)

Figure 13: Frame Level Pain Prediction on Two Sample Video Sequences

Pain expression intensity estimation using the PSPI scale and pain localization based on temporal analysis are two topics covered in past research (Wang et al., 2017b)(Majumder et al., 2016)(Egede et

al., 2017)(Sikka et al., 2014). Although this study focuses on modeling relations between patients’ self-report pain and observable pain expressions, we are able to address both problems to certain extent. The AFER system in the proposed decoupled framework is trained on binary AU labels. However, the summation of corresponding AUs from its output may be used to simulate the PSPI scale at frame level. We plot the ground truth of PSPI vs. its estimation using the single AU predictions from CNN, and the posterior probability $p_{f_i^k}$ derived from the MIL predictions for two sample video sequences in Fig. 13(a),(b). All three types of scores have been normalized between 0 and 1 for the comparison purpose. The results indicate that using the AFER output directly is more efficient than using MIL predictions for pain localization task in terms of the precision to follow the PSPI variation both in time and amplitude.

4.3 Trans-dataset Validation For Human Coders With Wilkie’s Dataset

4.3.1 Pain Expression Evaluation Via AU Combinations

Wilkie’s dataset was created in 1990 with videos captured by analog cameras, and was not targeted for automated analysis. The issues of video quality, illuminations and rigid motions pose significant challenges to build a robust AFER system. The highly imbalanced dataset with insufficient pain expression samples also prevent training an end-to-end APD system using this dataset alone with most available state-of-the-art CVML techniques. We solved this problem with the decoupled structure in the proposed framework, where person-specific features are isolated from the pain prediction task. The decoupled framework also facilitates consolidation of newly acquired video data with patients’ self report to train a more robust pain detector. Results of the experiment on the UNBC-McMaster datasets suggest that identifying pain from reliable predictions of pain expression improved performance compared with that obtained with high-dimensional features extracted directly from video frames. We applied the

TABLE V: VALIDATION WITH SELECTED PATIENTS FROM WILKIE DATASET

Patient No (Segment 1,2,3)	Human Coder Scored AU Combinations	System Scored AU Combinations
P21	6/7	6/7 , 20, 4+6/7/43
P24	6/7 , 20, 27	6/7 , 4+6/7/43, 4+9/10
P27	6/7 , 20	6/7
P41	6/7 , 4+6/7/43, 4+9/10	6/7 , 4+6/7/43
	4+26/27, 9/10+26	4+26/27
P44	6/7 , 20	6/7 , 20, 4+6/7/43
	4+6/7/43	4+9/10, 4+26/27

deep learning-based AFER that was trained on the combined data from the UNBC-McMaster and CK+ datasets to perform a trans-dataset validation of pain expressions on Wilkie’s dataset.

Our previous study on Wilkie’s dataset (Wilkie, 1995) supports the findings that 1) types of facial display in people with cancer pain are similar to those with acute and chronic pain; and 2) pain intensity was weakly correlated ($r = 0.30$, $p < 0.05$) and state anxiety was moderately correlated ($r = 0.45$, $p < 0.05$) with more than one action unit present in an interval. In addition, there was strong support for the notion that people with cancer pain are stoic in their facial expression of pain. Such stoicism may contribute to undertreatment of cancer pain. Therefore, we employ the Wilkie’s metric based on the

occurrence of AU combinations instead of the PSPI metric based on intensity of pain expression in this study.

Each test sequence in Wilkie's dataset is divided into 30 subsequences and each subsequence has a duration of 20 seconds. Three human expert coders performed FACS coding on each subsequence and pain is identified if any pain-related AU combination is detected by at least two human expert coders in the original research. The FACS coders recorded the observed pain-related AU combinations in a score sheet, and a sample score sheet is shown in Fig.14. However, the AU combinations were scored at subsequence level such that frame-level FACS label is not available. We selected the first three subsequences of videos from 5 patients (P21,P24,P27,P41 and P44) for validation purpose, where the patients displayed close-to-front-view faces. The face images of the 5 selected patients are shown in Fig.15. The evaluation is based on comparison of pain-related AU combinations scored by FACS coders (ground truth) and that scored by our system in the video of 1-minute duration. The result is summarized in Table V, which indicates the proposed system followed FACS coders' decisions well in general.

4.3.2 Consistency Evaluation Between The Automated System And FACS Coders

Next, we employ the Clustered-MCIL settings with GM approximation to train a pain detector on UNBC-McMaster dataset and test it on video sequences from selected patients in Wilkie's dataset. 393 subsequences from 27 patients are selected for this experiment, where at least 50% of each subsequence is analyzable by Emotient. A subsequence is considered as positive (pain) if AU combinations are coded by at least two coders (more credible) or only one coder (less credible). A subsequence is negative (no pain) if no AU combination is scored by any coder. As a result, 82 subsequences are identified as positive

SCORE SHEET -- LONG CANCER FACIAL PAIN EXPRESSIONS

Code #: P17 Date: 1-22-93 Scored By: SDI

Interval

PACS Action Units:	1	2	3	4	5	6	7	8	9	10	11
Pacial Expressions											
6/7		✓	✓	✓	✓	X	✓	✓	✓	✓	✓
20						✓					X
27											
4 + 6/7/43	X	X		✓	X	✓			X		
4 + 9/10											
4 + 26/27						✓					
9/10 + 26											
9/10 + 27											
None	✓		X		✓	X	X	X		✓	X

