Deep learning for automatic facial expression assessment with application to the expression of pain

Manh Tu Vu

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR

DE L'UNIVERSITÉ DE BORDEAUX

ECOLE DOCTORALE MATHEMATIQUES ET

INFORMATIQUE

DOCTORAT EN INFORMATIQUE

Par **Manh Tu VU**

Apprentissage profond pour l'évaluation automatique des expressions faciales avec application à l'expression de la douleur

Sous la direction de : **Marie BEURTON-AIMAR**

Soutenue le 13 décembre 2022

Membres du jury :

M. Kévin BAILLY	Maître de conférences	Sorbonne Université	Rapporteur
Mme. Marie BEURTON-AIMAR	Maîtresse de conférences	Université de Bordeaux	Directrice
M. Pascal BALLET	Maître de conférences	Université de Brest	Examinateur
M. Renaud SEGUIER	Professeur	CentraleSupélec	Rapporteur
M. Serge MARCHAND	Professeur	Université de Sherbrooke	Examinateur
Mme. Virginie RONDEAU	Directrice de Recherche	INSERM U1219	Président

université de BORDEAUX

THESIS PRESENTED

TO OBTAIN THE DIPLOMA OF

DOCTOR

OF BORDEAUX UNIVERSITY

DOCTORAL SCHOOL OF MATHEMATICS AND COMPUTER SCIENCE

SPECIALITY: COMPUTER SCIENCE

Presented by **Manh Tu VU**

Deep learning for automatic facial expression assessment with application to the expression of pain

Supervisor : **Marie BEURTON-AIMAR**

Defended in December 13th 2022

Committee in charge :

M. Kévin BAILLY	Associate professor	Sorbonne University	Reviewer
Mme. Marie BEURTON-AIMAR	Associate professor	Bordeaux University	Supervisor
M. Pascal BALLET	Associate professor	Brest University	Examiner
M. Renaud SEGUIER	Professor	CentraleSupélec	Reviewer
M. Serge MARCHAND	Professor	Sherbrooke University	Examiner
Mme. Virginie RONDEAU	Research director	INSERM U1219	President

To Mom and Dad

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor, Assoc. Prof. Marie BEURTON-AIMAR, for her supports during the past three years. Since the beginning of this thesis, I have always felt supported. From her mentoring, I have appreciated her scientific rigour, her ideas and human qualities. She gave me the liberty to explore my ideas, the motivation of always digging more and the opportunity to discover and trust myself, and I am grateful to her for this.

I would also like to thank the jury members of my defence. I express my gratitude to Prof. Renaud SEGUIER, University of Rennes I and Assoc. Prof. Kevin BAILLY at Sorbonne University, who agreed to review my manuscript. I would extend my gratitude to Prof. Serge MARCHAND at Sherbrooke University, Assoc. Prof. Pascal BALLET at Brest University, and Mme. Virginie RONDEAU, research director at INSERM institute, for being part of the thesis jury committee. Ultimately, I thank all of them for the attention they have given to my work as well as their valuable comments.

I would like to send my sincere thank to the people in LaBRI, who have supported from technical to administration during my stay. Thank to my friend Van Linh LE, whose dissertation I shamelessly took some inspiration from for formatting and organising this work.

Finally, I am very grateful to those who accompanies me outside of work, and through their support, have largely contributed to this thesis. Thanks to my family of course, that has supported and placed their trust in me all my life, and to my friends for making life more spicy. Special thanks to NDC Zeng, for her daily support and for boosting me in difficult times.

Title: Deep learning for automatic facial expression assessment with application to the expression of pain

Abstract: Facial expression is a major non-verbal means of expecting intentions in human communication. It is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intention. Thus, analysing and understanding human facial expression is crucial for many different applications in multiple domains, including health care and medical fields, virtual reality and augmented reality, education and entertainment. To measure facial expression intensity, the most popular and widely accepted way is using Facial Action Coding System (FACS). FACS defines a set of different facial Action Units (AUs), which are considered to be the smallest visually discernable facial movements. As any human facial expression can be decomposed into a set of facial AUs and their intensities, automatic measuring facial AUs intensity seems to be the key step towards better understanding human facial expressions. Hence, in this thesis, we study about automatic measuring facial expressions by utilising facial AUs, with an application on automatic pain intensity estimation.

Recently, deep learning techniques have emerged as powerful methods for learning feature representations directly from data and have achieved some major improvements in various face-related computer vision tasks. The main advantage of deep learning approaches is their ability to learn from experiences and generalise well on newly unseen data. However, to do so, these deep models require to be trained on a massive amount of data, which is difficult to obtain for the domains of facial AUs and pain intensity estimation. The reason for that is because it requires a costly and time-consuming labeling effort by trained human annotators. Moreover, the data distribution of AUs intensities is generally unbalanced, the performance of deep methods training on these databases are being negatively affected by insufficient data. Hence, in this thesis, we have proposed several approaches that are capable of exploiting better features of facial images on a limited amount of data. In particular, we present a learning approach that focuses

on the pain-related facial AU regions on the face image (regions-of-interest) for better extracting pain-related features from the image. By integrating the locations of regions-of-interest in face images into the network training process, we explicitly tell to the network where to focus in the face image and ignore other irrelevant or non-important regions. As deep learning in general requires to have massive amount of data to be able to extract correct features from images, our approach is expected to work effectively on a limited amount of data, as it has already know where to extract the important information on the image. Experiments on a benchmark database for pain estimation, i.e., the UNBC McMaster database, show that our approach outperforms other works in term of pain intensity estimation problem.

Realising the importance of focusing on regions-of-interest in extracting relevant feature information from limited amount of data, we have further improved the performance of the network in our second approach by not just focusing but isolating the regions-of-interest in face images. Based on the concept of *divide* and *conquer* paradigm, we utilise the Faster RCNN object detection network to locate the AU regions-of-interest (*divide*) before put them through a set of AU regressor networks for AU intensity estimation (*conquer*). By isolating each AU region, we are able to estimate correctly its intensity without worrying about learning incorrect features from other non-related regions, reducing the change of being overfitted. Besides, when analysing a face image, we also take into account the head pose factor to ensure that the network pay more attention on the visible parts instead of the obscured parts of the face image. By conducting extensive experiments on two well known benchmark databases of spontaneous facial expression, i.e., the DISFA and UNBC McMaster databases, our proposed approachs achieved state-of-the-art performance on both facial AUs intensity estimation and pain expression measurement domains.

Keywords : **Deep learning, Convolutional Neural Network, Facial expression analysis, Pain intensity estimation, Facial Action Units intensity estimation, Three-stages training network, FFAU network.**

Titre : Apprentissage profond pour l'évaluation automatique des expressions faciales avec application à l'expression de la douleur

Résumé : L'expression faciale est l'un des principaux moyens non verbaux d'exprimer des intentions dans la communication humaine. Il s'agit de l'un des signaux les plus puissants, naturels et universels permettant aux êtres humains de transmettre leurs états émotionnels et leurs intentions. Ainsi, l'analyse et la compréhension de l'expression faciale humaine sont cruciales pour de nombreuses applications dans de multiples domaines, notamment les soins de santé et les domaines médicaux, la réalité virtuelle et la réalité augmentée, l'éducation et le divertissement. Pour mesurer l'intensité des expressions faciales, la méthode la plus populaire et la plus largement acceptée est l'utilisation du Facial Action Coding System (FACS). Le FACS associe les changements d'expression faciale aux actions et aux intensités des muscles qui les produisent. Il définit un ensemble d'unités d'action (UA) faciales différentes, qui sont considérées comme les plus petits mouvements faciaux visuellement discernables. Comme toute expression faciale humaine peut être décomposée en un ensemble d'UA faciales et de leurs intensités, la mesure automatique de l'intensité des UA faciales semble être l'étape clé vers une meilleure compréhension et évaluation de l'expression faciale humaine. Par conséquent, dans cette thèse, nous étudions la mesure automatique des expressions faciales en utilisant les UA du visage, avec une application sur l'estimation automatique de l'intensité de la douleur.

Récemment, les techniques d'apprentissage profond sont apparues comme des méthodes puissantes pour apprendre des représentations de caractéristiques directement à partir de données et ont permis de réaliser des améliorations majeures dans diverses tâches de vision par ordinateur liées aux visages. Le principal avantage des approches d'apprentissage profond est leur capacité à apprendre à partir d'expériences et à généraliser sur de nouvelles données non vues. Cependant, pour ce faire, ces mod-

èles profonds doivent être entraînés sur une quantité massive de données, ce qui est difficile à obtenir pour le domaine des UA faciales et de l'estimation de l'intensité de la douleur. La raison principale est que cela nécessite un effort d'étiquetage coûteux et long par des annotateurs humains formés. Par exemple, il peut falloir plus d'une heure à un annotateur expert pour coder l'intensité des UA dans une seconde d'une vidéo de visage. De plus, le codage de l'intensité des UA nécessite une connaissance approfondie de FACS et une formation supplémentaire par des experts de FACS pour être en mesure d'étiqueter correctement les données. Environ 100 heures de formation FACS sont nécessaires pour un seul codeur FACS. Par conséquent, il est difficile d'obtenir un ensemble de données annotées de haute qualité à grande échelle. De plus, étant donné que la distribution des données de l'intensité de l'UA est généralement déséquilibrée vers une expression neutre (niveau d'intensité 0), la performance des méthodes d'apprentissage profond sur ces bases de données est affectée négativement par des données insuffisantes. Par conséquent, dans cette thèse, nous avons proposé plusieurs approches d'apprentissage qui sont capables d'exploiter de meilleures représentations des caractéristiques de l'image du visage sur une quantité limitée de données, améliorant ainsi les performances du réseau par rapport aux approches de l'état de l'art. La première approche que nous proposons consiste à apprendre à se concentrer sur les régions liées à la douleur dans l'image du visage (région d'intérêt) pour une meilleure estimation de l'intensité de la douleur. L'idée principale de cette approche repose sur le fait que les humains n'ont pas tendance à traiter tout ce qu'ils voient dans son intégralité en une seule fois. Il a plutôt tendance à se concentrer de manière sélective sur une partie de l'information au moment et à l'endroit où il en a besoin, tout en ignorant les autres informations perceptibles au même moment. Par conséquent, se concentrer sur les bons endroits et ignorer les autres informations non pertinentes semble être un aspect important non seulement pour les humains mais aussi pour les machines afin de se concentrer sur les informations révélatrices et d'extraire les caractéristiques correctes. Notre approche imite ce comportement cognitif des humains en intégrant les emplacements des régions d'intérêt dans les images de visages dans le processus de formation

du réseau. Ce faisant, nous indiquons explicitement au réseau où il doit se concentrer dans l'image du visage et ignorer les autres régions non pertinentes ou non importantes. Comme l'apprentissage profond en général nécessite une quantité massive de données pour être en mesure d'extraire les caractéristiques correctes des images, notre approche devrait fonctionner efficacement sur une quantité limitée de données, car il a déjà su où extraire les informations importantes sur l'image. En plus d'apprendre à se concentrer sur les régions d'intérêt, notre approche apprend également à modéliser l'information temporelle entre les images consécutives d'une vidéo. En reliant les caractéristiques spatiales extraites de chaque image à un réseau neuronal récurrent (RNN), notre réseau est capable de modéliser l'évolution dans le temps de chaque caractéristique faciale dans une séquence d'images, ce qui améliore encore les performances de notre réseau. Les expériences menées sur une base de données de référence pour l'estimation de la douleur, à savoir la base de données McMaster de l'UNBC, montrent que notre approche surpasse les autres travaux sur le problème de l'estimation de l'intensité de la douleur.

Conscients de l'importance de se concentrer sur les régions d'intérêt pour extraire des informations pertinentes à partir d'une quantité limitée de données, nous avons encore amélioré les performances du réseau dans notre deuxième approche en ne se contentant pas de se concentrer sur les régions d'intérêt dans les images de visages, mais en les isolant pour mieux extraire les représentations des caractéristiques liées à l'expression. Sur la base du concept du paradigme diviser pour mieux régner, nous utilisons le réseau de détection d'objets Faster RCNN pour localiser les régions d'intérêt de l'UA (diviser) avant de les faire passer par un ensemble de réseaux régresseurs de l'UA pour l'estimation de l'intensité de l'UA (régénérer). En isolant chaque région de l'UA, nous sommes en mesure d'estimer correctement son intensité sans nous soucier de l'apprentissage de caractéristiques incorrectes à partir d'autres régions non liées, ce qui réduit le risque de surajustement. En plus de la localisation et de l'estimation de l'intensité de l'UA, l'approche que nous proposons a également abordé un autre problème crucial des patients lors du tournage, à savoir le problème de la pose de la tête. Comme le patient ne regarde pas toujours directement la caméra lors de

l'enregistrement, il est important que notre réseau soit capable de traiter correctement les parties visibles, semi-visibles et obscures de l'image du visage. En entraînant explicitement le réseau à ne prendre en compte que les parties visibles du visage pendant l'entraînement et à ignorer les parties obscures, notre réseau est capable de détecter les caractéristiques correctes même dans les cas extrêmes de pose de la tête. En menant des expériences approfondies sur deux bases de données de référence bien connues sur les expressions faciales spontanées, à savoir les bases de données DISFA et UNBC McMaster, nos approches proposées ont atteint des performances de pointe dans les domaines de l'estimation de l'intensité des UA du visage et de la mesure de l'expression de la douleur. D'après les approches que nous proposons et les résultats des expériences, on peut constater que les UA du visage jouent un rôle important dans la description des expressions humaines, et qu'un système qui mesure correctement les UA du visage mesurera donc aussi correctement tous les types d'expressions faciales humaines, y compris l'expression faciale de la douleur. De plus, comme de nombreux chercheurs en psychologie ont indiqué que l'état affectif sous-jacent est linéairement lié à l'intensité physique des expressions faciales émotionnelles, nos approches de mesure de l'intensité des UA faciales peuvent donc être adoptées pour mesurer tout état affectif et physiologique humain de haut niveau.

Mots clés : Apprentissage automatique profond, Réseaux de neurones convolutionnels, Reconnaissance des expressions faciales, Évaluation de l'intensité de la douleur, Évaluation de l'intensité des unités d'action faciales, Technique d'entrainement de réseau en 3 phases, Réseau de neurones FFAU.

UMR 5800 – Laboratoire Bordelais de Recherche en Informatique (LaBRI)

Université de Bordeaux

351, cours de la Libération – F-33405 TALENCE

Contents

Acronyms

3D CNN 3D Convolutional Neural Network. 79–81, 91

AAM Active Appearance Model. 55, 84, 87, 88

AdaBoost Adaptive Boosting. 65, 87, 88

ANN Artificial Neural Network. 67, 68

BPTT Backpropagation Through Time. 77

CNN Convolutional Neural Network. 69, 70, 72–74, 78–81, 83, 90, 91, 135

CRF Conditional Random Field. 89, 90

DL Deep Learning. 69, 90

DNN Deep Neural Network. 29, 30, 67–71, 74, 78, 79, 83, 91

EEG Electroencephalography. 47

FACS Facial Action Coding System. 29, 43–45, 49–55, 57, 58, 156

FC Fully-connected. 72, 73

FLL Facial Landmark Localisation. 84

fMRI Functional magnetic resonance imaging. 47

FPN Feature Pyramid Network. 83, 125

FPS Frames Per Second. 56, 121

Gabor Gabor wavelet. 61, 62, 87, 89

GPA Generalised Procrustes Analysis. 23, 86, 105

GRU Gated Recurrent Unit. 22, 75–78

Haar Haar wavelet. 62, 83

HMM Hidden Markov Model. 89

HOG Histogram of Oriented Gradients. 62

HRV Heart Rate Variability. 47

ICA Independent Component Analysis. 63, 64, 87

ICC Intraclass Correlation Coefficient. 55, 57, 107, 112, 135, 138–141, 143, 145, 147

IoU Intersection Over Union. 126, 136, 137

KD Knowledge Distillation. 158, 160, 207

LBP Local Binary Pattern. 62, 88, 89

LDA Linear Discriminant Analysis. 64

LSTM Long Short-Term Memory. 22, 24, 26, 75–78, 90, 98, 100, 101, 104, 106–108, 131, 132, 141

MAE Mean Absolute Error. 107, 135, 138–141, 143, 145, 147

ML Machine Learning. 66, 68

MRF Markov Random Field. 88, 89

MSE Mean Squared Error. 106, 107, 135, 144, 145, 147

NMF Non-negative Matrix Factorisation. 64

NMS Non-Maximum Suppression. 126

PCA Principle Component Analysis. 63, 87

PCC Pearson Correlation Coefficient. 107, 112, 135, 138, 145, 147

PSPI Prkachin and Solomon Pain Intensity Scale. 23, 24, 27, 49, 50, 52, 53, 60, 86–88, 90, 91, 94, 97–100, 102, 103, 109, 111, 117, 118, 144–150, 153–155, 159

RNN Recurrent Neural Network. 21, 69, 74, 75, 78, 79, 131, 135

RVM Relevance Vector Machine. 65, 88

SGD Stochastic Gradient Descent. 68

SIFT Scale-Invariant Feature Transform. 63

SOTA State Of The Art. 27, 83, 85, 86, 91, 107, 112, 118, 124, 135, 142–145, 147–150, 154, 155, 159

SR Spectral Regression. 64, 87

SVC Support Vector Classification. 64, 65, 88

SVM Support Vector Machine. 64, 65, 87–89

SVR Support Vector Regression. 64, 65, 87, 89

List of Figures

List of Tables

Publications

List of original publications

This dissertation is based on the following articles, which are referred in the text by their Roman numerals (I–III):

I. M. T. Vu, M. Beurton-Aimar, P. -y. Dezaunay and M. C. Eslous, "Automated Pain Estimation based on Facial Action Units from Multi-Databases," 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2021, pp. 1-8, doi: 10.1109/ICIEVicIVPR52578.2021.9564244.

II. M. T. Vu, M. Beurton-Aimar and S. Marchand, "Multitask Multi-database Emotion Recognition," 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 3630-3637, doi: 10.1109/ICCVW54120.2021.00406.

III. M. T. Vu, M. Beurton-Aimar and K. TRAN, "FFAU: Faster-RCNN for Facial Action Unit intensity estimation," 2022 Pattern Recognition Journal (Submitted).

The author of the dissertation is the first author in all the above articles (i.e., I–III). The main responsibility for defining the research questions, developing ideas and methodologies, implementation and experiments, along with writing was carried out by the present author, while valuable comments and suggestions were given by the co-authors.

Introduction

Facial expression is a major non-verbal mean of expecting intentions in human communication. It is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intention. Thus, analysing and measuring human facial expression is crucial for many different applications in multiple domains, including health care and medical fields, human-computer interaction, virtual reality and augmented reality, advanced driver assistance systems, education, and entertainment. To measure human facial expression, the most popular and widely accepted way is using Facial Action Coding System (FACS). FACS associates facial expression changes with actions and intensities of the muscles that produce them. It defines a set of different facial Action Units (AUs), which are considered to be the smallest visually discernable facial movements. By using FACS, human coders can manually code nearly any anatomically possible facial display, decomposing it into the AUs and their intensities that produced the display. Hence, it can be seen that FACS plays a crucial role in analysing and understanding human affective and physiological states. Yet, as the work of FACS annotating is costly both in terms of time and effort of human expert annotaters, the amount of labelled publicly available data on this domain are generally limited. In this work, we primarily seek to address the problem of automatic facial AUs intensity estimation by relying on Deep Neural Network (DNN) techniques. Specifically, we try to construct different learning systems that are able to learn from limited amount of data to automatically measure human facial AUs intensity from face image or video. As facial AUs is the most basic building block for describing human facial expressions, this work of automatic facial AUs intensity estimation is one important step for better understanding the underlying affective state of human emotion.

Besides automatic estimating facial AUs intensity, we also aim to automatically measure one prototype of facial expressions, i.e., the facial expression of pain. Since pain

can be considered as the symptom of numerous diseases and tissue damage, it is crucial for clinicians to understand the pain of patients in order to inform them of their condition and recommend the best course of treatment. Hence, there is a need of developing an automatic pain assessment system that can infer the facial information and provide complementary objective information to the clinicians for better measuring and understanding the pain experience of the patients. As any facial expression can be decomposed into a set of facial AUs and their intensities, including the pain expression, in this work we try to construct a Deep Neural Network that capable of accurately measuring pain intensity level by relying on the measurement of facial AUs.

In summary, the objectives of this thesis are about studying human spontaneous facial behaviors in image and video for identifying the intensity levels of (1) facial Action Units (AUs) and (2) facial expression of pain. Between these two objectives, we focus primarily on automatic facial AUs intensity estimation, as it is the most basic building block for describing human facial expressions. A system that correctly measures facial AUs will therefore also correctly measures any types of human facial expressions, including facial expression of pain.

Thesis Organisation

This thesis is organised into two parts. The first part includes two chapters (Chapter 1-2), which provides an overview of the background, state of the art in automatic facial AUs and pain intensity estimation. The second part consists of the main contributions of this thesis (Chapter 3-5).

Chapter 1 focuses on the definition of the problems targeted by this work, i.e., estimating the intensity of facial Action Units (AUs) and the expressions of pain. The background and rationale behind the study, as well as the descrition of publicly available databases used in this thesis are mentioned in this chapter.

Chapter 2 provides an overview of different techniques that can be used in automatic image/video processing. The pipeline of processing face data and the literature

review of the state of the art methods in the objective domains of the thesis are also provided in this chapter.

Chapter 3 and **4** present our proposed methods for automatic facial AUs and pain intensity estimation problems. Chapter 3 introduces our novel method of learning to focus on regions-of-interest in face images for pain intensity estimation. Chapter 4 proposed our new method for isolating regions-of-interest in face images for better facial AUs and pain intensity estimation.

Chapter 5 concludes the thesis by discussing the contributions of the thesis, the opening challenges and future research direction on automatic facial AUs and pain measurement.

Context: CIFRE Thesis

This thesis is part of the Digital Therapeutics (DTx) project at Lucine[1] company. The aims of the DTx project is to provide software-based therapeutic solutions for chronic pain patients to prevent, manage or treat a wide range of physical, mental and behavioral symptoms. According to the SFETD[2], in 2017, there are 150 millions people suffer from chronic pain in Europe. In France, 70% of pain patients do not receive appropriate treatment and only 3% of them receive personalised care. Also, pain is the most common reason for medical doctor consultation and poses great challenges in terms of its treatment. Therefore, the objective of Lucine is to create new digital solutions that help patients to relieve chronic pain.

This CIFRE[3] thesis, as a part of the DTx project at Lucine company and under the supervision of the LaBRI[4] laboratory, aims to automatise the measurement of human facial expressions in general and pain expression in particular by using machine learning techniques on face image and video. These facial expressions measurement information will be a valuable data for better understanding human expressions and improving pain management of the patients.

[1] lucine.io

[2] Société Française d'Étude et de Traitement de la Douleur

[3] CIFRE means industrial training contract by research, and corresponds to a particular French type of thesis contract, supported by a company.

[4] Laboratoire Bordelais de Recherche en Informatique

Chapter 1

Background

Contents

In this chapter, we first give a general introduction to human facial expression in general and pain expression in particular. Then, we expose the current measurement approaches to measure each of these two domains. Finally, we review publicly available datasets on both of these two domains.

1.1 Overview

The face is a window with a view opening onto our emotions. The expressions on human face provide rich information in understanding the emotional state of the person, feeling and attitude. Although there are only a few words to describe different facial behaviors (smile, frown, furrow, squint, etc), human facial muscles are sufficiently complex to allow more than a thousand different facial appearances [RFD97]. These facial expressions can sometime provide much more information than any words can do. In 1872, Charles Darwin [DP98] once said:

> *" They (the movements of expression in the face and body) reveal the thoughts and intentions of others more truly than do words, which may be falsified. "*

It can be seen that facial expressions are an important non-verbal communication channel, which can reveal our true inner feelings and thoughts. In another research, Russell and Fernández-Dols [RFD97] also have said:

> *" When we turn our eyes to the face of another human being, we often seek and usually find a meaning in all that it does or fails to do so. "*

It is clear that facial expressions play an important role in human communication as revealing one's emotional states and intentions. Besides facial expressions, there are other means of communication that express emotions, including vocal intonation, hand gesture, head movement, body movement and posture, and more. Despite the available range of cues and modalities in human-human interaction, facial expressions is still the primary way to express people's feelings [AA21]. In 1967, psychologist Mehrabian

observed that 7% of knowledge moves between people through writing, 38% through voice, and 55% through facial expression [MF67]. Mehrabian's observations once again justify the interest and importance of facial expressions in human-human communication and in the expression of emotional feelings. In fact, facial expression figures prominently in research on almost every aspect of emotion, including psychophysiology [LEF90], neural correlates [EDF90], development of emotion [MCT+89], perception [ASC05], addiction [GS08], social processes [HCR92], depression [CKM+09], and other emotion disorders [TMD+05]. From these researches, we can see that facial expression possesses a lot of valuable information that gives an effective way to the perceive person's consciousness and mental activity. The analysis of facial expression shows an important theoretical research value, practical value and the life application value. Consequently, since the last quarter of the 20th century, with the advances in the field of computer graphics and computer vision, computer scientists have been starting to show interest in the study of human facial expressions. Automatic human facial expression analysis is thus becoming an important research topic and attracting a lot of attention from researchers, as it is applicable in many different domains. Pain assessment [VBADE21], telenursing [DSI+01], drowsy driver detection [JNK20], analysing mother-infant interaction [FCAL04], human-robot interaction [FVFP17], and expression mapping for video gaming [HT10] are among the domains that benefits from machine understanding of human facial expressions.

There are two main stream researches in automatic facial expressions analysis, one is about detecting the presence or absence of a certain facial expressions (e.g., detecting the facial expression of happiness), the other is about measuring the facial behavior intensity of an expression (e.g., pain intensity estimation). The intensity of a facial expression can be seen as the relative degree of displacement, away from a neutral or relaxed facial expression, of the pattern of muscle movements involved in emotional expressions of a given sort [HBK97].

Many of previous works have focused on detecting facial expressions due to its pioneering investigations along with the direct and intuitive definition of facial expressions

[LD20]. However, as facial expressions are complex and subtle, the meaning and function of these expressions depends largely on their intensity, rather than just the binary selection of presence and absence. For example, the smiles of enjoyment are full-blown smiles, while the "fake happiness smiles" may be asymmetric and are usually less in intensity when observed in naturalistic social settings [EF82]. In 2013, Gunnery *et al.* [GHR13] noted:

> *" Most of the smile genuineness impression is created by the intensity of the smile. "*

It can be seen that the intensity of a facial expression behavior plays a crucial role in defining the meaning and function of the expression. This is inline with many other psychology findings [RE05, HBK95, HBK97], in which they have found that the underlying affective state is linearly related to the physical intensity of the emotional facial expressions. Hence, the intensity of human affective and physiological states (e.g., pain emotion), which cannot be directly measured, can be effectively estimated by measuring the intensity of facial expression. In this work, we seek primarily to address the problem of automatic measuring the intensity of human facial expressions. Besides automatic estimating the intensity of facial expressions, we also aim to automatically measure the intensity of one prototype of facial expressions, i.e., the facial expression of pain.

Measurement of pain is a crucial requirement for many applications in health care and medical fields [TBLH17, AEAKAS20]. Since pain can be considered as the symptom of numerous diseases and tissue damage, understanding the patient's pain is very important for clinicians to provide information about the condition of the patient, and to advise the right course of treatment. In clinical trials and clinical practice, pain is usually diagnosed through the patient' self-report based on several factors including severity, sensory quality, location, and duration of the pain. Self-report is often referred to as the gold standard and the primary tool to measure the pain experience [Cra09], in which the patient is asked to quantify the level of pain that they are experiencing.

However, self-report is not applicable for population who are unable to articulate their pain experience [CPG11], e.g., unconscious or newborn patients. When assessing the distress of others, self-report is often considered as less weighted than non-verbal activities [Cra92, PC92]. In a review of research, von Baeyer *et al.* [vBJM84] have noted that "nonverbal behaviors may be a more accurate source of information than verbal reports because they are less subject to *'motivated dissimulation'*". Due to subjectivity of pain experience, self-report may not be a reliable assessment technique because it is a controlled and goal-oriented response to pain [SC10], which might be affected by reporting bias and variances in memory and verbal ability [Cra92].

Another approach to measure pain experience of a patient is observer rating, in which the medical staffs (e.g., professional nurses) examine the conditions of the patient and rate the pain intensity accordingly. However, many variables of the patient in pain, i.e., physical attractiveness, sympathy, gender, and age, are known to influence clinical judgments [HRVB90, HLHM00, RW04, DRGP+11]. Moreover, as the medical staffs exposed to a high number of painful facial expressions for a long period, they may develop an exaggerated bias over time [PKB15], which could have a negative impact on the accuracy of the pain assessment. To overcome these limitations, it is desirable to develop automatic pain assessment systems that can infer the facial information and provide complementary objective information to the clinicians for better measuring and understanding the pain experience of the patients. Hence, in this work, besides automatic measuring the intensity of human expressions, we also aim to construct a pain measurement system which could correctly measure pain intensity level from facial expressions of patients. This system will act as a computer-aided health management system to continuously monitor the pain condition of the patients, providing more insights of the patient's conditions to the clinicians.

In the next sections, we discuss deeply about facial expression and present different ways to measure it. Sections 1.2-1.4 introduce two approaches to measure human facial expressions, while Section 1.5 discusses about pain emotion and the measurement of pain.

1.2 Facial expression analysis

In order to analyse facial expression, we need to know the different ways of describing facial expression and the existing approaches to measuring it. There are two main ways for describing a facial expression: *judgement* and *sign* based approaches [CAE07, CE05]. Both of them are grounded on the non-verbal communication model proposed by Rosenthal [Ros05]. The model assumes communication between two human entities: the subject and the observer. The subject experiences an internal state (e.g., pain or other emotions), then expresses through his external features (e.g., facial muscles, body gestures, etc). These features are then recognised and interpreted by the observer.

The *judgement* based approach takes the role of the observer and also the way he interprets the expression. It tries to decode the meaning of the behaviour, e.g., by assigning one of the six basic emotions [EF71] or by giving an emotion intensity score, such as valence or arousal (see Section 1.3.2 for more information about valence and arousal). Contrastingly, the *sign* based approach uses the physical communication channel, e.g., the facial muscles. It analyses how each part of the face move, e.g., lowering of the brows or stretching of the mouth. As an example, on seeing a smiling face, an observer with a judgment-based approach would make judgments such as "happy," whereas an observer with a sign-based approach would code the face as having an upward, oblique movement of the lip corners. Compared to the *judgement* based approach, the *sign* based approach is better in term of objectivity, since it is purely just the description of each face part's movement. On the other hand, as each part of the face needs to be analysed separately, this *sign* based approach is harder and takes longer time for the observer to interpret compared to the *judgement* based approach.

In term of *judgement* based approach, we focus on the Discrete and Dimensional emotion measurement approaches in Section 1.3. In term of *sign* based approach, we focus on the Facial Action Coding System, which is described in Section 1.4.

1.3 Discrete and Dimensional emotion assessment

In the field of facial expression analysis, the discrete and dimensional emotion models [MG14] are the two well-known approaches for describing human affective states. Both of them are *judgement* based approach. This section briefly introduces the two emotion models and discusses the advantage and disadvantage for each of them.

1.3.1 Discrete emotion model

Discrete emotion model is based on the assumption that there is a limited set of basic emotions categories whose expression and recognition are fundamentally the same for all individuals regardless of ethnic or cultural differences. The model suggests that an independent neural system subserves every discrete basic emotion. However, neuro-imaging and physiological studies have failed to establish reliable, consistent evidence to support this theory (see [BW06, CBL+00]), and the matter remains under debate.

In studies about human emotions, the six basic emotions proposed by Ekman [EF71] is one of the most commonly used facial expressions measurement, which is based on this discrete emotion model. The joy, sadness, fear, anger, disgust and surprise are the six expressions included in the measurement (see Figure 1.1 for the visualisation of these expressions). These expressions are referred as "universal" as they were found to be universal across human ethnicities and cultures [EF71]. However, advanced research on neuroscience and psychology argued that the model of six basic emotions are culture-specific and not universal [JGY+12, GRvdVB14]. Psychology studies have found that affect expression patterns in face and eye movements vary within cultures and vary even more across cultures [CBC+06, JGY+12]. They suggest that the cultural factors need to be taken into account when measuring facial expressions [SB09].

The discrete emotion model is still the most popular perspective for facial expressions assessment due to its pioneering investigations along with the direct and intuitive definition of facial expressions. However, in the downside, this approach cannot express

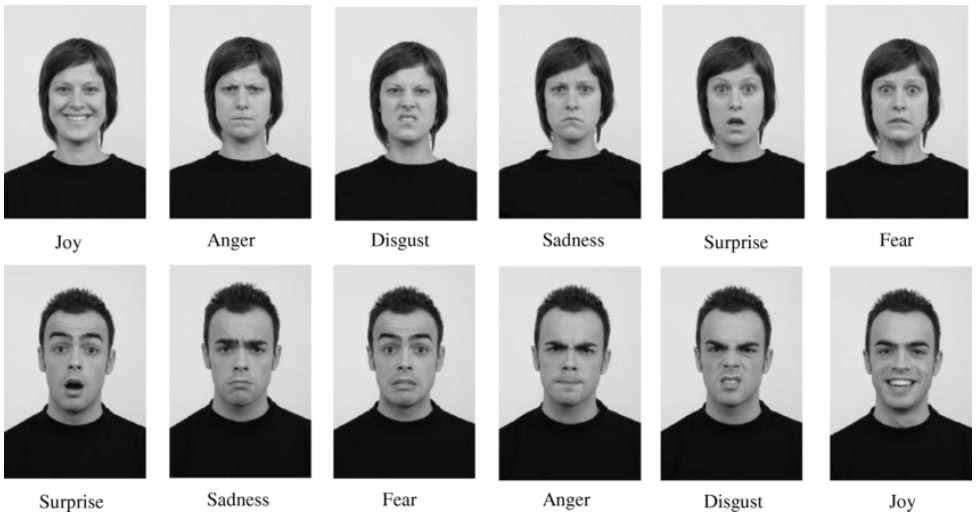


Figure 1.1: The facial expression visualisations of six basic universal emotions. Image from [SDF+21].

complex affective emotion states. E.g mixed emotions cannot adequately be transcribed into a limited set of categories [KMK06, DTM14]. Some researchers tried to define multiple distinct compound emotion categories, e.g., happily surprised, sadly fearful [DTM14], to overcome this limitation. However, the set is still limited, and the intensity of the emotion also cannot be defined in the categorical set of emotions.

1.3.2 Dimentional emotion model

Many cognitive scientists oppose the theory of a set of discrete, basic emotions [Man84, Rus95]. Some of these opponents instead take a dimensional view of the problem [SBR10]. In their view, affective states are not discrete and independent of each other, instead they are systematically related to one another [MR73].

Several dimentional emotion models have been proposed [RM77, Rus80, Tha90]. Yet, the valence-arousal model proposed by Russell [Rus80] seems to be the most famous and have gained great support among emotion researchers [Sch99, PRP05]. Instead of an independent neural system for every basic emotion, Russell's model proposes that all affective states arise from two independent neurophysiological systems: one related to **valence** (i.e., level of pleasure) and the other to **arousal** (i.e., level of affective

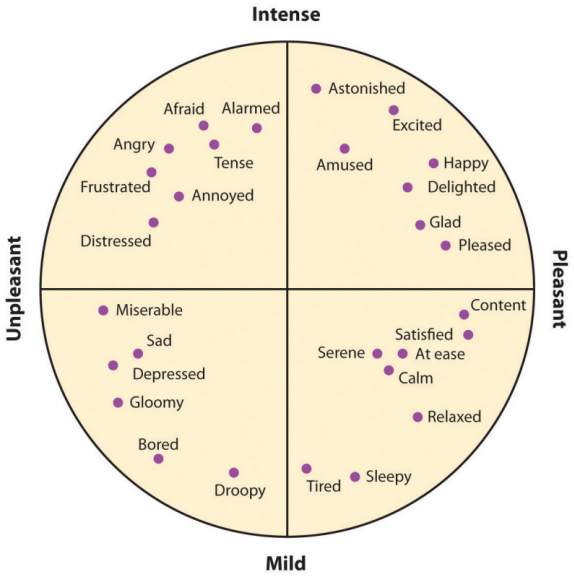


Figure 1.2: The 2D Emotion Wheel. Image from [KZ18].

activation). Each and every affective experience is the consequence of a linear combination of these two independent systems, which is then interpreted as representing a particular emotion. Values of valence and arousal form a 2D emotion wheel, where each point represents an emotional state, as can be seen in Figure 1.2. In term of quantifying the facial expressions, as the value of both valence and arousal is on a continuous scale, this dimensional model can distinguish between subtly different displays of affect and encode small changes in the intensity of each emotion, such as *low happy*, *happy* or *very happy* emotions [EA12].

Since the dimensional model covers both intensity and different emotion categories in the continuous domain, it is more robust compared to the discrete emotion coding approach. However, this model also has some limitations, it has been criticised for their lack of differentiation when it comes to emotions that are close neighbours in the valence-activation space, such as anger and fear [TWC99]. It is also unclear how a facial expression should be mapped to the space or, vice versa, how to define regions in the valence/arousal space that correspond to a certain facial expression. Being a *judgement* system that is based on feeling, it is again problematic to use this system

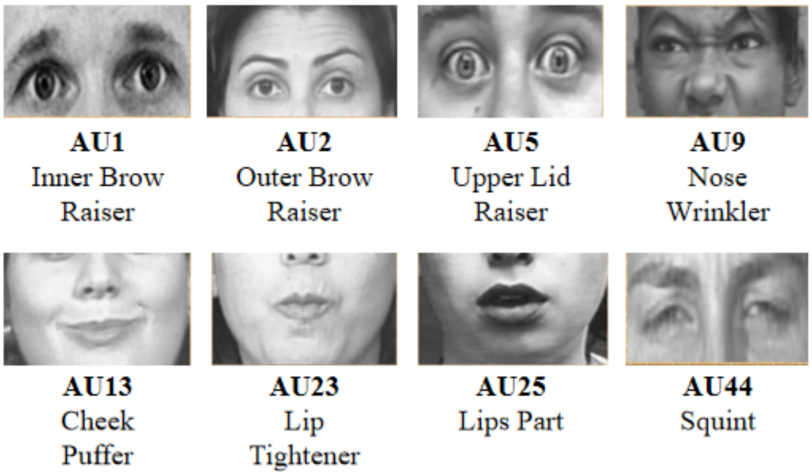


Figure 1.3: The visualisation of some AUs. Image from [HCLW19].

to describe non-emotional communicative signals (e.g., brow-flash used in greetings) [Val08]. Additionally, this dimensional emotion model is subjective and therefore requires experienced annotators to ensure consistency when constructing datasets.

1.4 Facial Action Coding System

The FACS [EFH02] is the most well-known, standardised and widely used *sign* based approach [CAE07], which was initially developed in the 1970s [EF78] and was informed by earlier research by [Hjo69, DP98]. FACS defines a unique set of anatomically based facial actions called Facial Action Units (AUs). Each AU is based on one or at most a few facial muscles and may occur individually or in combinations, e.g. AU1 (inner brow raiser) codes contractions of both the *frontalis* and *pars medialis* facial muscle, while AU23 (lip tightener) codes contractions of the only the *orbicularis oris* muscle [EFH02] (see Figure 1.3). Table 1.1 lists the main AUs along with their description and the facial muscle(s) involved.

In addition to the presence or absence of AUs, FACS also defines intensity codings on a five point scale from A to E for representing the intensity variation from barely detectable or trace (A) to maximum (E) [EFH02]. Recent works [LCP+12, MMB+13]

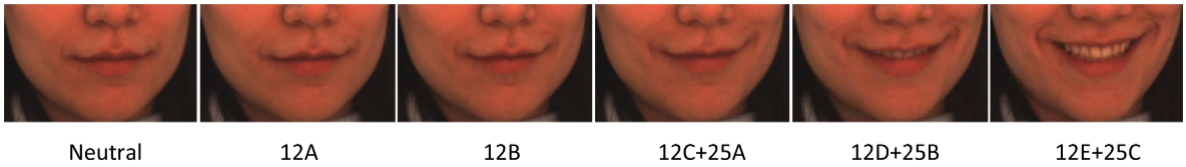


Figure 1.4: Sample facial images with AU intensity variations. Image from [MMB+13].

adapted the FACS scale and proposed to use a six-point ordinal scale $[0-5]$ to represent the AU intensity score, which seems to be easier to interpret. With this six-point ordinal scale, score 0 denotes the AU fails to occur and the $[1-5]$ scale is corresponding to the variation $[A-E]$ in FACS scale. Figure 1.4 shows examples of intensity variations of AU12 and AU25.

Numerous previous studies have relied on FACS for determining a patient's psychological and physiological state. Archinard *et al.* [AHRH00] used FACS to identify which depressed patients are at greatest risk for reattempting suicide. Prkachin and Solomon [PS08] constitute an index of physical pain with desirable psychometric properties by ultising the combination of different facial AUs. Keltner *et al.* [KMSL95] used FACS to distinguish different types of adolescent behavior problems. Cohn *et al.* [CKM+09] rely on FACS to detect depression of patients from face video. These are just a few of the numerous works that benefit from automatic facial AUs intensity estimation. It can be seen that FACS plays an important role in human facial analysis, which set the step stones for many high order decision making and applications that related to emotion, social interaction, psychological disorders and health. Yet, on the downside, the work of FACS annotating is not easy as it requires a costly and time-consuming labeling effort by trained human annotators. For instance, it may take more than an hour for an expert annotator to code the intensity of AUs in one second of a face video [RE20]. Furthermore, FACS coding requires profound knowledge of the FACS and additional training by FACS experts to be able to correctly label data. Approximately, it requires about 100 hours of time involved in this FACS training [Prk09]. Therefore, in this work, we attempt to construct a learning system that has the capability of automatic measuring

AU	Description	Facial Muscle
0	Neural face	-
1	Inner Brow Raiser	Frontalis, pars medialis
2	Outer Brow Raiser	Frontalis, pars lateralis
4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli, Currugator
5	Upper Lid Raiser	Levator palpebrae superioris
6	Cheek Raiser	Orbicularis oculi, pars orbitalis
7	Lid Tightener	Orbicularis oculi, pars palpebralis
9	Nose Wrinkler	Levator labii superioris alaquae nasi
10	Upper Lip Raiser	Levator Labii Superioris, Caput infraorbitalis
11	Nasolabial Deepener	Zygomatic Minor
12	Lip Corner Puller	Zygomatic Major
13	Cheek Puffer	Levator anguli oris
14	Dimpler	Buccinator
15	Lip Corner Depressor	Depressor anguli oris (Triangularis)
16	Lower Lip Depressor	Depressor labii inferioris
17	Chin Raiser	Mentalis
18	Lip Puckerer	Incisivii labii superioris and Incisivii labii inferioris
20	Lip stretcher	Risorius
22	Lip Funneler	Orbicularis oris
23	Lip Tightener	Orbicularis oris
24	Lip Pressor	Orbicularis oris
25	Lips part	Depressor Labii, Orbicularis Oris
26	Jaw Drop	Masetter; Temporal and Internal Pterygoid relaxed
27	Mouth Stretch	Pterygoids, Digastric
28	Lip Suck	Orbicularis oris
41	Lid droop	Relaxation of Levator Palpebrae Superioris
42	Slit	Orbicularis oculi
43	Eyes Closed	Relaxation of Levator Palpebrae Superioris

Table 1.1: List of common AUs with their description and the involved facial muscles.

FACS intensity with high precision when training on a limited amount of data. As facial AUs are independent of interpretation and can be used for any higher order decision making process, this work of automatic measuring facial AUs intensity is one of the key steps towards better understanding human facial expression and assessment.

In the next section, we discuss about one specific type of human facial expressions: the facial expression of pain. Since FACS can be used to describe any facial expression, we show how can we utilise FACS to measure the intensity of pain.

1.5 Pain expression and measurement

Pain is a complex phenomenon that affects millions of people around the world, and is a common cause of agony and suffering. In order to quantify the expressions of pain, we first explain what is pain in general and then how to measure it.

1.5.1 Pain emotion and expression

Pain is an inner feeling that draws attention, alerts individuals to possible bodily danger and subsequently prompts to escape from the dangerous situations, recovery and heal [Wil02]. According to the most widely accepted definition, pain is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage [Mer79]. The experience of pain is constructed in the brain based on information from multiple sources, including incoming nociceptive or danger signals, information form the senses (vision, touch, hearing), and other modulating factors such as attention, distraction, expectations, anxiety, stress, the physical and social context, and past experience [Mar08, Fit13]. As a result, the same pain stimulus (e.g., electric shock) may lead to different pain experiences, i.e. people differ in their pain sensitivity and also the same person can have different experiences to the same stimulus depending on many factors.

In order to describe pain, an extension of the Rosenthal non-verbal communication model [Ros05] is developed and presented by Prkachin and Craig [PC95]. The model begins with an experience of pain, then an encoding process encodes information about the experience into external features. Finally, these features are then decoded by an observer. According to the model, the perception of pain experience is influenced by three factors including: the pain stimulus (e.g., the severity of tissue damage), the intrinsic factors (e.g., aging) and the extrinsic factors (e.g., stress-induced analgesia). The intrinsic and extrinsic factors may amplify or attenuate the effects of the noxious pain stimulus, leading to higher or lower level of perceived pain experience. Hence,

different level of pain stimulus need to be applied for each individual to be able to have roughly the same sense of perceived pain experience [WGE+13].

Regarding the encoding process, pain experience can be encoded via different channels including: facial expressions, body gestures, non-verbal vocalisations, speech or different physiological signals (i.e. Electroencephalography (EEG), Functional magnetic resonance imaging (fMRI) or Heart Rate Variability (HRV)). Of these cues, only physiological signals cannot be naturally recognised and interpreted by human observers. Moreover, the measurement procedure of these signals is also complicated and intrusive, it requires patients to attach bulky sensors, chest straps, or stick electrodes. Hence, more and more works have been focusing on exploiting other unobtrusive expressing channels, such as facial expressions, body gestures or other vocal-related channels. In this work, we focus on exploiting the facial expressions, since it is unobtrusive and have been shown to be highly informative with regard to pain [PC95, SHW+07].