Figure 14: A sample score sheet filled by human coders for pain analysis

by at least two coders, 121 subsequences are identified as positive by only one coder, and the remaining 190 subsequences are identified as negative samples. However, since the human coding does not reveal pain intensity information and coders do not always agree with each other, this ground truth is more suitable for qualitative analysis. The automated system is used as an independent coder and it checks the consistency between machine prediction and human coder decisions and the results are summarized in Table VI. Note that the system has no prior knowledge about the test dataset. In general, observe that the decisions of automated system are highly correlated with that of a majority of human coders on both pain and no pain videos. Additionally, a 68.6% consistency rate is observed between the system and the case where a single coder scored pain. However this is more likely due to the ambiguity in less

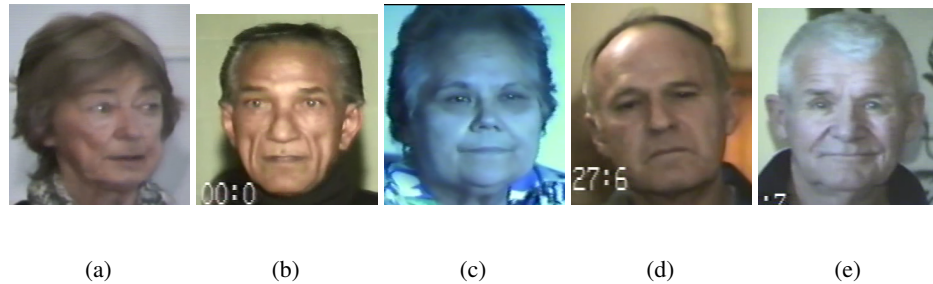


Figure 15: Selected patients for test (a) P21, (b) P24, (c) P27, (d) P41 (e) P44

credible videos rather than to the accuracy degradation. As a result, the system shows its potential in pain assessment for long videos and may be applicable to patient monitoring task under clinical settings.

Patients' self report pain intensity was collected via *PAINReportIt*[®] (Wilkie et al., 2003) after video recording using the VAS scale in Wilkie's dataset, which ranges from 0-9.8. Therefore, each video sequence carries only one VAS label for the entire 10mins video, which is insufficient to support training an automated system to detect pain intensities at the current stage of our study. Considering the variation in pain intensity distribution of different datasets as well as of different pain types, we believe binary pain detection is a good starting point to evaluate patients' self-report pain in the context of trans-dataset learning. The preliminary binary pain detection experiment on Wilkie's dataset using trans-dataset learning paves the way for the task of automated detection of cancer pain. We will advance our study with more focus on problems in clinical settings, including the influence of stoicism on pain expression intensity for pain estimation.

TABLE VI: COMPARISON OF MACHINE PREDICTION WITH HUMAN CODER DECISION ON SUBSEQUENCES IN WILKIE DATASET (Wilkie, 1995)

Subsequences Label	Pain (2+ coders scored)	Pain (1 coder scored)	No Pain (3 coders agreed)
Human Coder	82	121	190
Automated System	68	83	169
Consistency Rate	82.9%	68.6%	88.9%

CHAPTER 5

DISCUSSION AND FUTURE WORK

A highly customizable FACS automated pain detection system is proposed based on a decoupled structure. In our design, the end-to-end pain detection from patients' videos are decoupled into two consecutive tasks, namely FACS based pain express detection and pain estimation from the estimated pain expressions. The system for each task is developed independently, which maximizes utilization of existing data and facilitates fusion of new data collected in the future. The customizable design allows us to take advantage of its flexible configuration and try various combinations of state-of-the-art CVML and deep learning techniques to tailor the system for our need. We followed an end-to-front research strategy by dividing the system development and optimization into three phases, which allow us to focus on the high priority design goals in each stage while minimizing risks. However, given the very limited accessibility to pain-oriented video data containing patients' facial expressions, the space to further adapt the system to clinical application is limited in practice. With the recently approved NIH grant 1R43DA046973-01, we are proceeding to develop a commercialized product based on the decoupled APD system, and four targets have been identified for the future work.

First, we will integrate all the modules in the decoupled framework to a user interface (UI). The UI is expected to provide convenient visualizations of intermediate results from the APD system, which is important for the prototype demonstration and future product deployment.

Second, although patients' self-report pain is typically represented by OPI and VAS, the *PAINReportIt*[®] tool provides much richer description about pain associated with each patient, such as Pain Rating Index

(of Sensory, Affective, Evaluative), Number of Words Selected (NWS), comparable pain experience (worst pain intensity for toothache the patient has had). We will integrate the text information from *PAINReportIt*® as an important auxiliary input to the final stage of the APD system to reinforce the bond between pain expressions and self-report pain. To our best knowledge, there is no peer research addressing this problem so far.

Third, if our phase II proposal of the SBIR project is granted, we will establish a new pain-expression dataset in clinical settings with specific considerations of data acquisition protocol, multi-camera deployment, FACS labeling and *PAINReportIt*® recording etc, which is oriented to automated video analysis. The new dataset is expected to enrich the availability of pain data, and will further improve our proposed system via data fusion.

Finally, one of the most critical challenges in facial expression recognition is the performance impairment from person-specific features. The decoupled structure is capable of isolating person-specific features from the decision of pain, because pain detection is based on low-dimensional features of AU combinations. This conclusion is build on the premise that the predictions of pain expression must also be reliable. Therefore we are also keen on seeking a way to isolate person-specific features from the feature space of facial expressions. As part of our preliminary work, we investigated a novel 3D face tracking and alignment system called *OptiTrack*® that captures face fiducial points in real time with IR camera arrays, as shown in Fig.16. The output of *OptiTrack*® known as control points flows (as shown in Fig.17a) is fed in the *AutoDesk SoftImage*® software to drive an avatar to replicate human facial expressions (as shown in Fig.17b). Now that the facial expressions are produced from the identical subject, all person-specific features are suppressed. The *OptiTrack*® system cannot be applied to existing

facial expression datasets and the system settings are also complicated so that we did not pursue this idea initially. However, with the FAN framework, we are able to reliably track fiducial points in 3D even from existing 2D videos. Therefore it is very plausible for us to integrate the Avatar-based AFER in our decoupled system.

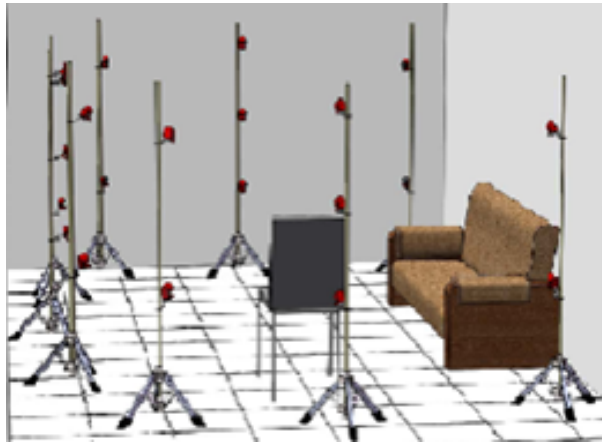


Figure 16: OptiTrack IR camera arrays setup

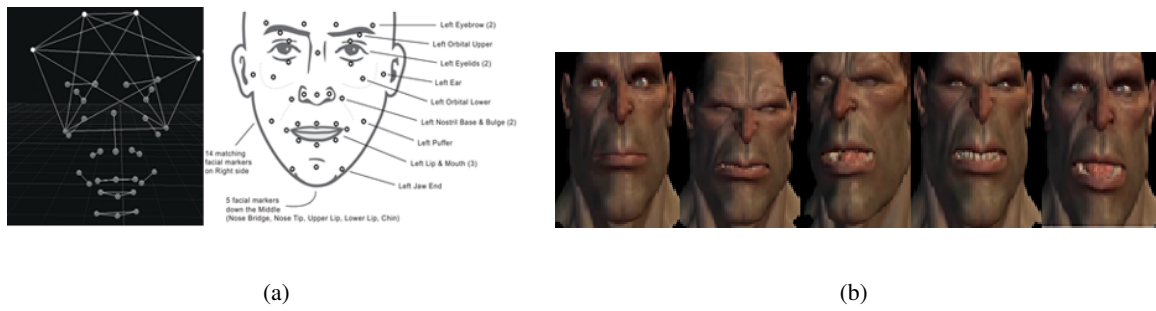


Figure 17: (a) SoftImage compatible 33 control points layout in Optitrack. (b) Facial expression replication by Avatar.

CHAPTER 6

CONCLUSION

This thesis proposed and investigated the performance of automated pain detection via spontaneous facial expressions in the context of clinical applications. The work was motivated to a large extent by finding solutions for automated pain analysis of Wilkie’s dataset. By conducting an adequate literature survey of state-of-the-art research, we are able to identify the most relevant techniques and challenges for the automated pain detection problem. We proposed a decoupled pain detection framework that mimics the decision strategy of FACS-certified human coders. The architecture of two independent machine learning networks facilitates data fusion from different pain-oriented video datasets, which helps to develop a robust and generic automated pain analysis system. We followed an end-to-front research strategy to develop three different implementations of the decoupled framework, each corresponding to a research phase. The proposed system not only demonstrates improvement over existing state-of-the-art work, but also shows flexibility on framework customization. In future work, we will conduct comprehensive test on new pain-oriented video dataset and we expect to commercialize our work to facilitate clinical pain analysis.

APPENDIX

COPYRIGHT PERMISSIONS

This appendix presents the copyright permissions for the articles, whose contents were utilized in our thesis. The list of the articles include several conference and symposium publications: IEEE Global Conference on Signal and Information Processing (GlobalSIP'19) (Chen et al., 2019), Medical Imaging 2012: Image Processing (SPIE 2012) (Chen et al., 2012). and a paper in IEEE Transactions on Affective Computing (Chen et al., 2019). The copyright permissions for reusing the published materials are presented in this Appendix. Note that IEEE does not require individuals working on a thesis to obtain a formal reuse license.

APPENDIX (Continued)



RightsLink



Home

Help

Email Support

Sign in

Create Account



Learning Pain from Action Unit Combinations: A Weakly Supervised Approach via Multiple Instance Learning

Author: Zhanli Chen

Publication: Affective Computing, IEEE Transactions on

Publisher: IEEE

Date: Dec 31, 1969

Copyright © 1969, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis online.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

APPENDIX (Continued)



Pain Detection from Facial Videos Using Two-Stage Deep Learning

Conference Proceedings:

2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)

Author: Guglielmo Menchetti

Publisher: IEEE

Date: Nov. 2019

Copyright © 2019, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

- 1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
- 2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
- 3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

- 1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
- 2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
- 3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK

CLOSE

CITED LITERATURE

- [Arif-Rahu and Grap, 2010] Arif-Rahu, M. and Grap, M. J.: Facial expression and pain in the critically ill non-communicative patient: state of science review. Intensive and critical care nursing, 26(6):343–352, 2010.
- [Ashraf et al. , 2009] Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., and Solomon, P. E.: The painful face–pain expression recognition using active appearance models. Image and vision computing, 27(12):1788–1796, 2009.
- [Asthana et al. , 2013] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M.: Robust discriminative response map fitting with constrained local models. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 3444–3451. IEEE, 2013.
- [Asthana et al. , 2014] Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M.: Incremental face alignment in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1859–1866, 2014.
- [Aung et al. ,] Aung, M., Kaltwang, S., Tyler, N., Watson, P., Williams, A., Pantic, M., Berthouze, N., Romera-Paredes, B., Martinez, B., Singh, A., et al.: The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. IEEE transactions on affective computing, (1):1–1.
- [Baker and Matthews, 2004] Baker, S. and Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International journal of computer vision, 56(3):221–255, 2004.
- [Bänziger and Scherer, 2010] Bänziger, T. and Scherer, K. R.: Introducing the geneva multimodal emotion portrayal (gemep) corpus. Blueprint for affective computing: A sourcebook, pages 271–294, 2010.
- [Bartlett et al. , 2006] Bartlett, M. S., Littlewort, G., Frank, M. G., Lainscsek, C., Fasel, I. R., and Movellan, J. R.: Automatic recognition of facial actions in spontaneous expressions. Journal of multimedia, 1(6):22–35, 2006.
- [Bassili, 1979] Bassili, J. N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of personality and social psychology, 37(11):2049, 1979.

- [Bousmalis et al. , 2013] Bousmalis, K., Mehu, M., and Pantic, M.: Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. Image and Vision Computing, 31(2):203–221, 2013.
- [Bulat and Tzimiropoulos, 2017] Bulat, A. and Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision, pages 1021–1030, 2017.
- [Chen et al. , 2018] Chen, Z., Ansari, R., and Wilkie, D.: Automated pain detection from facial expressions using facs: A review. arXiv preprint arXiv:1811.07988, 2018.
- [Chen et al. , 2019] Chen, Z., Ansari, R., and Wilkie, D.: Learning pain from action unit combinations: A weakly supervised approach via multiple instance learning. IEEE Transactions on Affective Computing, 2019.
- [Chen et al. , 2012] Chen, Z., Ansari, R., and Wilkie, D. J.: Automated detection of pain from facial expressions: a rule-based approach using aam. In SPIE Medical Imaging, pages 83143O–83143O. International Society for Optics and Photonics, 2012.
- [Chen et al. , 2019] Chen, Z., Guglielmo, M., Ansari, R., Wilkie, D., Cetin, E., and Yardimci, Y.: Pain detection from facial videos using Two-Stage deep learning. In 2019 IEEE Global Conference on Signal and Information Processing(GlobalSIP 2019), 2019.
- [Chetverikov and Péteri, 2005] Chetverikov, D. and Péteri, R.: A brief survey of dynamic texture description and recognition. In Computer Recognition Systems, pages 17–26. Springer, 2005.
- [Chew et al. , 2012] Chew, S. W., Lucey, P., Lucey, S., Saragih, J., Cohn, J. F., Matthews, I., and Sridharan, S.: In the pursuit of effective affective computing: The relationship between features and registration. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(4):1006–1016, 2012.
- [Chu et al. , 2017] Chu, W.-S., De la Torre, F., and Cohn, J. F.: Selective transfer machine for personalized facial expression analysis. IEEE transactions on pattern analysis and machine intelligence, 39(3):529–545, 2017.
- [Cootes et al. , 2001] Cootes, T. F., Edwards, G. J., and Taylor, C. J.: Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6):681–685, 2001.
- [Corneanu et al. , 2016] Corneanu, C. A., Simón, M. O., Cohn, J. F., and Guerrero, S. E.: Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-

- related applications. IEEE transactions on pattern analysis and machine intelligence, 38(8):1548–1568, 2016.
- [Craig et al. , 1991] Craig, K. D., Hyde, S. A., and Patrick, C. J.: Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain. Pain, 46(2):161–171, 1991.
- [Craig et al. , 1992] Craig, K. D., Prkachin, K. M., and Grunau, R. V.: The facial expression of pain. Handbook of pain assessment, 2:257–276, 1992.
- [Cristinacce and Cootes, 2008] Cristinacce, D. and Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition, 41(10):3054–3067, 2008.
- [Cristinacce and Cootes, 2006] Cristinacce, D. and Cootes, T. F.: Feature detection and tracking with constrained local models. In Bmvc, volume 1, page 3, 2006.
- [Deyo et al. , 2004] Deyo, K. S., Prkachin, K. M., and Mercer, S. R.: Development of sensitivity to facial expression of pain. Pain, 107(1):16–21, 2004.
- [Dhall et al. , 2011] Dhall, A., Asthana, A., Goecke, R., and Gedeon, T.: Emotion recognition using phog and lpq features. In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 878–883. IEEE, 2011.
- [Ding et al. , 2016] Ding, X., Chu, W.-S., De la Torre, F., Cohn, J. F., and Wang, Q.: Cascade of tasks for facial expression analysis. Image and Vision Computing, 51:36–48, 2016.
- [Egede et al. , 2017] Egede, J., Valstar, M., and Martinez, B.: Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 689–696. IEEE, 2017.
- [Ekman, 2002] Ekman, P.: Facial action coding system (facs). A human face, 2002.
- [Ekman and Friesen, 1969] Ekman, P. and Friesen, W. V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. semiotica, 1(1):49–98, 1969.
- [Ekman and Friesen, 1978] Ekman, P. and Friesen, W. V.: Manual for the facial action coding system. Consulting Psychologists Press, 1978.
- [Fawcett, 2006] Fawcett, T.: An introduction to roc analysis. Pattern recognition letters, 27(8):861–874, 2006.