1.5.2 Pain measurement

To measure the pain emotion, there are serveral possibilities including: stimulus measurement, self-report and observer rating. Regarding the stimulus measurement, this is an easy approach to measure the pain experience, since the intensity level can be obtained directly from the pain stimulus device, e.g. the voltage of an electro-shock stimulus or the temperature of an heating stimulus. However, as the perception of pain is influenced not only by the pain stimulus but also by the intrinsic and extrinsic factors, pain measurement should take into account both of these factors. Yet, intrinsic factors like mood, beliefs or personality seem impossible to be quantified and thus do not lead to a reliable measurement of pain. Self-report is another approach for measuring the pain experience, which refers to conscious communication of pain-related information by the patient. The approach is often referred to as the gold standard in pain assessment, and has been widely used in many different clinical applications. However, self-report is also being considered as a controlled and goal-oriented response approach

[CPG11], which can be affected by reporting bias and variance in memory and verbal ability [Cra92]. Furthermore, self-report is difficult to collect for dynamic situations that require a continuous intensity measurement over time. Regarding observer rating measurement, there is good evidence that facial pain expressions, which can be observed by an observer, is not only sensitive but also specific to pain and can be distinguished from expressions of basic emotions [Wil02, CPG11]. Hence, this work focuses on measuring the pain reaction of the subject in terms of facial expressions, based on the observer's observation. As observer rating is highly subjective, it is possible to combine several observers to obtain a more robust and reliable measurement result.

Figure 1.5: Example of some facial action units occur in painful experience. Image from [WLMW+22].

The facial expressions of pain, like any other type of facial expressions, can be described by a combination of several different facial AUs. It is shown that there are a certain number of AUs are displayed during the experience of experimental pain as well as in clinical pain conditions [KFL12]. These AUs are including: rising the brows (AU1, AU2), lowering the brows (AU4), cheek raise and lid tightening (AU6, AU7), nose wrinkling and upper lip raising (AU9, AU10), mouth opening (AU25, AU26, AU27), and eye closure (AU43) [PS08, KML19, KL14]. These pain-related facial AUs seems to encode the essential information about pain available in the face. An example of painful facial expression are shown in Figure 1.5. From this figure, we can see that these AUs can be triggered during a painful experience, and not all of them are triggered at the same time. In fact, individuals often display only parts of these AUs or combine these AUs

differently [CPG11, KML19]. Hence, the pain expressions can be measured by detecting and estimating the intensity of the right combination of AUs.

In order to formulate the measurement of pain, Prkachin and Solomon proposed a pain intensity scale [PS08] termed by PSPI. The PSPI pain intensity is defined as the sum of some pain-related facial AU intensities as follows:

$$PSPI = AU4 + \max(AU6, AU7) + \max(AU9, AU10) + AU43 \tag{1.1}$$

Each AU has its intensity score with the range of $[0-5]$, except for AU43 with the range of $[0-1]$ representing the present or absence of the AU. Thus, the PSPI pain scale has a range of $[0-16]$, whereas 0 denotes no pain and 16 is the maximum pain experience. Since PSPI is based on FACS, it can be calculated for each individual image or video frame. Prkachin and Solomon show that PSPI correlates well with observer rated pain intensity levels [PS08].

The main advantage of PSPI is that the subjective part of the *judgment* based pain rating is eliminated, and directly mapped to the *sign* based FACS, making the results easily reproducible. On the downside, it was found that the PSPI pain score can go up and down with tension and relaxation of facial muscles despite the felt pain is steadily increasing [WAHLE+17]. Thus, the temporal resolution of PSPI could be misleading if the pain persists for long time. Moreover, the PSPI may be zero in some cases, even though the person is actually experiencing pain. There may be no facial reaction at all due to low pain intensity or expressiveness [PC95, KL14]. Additionally, several factors that are relevant for pain rating are missed in the calculation of PSPI score including eyebrow raiser (AU1, AU2) or mouth opening (AU25, AU26, AU27) [KL14], which could potentially lead to an incorrectly pain measurement in some cases. Further, several facial expressions of emotions share AUs with PSPI [ZYC+14], e.g. disgust (AU9, AU10), fear and sadness (AU4), and happiness (AU6). If PSPI is used in a wider context, many frames are labelled as painful by mistake.

Despite having these shortcommings, PSPI score is still a valuable pain measurement and have been widely used as ground truth to approach pain recognition and

assessment. As PSPI is a FACS based approach, information regarding both the name and intensity of each facial muscle movement is encoded at the frame level, which is a great source of information to exploit. E.g. a learning approach can learn to model precisely the appearance, location and intensity of each facial muscle that caused the pain emotion. Additionally, the high correlation of PSPI score with high pain intensities [WAHLE+16] also seems to be an important aspect of the measurement. E.g. an automated system could be built to continuously monitor a patient and raise the alarm if an abnormal pain experience is recognised. Yet, PSPI score should still be used with caution or supplemented with other ground truth.

Another way to approach this problem is constructing a better frame-level pain measure, e. g. by subtracting AUs that do not occur during pain and by considering multiple "faces of pain" in a non-linear combination [WAHLE+16]. This seems to be a promising approach as it takes the full advantages of FACS based approach while still correctly measure pain expressions at frame level. Yet, this approach requires extensive work by neuroscientists and pain specialists to construct a new FACS based pain measure, which is out of the scope of this work. However, as FACS is always blind to the research hypothesis [PCBH17], a system that correctly measures AU intensity is always useful as it is not only able to compute the PSPI score to measure pain for now, but also able to adapt to the new better FACS based pain measure in the future. Hence, in this work, we focus on constructing a system to measure AU intensity in general and subsequently measure PSPI score for estimating pain intensity in particular.

1.6 Databases

Representative data are essential for developing a facial expression assessment system and proving its usefulness. In this section, we provide an overview of the publicly available face image/video databases which consist of AU intensity and pain estimation. We introduce the two mainly used databases in our experiments: DISFA (Section 1.6.2) and UNBC McMaster (Section 1.6.1). Besides these two databases, we also briefly summary

other publicly available databases with AU intensity annotated, as can be seen in Table 1.2. Since the process of AU intensity annotation is time consuming and requires trained experts, only a few databases exist.

Database name	Year	N. Subjects	N. Images	N. AUs
UNBC McMaster [LCP+12]	2012	25	48,398	11
DISFA [MMB+13]	2013	27	130,754	12
BP4D-Spontanous [ZYC+14]	2013	41	146,847	2
FERA 2015 [VAG+15]	2015	41	146,847	5
BP4D+ [ZGW+16]	2016	140	197,875	5
GFT [GCJC17]	2017	96	172,800	5

Table 1.2: List of publicly available databases that are annotated with facial action unit intensities.

Regarding the problem of pain intensity estimation, as our work is mainly focus on FACS based approaches (see Section 1.5.2), the UNBC McMaster appears to be the only FACS based pain database that available in research community.

1.6.1 UNBC McMaster database

The UNBC-McMaster Shoulder Pain Expression Archive Database (UNBC McMaster) [LCP+12] contains face videos of patients who suffer from shoulder pain while performing different range-of-motion tests on their arms to elicit the pain emotion. The participants were recruited from 3 physiotherapy clinics and by advertisements posted on the campus of the McMaster University. The inclusion criterium was self-identification with shoulder pain, which included different medical conditions such as arthritis, bursitis, tendonitis, subluxation, rotator cuff injuries, impingement syndromes, bone spur, capsulitis and dislocation. Two different movement tests are recorded, including active test and passive test. For active test, the subject moves the arm himself, while for passive test, the subject's arm is moved by a physiotherapist. Only one of the arms is affected by pain, but movements of the other arm are recorded as well for comparison. Of all the recorded sequences, 200 sequences of 25 subjects (13 women and 12 men) were selected

Figure 1.6: Sample images from the UNBC McMaster database.

for distributing to the research community. Totally, there are $48,398$ frame images in this publicly distributed portion of the database. Figure 1.6 shows some sample images of the database. It can be seen that images inside the dataset were cropped around the upper body of the participants.

For each frame, the intensities of pain related AUs, including AU 4, 6, 7, 9, 10, 12, 20, 25, 26, 27 and 43 are provided on a $[0-5]$ discrete intensity scale, except for AU 43, which is binary. The AU annotations were obtained by one of three certified FACS coders and an inter-observer agreement of 95% (according the Ekman-Friesen formula [EF78]) was reached on a small subset of the data ($1,738$ frames) which was annotated by all three coders. The intensity distribution of these AUs is shown in Figure 1.7. From this figure, we can see that the database is highly imbalance with more than 85% of the data is labelled as zero. Obviously, without rebalancing the dataset, any leading-based AU intensity estimation system derived from this dataset will be biased towards the zero intensity, as it is too dominant the dataset [JCDLT13]. Therefore, rebalancing is an important step before any further analysis on this data set.

Besides the AUs annotations, the database creators also provided discrete pain intensities according to Prkachin and Solomon method [PS08]. The pain score is basically the aggregation of 6 different AUs together to form a single number called PSPI whose value ranging between $[0-16]$ (see Section 1.5.2). As the distribution of AU intensities is unbalanced, so is the distribution of the PSPI score, as shown in Figure 1.8. In

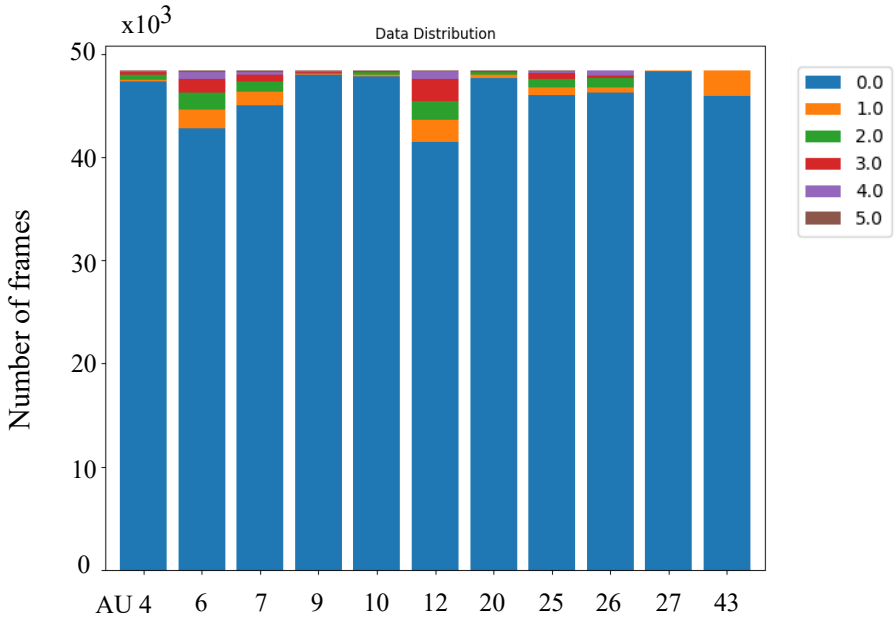


Figure 1.7: The distribution of facial AU intensity of the UNBC McMaster database.

addition to the PSPI and AUs annotations, 66 facial landmark points have also been extracted and provided in the database (see Section 2.2.2 for more information about facial landmarks).

The reason we chose the UNBC McMaster database is because it is the only database that focuses solely on pain and that provides detailed for both AU intensities and pain intensities per frame. Apart from that, in this database, the expressions of pain are spontaneous and come from people of different genders and ages. Furthermore, as the agreement score between the three FACS coders in the database has reached the score of 95%, the annotated AUs of the database appear to be highly reliable.

1.6.2 DISFA database

The Denver Intensity of Spontaneous Facial Action Dataset (DISFA) [MMB+13] is a non-posed facial expression database, which contains spontaneous facial expressions of subjects while watching a stimulating video. The video contains 9 short clips from youtube which are related to five emotions including: happiness, surprise, fear, disgust

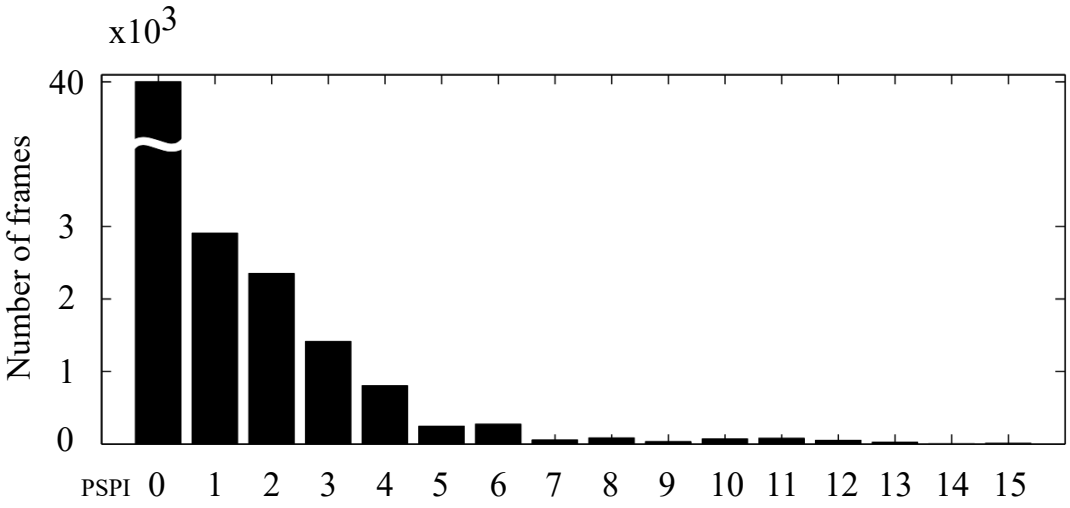


Figure 1.8: UNBC McMaster: Frame distribution of the PSPI intensity levels $[0-16]$.

and sadness. The participants were 27 adults (12 women and 15 men) of different ethnicities: three were Asian, 21 Euro-American, two Hispanic, and one African-American. Their age is between 18 and 50 years. Their facial behavior was recorded with a resolution of 1024×768 pixels, at 20 fps under uniform illumination. Figure 1.9 shows some sample frame images from the database.

Figure 1.9: Sample images from the DISFA database. It can be seen that the database contains both men and women, of different ethnicities and ages.

For each participant, $4,845$ video frames were recorded, resulting in a total number of $130,754$ frames for the whole database. Each of these frames has been annotated with AUs and their corresponding intensity on a $[0-5]$ discrete scale by a single expert FACS

rater. In order to validate the inter-observer reliability, 10 randomly selected videos were annotated by a second FACS rater and the inter-rater Intraclass Correlation Coefficient (ICC) for different AUs ranges from 0.80 to 0.94. For the record, ICC value of 0.80 and higher is considered as high reliability [Coh88]. Hence, the annotated AUs of the database seem to be reliable. The annotated AUs include AU 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, and 26. The intensity distribution of these AUs is shown in Figure 1.10. From this figure, we can see that the DISFA database also has the imbalance problem, as does the UNBC McMaster database. Hence, we also need to work on rebalancing the dataset before any further training or analysing. In addition to the annotations, the database creators also have provided 66 Active Appearance Model (AAM) tracked facial landmark points.

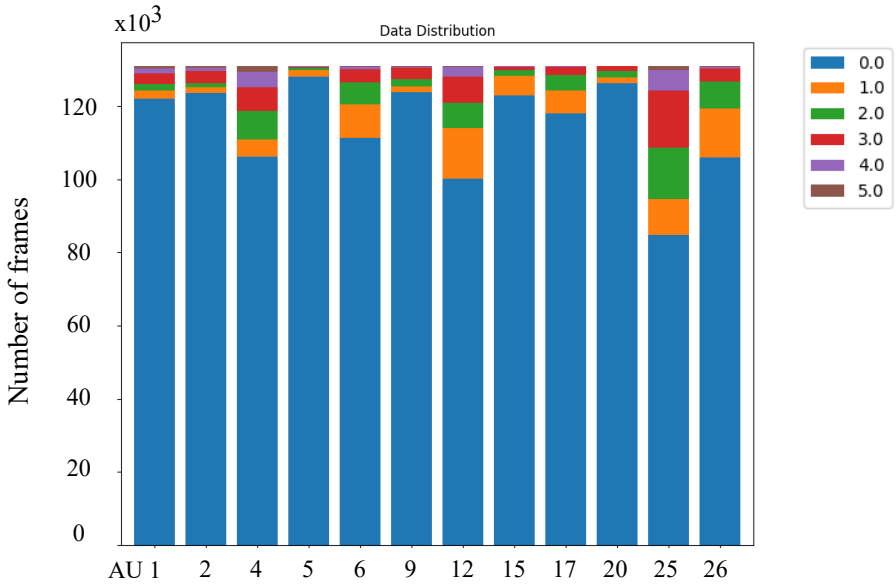


Figure 1.10: The distribution of facial AU intensity of the DISFA database.

We chose the DISFA as one of the main evaluating databases in our work since it is one of the few naturalistic databases which provide per-frame annotated videos for quite a lot of differences AUs (12 AUs) in all 6 intensity levels. Other databases only contain very few AUs or only posed facial expressions. More importantly, the annotated AUs in the database are highly reliable, with the ICC between expert raters

being greater than or equal to 0.80. As the correctness of the database will directly impact the performance of any system derived from the database, the high reliability of the database is one of the main points when we select a database for training and evaluation of our proposing approaches.

1.6.3 BP4D-Spontanous database

The Binghamton–Pittsburgh 4D spontaneous expression database (BP4D-Spontaneous) [ZYC+14] contains 328 sequences of facial 3D images and 2D texture recorded at 25 Frames Per Second (FPS) from 41 subjects. These subjects are including 18 males and 23 females of different ethnicities, which are including 11 Asian, 6 African-American, 4 Hispanic and 20 Euro-American. Their age is between 18 and 25 years. During each sequence, the subject is recorded while performing one of 8 interaction tasks with an experimenter. Each task is designed to elicit one of the emotions: happiness, sadness, surprise, embarrassment, fear, pain, anger and disgust. Each sequence has been AU annotated for the most expressive 20 sec period. The onset and offset are annotated for 27 AUs, including AU 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 27, 28, 30, 32, 38 and 39. Despite having quite a lot of different AU annotated, only two AUs are annotated as 6 level discrete intensities $[0-5]$, including AU 12 and AU 14.

Having the pain expression annotated is a plus point of this database. However, as the number of AUs that annotated with intensity are extremely low (only two AUs) and AU intensity is the main focus of our work. Hence, this database is not suitable for our purpose.

1.6.4 FERA 2015 database

The Facial Expression Recognition and Analysis challenge 2015 database (FERA 2015) [VAG+15] is the main database built for the FERA 2015 competition. The database is drawn from BP4D+ [ZGW+16] and SEMAINE [MVC+11] databases for the task of AU occurrence and intensity estimation. For the task of AU occurrence detection, 14 AUs

are drawn from both the two SEMAINE and BP4D+, including AU 1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, 25, 28, and 45. For the task of AU intensity estimation, 5 AUs are drawn from the BP4D+ database, including AU 6, 10, 12, 14, and 17. The training set of the FERA 2015 database consists of 146,847 images from the BP4D+ database and 48,000 images from the SEMAINE database. The testing set is kept private and contains 75,726 images from the BP4D+ and 37,695 images from the SEMAINE database. The inter-rater ICC for different AUs of the database ranges from 0.79 to 0.92, which indicates a strong to very strong inter-rater reliability for intensity.

FERA 2015 is a large database with a wide range of different AUs annotated. However, most of the annotated AUs are about the occurrence of AUs, only a few of them are annotated with intensities. Furthermore, as the database is no longer accessible[2], we not include this database in our work.

1.6.5 BP4D+ database

The Multimodal Spontaneous Emotion database (BP4D+) [ZGW+16] is a large-scale multimodal spontaneous emotion database, which has a similar style than the BP4D [ZYC+14] database but larger scale and variability. The database consists of 140 subjects (82 females and 58 males) of different ages and ethnicities. These subjects were asked to complete 10 tasks to elicit 10 different emotions, during which 2D RGB images, 3D model sequences, thermal videos and 8 physiological signal sequences with 1.4 million frames were captured by different sensors. Despite of having a large number of frames recorded, only 197,875 frames are FACS coded. The onset and offset are annotated for 34 AUs, including AU 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, and 39. Among these AUs, only 5 AUs are selected for intensity coding, including AU 6, 10, 12, 14, and 17. The inter-rater reliability of these AUs annotations ranges from 0.70 to 0.84, which indicates a good to strong reliability of the annotations.

[2]We contacted the person who manages FERA 2015 database and learned that the database is no longer accessible (10/05/2022)

1.6.6 GFT database

The Sayette Group Formation Task database (GFT) [GCJC17] is the first to address the need for a well-annotated facial expression database of multiple participants during unscripted interactions. The database consists of $172,800$ video frames from 96 subjects, spontaneously interacting with each other in group settings (from 2 to 3 persons per group). These subjects are including 54 males and 42 females with their age is between 21 and 28 years. They were drawn from a larger study on the impact of alcohol on group formation processes. The occurrence of 20 AUs was annotated in the database, which are including AU 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 17, 18, 19, 22, 23, 24, 28, and 99. Among them, five AUs were selected for intensity coding, including AU 1, 6, 10, 12, and 14. Regarding the reliability of the annotations, the inter-rater agreement between their FACS coders ranges from 0.72 to 0.88, which indicates a good to strong reliability for the AU annotations.

As the main purpose of the database is about studying the impact of alcohol, participants had have to drink some alcoholic beverages, which could lead to some differences between their expression and their true feelings. Moreover, the study were focusing on young people with their age is between 21 and 28 years, which is only a small portion of the population. Additionally, despite of having a large number of different AUs annotated, only a few are annotated with intensities. Hence, we do not include this database in our work.

Chapter 2

State of the art

Contents

In this chapter, we review prior works on automatic intensity estimation of facial expressions. We provide an overview about different techniques that can be used in face image analysis in Section 2.1. Then, in Section 2.2, we review the state of the arts preprocessing techniques for facial images, which are common to any face analysis approaches. Finally, we provide a literature survey on the fields of facial AUs and PSPI pain intensity estimation in Section 2.3.

2.1 Image processing techniques

In order to measure facial expressions from face image or video, we need to apply different image processing techniques to extract important information from the image. In this section, we briefly introduce these methods for analysing face images. We categorise these methods into two technical-groups including traditional and deep learning approaches. Traditional image processing approaches refer to the conventional hand design feature extracters, which requires a considerable amount of engineering skill and domain expertise, while deep learning based methods refer to the learning approaches that automatically learn to extract features by training on a large amount of data. We review the commonly used techniques in both of these two technical-groups in the next sections.

2.1.1 Traditional image processing methods

Traditionally, in order to measure facial expressions from an image, there are three main steps of image processing including feature extraction, dimensionality reduction, and feature estimation. Feature extraction reviews the techniques that can be used to extract a vector containing information about the face image, commonly called “feature” vector. The dimensionality of these features is then reduced by applying different dimensionality reduction techniques to remove irrelevant or redundant information. Finally, these features data are fed into feature estimation model for measuring the

intensity of the facial expression.

2.1.1.1 Feature Extraction

Feature extraction refers to the task of extracting important information (features) from an image. In this section, we introduce some popular techniques that can be used for extracting features from images, including spatial filter based and histogram based methods. For in-depth review, see Zhi *et al.* [ZLZ20].

Spatial filter based methods Spatial filter is refer to a multi-resolution technique which consists of a set of discrete wavelets that scan over the facial images to capture both frequency and location information. Generally, the spatial filter based methods adopt the two-dimensional form to deal with facial image processing.