- [Girard et al. , 2017] Girard, J. M., Chu, W.-S., Jeni, L. A., and Cohn, J. F.: Sayette group formation task (gft) spontaneous facial expression database. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 581–588. IEEE, 2017.
- [Girard et al. , 2015] Girard, J. M., Cohn, J. F., Jeni, L. A., Sayette, M. A., and De la Torre, F.: Spontaneous facial expression in unscripted social interactions can be measured automatically. Behavior research methods, 47(4):1136–1147, 2015.
- [Gudi et al. , 2015] Gudi, A., Tasli, H. E., Den Uyl, T. M., and Maroulis, A.: Deep learning based faces action unit occurrence and intensity estimation. In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, volume 6, pages 1–5. IEEE, 2015.
- [Hadjistavropoulos et al. , 2007] Hadjistavropoulos, T., Herr, K., Turk, D. C., Fine, P. G., Dworkin, R. H., Helme, R., Jackson, K., Parmelee, P. A., Rudy, T. E., Beattie, B. L., et al.: An interdisciplinary expert consensus statement on assessment of pain in older persons. The Clinical journal of pain, 23:S1–S43, 2007.
- [Hamm et al. , 2011] Hamm, J., Kohler, C. G., Gur, R. C., and Verma, R.: Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. Journal of neuroscience methods, 200(2):237–256, 2011.
- [Hammal and Cohn, 2012] Hammal, Z. and Cohn, J. F.: Automatic detection of pain intensity. In Proceedings of the 14th ACM international conference on Multimodal interaction, pages 47–52. ACM, 2012.
- [Hammal and Cohn, 2014] Hammal, Z. and Cohn, J. F.: Towards multimodal pain assessment for research and clinical use. In Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges, pages 13–17. ACM, 2014.
- [Herr et al. , 2006] Herr, K., Coyne, P. J., Key, T., Manworren, R., McCaffery, M., Merkel, S., Pelosi-Kelly, J., and Wild, L.: Pain assessment in the nonverbal patient: position statement with clinical practice recommendations. Pain Management Nursing, 7(2):44–52, 2006.
- [HH, 2000] HH, A.-S.: Challenge of pain in the cognitively impaired. Lancet, 2(356):1867–1868, 2000.
- [Hogg and Craig, 1995] Hogg, R. V. and Craig, A. T.: Introduction to mathematical statistics.(5th edition). Upper Saddle River, New Jersey: Prentice Hall, 1995.

- [Hsu et al. , 2014] Hsu, K.-J., Lin, Y.-Y., and Chuang, Y.-Y.: Augmented multiple instance regression for inferring object contours in bounding boxes. IEEE Transactions on Image Processing, 23(4):1722–1736, 2014.
- [Huang et al. , 2011] Huang, D., Shan, C., Ardabilian, M., Wang, Y., and Chen, L.: Local binary patterns and its application to facial image analysis: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(6):765–781, 2011.
- [iMotions A/S, 2016] iMotions A/S: imotions biometric research platform 6.0. 2016.
- [Irani et al. , 2015] Irani, R., Nasrollahi, K., and Moeslund, T. B.: Pain recognition using spatiotemporal oriented energy of facial muscles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 80–87, 2015.
- [Jaiswal and Valstar, 2016] Jaiswal, S. and Valstar, M.: Deep learning the dynamic appearance and shape of facial action units. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, pages 1–8. IEEE, 2016.
- [Jeni et al. , 2013a] Jeni, L. A., Cohn, J. F., and De La Torre, F.: Facing imbalanced data—recommendations for the use of performance metrics. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pages 245–251. IEEE, 2013.
- [Jeni et al. , 2013b] Jeni, L. A., Girard, J. M., Cohn, J. F., and De La Torre, F.: Continuous au intensity estimation using localized, sparse facial feature space. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), pages 1–7. IEEE, 2013.
- [Jiang et al. , 2011] Jiang, B., Valstar, M. F., and Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 314–321. IEEE, 2011.
- [Kanade et al. , 2000] Kanade, T., Tian, Y., and Cohn, J. F.: Comprehensive database for facial expression analysis. In fg, page 46. IEEE, 2000.
- [Keefe and Block, 1982] Keefe, F. J. and Block, A. R.: Development of an observation method for assessing pain behavior in chronic low back pain patients. Behavior Therapy, 1982.
- [Kendall, 1938] Kendall, M. G.: A new measure of rank correlation. Biometrika, 30(1/2):81–93, 1938.

- [Kim and Pavlovic, 2010] Kim, M. and Pavlovic, V.: Hidden conditional ordinal random fields for sequence classification. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 51–65. Springer, 2010.
- [Koelstra et al. , 2010] Koelstra, S., Pantic, M., and Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE transactions on pattern analysis and machine intelligence, 32(11):1940–1954, 2010.
- [Krizhevsky et al. , 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
- [LeResche, 1982] LeResche, L.: Facial expression in pain: a study of candid photographs. Journal of Nonverbal Behavior, 7(1):46–56, 1982.
- [LeResche and Dworkin, 1984] LeResche, L. and Dworkin, S. F.: Facial expression accompanying pain. Social Science & Medicine, 19(12):1325–1330, 1984.
- [LeResche and Dworkin, 1988] LeResche, L. and Dworkin, S. F.: Facial expressions of pain and emotions in chronic tmd patients. Pain, 35(1):71–78, 1988.
- [Li and Deng, 2018] Li, S. and Deng, W.: Deep facial expression recognition: A survey. arXiv preprint arXiv:1804.08348, 2018.
- [Li et al. , 2017] Li, W., Abtahi, F., and Zhu, Z.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6766–6775. IEEE, 2017.
- [Littlewort et al. , 2011] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M.: The computer expression recognition toolbox (cert). In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 298–305. IEEE, 2011.
- [Littlewort et al. , 2007] Littlewort, G. C., Bartlett, M. S., and Lee, K.: Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In Proceedings of the 9th international conference on Multimodal interfaces, pages 15–21. ACM, 2007.
- [Littlewort et al. , 2009] Littlewort, G. C., Bartlett, M. S., and Lee, K.: Automatic coding of facial expressions displayed during posed and genuine pain. Image and Vision Computing, 27(12):1797–1803, 2009.