Gabor wavelet (Gabor) [Gab46] is a popular spatial filter technique for extracting features from images. Gabor features are conducted by convolving facial images with a specific set of Gabor filters of various orientations and scales. A Gabor filter can be viewed as a sinusoidal signal of particular frequency and orientation, modulated by a Gaussian wave. Gabor filters are represented as below:

$$
\begin{aligned}
g(x, y, w, \theta) &= \frac{1}{2\pi\sigma^2} e^{\frac{-(x'^2+y'^2)}{2\sigma^2}} [e^{iwx'} - e^{\frac{-w^2\sigma^2}{2}}] \\
x' &= xcos\theta + ysin\theta \\
y' &= -sin\theta + ycos\theta
\end{aligned}
\tag{2.1}
$$

where x and y represents the coordinates of the pixel value in spatial domain, w represents the radial center frequency, θ represents the orientation of the Gabor filter, and σ is the standard deviation of the round Gaussian function along the $x - y$ axis.

Gabor features provide multi-scale characteristics of the facial images, reflecting the local neighboring relationship among pixels. The features are tolerant to illumination variations, small translation and rotations, and robust to registration errors to a degree [SGC14]. However, Gabor filtering is known to be computationally expensive and

suffers from identity bias [ZLZ20].

Haar wavelet (Haar) [Haa11] is another spatial filter technique which is faster than Gabor wavelet in feature extraction from images. Haar wavelet exploits the pair of low-pass filters and high-pass filters in facial image columns and rows independently, where the mean and difference of two adjacent pixel values are figured out for low-pass and high-pass filtering individually [ZLZ20]. Haar features is robust to illumination variations, and an acceptable extent registration error. Haar wavelet is the simplest possible wavelet, and it is suitable for use in a real-time application system.

Histogram based methods Besides spatial filter based methods, there are another type of feature extraction techniques that are based on histograms of quantised local descriptors. A local descriptor uses the image intensities within a small neighbourhood, with only a few pixels in diameter. The quantised local descriptor response is accumulated over a larger image region within a histogram. This process discards spatial information and thus provides a compressed descriptor that is invariant regarding small translations.

Local Binary Pattern (LBP) [OPM02] is one of the widely-known histogram based feature extraction technique. It uses the sign of the intensity difference between the center pixel and circular surrounding pixels as local descriptor. The value of each center pixel is converted to an integer, which forms the LBP histogram with counting all the integers. Perhaps the most important property of LBP is its robustness to monotonic gray-scale changes caused, for example, by illumination variations [ZLZ20]. Another important properly of LBP is its computational simplicity, which makes it possible to analyse images in challenging real-time configurations.

Histogram of Oriented Gradients (HOG) is another histogram based method, uses the intensity gradients as local descriptor. It is simplicity in computing and represents both texture and shape-skin information. The use of orientation histograms has many precursors, but it only reached maturity when combined with local spatial histogram and normalisation in Lowe's Scale Invariant Feature Transformation approach [DT05],

in which it provides the underlying image patch descriptor for matching scale-invariant keypoints [ZLZ20].

Scale-Invariant Feature Transform (SIFT) [Low04] is a histogram based method that uses weighted 3D histogram of gradient locations and orientations as local descriptor. The SIFT features are robust to rotation and scale, meanwhile they are tolerant to illumination variations and small registration errors.

2.1.1.2 Dimentionality Reduction

Since the features extracted from face images can have many dimensions, sometimes more than several thousand, there is a need of reducing the number of dimensions of these features. Dimentionality reduction (DR) methods provide a mapping from the original features to a feature subspace, either by selecting a subset of dimensions or by mapping to a new space of reduced dimensionality. This section briefly reviews some common DR methods that are widely used in face image analysis. Further details can be found in the comparative review by Van der Maaten *et al.* [VDMPVdH+09].

Principle Component Analysis (PCA) [Jol05] is one of the oldest and most studied DR method. It is based on the extraction of the important and relevant information as new orthogonal feature vectors called principal components from a set of input observations. These principal components are linear combinations of the original variables, with the first principal component having the largest variance, the second principal component having the second largest variance, and so on. It is thus possible to select a number of significant components, so that data dimension is reduced by preserving the systematic variation in the data retained in the first selected components, while noise is excluded, being represented in the last components [Bal15]. As a result, PCA can drastically reduce the dimensionality of the original feature vectors without loss of much information in the sense of representation.

There are some other matrix factorisation methods which also calculate linear projections of the data like PCA, but imposing different constraints. Independent Compo-

nent Analysis (ICA) [HO00] constrains the new dimensions to be not only uncorrelated but statistically independent. This characteristic allows ICA to reveal hidden factors that underlie sets of random variables, measurements, or signals. Non-negative Matrix Factorisation (NMF) [LS99] is another matrix factorisation method that constraints all values to be greater or equal to zero, thus well suited for pixel intensities. NMF provides a part-based decomposition of the data, i.e., most of the new component weights are zero. Both, ICA and NMF lead to sparse subspace weights.

Linear Discriminant Analysis (LDA) [Fis38] is another DR approach, which can be defined as a supervised learning method that works by transforming the data onto a subspace that maximises the ratio between-class variance to within-class variance in order to increase the separation between classes. A more recent approach is Spectral Regression (SR) [CHZH07] , which first performs spectral analysis on the Laplacian matrix, followed by learning a linear projection through least squares regression.

2.1.1.3 Feature estimation

In the domain of traditional facial expressions measurement, after extracting features from a face image, we need to apply a feature estimation method to estimate the intensity of an expression. Normally, this feature estimation method belongs to supervised learning approaches because it works best with vast amounts of fuzzy data. In this section, we briefly review some common supervised learning techniques that can be used to measure the intensity of facial expressions.

Support Vector Machines Support Vector Machines (SVMs) are powerful statistical classifier for binary linear classification. The main idea behind SVMs is the creation of distinct borders between partitions of given data, in order to break the data into multiple sections that could be used for classification purposes with the future input [Bur98]. SVM was introduced in 1992 by Vapnik *et al.* [BGV92] and can be used for classification (SVC) [Vap99] or regression (SVR) [DBK+96]. Support Vector Classification (SVC) is a max-margin classifier, i.e., it learns the decision boundary by maximising the margins

between the classes. Support Vector Regression (SVR) maps regression to a classification problem, by defining a tube around the target function as the correct class and then applying the same max-margin framework as SVC.

Relevance Vector Machines Relevance Vector Machines (RVMs) [Tip01] is a machine learning technique based on Bayesian formulation of a linear model with an appropriate prior that results in a sparse representation. As a consequence, they can generalise well and provide inferences at low computational cost. RVMs is kernel based machine learning approach and can be used as an alternative to SVMs for both regression and classification problems. The advantages of the RVMs over the SVMs is the availability of probabilistic predictions, using arbitrary kernel functions and not requiring setting of the regularisation parameter [Pal11].

Adaptive Boosting Adaptive Boosting (AdaBoost) [FS97] is an optimisation method and can aslo be used for both classification and regression problems. It is based on the idea of creating a highly accurate prediction rule by combining many relatively weak and inaccurate rules. AdaBoost creates a collection of component classifiers by maintaining a set of weights over training samples and adaptively adjusting these weights after each toosting iteration: the weights of the samples which are misclassified by current component classifier will be increased while the weights of the samples which are correctly classified will be decreased. As AdaBoost consists of many classifiers, this optimisation method appears to be slower compared to other learning approaches. On the other hand, AdaBoost seems to have a great resisting against the overfitting problem [WLH+19].

2.1.1.4 Conclusion

In this section, we have reviewed some traditional image processing methods that can be used for analysing face images. These methods requires a considerable amount of engineering skill and domain expertise to be able to correctly analyse facial expressions

from face image or video. Besides these traditional approaches, there is another group of image processing techniques that are able to automatically learn to extract rich information from large amount of data, which we will discuss in the next section.

2.1.2 Machine Learning and Deep Neural Network

Machine Learning (ML) is a field of study which allows machines to learn from data or experience and make a prediction based on the experience. Instead of trying to program knowledge into computers, Machine Learning (ML) seeks to automatically learn meaningful relationships and patterns from examples and observations [BN06]. Depending on the approach, type of input/output data, and kind of tasks to achieve, ML can be categorised into three categories, including *supervised*, *unsupervised learning*, and *reinforcement learning*.

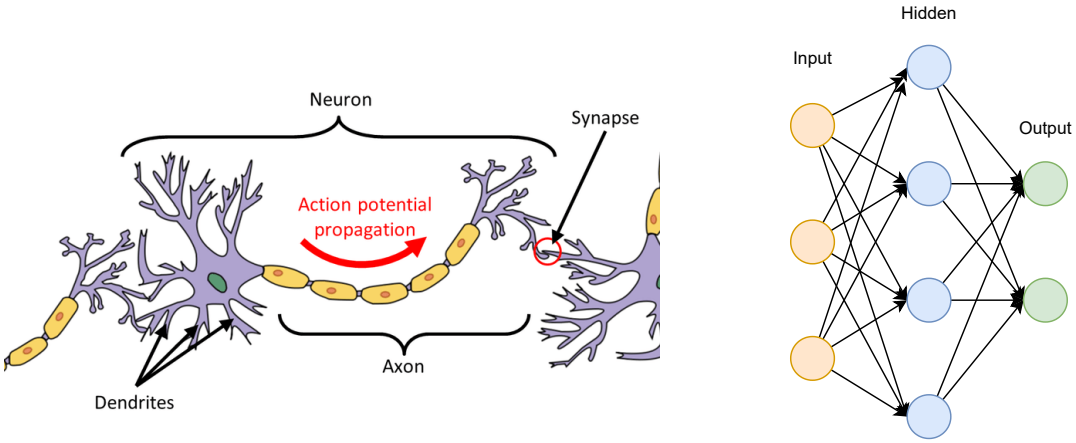
In *supervised learning* [LBH15], the ML model learn from examples. In the training process, each pair of input data and its ground truth label is used to calibrate the open parameters of the ML model. Once the model has been successfully trained, it can be used to predict the label of newly unseen data.

In *unsupervised learning* [Fri98], the ML model track operations to describe the structure of unlabelled data. For example, clustering analysis is a branch of this group that proposes to classify the unlabelled data. The algorithm tries to identify the common features of data belonging to a group. When a new piece of data appears, it will be assigned to the group which exhibits the same common features.

In *reinforcement learning* [JZH21], instead of providing input and ground truth pairs, we describe the current state of the system, specify a goal, provide a list of allowable actions and their environmental constraints for their outcomes. Then, we let the ML model to experience the process of achieving the goal by itself using the principle of trial and error to maximise a reward concerning how to map situations to actions.

An algorithm which is built for tasks of a ML system and able to learn from data is called a *ML algorithm*. Depending on the learning task, there are various classes of

ML algorithms, each of them coming in multiple specifications and variants, including regressions models, instance-based algorithms, decision trees, Bayesian methods, and Artificial Neural Networks (ANNs). In the next sections, we will briefly introduce some *ML algorithms* that have been widely used for analysing facial expression from image.



a) Biological neuron b) Artificial neural network

Figure 2.1: The visualisation of (a) the Biological neuron from [Vod17] and (b) the Artifical neural networks.

Among these classes, ANNs is of particular interest since their flexible structure allows them to be modified for a wide variety of contexts. Inspired by the principle of information processing in biological systems, ANNs consist of mathematical representations of connected processing units called artificial neurons. Like synapses in a brain (see Figure 2.1a), each connection between neurons transmits signals whose strength can be amplified or attenuated by a weight that is continuously adjusted during the learning process. Signals are only processed by subsequent neurons if a certain threshold is exceeded as determined by an activation function [JZH21]. Typically, neurons are organised into networks with different layers. An input layer usually receives the data input , e.g., face image, and an output layer produces the final result. In between, there are zero or more hidden layers (see Figure 2.1b) that are responsible for learning a non-linear mapping between input and output [BN06, GBC16]. Deep Neural Networks (DNNs) typically is refer to ANNs which consists of more than one hidden layer,

organised in deeply nested network architectures. Each level of these network learns to transform its input data into a slightly more abstract and composite representation. With the composition of enough such transformations, very complex functions can be learned. The key aspect of DNNs is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure [LBH15]. The hierarchical relationships between ML, ANNs, and DNNs is summarised in Venn diagram of Figure 2.2.

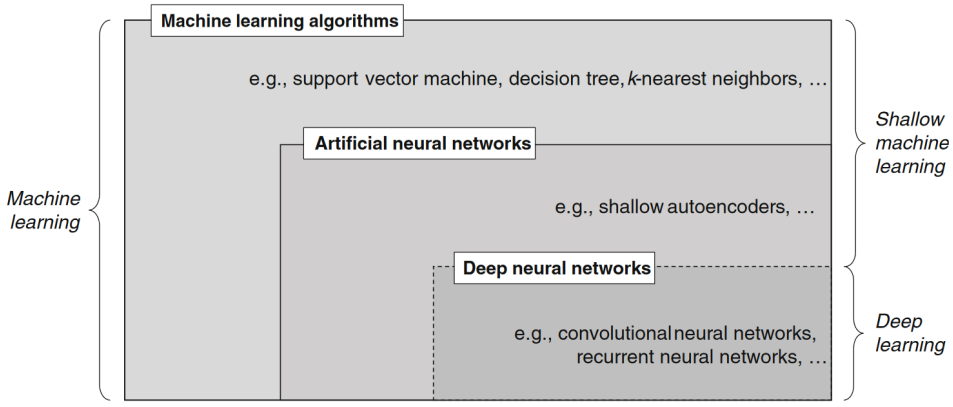


Figure 2.2: Venn diagram of machine learning concepts and classes. Image from [JZH21].

Deep Neural Networks consist of multiple layers with thousands or millions of adjustable parameters. These adjustable parameters, often called weights, are real numbers that can be seen as 'knobs' that define the input-output function of the network. In order to adjust the weights of the network appropriately and automatically, we use learning algorithms such as Stochastic Gradient Descent (SGD). Basically, the learning algorithm computes a gradient vector with respect to the weights of the network though a process called gradient backpropagation [LBH15]. This gradient vector, for each weight of the network model, indicates by what amount the prediction error would increase or decrease if the weight were increased by a tiny amount [LBH15]. The weight vector is then adjusted in the opposite direction to the gradient vector. By slowly moving with each of these tiny steps, it will eventually reach to the point where the error is minimal, the model is then fully trained and can be used for predicting newly unseen

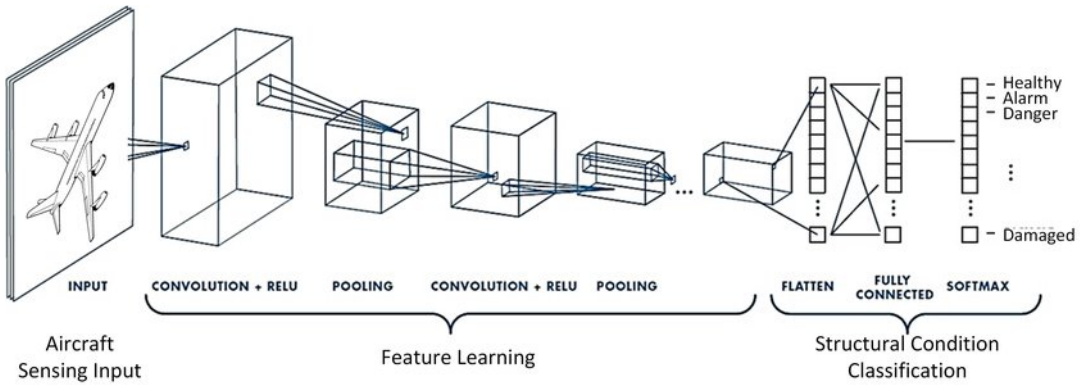


Figure 2.3: A visualisation of a typical Convolutional Neural Network. Image from [TFSK19].

samples.

In the early ages, Deep Neural Network encounters several problems to take into account real-world cases because of the limitation of the memory size or computing power. When applying to images, DNNs model can easily have tens of millions of free parameters, which can take weeks to train using a conventional implementation on a single CPU [RMN09]. However, with advent of fast and convenient programming on Graphics Processing Units (GPUs), researchers were able to train their DNNs model at about 100 times faster compared to training on CPU [RMN09], which opens a new era in the revolution of Deep Learning.

Deep Learning has been making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. With the invent of many powerful variants of DNNs such as Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN), DNNs is beating records in many domains including image recognition and analysis [KSH12, FCNL12, LGTB97], speech recognition [Zue90, HDY+12], object detection [RHGS15, BWL20], and more. In the next sections, we focus on Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) which variants of DNNs useful for the analysis of face images and videos.

2.1.2.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is a special type of DNNs which is designed to process data that comes in the form of multiple arrays, e.g., a colour image composed of three 2D arrays contaning pixel intensities in the three colour channels. Many data modalities are in the form of multiple arrays: 1D for signals and sequences, including language; 2D for images or audio spectrograms; and 3D for video or volumetric images. There are four key ideas behind CNNs that take advantage of the properties of natural signals: local connections, shared weights, pooling and the use of many layers [LBH15].

A typical CNN is shown in Figure 2.3. The network requires to have a *Convolutional layer* but can have other types of layers, such as *Activation function* (e.g., ReLU layer), *Pooling*, and *Fully connected layers*, to create a deep Convolutional Neural Network. Sometimes, *Dropout layers* are added, for example, between the *Fully connected layers* to prevent the overfitting of the network. In CNN, convolutional filters are trained using the backpropagation method. The values of these filters learn automatically through training to extract relevant features from input data, depending on the given task. For example, in an application such as face detection, one filter can perform edge extraction, whereas another can carry out eye extraction. In the next paragraphs, we briefly introduce some important layer types in CNN.

Convolutional layer Convolutional layer of CNN consists of multiple learnable filters which slide over the layer for the given input data (see Figure 2.4a). A summation of an element-by-element multiplication of the filters and receptive field of the input is then calculated as the output of this layer. The weighted summation is placed as an element of the next layer. Each of the convolutional operation is specified by stride, filter size, and padding. Stride, which is a positive integer number, determines the sliding step. For example, stride 1 means that we slide the filter one place to the right each time and then calculate the output. Filter size (also called receptive field) must be fixed across all filters used in the same convolutional operation. Padding configuration adds a number

of rows and columns with zero values to the original input matrix to control the size of the output feature map [Wu17]. Without using padding, the convolution output is smaller in size than the input. Therefore, the network size shrinks by having multiple layers of convolutions, which limits the number of convolutional layers in a network. Padding prevents this shrinking effect of convolutional layer and provides the ability to have unlimited deep layers in our network architecture.

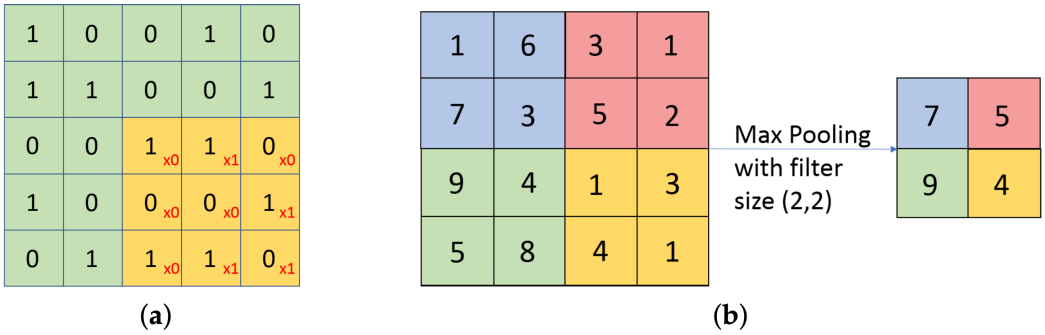


Figure 2.4: (a) Convolution operation. The amber squares represent the position of the kernel as it slides through the green input slice. (b) Max Pooling Operation with a filter size of (2,2). Image from [TFSK19].

Activation function Activation functions are functions used in neural networks to computes the weighted sum of input and biases, of which is used to decide if a neuron can be fired or not. Activation function can be either linear or non-linear depending on the function it represents, and are used to control the outputs of out neural networks. The choice of which type of activation function to be used in a neural network varies depends on each architecture. However, if we only use linear activation function, the output of each layer in DNN is a linear function of the upper layer. Then, no matter how many layers the DNN has, the outputs are linear combination of the inputs. Hence, in order to get access to a much richer hypothesis space that would benefit from deep representations, we need non-linearity activation functions [Cho21]. Non-linearity activation functions such as sigmoid [LDH19] or hyperbolic tangent [LDH19] have been widely used in the convolution classification model during the beginning of

deep learning research, but all of them are easy to make the convolution model appear the phenomenon of gradient diffusion [WLSR20]. The coming of ReLU function [ESH19] has effectively solved the above problem and becomes one of the most common non-linearity activation function applied in various fields, such as image processing [Wu17]. That is being said, other activation functions such as linear, sigmoid or hyperbolic tangent are still useful depend on each situation, e.g., linear activation function can be used as the output layer of a neural network model to solve a regression problem.

Pooling layer Pooling layer is mostly used to down-sampling the size of inputs with the purpose to reduce the spatial resolution of the feature map and so to reduce the computation cost. There are two major types of pooling: max pooling and average pooling [SMB10]. The pooling operation is also based on a sliding window which goes through the input feature map, and the pooling operation is conducted in the overlapping area of the pool. For max pooling layer, as can be seen in Figure 2.4b, the pooling operation outputs the maximum value of the given matrix while it is obviously the average value for the average pooling layer. Regarding the performance of these two layers, Boureau *et al.* [BPL10] provided a detailed theoretical analysis of their performances in selecting features. Scherer *et al.* [SMB10] further conducted a comparison between the two pooling operations and found that max-pooling can lead to faster convergence, select superior invariant features and improve generalisation of the whole CNN network.

Fully Connected layer In the CNN, Fully-connected (FC) layer usually follows the group of convolutional and pooling layers, as can be seen in Figure 2.3. The main use of this layer is to extract the abstract feature representations of the input data. Depending on the problem, an activation function can be added to promote the output for the network. For example, we use linear activation function for a regression problem, and for binary classification we use sigmoid activation function [LDH19]. A CNN may have one or more FC layers, and most of the time these layers dominate the number of parameters in a CNN [HMD15]. Figure 2.5 shows an example of the comparison

Layer	#Weights	Weights% (P)	Weight bits (P+Q)	Weight bits (P+Q+H)	Index bits (P+Q)	Index bits (P+Q+H)	Compress rate (P+Q)	Compress rate (P+Q+H)
conv1	35K	84%	8	6.3	4	1.2	32.6%	20.53%
conv2	307K	38%	8	5.5	4	2.3	14.5%	9.43%
conv3	885K	35%	8	5.1	4	2.6	13.1%	8.44%
conv4	663K	37%	8	5.2	4	2.5	14.1%	9.11%
conv5	442K	37%	8	5.6	4	2.5	14.0%	9.43%
fc6	38M	9%	5	3.9	4	3.2	3.0%	2.39%
fc7	17M	9%	5	3.6	4	3.7	3.0%	2.46%
fc8	4M	25%	5	4	4	3.2	7.3%	5.85%
Total	61M	11%(9×)	5.4	4	4	3.2	3.7% (**27**×)	2.88% (**35**×)

Figure 2.5: Compression statistics for AlexNet. P: pruning, Q: quantization, H:Huffman coding. Image from [HMD15].

between the number of parameters of FC layers compare to other layers in AlexNet [KSH12]. It can be seen that these FC layers contain more than 90% parameters of the network. As these layers contain a large number of parameters, which results in a large computational effort for training them. Therefore, a promising and commonly applied direction is to remove these layers or decrease the connections with a certain method. For example, GoogLeNet [SLJ+15] designed a deep and wide network while keeping the computational budget constant, by switching from fully connected to sparsely connected architectures.

The four types of layers that we have mentioned above are not the only ones that exist. A large number of different layers can be found in the related documents of deep learning [LBH15, GWK+18]. Over the past few years, storage has become more affordable, datasets have grown far larger, and the field of parallel computing has advanced considerably. All these conditions give wings to CNNs to become dominant method in a variety of computer vision problems such as image classification [KSH12, FCNL12, LGTB97], object detection [RHGS15, BWL20], semantic segmentation [LSD15, GDDM13, KS14], computational creativity [GPAM+20, CWD+18], and many more. However, despite their power, CNNs also have limitations. For example, CNN works with each training sample independently, without considering whether there are relationships between each of the training samples. In many cases where

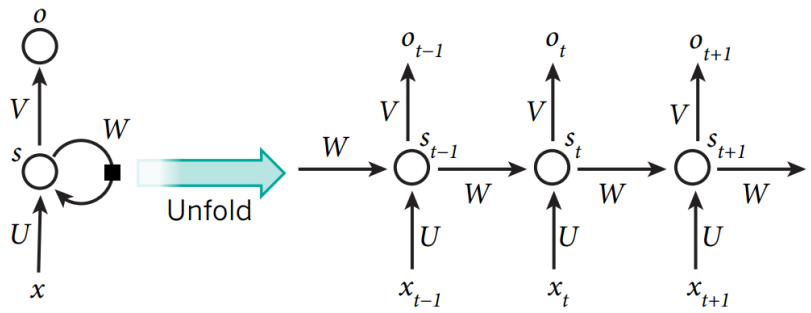


Figure 2.6: The visualisation of a simple Recurrent Neural Network with its unfolding in time calculations. Image from [LBH15].

each sample is independent (e.g., image classification), this presents no problem. But if data points are related in time or space (e.g., natural language processing), this is unacceptable. There is a need of a different type of DNNs that takes into account the relationships between each data point in a sequence. That is when Recurrent Neural Network comes into the picture.

2.1.2.2 Recurrent Neural Network

Recurrent Neural Network (RNN) is a special type of DNNs which is designed to process data that comes in the form of sequences, e.g., frames from video, snippets of audio, or words in sentences. Traditional DNNs such as CNNs rely on the assumption of independence among the training and test examples, which is not the case for these sequential data. For this kind of data, there are relationships between the data points in a sequence. Hence, DNNs such as RNNs are designed to exploit these relationship information. RNNs are connectionist models with the ability to selectively pass information across sequence steps, while processing sequential data one element at a time. Thus they can model input and/or output consisting of sequences of elements that are not independent. Further, RNNs can simultaneously model sequential and time dependencies on multiple scales [LBE15]. The visualisation of a simple RNN with its unfolding can be seen in Figure 2.6. Each step in the unfolding is referred to as a time step, where x_t is the input at time step t. RNNs can take an arbitrary length sequence as input, by

providing the RNN a feature representation of one element x of the sequence at each time step. s_t is the hidden state at time step t and contains information extracted from all time steps up to t. The hidden state s is updated with information of the new input x_t after each time step:

$$s_t = f(Ux_t + Ws_{t-1}) \tag{2.2}$$

where U and W are vectors of weights over the new inputs and the hidden state, respectively. Function f, known as the activation function (see Section 2.1.2.1), is usually either the hyperbolic tangent or the sigmoid function. RNNs, once unfolded in time, can be seen as very deep feedforward networks in which all the layers share the same weights. Although their main purpose is to learn long-term dependencies, theoretical and empirical evidence shows that it is difficult to learn to store information for very long [LBH15]. Standard RNNs fail to learn in the presence of time lags greater than 5 - 10 discrete time steps between relevant input events and target signals [GSC00] due to gradent vanishing problem. Recently, researchers have proposed two variants of RNNs that are capable of solving the problem, including Long Short-Term Memory and Gated Recurrent Unit. We will briefly introduce these two networks in the next paragraphs.

Long Short-Term Memory LSTM network is a variant of RNN which is designed to learn long-term dependencies in sequence prediction problem. The network consists of three layers including input, hidden and output layers. Unlike the standard RNN, the basic unit of LSTM's hidden layer is a memory block, and LSTM adds a 'processor' in the algorithm to decide whether the information is useful or not, which is called a cell [GBC16]. The typical structure of a LSTM cell is shown in Figure 2.7a. A LSTM cell is configured mainly by three gates: *Input gate*, *Forgot gate*, and *Output gate*. These gates regulate the flow of information into and out of the cell, indicate which information to keep and to be discarded. Let x_t and h_t to be the input vector and the hidden state

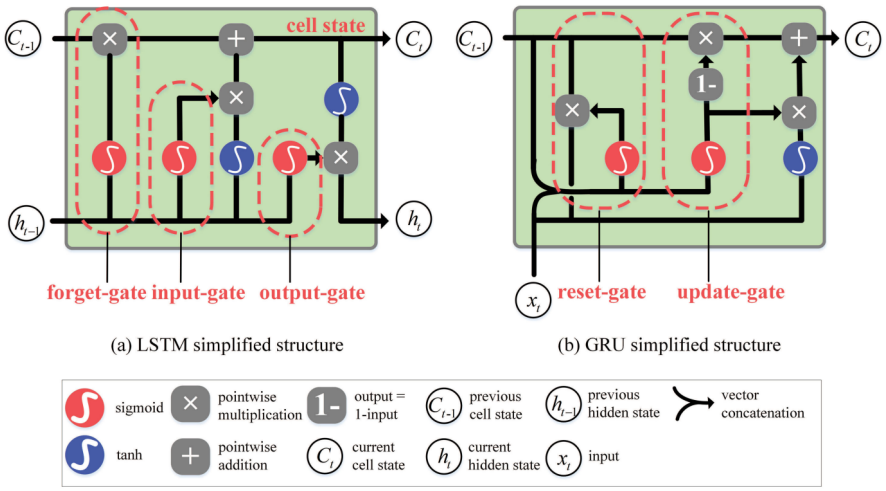


Figure 2.7: The visualisation of the simplified architecture of (a) LSTM and (b) GRU layers. Image from [ZNNS20].

at time step t, respectively. Then, the *Forget gate* f_t determines whether x_t should be retained or not by applying sigmoid activation function on the combination of x_t and the previous hidden state h_{t-1} as below:

$$f_t = \sigma\big(W_f \cdot \big[h_{t-1}, x_t\big] + b_f\big) \tag{2.3}$$

Where W_f and b_f denote the weight and bias of the forget gate. The output value f_t ranges between $[0, 1]$ to make decisions for filtering non-significant information. Next, the *Input gate* i_t determines the extent to which new memories should affect old memories (Equation 2.4). Meanwhile, this unit determines how much new information should be delivered to the next cell (Equation 2.5). Then, the cell state is updated through discarding the information that needs to be discarded and adding new information (Equation 2.6).

$$i_t = \sigma\big(W_i \cdot \big[h_{t-1}, x_t\big] + b_i\big) \tag{2.4}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.5}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{2.6}$$

Finally, we need to decide what we are going to output. This output will be based on our cell state, but will be a filtered version. In *Output gate*, we run a sigmoid layer which decides what parts of the cell state we are going to output (Equation 2.7). Then, we put the cell state through $tanh$ (to push the values to be between $[-1, 1]$) and multiply it by the output of the sigmoid gate (Equation 2.8), so that we only output the parts we decided to.

$$o_t = \sigma\big(W_o \cdot \big[h_{t-1}, x_t\big] + b_o\big) \tag{2.7}$$

$$h_t = o_t \times tanh(C_t) \tag{2.8}$$

The learning process of LSTM mainly includes the error backpropagation process and optimization algorithm. The Backpropagation Through Time (BPTT) algorithm [WZ95] is applied in the error backpropagation process of LSTM.

Gated Recurrent Unit GRU was proposed by Cho *et al.* [CvMG+14] in 2014, similar to LSTM, but simpler to compute and implement. The typical structure of GRU cells is shown in Figure 2.7b. Instead of having three gates: *Input gate*, *Forget gate*, and *Output gate* to control the flow of data as in LSTM, the GRU cell has only two gates including *Update gate* and *Reset gate*. The *Update gate* is in charge of inputting and discarding information, which covers the work of the *Input gate* and the *Forget gate* in LSTM. The *Reset gate* focuses on how much previous information to be discarded. In GRU, the fewer parameters are computed and processed, and the hidden state is propagated directly among the network cells instead of being controlled by the *Output gate*. Hence, GRU is simpler in implement but faster in training and evaluating compared to LSTM neural network.

2.1.2.3 CNN-RNN hybrid neural network

There are some circumstances where we need to extract both local features and temporal features from the input data. For example, in video processing, we need to extract

the spatial features from each frame image to see what we have in a single frame and the temporal features to see the changing over time of these spacial features. When CNN is great for extracting spacial features and RNN is great for extracting temporal features, we can combine these two type of DNNs to effectively learn to model both spacial and temporal information. In fact, the combination of CNN-RNN has been proven successful in several classification and regression tasks for modeling spacial-temporal information in many previous works. For example, they have been used for handwriting recognition [DKMJ18], speech recognition [HZZ+20] from audio streams. In the domain of video analysing, CNN-RNN architectures have also been used for emotion detection [KZ20], sign language recognition [MSTA18] or action recognition [UAM+17], taking advantage of their ability to learn scene features using the CNN and sequential features using the RNN.

There are two main types of CNN-RNN hybrid neural network in the domain of video analysing: sequence-based and frame-based networks. Sequence-based network is basically the network designed to work on sequential data where we have only one label per sequence. The network takes a sequence of frame images and tries to predict the label for the whole sequence. For this sequential data, sign language recognition and action recognition are the two well-known sequence-related problems when we have only one gesture or action labelled per video. To solve this problem, in sequence-based CNN-RNN hybrid network, features extracted by CNN network are fed into RNN network, and only the RNN hidden states of the last frame are extracted for the classification or regression tasks. Since the RNNs such as LSTM or GRU have the capacity of remembering long-term dependencies, hence features from the last frame of a sequence should contain the important information of the whole sequence. Figure 2.8 shows the visualisation of an example of sequence-based CNN-RNN architecture.

Besides the sequence-based network, we also have the frame-based CNN-RNN hybrid network. Frame-based network is basically the CNN-RNN hybrid network that is designed to work on per-frame labelled sequential data. Instead of using only the RNN hidden states of the last frame as in sequence-based network, this frame-based network

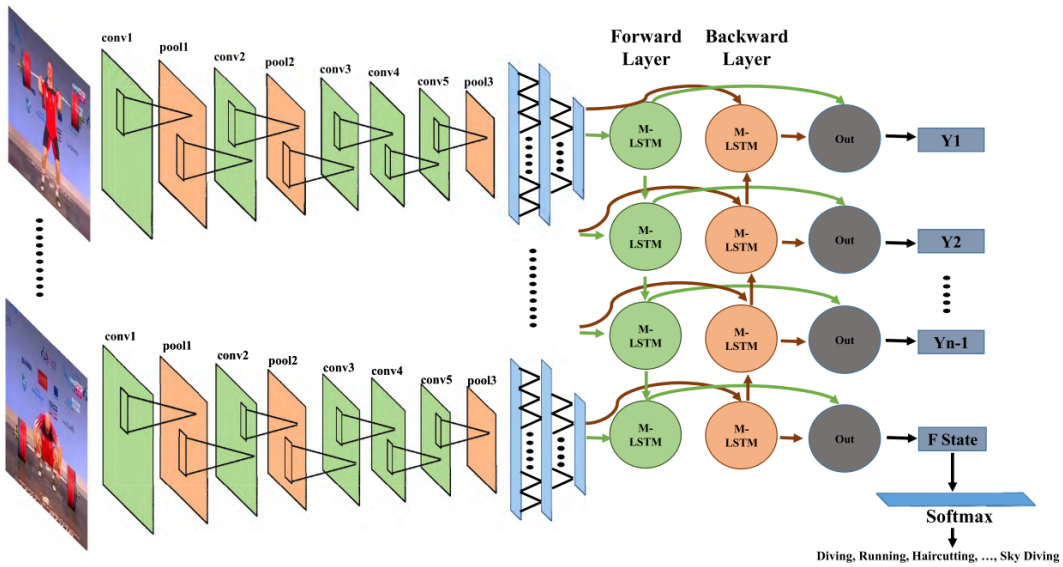


Figure 2.8: The visualisation of an example sequence-based CNN-RNN architecture. Image from [UAM+17].

uses the hidden states of all time steps in a sequence of frame images, each time step is corresponding to the label of each input frame. In fact, for frame-based type of data, as we have a label for each frame image, we can simply train a CNN-only neural network to perform the task, without the need of RNN. However, in order to further improve the performance of the whole network, one can integrate the RNN network to model the temporal dynamics information of the data. Figure 2.9 shows the visualisation of a frame-based CNN-RNN hybrid neural network.

CNN-RNN hybrid neural network is a great way to model both spacial and temporal information of the input data. In the domain of face video analysis, CNN-RNN networks have been widely used and achieved great successes in many domains such as emotion recognition [LZH+19, VBA21], micro-expressions recognition [ASH+22], valence-arousal estimation [KZ20, DCS20], pain intensity estimation [VBADE21, RCG+17]. However, CNN-RNN hybrid networks are not the only way to model spatial-temporal information, there is another type of DNNs called 3D Convolutional Neural Network (3D CNN) which have the same capacity to some extent. In the next section, we will briefly introduce 3D CNN and also discuss about their use in our thesis.

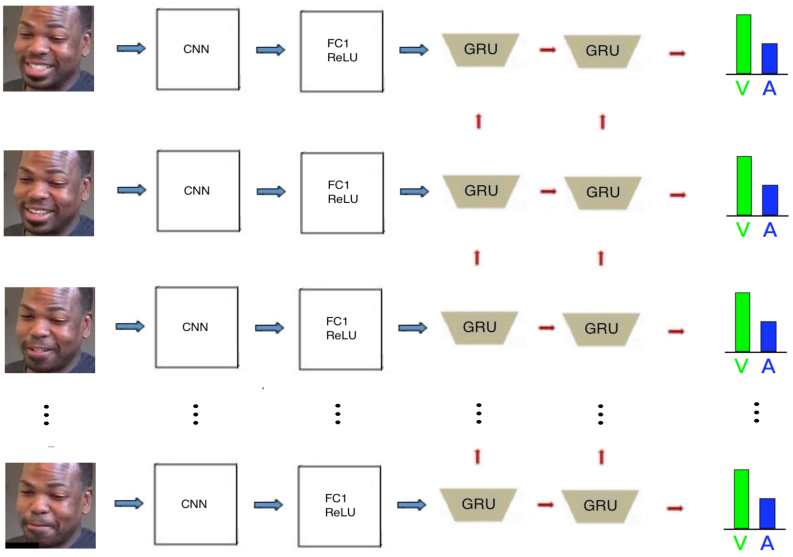


Figure 2.9: The visualisation of a frame-based CNN-RNN architecture. Spacial features are extracted from each images by a CNN network, then these features are fed to a GRU layer for extracting temporal information. Image from [KTN+19].

2.1.2.4 3D Convolution Neural Network

3D Convolutional Neural Network (3D CNN) [JXYY12] is a logical extension of CNN which works with three dimensional data like video which has an additional temporal dimension in addition to the X and Y co-ordinates [RM19]. Figure 2.10 shows an example of 3D CNN architecture for video classification problem. It can be seen that the 3D convolution of 3D CNN is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frame images together. By this construction, the feature maps in the convolution layer is connected to multiple contiguous frames in the previous layer, thereby capturing temporal information [JXYY12]. Because of this characteristic of 3D CNN, in terms of temporal information modelling capability, while CNN-RNN hybrid networks focus on learning global temporal information, 3D CNNs focus on modelling local temporal information of the input data. Also because of this characteristic of coupling spatial and temporal signals with each other through each 3D convolution, when training these 3D CNNs, it becomes much more difficult to optimise the network's parameters because of the exponential growth of the solution space with respect to the

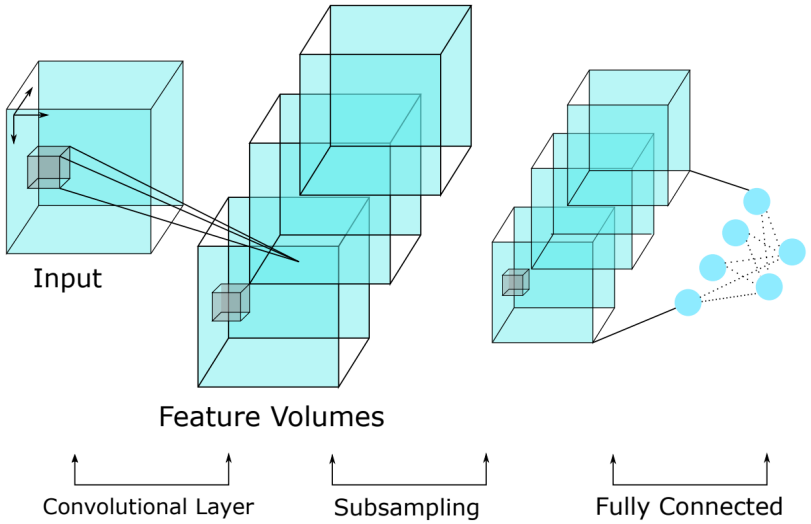


Figure 2.10: Basic 3D CNN architecture: the 3D filter is convolved with the video in three dimensions as indicated by the arrows to produce feature volumes. After subsampling and flattening the features are fed to a fully connected layer for classification. Image from [RM19].

case of 2D CNNs [ZSZZ18]. Besides, the memory cost and model size of 3D CNN are much higher compared to 2D CNN [KR19]. For example, an 11-layer 3D CNN requires nearly 1.5 times as much memory as a 152-layer Residual Network [ZSZZ18]. Furthermore, as for model/computation complexity, 3D CNNs are much more expensive than 2D CNNs and prone to overfit [XSH+18]. Meanwhile, as this work focuses on facial AU and pain intensity estimation, the amount of well-labelled publicly available data on these domains are generally limited (see Section 1.6). As the size of these 3D CNNs is generally huge and the amount of training data is small, poor generalisation is to be expected [L+89]. Thereby, in this thesis, when modeling spacial-temporal information from face video, we have chosen to use CNN-RNN hybrid network instead of 3D CNNs.

To conclude, in this section, we have introduced different image processing techniques for analysing face images, including both traditional and deep learning approaches. These techniques are fundamental for any facial expression measurement application to be able to work properly. In the next section, we discuss about how to apply these techniques to measure human facial expressions from image and video. We review the state-of-the-arts approaches that have been used in this particular face

analysis domain.

2.2 Face image pre-processing

Typically, in the domain of face image analysis, the first and foremost task we need to do is preprocessing, which is basically localising and normalising the face(s) within an image frame. This preprocessing step includes face detection, facial landmark localisation, and face registration, as visualised in Figure 2.11. The following sections review different preprocessing methods that are commonly used in the field of human face analysis.

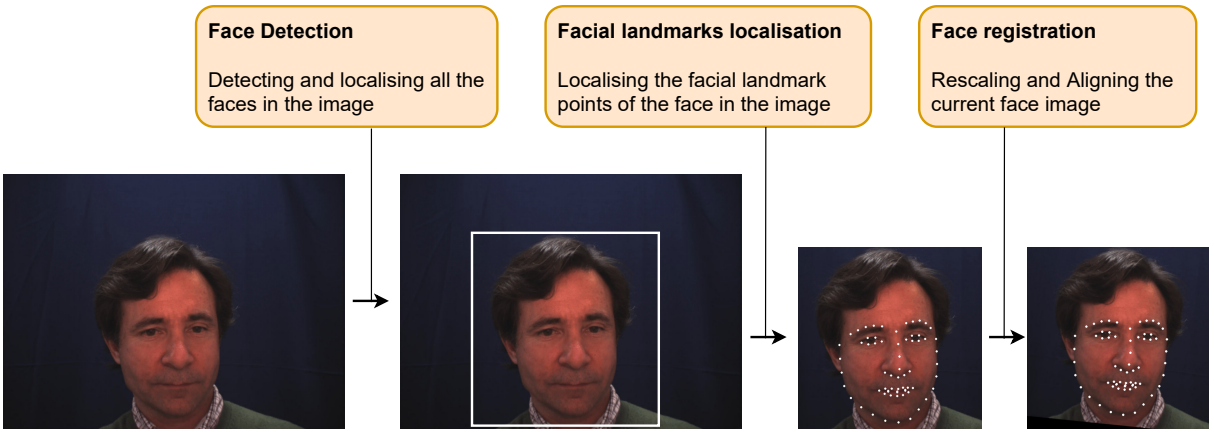


Figure 2.11: Generic overview of the face image pre-processing pipeline.

2.2.1 Face detection

Automatic face detection is the cornerstone of almost all applications revolving around automatic face image/video analysis including face recognition and verification, face tracking for surveillance, facial expressions assessment, gender/age recognition [SS18, YKA02]. The goal of face detection is to determine whether or not there are any faces in the image and if so, then return the location and the extent of each face in the image. While this appears as a trivial task for human beings, it is difficult for computers, and has been one of the most studied research topics in recent decades.

Face detection is a relatively mature problem in computer vision, i.e., many algorithms exist that solve the problem robustly and efficiently. One of the first widely adopted algorithms was the Viola-Jones object detection framework [VJ01]. The framework introduces the idea of computing an integral image over the greyscale input to enable fast evaluation of boosted weak classifiers based on Haar-like features [MBPG14]. Following the pioneering work of Viola-Jones, numerous methods have been proposed for face detection in the past decade. Early research studies in the literature were mainly focused on extracting hand-crafted features with domain experts in computer vision, and training effective classifiers for detection and recognition with traditional machine learning algorithms. Such approaches are limited in that they often require computer vision experts in crafting effective features and each individual component is optimised separately, making the whole detection pipeline often sub-optimal.

With the great success of CNNs in computer vision, researchers have proposed several promising model architectures for face detection problem over the past few years. Deep learning based approaches are getting better and better in the task of detecting face from images thanks to the introduction of many large face databases such as Wider Face [YLLT16], MALF [YYLL15], or VGGFace2 [CSX+18]. Cascade-CNN [LLS+15, QJM+19], Single-short Detection [NSCD17, CHP+21] , RCNN based architectures [CHWS16, ZZLS17, COG19], Feature Pyramid Network (FPN) models [ZWHZ20, TDHL18, NSD19] are among the most well-known Deep Neural Networks for face detection. Performance of these deep learning based face detectors is much better than that of feature hand-crafted based methods, which is once again confirming the advantages of DNNs in learning from data. Further details about different techniques in face detection can be found in the review of the SOTA paper provided by Minaee *et al.* [MLLB21].

2.2.2 Facial landmark localisation

Facial Landmark Localisation (FLL) algorithm is defined as the detection and localisation of certain points characteristic on face images, which have an impact on subsequent task focused on the face, like animation, face recognition, gaze detection, face tracking, expression recognition, gesture understanding, etc. Commonly used facial landmarks usually include points around the eyebrows, eyes, nose, mouth and the face contour. According to various application scenarios, different numbers of facial landmark points are labelled as, for example, a 5-point, 17-point, 29-point, 66-point, or 68-point model. Generally speaking, more points indicate richer information, although it is more time-consuming to detect all the points. FLL methods could be divided into four groups: Constrained Local Model (CLM) based methods, AAM based methods, regression based method, and others [WGT+18]. CLM based methods consist of a shape model and a number of local experts, each of them is utilised to detect a facial feature point [CILS12, LBL+12]. AAM based techniques fit a shape model to an image my minimising texture synthesis errors [MCB13]. Regression based methods directly learn a mapping function from facial image appearance to facial feature points [MVBP12, BAPD13]. Besides the three main categories that we have mentioned, there are also other methods, such as graphical model-based methods [ZSCC13], or independent facial feature point detectors [SLBW13].