- [Liu et al. , 2017] Liu, D., Peng, F., Shea, A., Picard, R., et al.: Deepfacelift: interpretable personalized models for automatic estimation of self-reported pain. arXiv preprint arXiv:1708.04670, 2017.
- [Lucey et al. , 2009] Lucey, P., Cohn, J., Lucey, S., Sridharan, S., and Prkachin, K. M.: Automatically detecting action units from faces of pain: Comparing shape and appearance features. In Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, pages 12–18. IEEE, 2009.
- [Lucey et al. , 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 94–101. IEEE, 2010.
- [Lucey et al. , 2011] Lucey, P., Cohn, J. F., Matthews, I., Lucey, S., Sridharan, S., Howlett, J., and Prkachin, K. M.: Automatically detecting pain in video through facial action units. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(3):664–674, 2011.
- [Lucey et al. , 2012] Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., Chew, S., and Matthews, I.: Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. Image and Vision Computing, 30(3):197–205, 2012.
- [Lucey et al. , 2011] Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., and Matthews, I.: Painful data: The unbc-mcmaster shoulder pain expression archive database. In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 57–64. IEEE, 2011.
- [Lucey et al. , 2008] Lucey, P., Howlett, J., Cohn, J., Lucey, S., Sridharan, S., and Ambadar, Z.: Improving pain recognition through better utilisation of temporal information. In International conference on auditory-visual speech processing, volume 2008, page 167. NIH Public Access, 2008.
- [Lundtoft et al. , 2016] Lundtoft, D. H., Nasrollahi, K., Moeslund, T. B., and Escalera, S.: Spatiotemporal facial super-pixels for pain detection. In International Conference on Articulated Motion and Deformable Objects, pages 34–43. Springer, 2016.
- [Majumder et al. , 2016] Majumder, A., Behera, L., and Subramanian, V. K.: Gmr based pain intensity recognition using imbalanced data handling techniques. In 2016 International Conference on Signal and Information Processing (IconSIP), pages 1–5. IEEE, 2016.
- [Manfredi et al. , 2003] Manfredi, P. L., Breuer, B., Meier, D. E., and Libow, L.: Pain assessment in elderly patients with severe dementia. Journal of Pain and Symptom Management, 25(1):48–52, 2003.

- [Mavadati et al. , 2013] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F.: Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 4(2):151–160, 2013.
- [McDuff et al. , 2013] McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., and Picard, R.: Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 881–888, 2013.
- [McGuire et al. , 2010] McGuire, B., Daly, P., and Smyth, F.: Chronic pain in people with an intellectual disability: under-recognised and under-treated? Journal of Intellectual Disability Research, 54(3):240–245, 2010.
- [McKeown et al. , 2012] McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M.: The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing, 3(1):5–17, 2012.
- [Neshov and Manolova, 2015] Neshov, N. and Manolova, A.: Pain detection from facial characteristics using supervised descent method. In 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), volume 1, pages 251–256. IEEE, 2015.
- [Pantic and Patras, 2006] Pantic, M. and Patras, I.: Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36(2):433–449, 2006.
- [Pantic et al. , 2005] Pantic, M., Valstar, M., Rademaker, R., and Maat, L.: Web-based database for facial expression analysis. In 2005 IEEE international conference on multimedia and Expo, page 5. IEEE, 2005.
- [Parkhi et al. , 2015] Parkhi, O. M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In BMVC, volume 1, page 6, 2015.
- [Payen et al. , 2001] Payen, J.-F., Bru, O., Bosson, J.-L., Lagrasta, A., Novel, E., Deschaux, I., Lavagne, P., and Jacquot, C.: Assessing pain in critically ill sedated patients by using a behavioral pain scale. Critical care medicine, 29(12):2258–2263, 2001.
- [Prkachin, 2009] Prkachin, K. M.: Assessing pain by facial expression: facial expression as nexus. Pain Research and Management, 14(1):53–58, 2009.

- [Prkachin and Mercer, 1989] Prkachin, K. M. and Mercer, S. R.: Pain expression in patients with shoulder pathology: validity, properties and relationship to sickness impact. Pain, 39(3):257–265, 1989.
- [Puntillo et al. , 2004] Puntillo, K. A., Morris, A. B., Thompson, C. L., Stanik-Hutt, J., White, C. A., and Wild, L. R.: Pain behaviors observed during six common procedures: results from thunder project ii. Critical care medicine, 32(2):421–427, 2004.
- [Rodriguez et al. , 2017] Rodriguez, P., Cucurull, G., González, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B., and Roca, F. X.: Deep pain: Exploiting long short-term memory networks for facial expression classification. IEEE transactions on cybernetics, 2017.
- [Rojo et al. , 2015] Rojo, R., Prados-Frutos, J. C., and López-Valverde, A.: Pain assessment using the facial action coding system. a systematic review. Medicina Clínica (English Edition), 145(8):350–355, 2015.
- [Rudovic et al. , 2012] Rudovic, O., Pavlovic, V., and Pantic, M.: Kernel conditional ordinal random fields for temporal segmentation of facial action units. In European Conference on Computer Vision, pages 260–269. Springer, 2012.
- [Rudovic et al. , 2015] Rudovic, O., Pavlovic, V., and Pantic, M.: Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. IEEE transactions on pattern analysis and machine intelligence, 37(5):944–958, 2015.
- [Sandbach et al. , 2012] Sandbach, G., Zafeiriou, S., Pantic, M., and Yin, L.: Static and dynamic 3d facial expression recognition: A comprehensive survey. Image and Vision Computing, 30(10):683–697, 2012.
- [Sayette et al. , 2012] Sayette, M. A., Creswell, K. G., Dimoff, J. D., Fairbairn, C. E., Cohn, J. F., Heckman, B. W., Kirchner, T. R., Levine, J. M., and Moreland, R. L.: Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding. Psychological science, 23(8):869–878, 2012.
- [Shi and Malik, 2000] Shi, J. and Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence, 22(8):888–905, 2000.
- [Sikka, 2014] Sikka, K.: Facial expression analysis for estimating pain in clinical settings. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 349–353. ACM, 2014.