Recently, deep learning methods become popular tools for computer vision problems and they also have achieved great successes in this domain of Facial Landmark Localisation. In fact, for facial landmark detection and tracking, there is a trend to shift from traditional methods to deep learning based methods [WJ19], indicating the superiority of these data-driven learning approaches. Since deep learning based methods also perform regression to locate the facial landmarks, they fall into the category of regression based methods. Recent deep learning methods [RPC17, RSCC17] can jointly perform face detection, facial landmark localisation, pose estimation, and gender recognition, all in a single neural network. Another approaches [ZLL+16, JL16] went to different di-

Figure 2.12: Example of large-pose face landmark localisation. From left to right: initial landmarks, fitted 3D dense shape, estimated landmarks with visibility. The green/red/yellow dots in the right column show the visible/invisible/cheek landmarks, respectively. Image from [JL16].

rection by fitting a dense 3D model to face image, solving the problem of large-pose face landmark localisation, as can be seen in Figure 2.12. Many more different approaches in facial landmark localisation can be found in the review SOTA paper proposed by Wo *et al.* [WJ19].

2.2.3 Face registration

Face Registration is an intermediate step to prepare the shape or appearance of the face image for further feature extraction. It aims to find the transformation (or the deformation) which reduces the discrepancies between two or more faces. These registration approaches modify facial characteristics (texture, geometry, motion) while reducing variations in translation, rotation and scale changes [AMBD18].

There are two main approaches for face registration in 2D images, including eyes registration and shape registration. Eyes registration is the most simple and also the most popular strategy in near frontal-view databases. Eyes are detected and images are

aligned and scaled with regard to the inter-pupilar distance and orientation [AMBD18]. The reason of using the eyes instead of other facial components for registration is that they are the most reliable facial component that can be detected and they hardly change in the presence of expressions. The limitation of this approach is that eyes must be well detected. Usually, when out-of-plane rotations occur, the eyes disappear quickly and additional deformations are induced, avoiding the detection of eyes [AMBD18].

Shape registration is another more robust approach, which takes all facial landmarks points into account for alignment. Extensions considering more landmarks is supposed to provide greater stability in case of individual poor landmark detections. Among the shape registration methods, Generalised Procrustes Analysis (GPA) [Gow75] seems to be the most famous and widely adopted one [KGAS15, RCG+17, VBADE21]. Procrustes algorithm iteratively estimates the reference shape and the frame-wise alignment transform until convergence. Then, the reference shape is initialised by the mean of all points and then iteratively updated with the mean of all aligned points. Finally, the transformations are obtained by minimising the squared differences between the actual shapes and the mean shape. After these transformations, face images are aligned and scaled with regard to the GPA reference shape, hence reduces the varies of face pose, camera position, or anthropomorphic differences between subjects.

2.3 Automatic facial expression measurement

Automatic facial expression measurement is an emerging topic in artificial intelligence due to its wide range of applications in many different domains, especially in health care and medical fields. Researchers have proposed a large number of different approaches to analyse human facial expressions. As mentioned in Chapter 1, this thesis mainly focuses on measuring facial AUs intensity as it is the basic building block of facial expressions in general. Hence, in this section, we jointly review the SOTA approaches in both facial AU intensity and PSPI pain intensity estimation domains, as approaches in these two domains share some common technologies. These methods

could be categorised into two groups, including feature hand-crafted (Section 2.3.1) and deep learning based methods (Section 2.3.2). For each groups, we discuss the underlying algorithms and existing works that utilise them.

2.3.1 Feature hand-crafted methods

Traditional hand-crafted feature-based methods have sought to address the problem of facial AUs and PSPI intensity estimation for quite a long time. One of the earliest works on these domains was proposed by Fasel and Luettin [FL00]. They have tried to extract features from face images by using PCA and ICA dimentionality reduction techniques for automatic facial AUs intensity estimation. They found that the PCA is more effective at extracting facial feature representations than ICA, since the ICA is quite sensitive to noise. In 2006, Bartlett *et al.* [BLF+06] revisited the problem of facial AUs intensity estimation with real-time application. They used Gabor wavelet for feature extraction, AdaBoost for feature selection and dimensionality reduction. Both SVMs and regression analysis were used for classification, resulting the mean rate of 93% recognition with 20 AUs for the task of facial AUs detection. SVMs have also been used by Hammal and Cohn [HC12] for one of the first methods on pain intensity estimation. They tracked AAM landmarks on the UNBC McMaster database and used log-normal filter features to recognise 4 discrete pain levels.

In 2009, Mahoor *et al.* [MCMC09] conducted a research on detection and intensity estimation of AUs in non-posed mother-infant face-to-face interactions. They used AAMs for facial representation, and applied Spectral Regression for dimensionality reduction. They trained a SVM based system for automatic recognition of intensities for AU6 and AU12. In 2012, Savran *et al.* [SSB12] applied AdaBoost feature selection and SVR to Bosphorus 3D data, combining 3D shape and Gabor filters. They showed that for 3D images, AdaBoost shows superior efficiency for feature selection. Their regression based method showed better performance than previously suggested SVM based methods. They also showed that it is both possible and feasible to generate an intensity

estimation system that uses facial grid and/or facial fiducial point detection. In conclusion, they showed that using the combination of 3D and 2D images gives better results compared to either one individually. Still in 2012, Lucey *et al.* [LCP+12] conducted some research on the UNBC McMaster database for automatic pain analysis. They have applied SVCs for binary detection of AUs to detect the existence of pain in a human face. Their research was also based on frame-level and sequence-level pain detection. They showed that both frame-level and sequence level pain detection is successful, but the speed of the system is faster with sequence based analysis.

Kaltwang *et al.* [KRP12] used AAMs for feature extraction and RVM for classification. They have reported that pain intensity detection from facial AUs intensities received a higher classification rate compared to direct pain intensity detection from PSPI. Hammal and Cohn [HC12] proposed to use the normalised of *canonical normalised appearance* (CAPP) of the face on top of the AAM landmark points, and four different SVMs to classify between pain intensities. They achieved a classification success between 40% and 67% with 5-folds leave-one-out evaluation method.

Another research on facial AUs intensity estimation was conducted by Jeni *et al.* [JGCDLT13]. Their system consists of four parts, including fiducial point detection by constrained local models, local patch removal using fiducial points, application of non-negative matrix factorisation, and training the SVM using the extracted features. They have conducted their experiments on the CK-Enhanced [LCK+10], BU-4DFE and BP4D-Spontaneous databases, achieved some great improvements. Sandbach *et al.* [SZP13] applies LBP features, GentleBoost (a more stable version of the AdaBoost algorithm) feature selection and a Markov Random Field (MRF) [Kin80] on the DISFA database for inferring upper facial AUs intensities. A MRF is a graph consisting of nodes and edges, where each node corresponds to a random variable and each edge corresponds to a parameterised potential function. Through the MRF graph, all AU intensities are estimated jointly. This is different from most other previous approaches, which train a separate model for each AU. Another research that focused on pain intensity was done by Rudovic [RPP13]. They proposed to use Conditional Ordinal Random Fields (CORF)

for classifier and LBP for feature extraction. CORF is a special case of MRF where all the clique potentials are conditioned on input features. Through experimentation, they came to the conclusion that accounting for heterogeneity in the dataset would give better results.

Khan *et al.* [KMKB13] tried to extract shape and appearance information to obtain discriminative representations of facial expressions. Four common classifiers including SVM, Decision Tree (DT), Random Forest (RF), and 2 Nearest Neighbor (2NN) have been used to test the performance of this method, and the nonparametric classifier 2NN achieves the best experimental results. Mavadati *et al.* [MM14] proposed to use Gabor feature to represent the facial AUs and Hidden Markov Model (HMM) to model the temporal patterns of them. HMMs are a specific type of MRF and exact inference is achieved by the forward-backward algorithm. Florea *et al.* [FFV14] proposed to use histogram of topographic features (HoT) to describe faces with different pain levels, and to use transfer learning to enhance the robustness of the model.

Zhang *et al.* [ZZH15] proposed a method that extracts the dynamic motion-based facial features which were measured through the facial landmark points' displacement between natural and expressive frames on 3D facial video. These facial features are then fed to SVRs regressors for AU intensity estimation. Kaltwang *et al.* [KTP15] formulated a Latent Tree (LT) where fiducial points were set as part of leaf nodes accompanying by several other leaf nodes of AU targets and hidden variables. This graphical model represents the joint distribution of targets and features that was further revised through conducting graph-edits for final representation. Walecki *et al.* [WRPP16] proposed a Copula Ordinal Regression (COR) framework to model the co-occurring AU intensity levels with the power of copula functions and Conditional Random Fields (CRFs). Ruiz *et al.* [RRBP16] proposed Multi-instance Dynamic Ordinal Random Fields by exploiting the idea of multi-instance learning for automatic facial AUs intensity estimation. They treated each sequence as a bag (set of training samples) and treated the maximum intensity of a sequence as the bag label. Hong *et al.* [HZZ+16] proposed a second-order pooling framework for medical image analysis, texture classification, micro-expression

recognition and pain intensity assessment. Zhao *et al.* [ZGWJ16] proposed the peak-piloted method that uses the peak samples to supervise the feature responses for the non-peak frames of the same emotion and the same subject, achieving a competitive performance in prediction of PSPI scores. Zhang *et al.* [ZZD+18] proposed a weakly supervised regression model so called Bilateral Ordinal Relevance Multi-Instance Regression (BORMIR), which exploits the relationships among instances and incorporated domain knowledge to learn AU intensity regression.

2.3.2 Deep learning based methods

Deep Learning (DL) in general and CNN in particular have shown some great improvements in many computer vision tasks, including facial expressions analysis. The superior performance of deep models is largely due to their ability to learn from experience and generalise well on newly unseen data. Hence, more and more works focus on analysing human facial expressions using these deep learning CNN techniques. Zhou *et al.* [ZHSZ16] proposed to use Recurrent Convolutional Neural Network (RCNN) for PSPI intensity estimation problem. Their RCNN uses recurrent connections in the convolution layers to capture the temporal information without increasing the overload of parameters to avoid over fitting. Walecki *et al.* [WOR+17] placed a CRF graph on top of a CNN to exploit the spatial relations between different AUs, improving the performance of deep network in estimating facial AUs intensity. Rodriguez *et al.* [RCG+17] trained a VGG-16 network to extract features from each frame of the dataset video sequences. Those features are fed into a LSTM network to model the temporal dynamics information between consecutive frame images. They have shown that their CNN-RNN hybrid network is capable of capturing both spatial and temporal information from sequence frames, outperforms hand-crafted approaches by a large margin in the domain of PSPI intensity estimation problem.

In 2018, Zhang *et al.* [ZDHJ18] tried to exploit the temporal dynamics information between consecutive frames by proposing a knowledge-based CNN for AU intensity es-

timation with peak and valley frames in training sequences. Tavakolian *et al.* [TH18a] extracted the features of the image by using the deep neural network and then processed them to obtain the binary code. Sánchez *et al.* [SLTV18] proposed a network based on Hourglass architecture [NYD16], which directly regress both AU locations and intensities. They have shown that integrating AU locations into the training process has greatly improved the performance of deep neural network. Zhang *et al.* [ZJW+19] and Chu *et al.* [CTC17] both proposed deep CNN-RNN hybrid networks for automatic facial AUs intensity estimation. Fan *et al.* [FLL20] stacked a Semantic Correspondence Convolution (SCC) module on top of a heatmap regression-based network to perform both AU location regression and AU intensity estimation. Huang *et al.* [HQX+21] proposed HybNet, which is a fusion of three sub-networks including 3D CNN, 2D CNN, and 1D CNN for PSPI intensity estimation. Song *et al.* [SCW+21] integrated AU locations and probabilistic graphs into the training of ther deep neural network, resulting some great improvements in facial AUs intensity estimation problem.

To conclude, from the SOTA approaches that we have introduced in this section, it can be seen that Deep Neural Networks have a great advantages over traditional image processing approaches because their ability to automatically learn complex patterns from training data. However, as the amount of data for the domain of facial AUs and PSPI intensity estimation are quite limited (see Section 1.6), we need to design new learning approaches that are able to learn to extract correct features from a limited amount of data. In the next chapter, we introduce our proposing approach for better pain intensity estimation when learning from a limited amount of data.

Chapter 3

Learning to focus on regions-of-interest for pain estimation

Contents

Since human inner feelings and physiological states are typically characterised by subtle movements of facial parts, the analysis of the facial details could improve the overall quality of automatic expression assessment system. Thus, the development of a powerful feature extraction system which is able to focus on these facial detail and extract features from them is a crucial task and has been receiving attention from research community. In this chapter, we summarise our results in paper I and provide some extended findings that we did not mention in this paper due to the limitation of both the scope-of-the-work and the maximum number-of-pages allowed. These findings demonstrate the importance of learning to focus on regions-of-interest for pain intensity estimation.

Paper I: M. T. Vu, M. Beurton-Aimar, P. -y. Dezaunay and M. C. Eslous, "Automated Pain Estimation based on Facial Action Units from Multi-Databases," 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2021, pp. 1-8, doi: 10.1109/ICIEVicIVPR52578.2021.9564244.

3.1 Context

Facial expression provides sensitive cues about emotional response and plays a major role in human interaction and nonverbal communications. It can complement verbal communication, or can convey complete thoughts by itself [Lie98]. Facial expression can easily be perceived and processed by human observer in a concise manner. The ability to recognise other's facial expression seems to be innate and universal across cultural and racial borders [SBB+03]. As result, humans can easily recognise a wide range of different expressions, even though different people may look different for different expressions. While it is natural for humans to recognise other's facial expressions, it is a challenging task for a computer vision system to imitate the same cognitive behaviour. One of the main reasons for this, particularly in the context of pattern recognition appli-

cations, is the so-called *curse of dimensionality* [Bel66]. The learning complexity grows exponentially with linear increase in the dimensionality of the data. For human, it is effortlessly to receive and process a myriad of sensory data and capture critical aspects of this data in a way that allows for its future use [DZR12]. Contrastingly, high dimensionality of data is a fundamental hurdle in many science and engineering applications [Bel66]. The typical approach to overcome the problem of *the curse* is to extract only the important features from the data for reducing its dimensionality to that which can be effectively processed, e.g., by a regression algorithm for pain intensity estimation. For instance, a computer vision algorithm can process thousands of images per second to estimate pain intensity from face image. However, the feature extraction system to find and extract important features from this high dimensional data is only possible through highly engineering systems, which are well designed and trained through specialised image processing and pattern recognition algorithms. Traditional hand-engineered feature extraction methods have been around for decades and have been used for extracting features for many facial expression recognition and analysis systems [KRP12, BLF+06, SZP13]. However, these hand-engineered methods at times can be challenging, highly application-dependent, time consuming, brittle and not scalable in practice [SA19].

Recently, deep learning techniques have emerged as powerful methods for learning feature representations directly from data and have achieved some major improvements in various face-related computer vision tasks [SKP15, VBADE21, ZPS17, STE13]. Because these learned feature representations are extracted automatically to solve a specific task, they are extremely effective at it. In fact, deep learning models that perform feature extraction and classification outperform models that classify manually extracted features by a large margin, in many different domains [SBAO18, ABRD15, XLW+16]. The main advantage of deep learning approaches is their ability to learn from experience and generalise well on newly unseen data [TSV20]. However, to do so, these deep models require to be trained on massive amount of data, which is difficult to obtain for the domain of facial expressions, especially the facial AU and PSPI pain intensity esti-

mation. The reason for that is because it requires a costly and time-consuming labeling effort by trained human annotators. For instance, it may take more than an hour for an expert annotator to code the intensity of AUs in one second of a face video [LTWE+17]. In addition to limiting the amount of data, the distribution of AU intensities in these databases is also highly unbalanced. Consequently, the performance of deep methods training on these databases are being negatively affected by insufficient data. Therefore, it is necessary to develop a learning approach that is capable of exploiting better feature representations of facial image features on a limited amount of data. In this work, we propose a new three-stages training approach which can combine multi-database together for more training data and, at the same time, learn to focus on the right regions on the face (regions-of-interest) for better exploiting the data. We demonstrate the effectiveness of our three-stages training approach on the UNBC McMaster database, showing some promising results.

In the next sections, we explain step-by-step our approach. In section 3.2, we discuss about the idea of learning to focus on regions-of-interest. Section 3.3 mentions about the problem of multi-database combination. Section 3.4 explains the architecture of our proposed 3-stages approach. Finally, the experiment results and discussion are explained in sections 3.5–3.6.

3.2 Learning to focus on Regions-Of-Interest

One important property of perception is that humans do not tend to process whole information in its entirety at once. Instead, humans tend to selectively focus on a part of the information when and where it is needed, but ignore other perceivable information at the same time [NZY21]. Hence, focusing on the right places and ignoring other irrelevant information appears to be an important aspect for not only human but also for machine to concentrate on the relevant information and extract the correct features. In the field of machine learning and deep learning, if we can tell the neural network where to focus in the image, it will ease the training process and improve the general-

isability of the network. Previous works such as Guan *et al.* [GHZ+18] and Tang *et al.* [TWH+18] have tried to integrate the location of lesion area in the Chest X-ray image to the training as heatmap regression, resulting some great performance improvements. Wo and Ji [WJ16] proposed a cascade regression approach that incorporated the location of facial landmarks into the training process, which improved the performance of the face AU recognition task. Li *et al.* [LAZ17] proposed a region-based network which integrated the information regarding the location of each AU into the training for better AU detection. Sánchez *et al.* [SLTV18] and Fan *et al.* [FLL20] tried to encode both the location and intensity of each facial AU as a heatmap (see Figure 3.1) and train the network as a per-pixel regression problem, resulting some great improvements in the task of facial AUs intensity estimation. These findings once again confirm that learning to focus on the right parts of the face image would definitely help to improve network performance.

In this work, inspired by [SLTV18, FLL20], we utilise the heatmap regression to force our deep neural network to focus on each facial AU's regions. However, different from [SLTV18, FLL20], we do not use the predicted heatmaps for AU intensity estimation, instead, we extract the embedded feature representations of the network for further pain intensity estimation training. Since the network have been trained to focus on each of the pain-related AU regions (regions-of-interest), hence these feature representations should contain the important information regarding each facial AU. Section 3.4 explains step by step our approach to utilise heatmap regression for boosting performance of our pain intensity estimation network.

3.3 Multi-database combination

Before presenting the learning approach and network architecture, in this section, we discuss about multi-database combination, which is a way to improve the coverage of the training data. As we have mentioned earlier, the key point for many deep learning related problems is to have a large amount of data for the training to improve the model

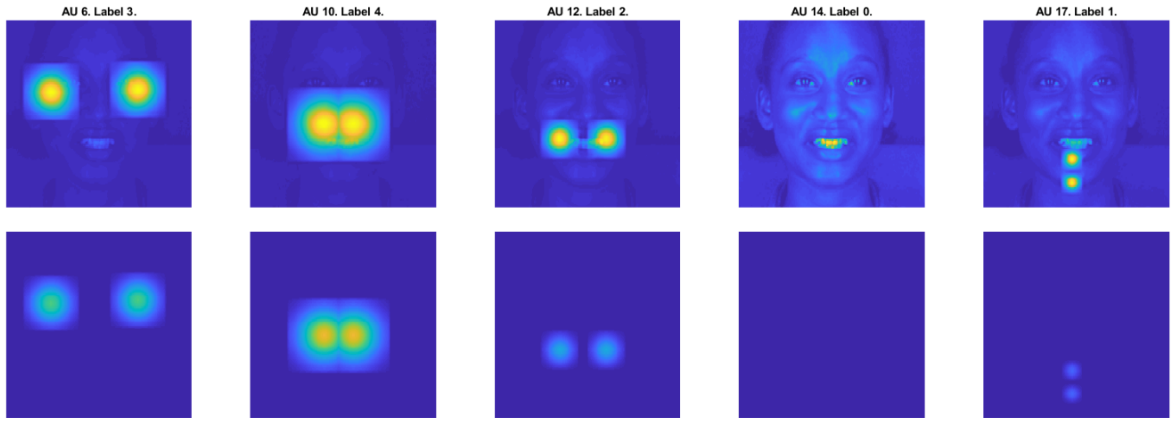


Figure 3.1: A visualisation example of target heatmaps for a given sample. The size and peak of the heatmaps are given by the corresponding labels, and are located according to the landmarks defining the AU locations. Image from [SLTV18].

generalisation. Deep neural networks trained on large supervised datasets have led to impressive results in many different domains [RVBS17, KSZQ20, TYRW14]. While constructing a new large database is costly and time consuming, combining the existing databases seems to be feasible. Lefter *et al.* [LRVL10] combined several emotional speech corpora within the training set to reduce the data scarcity problem and extending the variety of acoustic background, resulting improved performance compared to training on a single dataset alone. Schuller *et al.* [SZWR11] also found that fusing a variety of training data is on average better than relying on a single training corpus. In the domain of image processing, Dobrescu et al. [DVGT17] used a regression model based on ResNet-50 [HZRS15] architecture, training on the combination of multiple leaf datasets to produce a more generalized model, with excellent results. These researchs show a promising way to improve the performance of the deep learning model, while maintaining initial complexity level and memory size required.

In this work, we introduce the multi-database combination approach for better pain intensity estimation. We combine the training data of the two well-known facial AU intensity databases, including the UNBC McMaster and the DISFA databases (see Section 1.6 for more information about the databases). Although the DISFA database does not have the PSPI pain intensity annotation, these two databases share some pain-related

facial AU intensity. Hence, there is a possibility to combine these two databases using their common facial AUs and train a neural network as a feature extractor to extract the important features regarding these facial AUs. Then, we can finetune this trained network for pain intensity estimation. To the best of our knowledge, our approach is the first to combine these databases together, probably due to some differences in the annotation and the main purpose of each database.

This approach of combining multiple databases cooperates well with the learning to focus on facial regions-of-interest that we have mentioned earlier. One approach increases the amount of training data and its ethnic coverage by combining multi-database together, the other provides a better way to exploit these combined training data. The next section provides more details about how do we utilise these databases for better pain intensity estimation.