- [Sikka et al. , 2015] Sikka, K., Ahmed, A. A., Diaz, D., Goodwin, M. S., Craig, K. D., Bartlett, M. S., and Huang, J. S.: Automated assessment of childrens postoperative pain using computer vision. Pediatrics, 136(1):e124–e131, 2015.
- [Sikka et al. , 2014] Sikka, K., Dhall, A., and Bartlett, M. S.: Classification and weakly supervised pain localization using multiple segment representation. Image and vision computing, 32(10):659–670, 2014.
- [Sikka et al. , 2016] Sikka, K., Sharma, G., and Bartlett, M.: Lomo: Latent ordinal model for facial analysis in videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5580–5589, 2016.
- [Sikka et al. , 2012] Sikka, K., Wu, T., Susskind, J., and Bartlett, M.: Exploring bag of words architectures in the facial expression domain. In Computer Vision–ECCV 2012. Workshops and Demonstrations, pages 250–259. Springer, 2012.
- [Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
- [Sullivan et al. , 2004] Sullivan, M. J., Adams, H., and Sullivan, M. E.: Communicative dimensions of pain catastrophizing: social cueing effects on pain behaviour and coping. Pain, 107(3):220–226, 2004.
- [Sun et al. , 2014] Sun, B., Li, L., Zuo, T., Chen, Y., Zhou, G., and Wu, X.: Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In Proceedings of the 16th International Conference on Multimodal Interaction, pages 481–486. ACM, 2014.
- [Szegedy et al. , 2017] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.
- [Szeliski, 2010] Szeliski, R.: Computer vision: algorithms and applications. Springer Science & Business Media, 2010.
- [Taigman et al. , 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1701–1708, 2014.
- [Tian et al. , 2001] Tian, Y.-I., Kanade, T., and Cohn, J. F.: Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence, 23(2):97–115, 2001.

- [Tong et al. , 2007] Tong, Y., Liao, W., and Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE transactions on pattern analysis and machine intelligence, 29(10), 2007.
- [Tóssér et al. , 2016] Tóssér, Z., Jeni, L. A., Lőrincz, A., and Cohn, J. F.: Deep learning for facial action unit detection under large head poses. In European Conference on Computer Vision, pages 359–371. Springer, 2016.
- [Tzimiropoulos and Pantic, 2017] Tzimiropoulos, G. and Pantic, M.: Fast algorithms for fitting active appearance models to unconstrained images. International journal of computer vision, 122(1):17–33, 2017.
- [Valstar et al. , 2015] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M., and Cohn, J. F.: Fera 2015-second facial expression recognition and analysis challenge. In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, volume 6, pages 1–8. IEEE, 2015.
- [Valstar et al. , 2012] Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., and Scherer, K.: Meta-analysis of the first facial expression recognition challenge. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(4):966–979, 2012.
- [Valstar and Pantic, 2012] Valstar, M. F. and Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(1):28–43, 2012.
- [Valstar et al. , 2017] Valstar, M. F., Sánchez-Lozano, E., Cohn, J. F., Jeni, L. A., Girard, J. M., Zhang, Z., Yin, L., and Pantic, M.: Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 839–847. IEEE, 2017.
- [Viola et al. , 2001] Viola, P., Jones, M., et al.: Rapid object detection using a boosted cascade of simple features. CVPR (1), 1:511–518, 2001.
- [Viola et al. , 2005] Viola, P., Platt, J. C., Zhang, C., et al.: Multiple instance boosting for object detection. In NIPS, volume 2, page 5, 2005.
- [Walecki et al. , 2017] Walecki, R., Pavlovic, V., Schuller, B., Pantic, M., et al.: Deep structured learning for facial action unit intensity estimation. arXiv preprint arXiv:1704.04481, 2017.

- [Wang et al. , 2017a] Wang, F., Xiang, X., Cheng, J., and Yuille, A. L.: Normface: 12 hypersphere embedding for face verification. In Proceedings of the 2017 ACM on Multimedia Conference, pages 1041–1049. ACM, 2017.
- [Wang et al. , 2017b] Wang, F., Xiang, X., Liu, C., Tran, T. D., Reiter, A., Hager, G. D., Quon, H., Cheng, J., and Yuille, A. L.: Regularizing face verification nets for pain intensity regression. In IEEE International Conference on Image Processing (ICIP) 2017. IEEE, 2017.
- [Werner et al. , 2017] Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., and Traue, H. C.: Automatic pain assessment with facial activity descriptors. IEEE Transactions on Affective Computing, 8(3):286–299, 2017.
- [Werner et al. , 2013] Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., and Traue, H. C.: Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In Proceedings of the British Machine Vision Conference, pages 1–13, 2013.
- [Wilkie, 1995] Wilkie, D. J.: Facial expressions of pain in lung cancer. Analgesia, 1(2):91–99, 1995.
- [Wilkie et al. , 2003] Wilkie, D. J., Judge, M. K. M., Berry, D. L., Dell, J., Zong, S., and Gillespie, R.: Usability of a computerized painreportit in the general public with pain and people with cancer pain. Journal of Pain and Symptom Management, 25(3):213–224, 2003.
- [Williams, 2002] Williams, A. C. d. C.: Facial expression of pain: an evolutionary account. Behavioral and brain sciences, 25(4):439–455, 2002.
- [Wu et al. , 2015] Wu, B., Lyu, S., Hu, B.-G., and Ji, Q.: Multi-label learning with missing labels for image annotation and facial action unit recognition. Pattern Recognition, 48(7):2279–2289, 2015.
- [Wu et al. , 2017] Wu, S., Wang, S., Pan, B., and Ji, Q.: Deep facial action unit recognition from partially labeled data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3951–3959, 2017.
- [Wu et al. , 2010] Wu, T., Bartlett, M. S., and Movellan, J. R.: Facial expression recognition using gabor motion energy filters. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 42–47. IEEE, 2010.
- [Wu et al. , 2012] Wu, T., Butko, N. J., Ruvalo, P., Whitehill, J., Bartlett, M. S., and Movellan, J. R.: Multilayer architectures for facial action unit recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(4):1027–1038, 2012.

- [Xiao et al. , 2004] Xiao, J., Baker, S., Matthews, I., and Kanade, T.: Real-time combined 2d+ 3d active appearance models. In CVPR (2), pages 535–542, 2004.
- [Xiong and De la Torre, 2013] Xiong, X. and De la Torre, F.: Supervised descent method and its applications to face alignment. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 532–539. IEEE, 2013.
- [Xu et al. , 2014] Xu, Y., Zhu, J.-Y., Eric, I., Chang, C., Lai, M., and Tu, Z.: Weakly supervised histopathology cancer image segmentation and classification. Medical image analysis, 18(3):591–604, 2014.
- [Zamzmi et al. , 2016] Zamzmi, G., Pai, C.-Y., Goldgof, D., Kasturi, R., Sun, Y., and Ashmeade, T.: Machine-based multimodal pain assessment tool for infants: a review. arXiv preprint arXiv:1607.00331, 2016.
- [Zeng et al. , 2009] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1):39–58, 2009.
- [Zhang et al. , 2006] Zhang, C., Platt, J. C., and Viola, P. A.: Multiple instance boosting for object detection. In Advances in neural information processing systems, pages 1417–1424, 2006.
- [Zhang and Zhang, 2010] Zhang, C. and Zhang, Z.: A survey of recent advances in face detection. 2010.
- [Zhang et al. , 2017] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z.: s^3fd : Single shot scale-invariant face detector. 2017 IEEE International Conference on Computer Vision (ICCV), pages 192–201, 2017.
- [Zhang et al. , 2014] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing, 32(10):692–706, 2014.
- [Zhao et al. , 2015] Zhao, K., Chu, W.-S., De la Torre, F., Cohn, J. F., and Zhang, H.: Joint patch and multi-label learning for facial action unit detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2207–2216, 2015.
- [Zhao et al. , 2016] Zhao, K., Chu, W.-S., and Zhang, H.: Deep region and multi-label learning for facial action unit detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3391–3399, 2016.

- [Zhou et al. , 2017] Zhou, Y., Pi, J., and Shi, B. E.: Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 872–877. IEEE, 2017.
- [Zhu and Ramanan, 2012] Zhu, X. and Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2879–2886. IEEE, 2012.

VITA

Zhanli Chen

Education	Ph.D. Electrical and Computer Engineering University of Illinois at Chicago	2008 – 2020
	M.S. Electrical and Computer Engineering Hong Kong University of Science and Technology (HKUST)	2006 – 2007
	B.S. Electrical Engineering Nanjing University of Posts and Telecommunications (NUPT)	2002 – 2006
Publications	Chen, Zhanli Ansari, Rashid and Wilkie, Diana J, “Learning Pain from Action Unit Combinations: A Weakly Supervised Approach via Multiple Instance Learning”, IEEE Transactions on Affective Computing, IEEE, 2019.	
	Z Chen <i>et al</i> , “Pain Detection from Facial Videos Using Two-Stage Deep Learning”, 2019 IEEE Global Conference on Signal and Information Processing, pp. 1-5, Ottawa, 2019.	
	Zhifeng Luo Chen, Zhanli , “A Privacy Preserving Group Recommender Based on Cooperative Perturbation”, 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 106-111, Shanghai, 2014.	
	Chen, Zhanli Ansari, Rashid and Wilkie, Diana J, “Automated detection of pain from facial expressions: a rule-based approach using AAM”, (<i>appearing in</i>) SPIE Medical Imaging, International Society for Optics and Photonics, 2012.	
Patent	P. Chen, Z.Chen , “Horizontally Opposed Internal Combustion Generator”, Granted Patent, Patent Publication No.: CN201221408Y, published on Apr. 15th 2009.	
Grants Support	R43 DA046973: Device to Measure Pain using Facial Expression Recognition Integrated with Patient PAINReportIt Tablet	
	P30 NR010680: Center for End-of-Life Transition Research (CEoLTR)	

Presentations	Conference Presentations at IEEE Global Conference on Signal and Information Processing 2019	
	Conference Presentation at American Medical Informatics Association 2012	
	Poster Presentations at SPIE Medical Imaging 2012	
Memberships	Student Member <i>IEEE</i>	
Services	Journal Reviewer at Signal, Image and Video Processing	2017 – now
	Secretary and Web Master of SPS, Chicago Chapter IEEE	2013 – now
Experience	Adjunct Lecturer at University of Illinois at Chicago	2013-2014
	Research Assistant at University of Illinois at Chicago	2008-2019
	Teaching Assistant at University of Illinois at Chicago	2008-2019
	Research Specialist for the FOK YING TUNG Institute HKUST	2007-2008