3.4 The three-stages training approach

In this work, we strive to take a step towards the goal of automatic pain intensity estimation by introducing a novel three stages training approach. Similar to [RCG+17], we also perform regression using deep CNNs linked with LSTM model to predict PSPI score for each frame image. However, instead of fine-tuning a CNN model directly from the original database as [RCG+17], we partially train it on the combination of the UNBC McMaster [LCP+12] and DISFA [MMB+13] databases by using the learning to focus on regions-of-interest approach. Then, we utilise the knowledge that has been learned on the two databases for feature extraction and PSPI pain intensity estimation. Particularly, in the first stage, we train our CNN model with the combination of the two databases for predicting their common AUs intensities as heatmap regression. In the second stage, we freeze the first layers of the network to preserve the parameters that have been trained on the two databases, then fine-tune its mid and top layers for predicting PSPI scores. For the last stage, we link the features extracted from the fine-tuned CNN model to the LSTM recurrent network for exploiting the temporal axis information between

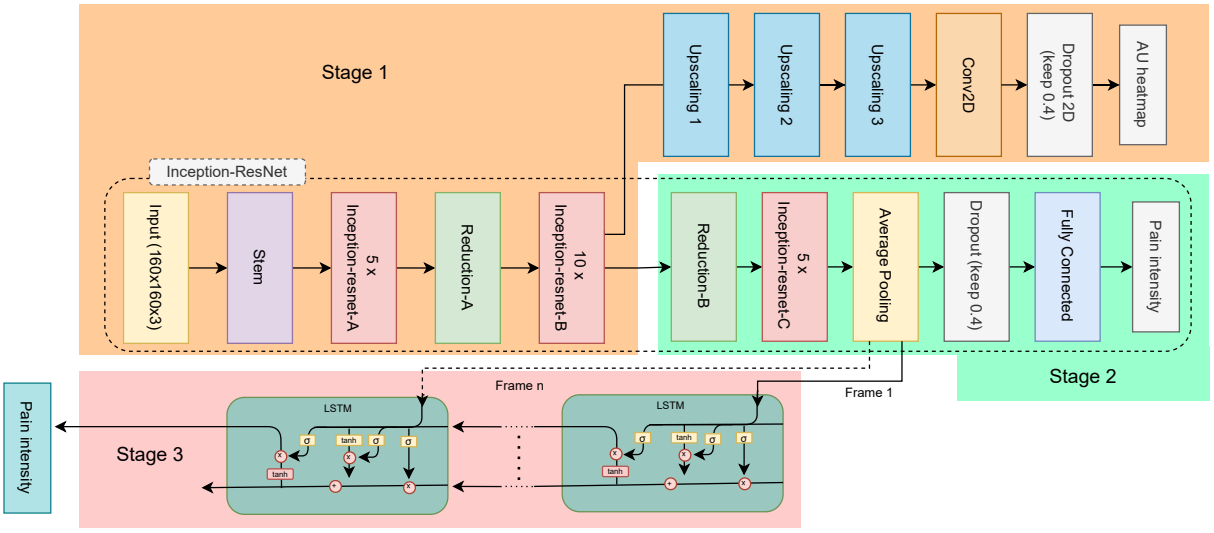


Figure 3.2: The overview of the proposed three-stages training approach. Several upscaling layers are added for heatmap regression training in the first stage (blue block). The mid and top layers of the base Inception Resnet network are trained as linear regression in the second stage (green block) and then, the output of the `Average Pooling` layer are extracted to train the LSTM network in the last stage (pink block).

the video frames. With this approach, we are able to effectively extract the important features from the combination of two different databases, boosting the performance of our pain intensity estimation network. As UNBC McMaster is the only database annotated with PSPI pain intensity levels and the number of subjects of this database is quite small with only 25 subjects, the addition of subjects from the second database is extremely valuable. For the base CNN model, the Inception Resnet v1 [SIVA16] model pretrained[1] on VGG-Face2 [CSX+18] database is selected as this architecture is proven to be computationally efficient while keeping high performance on different learning tasks. By fine-turning from this model, we also benefit from the VGG-Face2 database [CSX+18], which contains millions of faces that helps to improve the generalisability of our model.

[1]The pretrained Inception ResNet v1 model is taken from the work [SKP15] and can be found at https://github.com/davidsandberg/facenet

3.4.1 Model architecture

The backbone framework of our network is the Inception Resnet v1 [SIVA16] architecture, in which we add some different layers to train different parts of the network at each of the three-stages, as showing in Figure 3.2. Specifically, in the first stage, we add some upscaling layers, which are including Transposed Convolution [DV16] and ReLU [Aga18] layers, on top of the InceptionResnet-B blocks of the base network for reconstructing the AUs intensities as heatmaps from the original images. The reason of choosing the output of InceptionResnet-B blocks is because the output dimension at this layer is 8×8 per channel, which is small enough for reconstruction. Since the output dimension of the previous layer (Reduction-A) is 17×17 per channel, which is too large and can introduce noise while the output dimension of the next layer (Reduction-B) is 3×3 per channel, which is too small and does not provide enough information to reconstruct. Thus, the output of InceptionResnet-B is perfectly fit for our problem and is selected to reconstruct our AU heatmaps for training in this stage. The intuition behind this heatmap regression is to train the network to focus on the pain-related AU regions (regions-of-interest) for better extracting feature representations, the added layers will be discarded afterward. Since the network takes data from the combination of UNBC McMaster and DISFA database as input images for training, this first stage is used to improve the generalisability of the model.

The second stage is mainly used for dimension reduction. As the output dimension of the first stage is 8×8 per channel (after removing the added upscaling layers), it still contains some structural information that can be exploited. Moreover, as the LSTM network in the last stage requires to have a 1D vector of data to train, we have to reduce the output dimension of the first stage from 8×8 per channel to a 1D vector of features. A naive approach would directly flatten the output feature to 1D vector. However, this flattening approach may result of losing the structural information and making noise. Instead, we train the mid and top layers of the base network for PSPI scores estimation, which in the same time, has the effect of reducing the dimension of the input to be

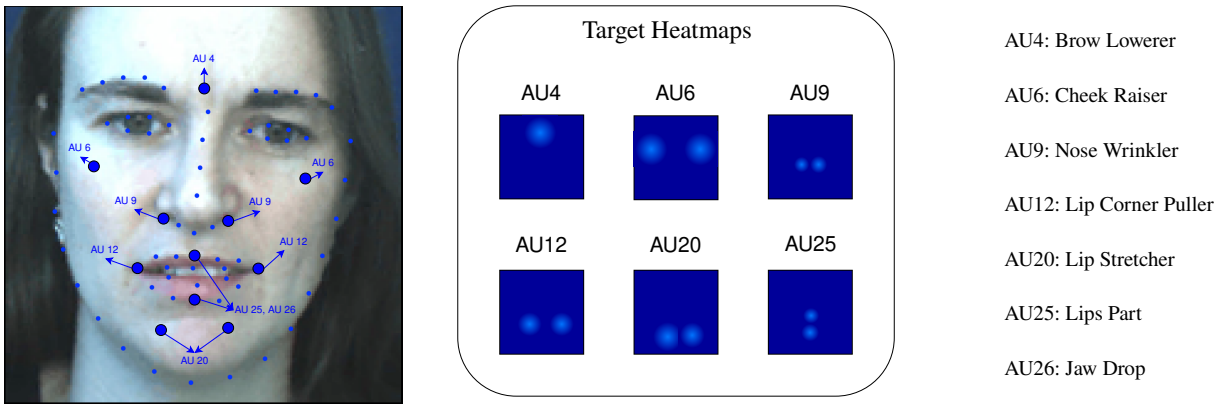


Figure 3.3: Central locations of the common AUs between the two databases and the visualization of the Target Heatmaps generated from the ground truth.

able to fit the requirement of the LSTM training in the last stage. Finally, the outputs of the `Average Pooling` layer are extracted to train the LSTM network in the last stage to exploit the temporal dynamics information between video frames. We don't take the output of the `Fully Connected` layer as input of the LSTM network because it has less temporal invariability than the one from `Average Pooling` layer, as can be seen in Table 3.2, the latter yields better performance when the LSTM network is fed by this layer outputs.

In the next sections, we explain step-by-step the network architecture, input and output data at each stage of our three-stages training approach.

3.4.2 First stage: Action Unit intensity estimation

The main goal of this stage is to improve the generalisability of the network by train the first layers' parameters of the base network by using the learning to focus on regions-of-interest approach on the combination of the two databases. Several upscaling layers are added on top of the InceptionResnet-B blocks for heatmap regression's training. Similar to [SLTV18, FLL20] we first generate the heatmaps ground-truth from our databases by applying Gaussian function on the predefined AU locations, as depicted in Figure 3.3. Each image frame generates a set of N heatmaps for our selected common AUs

between the two databases, where N is the total number of the predefined AU locations. For a predefined AU location $L_i(i = \{1, ..., N\})$, the ground-truth heatmap $g_i(x)$ is a 64×64 image generated by applying a Gaussian function centered on its corresponding coordinate $\hat{x}$ as follows:

$$g_i(x) = \frac{I}{2\pi\sigma^2} exp(-\frac{\|x - \hat{x}_i\|_2^2}{2\sigma^2}) \tag{3.1}$$

Where I is the labelled intensity of the specified AU, and σ is the standard deviation. Thus, the generated heatmap has the highest value at the centre AU location $\hat{x}$ and smoothly decrease when the pixel is farther away. This way, we can encode both spatial and intensity ground-truth information of AUs into heatmaps and then use it to train our model as heatmap regression. Because the output of our model is also AU heatmaps, the loss function should be a per-pixel loss function between the predicted heatmaps and the generated one from ground-truth. As we are doing heatmap regression, the per-pixel loss is defined as the L_2 norm, which is defined as:

$$\mathcal{L}_{i,j} = \|\hat{y}_{i,j} - y_{i,j}\|_2^2 \tag{3.2}$$

Where $\hat{y}_{i,j}$ is the output heatmap generated by the network at pixel i, j and $y_{i,j}$ is the corresponding ground-truth. The total loss is computed as the average of the per-pixel loss per AU. The model state with the lowest validation loss is selected to continue on the next stage. The summary of the whole process is depicted in Figure 3.4.

In this stage, we train this network using all of the common AUs between the two databases as heatmap regression, which are including AU4, AU6, AU9, AU12, AU20, AU25 and AU 26. We also report the results of the network when training with AU4, AU6 and AU9 as target heatmaps, since these common AUs are parts of the PSPI formula (Equation 1.1). We expect that learning to focus on regions-of-interest of these AUs will help the network to exploit better feature representations and the additional data from the secondary database will compensate the lacking of different face appearances of the UNBC McMaster database, thus, improving the generalisability of the network.

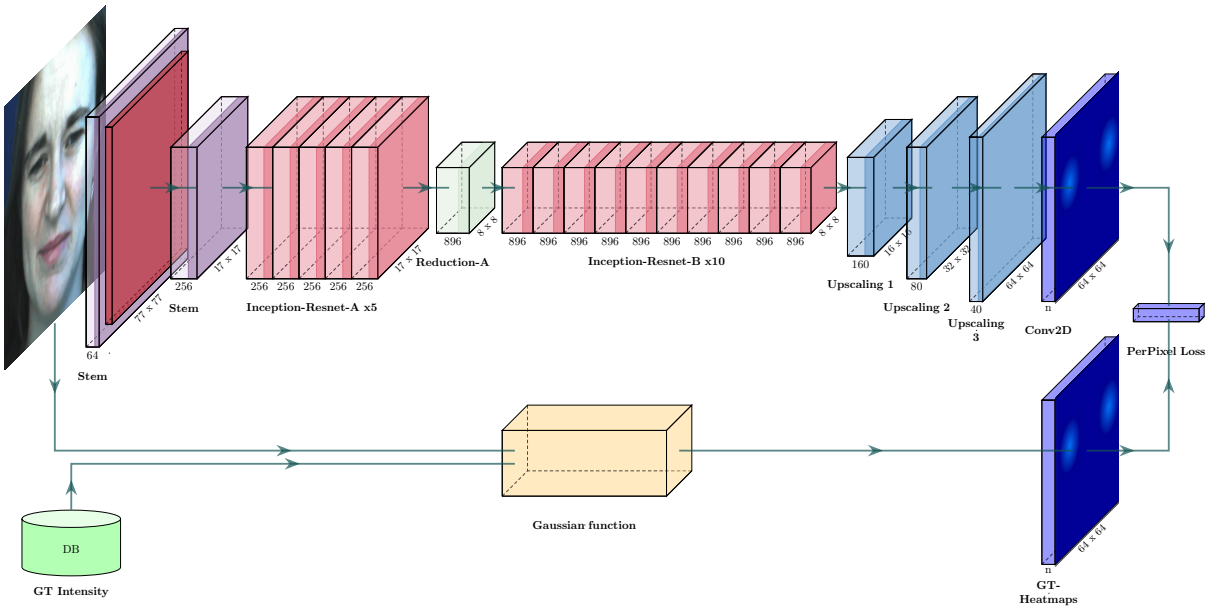


Figure 3.4: The visualisation of the Heatmap regression in the first stage. Several upscaling layers are added on top of the InceptionResnet-B blocks of the base network to reconstruct the AUs intensity as a set of n heatmaps, where n is corresponding to the number of AUs. These reconstructed heatmaps will then be compared with ground truth heatmaps to compute per-pixel loss function for optimising the model's parameters.

3.4.3 Second stage: Frame level pain intensity estimation

As stated, the main purpose of this second stage is for dimension reduction. The first layers of the base network (from the begining to the last of the InceptionResnet B blocks, which can be seen in Figure 3.2) are frozen to preserve the parameters that have been trained using the two databases on the first stage. All the upscaling layers that we added in the first stage are discarded as we don't need to use them anymore. Mid and top layers of the base network are trained as regression to predict pain intensity level. Data from the UNBC McMaster database are used to train at this stage as this database is the only one that has annotated with the PSPI score. L_2 objective function between the predicted label $\hat{y}$ and actual label y are used as loss function to optimise the network:

$$E = \frac{1}{N}\sum_{n=0}^{N}\|\hat{y}_n - y_n\|_2^2 \tag{3.3}$$

Where N is the total number of predictions. At this stage, the whole of InceptionRes-

net network is fully trained and can be used to predict pain intensity level from images or frames. However, we still can improve it by exploiting the temporal information between the frames in video sequence, in which we describe in the next stage.

3.4.4 Last stage: Sequence level pain intensity estimation

In this last stage, temporal information of the video sequence is exploited by linking the features extracted using the base model trained on the previous stages to a LSTM network. LSTM is a variant of Recurrent Neural Network (RNN) which is capable of keeping long-term information from previous inputs. By learning the changing of the facial expressions over time through the sequence of frames, we expect this network to be able to detect the trending of the expression using the past information, and to combine it with the information from the current frame to make a better decision. In order to train this network, outputs of the `Average Pooling` layer of the base network are extracted as sequences and fed to this LSTM network. L_2 objective function between the predicted label of the sequence and actual label will be used as loss function to optimise the model's parameters.

3.5 Experiments and results

3.5.1 Implementation details

The training and testing processes were performed using a NVIDIA Gerforce RTX 2080 Ti 11G GPU with Pytorch v1.6 [PGM+19]. During the training phase, Adam optimiser [KB14] were employed with initial batch size of 64 for all the three stages. Initial learning rate is set of $1e^{-5}$ for the first stage, $3e^{-4}$ for the second stage and $9e^{-6}$ for the last stage. For LSTM network, the number of layers and hidden units are set to 2 and 744, respectively. These configurations are set based on a large grid searching.

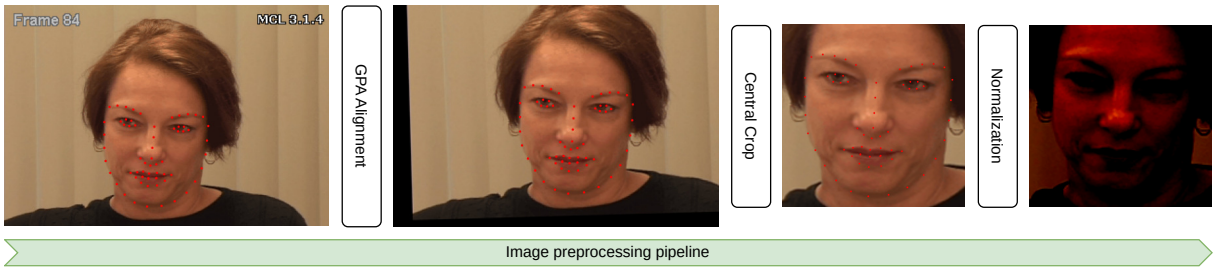


Figure 3.5: The preprocessing pipeline. First, the original image frame is aligned using GPA alignment, then it is cropped and resized on the face area based on its landmarks. Finally, fixed image normalisation is applied to ease the training process.

3.5.2 Data preprocessing

For the data preprocessing step, we want our images to be as similar to the pretrained VGG-Face2 [CSX+18] images as possible, which would make the model's activation functions to be activated in the same way. The transforming pipeline consists of three steps including face alignment, face cropping and image normalisation, as can be seen in Figure 3.5. Similar to [LCP+12, RCG+17] we also use the Generalised Procrustes Analysis (GPA) [Gow75] to align the face images based on the provided landmarks. Next, the aligned face images are cropped, resized to 160×160 pixels and then normalised using fixed image standardisation as it is the input format of the pretrained model. For a given tensor image X, the normalisation's formula[2] is denoted as follows:

$$X_{normalised} = \frac{X - 127.5}{128} \quad \forall i \in X, 0 \leq i \leq 255 \tag{3.4}$$

In training, we apply some data augmentation techniques to improve data scarcity limitation. Specifically, instead of using central crop, for each sample image, we randomly apply Image Translation and Horizontal Flip techniques, as these techniques have proven to be efficient in the task of facial expression recognition [PFA20]. We also have applied ColorJitter technique which randomly change the brightness, contrast, saturation and hue of the image.

[2]The equation 3.4 is the improved formula using in the preprocessing step of the pretrained model that is given by the authors of [SKP15] and can be found at https://github.com/davidsandberg/facenet

For generating sequence database to train LSTM network, similar to [RCG+17] we first extract the feature vector for each image using the model that has been trained on the second stage. This process produces a set of feature vectors v with $v \in \mathbb{R}^{1792}$ since the `Average Pooling` layer of the model return 1792 output numbers as 1D vector. Those vectors are grouped together in sequences of length p in a way that each frame is the last of a sequence once. E.g., if the first sequence is $s_0 = \{v_0, v_1, ..., v_{p-1}, v_p\}$, then the next sequence is $s_1 = \{v_1, v_2, ..., v_p, v_{p+1}\}$. Because we are building a sequence database for pain intensity estimation as regression task, each of those generated sequences is labelled as the pain intensity of its last frame. Hence, the prediction of a frame is done taking into account the past p frames. The value of p is set to 16 based on preliminary testing.

Facing imbalanced data

As stated, the UNBC McMaster is a huge imbalance dataset with about $8,000$ pain frames and about $40,000$ no-pain frames. So, we balance the training data for both the original and generated sequence databases by randomly under-sample the majority class, i.e. the no-pain class, so that both pain and no-pain categories have the same probability to be randomly picked by the training algorithm. For the DISFA database, since this database is also imbalance and is only used to train for the first stage, we keep only the frames that have minimum of two AUs with its intensity greater than zero.

3.5.3 Evaluation metrics

We conducted a series of experiments to evaluate the effectiveness of the proposed approach on the widely used UNBC McMaster [LCP+12] database. To compare our results with the other works, the leave-one-subject-out cross-validation is applied on all of our experiments. Data from one subject of the UNBC McMaster database is excluded for validation, the rest are combined with data from DISFA database for training phase, repeatedly. For comparing within the author's scheme, we use Mean Squared Error

Table 3.1: Evaluating the effectiveness of the learning to focus on regions-of-interest in the first stage training. CNN refer to the vanilla InceptionResnet network.

Model	Trained on databases	MSE	PCC
CNN model	UNBC	0.75	0.76
CNN stages 1 + 2 model	UNBC	0.67	0.78
CNN stages 1 + 2 model	UNBC & DISFA	**0.63**	**0.80**

(MSE) and Pearson Correlation Coefficient (PCC). For comparison with other SOTA approaches, we use MSE, Mean Absolute Error (MAE), PCC, and ICC. Between these evaluation metrics, for MSE and MAE: the lower the better; for PCC and ICC: the higher the better.

3.5.4 Experiments and results

Firstly, we would like to evaluate the effectiveness of training AU estimation task as heatmap regression for better exploiting regions-of-interest. Table 3.1 shows the evaluation results of CNN model training with and without the task. From this table, we can see that the model trained with AU estimation task (CNN stages 1 + 2) clearly outperforms the model trained without this task. This result is not surprising, as the heatmap regression task guides the network to focus on the right pain-related regions-of-interest, boosting performance of the whole network when utilising the learned features for pain intensity estimation training. Besides learning to focus on regions-of-interest, the AU estimation task also unlocks the ability to train our network on multi-database combination, which further improves the performance of the network. These results once again confirms the effectiveness of learning to focus on regions-of-interest in exploiting the appropriate feature representations from face image. Figure 3.6 shows the visualisation of heatmap prediction results of a subject in the UNBC McMaster database. It can be seen that our network have learned well to focus on the right location of each individual pain-related facial AU.

Next, we would like to test the effectiveness of the LSTM network by comparing the

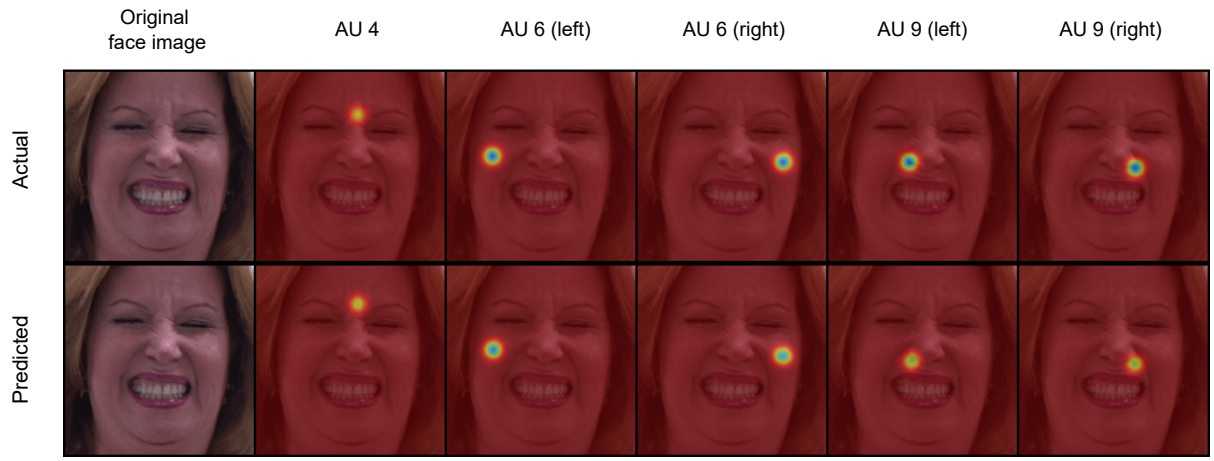


Figure 3.6: The visualisation of the heatmap outputs of the first stage of our network. It can be seen that our network predicted both the location and intensity of each facial AU quite correctly.

Table 3.2: Comparison the effectiveness of the LSTM and the layer to extract features. The models was trained on the UNBC McMaster only (without DISFA database).

Model	With LSTM	MSE	PCC
CNN model	No	0.67	0.78
CNN (Fully Connected)	Yes	0.74	0.78
CNN (Average Pooling)	Yes	**0.65**	**0.80**

performance of the CNN model alone and the CNN model that links to LSTM model. To eliminate the effect of other factors like the selection of AUs or the use of the secondary database, we trained these models on the UNBC McMaster database only, for pain intensity estimation. Table 3.2 shows the evaluation results of three different training configurations: CNN model alone, CNN model linked with LSTM at the first `Fully Connected` layer, and CNN model link with LSTM at the `Average Pooling` layer. From Table 3.2, we can see that when using the output of `Average Pooling` layer as input for training LSTM model, the performance of the whole network have been significantly improved compared to CNN model alone. The output of `Fully Connected` layer seems to contain less temporal information, as it gives a worse result than the CNN model alone.

We further investigated the effectiveness of the data contribution from the secondary database. As it can be seen in Table 3.3, results from the first two stages are already

Table 3.3: Comparison the performance of the model with and without the data contribution from the DISFA database.

Model	Database(s)	MSE	PCC
CNN model	UNBC	0.67	0.78
CNN stages 1 + 2 model	UNBC & DISFA	0.63	0.80
The three stages model	UNBC & DISFA	**0.60**	**0.82**

Table 3.4: Comparison the performance of the model at the second stage when using different AUs as target heatmaps for training in the first stage.

Model	AUs	MSE	PCC
The second stage	All common AUs	0.78	0.74
The second stage	AU4, AU6, AU9	**0.63**	**0.80**

better than the one trained without the secondary database. And when we put all the three stages together, it pushes the result even higher. This demonstrates the effectiveness of our approach when learning from the combination of two database instead of just a single one. However, to draw this advantage, the work of selecting AUs for the heatmap regression in the first stage is also important. As shows in Table 3.4, the results when we use all common AUs between the two databases are worse than if we use only AU4, AU6 and AU9. This could be happened because AU4, AU6 and AU9 are parts of PSPI formula (Eq. 1.1) , which make the network easier to learn in next stages. If we train with all common AUs, the non-related AUs can introduce noise, which may reduce the capacity of the model.

Finally, we would like to test the effectiveness of freezing the first layers's parameters of the base model. We hypothesised that the UNBC McMaster is a huge imbalanced database, so overfitting can easily occur when we fine-tune the whole network, i.e. no freezing. Furthermore, we believed that this freezing layers will preserve the parameters that have been trained with data from the two databases, which will make it better when predicting new cases. Table 3.5 shows the result of the first two stages of the model when applying and not applying the layers freezing. It is clear that the model works better when first layers' parameters are frozen.

Table 3.5: Comparison the performance of the model at the second stage when freezing and not freezing the first layers of the base network

Model	Freezing	MSE	PCC
The second stage	No	0.69	0.79
The second stage	Yes	**0.63**	**0.80**

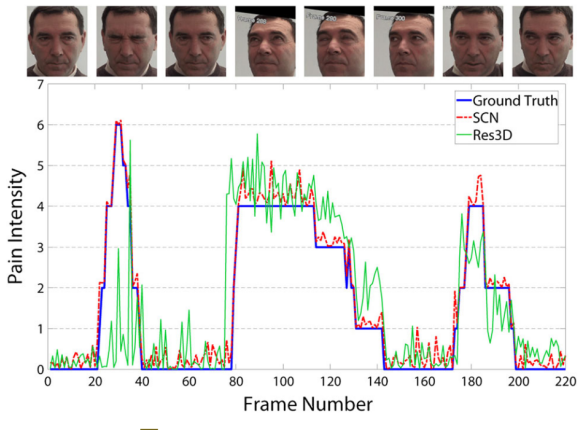
3.5.5 Reference to the works of Mohammad Tavakolian

Different from our original published paper I, in this thesis, we do not cite the works [TH18b, TH19] of Mohammad Tavakolian. The reason for that is because we see that their results are rather strange, as their deep learning model consists of 423.2 millions parameters, while the UNBC database contains only 48.398 images and only 8.369 images with $PSPI > 0$. The amount of training examples are way smaller than the size of their proposing network. Regarding this problem, Yan Lecun once said in [L+89]:

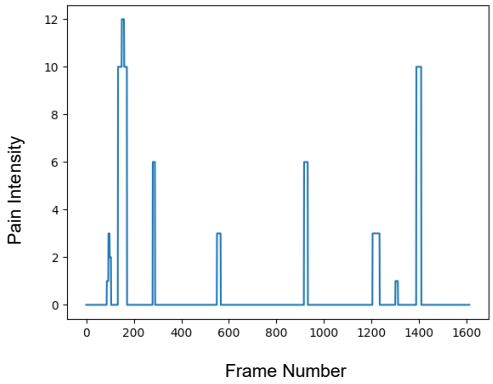
"*Theoretical studies (Denker et al , 1987) (Patarnello and Carnevah, 1987) have shown that the likelihood of correct generalization depends on the size of the hypothesis space (total number of networks being considered), the size of the solution space (set of networks that give good generalization), and the number of training examples. If the hypothesis space is too large and/or the number of training examples is too small , then there will be a vast number of networks which are consistent with the training data , only a small proportion of which will lie the true solution space, so poor generalization is to be expected*".

Generally speaking, the more parameters there are in a learning model, the more training examples are needed to avoid the problem of overfitting. And the size of Tavakolian's model is simply insane compared to the tiny size of the UNBC McMaster dataset. For reference, the model of Tran *et al.* [TBF+15] (the one that their model is based on) has only 17.5 millions of parameters, which is about $\frac{1}{24}$ of the size of Tavakolian's model. And they trained it on the Sport1M database, which is about 20 times bigger than the UNBC McMaster database.

Apart from the size of the training dataset, the imbalance of the data is also one of the strange things in their results. Since the UNBC McMaster database is an extremely



(a) Fig. 9 - Tavakolian and Hadid, 2019



(b) Ground truth of subject 064-ak064

Figure 3.7: Visualisation of pain intensity prediction in the paper [TH19] (a) and the PSPI ground truth of the subject 064-ak064. We can see that despite visualising the same subject (064-ak064), we can't find the same pattern in (a) compare to the ground truth PSPI (b).

unbalanced database, as can be seen in the Figure 1.8. Hence, when training a deep model on this database without applying any rebalancing technique, the learning model will certainly bias to have its prediction toward the dominant category [CJK04, GMS10]. So, it is strange that they did not apply any data re-balancing technique but still reach a high level of generalisation.

Next, Figure 9 in their paper [TH19] shows the prediction vs ground truth of subject 064-ak064 in the UNBC McMaster database. However, there are no such pattern in the ground truth PSPI visualisation, as can be seen in Figure 3.7, which could be a sign of data leakage or reading incorrect data.

Regarding these concerns, we have tried to send an email to the authors using the email addresses that they provided in their papers. However, when we was trying to send the email, it was rejected as incorrected email addresses. We have further contacted the head of their team at Oulu University and obtained a new email address. We have sent another email regarding these issues to the new email address of the author, but we have not received any response since then. Therefore, we decided to exclude these works out of this dissertation.

3.5.6 Comparison with State of the Art

Table 3.6: Comparison against Leave-One-Subject-Out method with MSE, MAE, PCC, and ICC on the UNBC McMaster database. The best results are shown in bold.

Model	MSE	MAE	PCC	ICC
Kaltwang *et al.* [KRP12]	1.39	-	0.59	0.50
Florea *et al.* [FFV14]	1.21	-	0.53	-
Zhao *et al.* [ZGWJ16]	-	0.81	0.60	0.56
Zhou *et al.* [ZHSZ16]	1.54	-	0.64	-
Rodriguez *et al.* [RCG+17]	0.74	0.50	0.78	0.45
Tavakolian *et al.* [TH18a]	0.69	-	0.81	-
Our 3Stages model	**0.60**	**0.35**	**0.82**	**0.80**

We compared our proposed three stages approach with other works that related to continuous pain intensity estimation, which including Kaltwang *et al.* [KRP12] with their shape and appearance features fusion network, Florea *et al.* [FFV14] with HoT and SVM classifier, Zhao *et al.* [ZGWJ16] with OSVR regression model and Zhou *et al.* [ZHSZ16] with Recurrent Convolution Neural Network, Rodriguez *et al.* [RCG+17] with VGG + LSTM network and Tavakolian *et al.* [TH18a] with their Deep binary representation model. Table 3.6 shows the comparative results of leave-one-subject-out cross-validation method for the above mentioned approaches evaluated on the UNBC McMaster database. From this table, we can observe that our proposed three stages approach outperforms other works in all evaluation metrics. Specifically, our method archives 30% higher than the previous SOTA in term of MAE (Rodriguez *et al.* [RCG+17]) and 13% higher than previous SOTA in term of MSE (Tavakolian *et al.* [TH18a]). For the correlation evaluations, our approach achieved 35% higher than previous SOTA in term of ICC and 1% higher than previous SOTA in term of PCC. These results once again confirm the effectiveness of our approach in focusing on the regions-of-interest for better exploiting data on the multi-database combination.

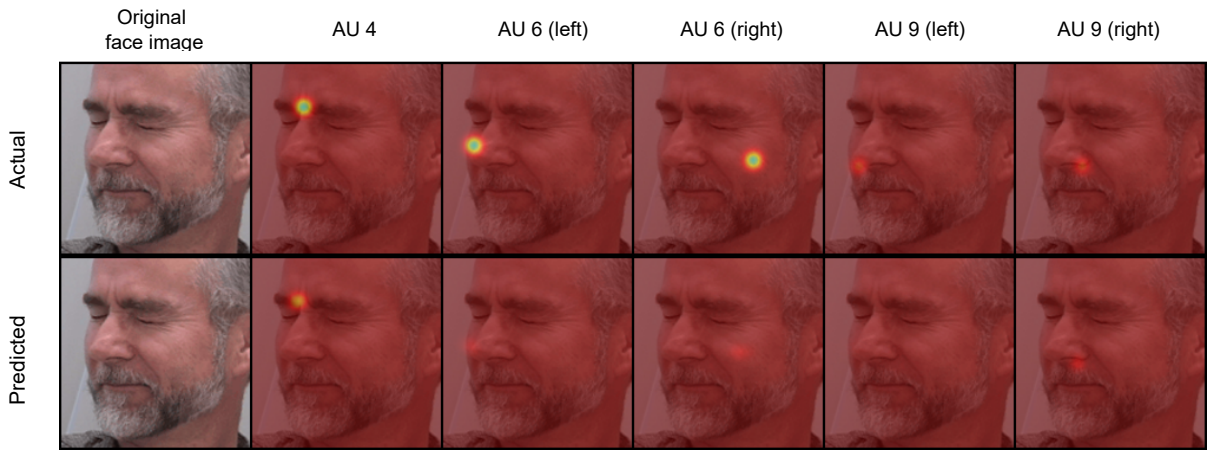


Figure 3.8: The visualisation of an incorrect AU intensity prediction of our network. Using the same weight for both the visible and the obscured parts of the face could be the reason for these incorrect prediction.

3.6 Conclusion

Facial expression of a subject changes spontaneously when experiencing an inner feeling and it is important to precisely extract the feature representations of these changes for better human expression understanding. In this work, we have proposed the learning to focus on regions-of-interest approach training on the combination of multi-database for better pain intensity estimation. Our extensive experiments demonstrate that integrating the locations of the regions-of-interest into the training process provides the deep network with valuable information regarding where to focus in the face image, thus improves the overall performance of the network when finetuning for different facial expressions related tasks, e.g. pain intensity estimation.

While the idea of focusing on regions-of-interest is great for boosting performance of deep neural network, our approach of integrating the location and intensity of the AUs into the training process in the form of heatmap regression still has some limitations. Firstly, we consider the left and right sides of the face with equal weight, which is not entirely correct in some cases when the patients turn their head due to enduring pain (see Figure 3.8). In these cases, some parts of their face are barely visible or completely obscured. Hence, we should pay more attention to the AUs in the visible parts

of the face and reduce the attention to the hard-to-see or obscured parts. Secondly, for generating the heatmap target of each AU, we defined its central location with highest value and gradually decreases as the distance of the pixel increases. This is not an optimal choice because this configuration emphasises the central location of the AU region, while reduces the attention when the pixel move further away from the centre. As the muscle movement of an AU is not only limited to the central of the AU region, but also to the entire muscle-related area of the AU. Therefore, instead of relying on AU target heatmap, an approach that can learn to automatically extract features from the entire AU region is much more appreciable.

In the next chapter, we will present a new approach that addresses both of these limitations. Inspired by the advantage of learning to focus on regions-of-interest, we take it to the next level by not just focusing, but isolating these regions for better extracting feature representations. Besides introducing the new deep learning network architecture, we also propose a new approach for better balancing the training data, which is also a critical problem for the domain of facial analysis in general and pain intensity estimation in particular.

Chapter 4

Learning to isolate regions-of-interest for better pain estimation

Contents

Deep learning is successful when massive amounts of annotated data are available, as evidenced by astonishing results in many different domains, including speech recognition, machine translation and image categorisation. For many problems, however, the precious annotated data may be scarce, hard to obtain or simply unavailable. Facial action unit and pain intensity estimation are among those domains that suffer from data-deficiency. Hence, there is a need of developing a deep learning approach that effectively exploits better feature representations from sparse data. In this chapter, we summarise our findings in paper III, which presents an approach to efficiently learn to isolate regions-of-interest from face image for better extracting feature representations, improving the performance of facial AU estimation and pain intensity assessment.

Paper III: M. T. Vu, M. Beurton-Aimar and K. TRAN, "FFAU: Faster-RCNN for Facial Action Unit intensity estimation," 2022 Pattern Recognition Journal (Submitted).

4.1 Context

Data scarcity has long been the major issue while building a deep learning model, as in many fields, sufficient amount of data is not available to train the deep model. The lesser amount of training data often leads to a phenomenon called over-fitting: the model performs well in training but not on newly unseen data. In fact, deep neural network overfits the training data by memorising small training data without learning underlying patterns [LKG19]. In such a situation, the model performs exceptionally well on the training data but fails miserably on the test data or in the real world. The typical approaches to overcome this problem are including data augmentations and transfer learning. Data augmentation techniques enrich training data by generating additional training examples using various label-preserving transformations, such as scaling, zooming, and random cropping of images. In the other hand, the transfer learning approach attempts to transfer the knowledge gained on a large labelled source dataset for a target task [HAG+17]. Both of these approaches are useful to fight against

the data scarcity problem. However, in cases where the dataset is extremely unbalanced and the number of samples is also very limited, applying these techniques alone is not sufficient. Facial AU intensity and PSPI pain intensity estimation are among these cases where labelled data are limited both in the number of samples and in the distribution of the intensity levels (see Section 1.6). To overcome this problem, prior studies [ZDHJ18, ZZD+18, LTWE+17, SCW+21] attempted to use a semi-supervised approach or leverage prior knowledge to have more training data. The works [KTP15, WRPP16, RRBP16] tried to exploit more information from a single image by utilising the co-occurrence of the AUs. Other works [ZJW+19, CTC17] tried to exploit temporal information between the consecutive frames of a video. The common point of these works is the fact that they tried to analyse all the AUs together, without pointing out explicitly where to find the information regarding these AUs on the face image. Despite the fact that deep learning has the capability to find these information automatically through learning [DT18, Kim10, dSP22], it requires a huge amount of data for model learning to avoid overfitting [dSP22, ZDHJ18], which is difficult to obtain in the domain of facial AU intensity estimation as mentioned earlier.

Inspired by the advantages of the learning to focus on regions-of-interest that we have presented in Chapter 3. Here we present a new deep neural network called FFAU network, which is not just focusing but isolating the regions-of-interest for better feature extraction. Based on the concept of *divide and conquer* paradigm, we utilise the Faster RCNN object detection network [RHGS15] to locate the AU regions of interest (*divide*) before put them through a set of AU regressor networks for AU intensity estimation (*conquer*). By isolating each AU region, we are able to estimate correctly its intensity without worrying about learning incorrect features from other non-related regions, reducing the chance of being overfitted. In addition to localising and estimating AU intensity, our FFAU network has also addressed the head pose problem of the patients when filming. By explicitly training the network to take into account only the visible parts of the face during training and ignoring the obscured parts, our network is able to pick up the correct features even in the extreme head pose cases when the

patients over turn their faces to the left or right. Experiments on the two widely known databases UNBC McMaster and DISFA databases show that our approach outperforms other SOTA approaches on both the two databases.

Besides improving network's performance, the explainability of the model behaviour is also an important aspect that we have considered when designing our network. As deep learning models are usually considered as black-box due to their complex mapping of millions of parameters inside these networks, it is difficult to obtain interpretations and explanations for the behaviour of the network. In this work, our approach takes a step towards bettter explicability in predicting PSPI pain intensity level. Previous SOTA approaches only give a final PSPI intensity level for a given image without giving any explanation. In the other hand, our approach not only tells the intensity value of each pain-related AUs that contributed to the PSPI score, but also shows where are the regions for each of these AUs. These are important information which can provides some insight about the behaviour of our model and help practitioner or medical doctor to see and evaluate the reliability of our predictions.

In the next sections, we explain step-by-step about our approach. In section 4.2, we review the object detection network, which is the base network that we used in our approach. We discuss the problem of data imbalance and how we have dealt with it in section 4.3. Section 4.4.2 explains the architecture of our proposed FFAU neural network. Finally, the experiment results and discussion are explained in sections sections 4.5–4.8.

4.2 Object detection network

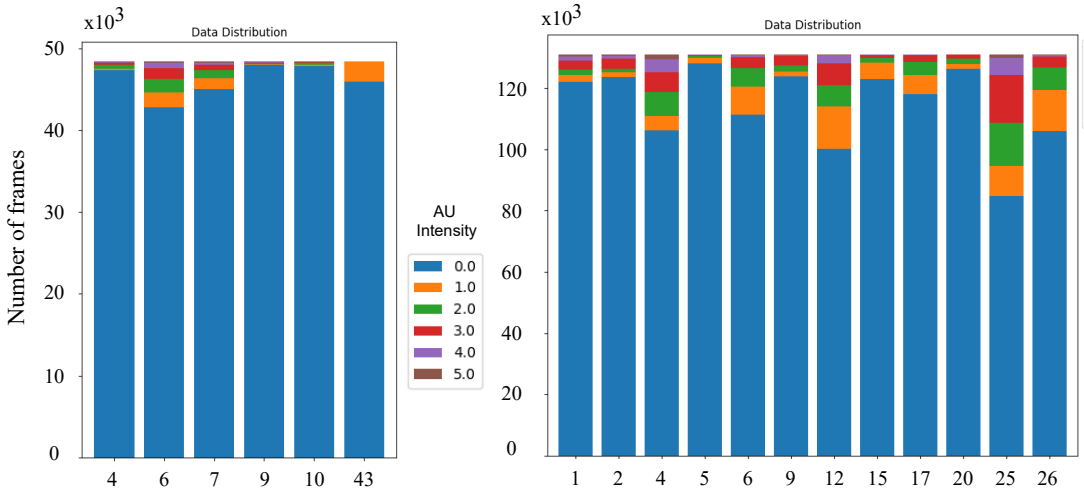
Object detection is an important task for many different computer vision problems [ZZXW18, JLM17, ZXT18, EV18]. The task involves locating and classifying objects in an image or video. There are two types of object detectors: single stage and two-stage. One of the first two-stage detector network is Selective Search [USGS13], in which the first stage generates a set of candidate proposals and the second stage clas-

sifies the proposals as one of the foreground (target) classes or as background. RCNN [GDDM13] replaces the second stage classifier by a CNN network, yielding large gains in accuracy. SPPNet [HZRS14] improves RCNN network by sharing feature extraction stage and use spatial pyramid pooling to extract fixed length feature for each proposal. Fast RCNN [Gir15] improves over SPPNet by introducing a differentiable RoI Pooling operation, enabling the network to be able to train end-to-end. Region Proposal Networks (RPN) is proposed in [RHGS15], which integrates proposal generation with the second-stage classifier of Fast RCNN into a single convolution network, forming the Faster RCNN framework. Besides the two-stage detector, we also have single-stage detector that is primarily aimed at detecting objects in real time. This type of detector network predict object classes and locations directly, hence much faster than two-stage detector networks. OverFeat [SEZ+13], SSD [FLR+17, LAE+16] and YOLO [RDGF15, RF17, BWL20] are some of one-stage methods. Yet, as the two-stage detectors are generally more accurate than single stage detectors [ZAA+21, CWS+18] and the inference speed is out of our concern, we have selected Faster RCNN as the object detector backbone of our network.

Regarding the Object detection in face analysis, since there are no directly relation between these two domains, there are only a few works in the literature. Faster RCNN was used in [JLM17, ZXT18] for face detection from images. Li *et al.* [LZZ+17] and Zaman *et al.* [ZSS+22] both utilised Faster RCNN to localise face-area and then classify facial expressions. The main objective of the object detection task in these approaches is still limited to localising the human head or face. To the best of our knowledge, our approach is the first to utilise object detection to localise and analyse human emotion at AU level.

4.3 Dataset re-balancing

Since our experimental databases are quite unbalanced, most of the samples are being labelled with zero intensity, which can be seen in Figure 4.1. This certainly will bias



(a) UNBC McMaster dataset

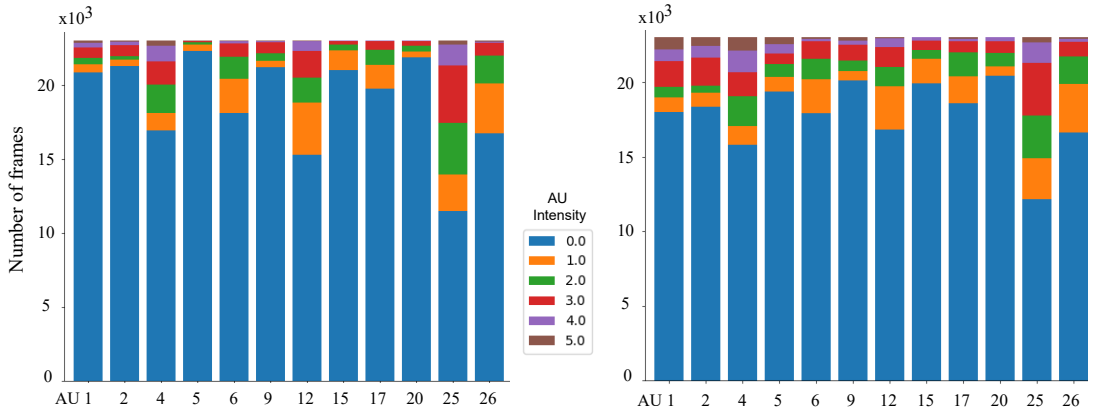
(b) DISFA dataset

Figure 4.1: The distribution of facial AU intensity of the UNBC McMaster dataset (a) and the DISFA dataset (b).

the training network to have its prediction toward the dominant category [JCDLT13, CJK04, GMS10], which obviously reduces the capability of the network. Therefore, it is crucial to rebalance the dataset before any further training or analysis. Since the samples in our datasets were annotated in one-to-many fashion, i.e. multiple AUs with multiple intensity levels for a single image, we have applied two popular rebalancing techniques with some modifications to rebalance our databases. These techniques are including under-sampling and over-sampling techniques, which are described in the next sections.

Under-sampling

Traditional under-sampling technique is about randomly drop samples in the majority category. However, since our datasets is quite small and deep learning model in general requires massive amount of data, dropping data should be used as little as possible. Instead, as our datasets are of the video type, for every k consecutive frame samples, we collapse those samples into a single representative sample. In training, when reading a representative sample, we unfold and randomly pick one of its collapsed samples. This



(a) DISFA dataset collapsed

(b) DISFA dataset re-balanced

Figure 4.2: The distribution of facial AU intensity of the DISFA database after collapsing (a) and re-balancing (b).

way, we are literally not dropping a single frame in our dataset, while consolidating the balancing of our dataset. Considering the FPS of these datasets are around 20 frames per second, for the majority category (i.e the intensity of all AUs are all zeroes), k is set to 20. Otherwise, k is set to 3. The reason for setting $k = 3$ for the minority categories is the fact that in video type of dataset, there is not much of difference between two consecutive frame samples. Furthermore, the model could be getting overfitted if the case of several consecutive images appearing in a single batch is repeated. Figure 4.2a shows the distribution of facial AU intensity of the DISFA dataset after this collapsing step.

Over-sampling

Over-sampling technique is about randomly duplicating samples in the minority category. However, if we duplicate too many times a particular sample, the training model could remember it instead of learning something from it. Therefore, selecting the duplicating rate wisely for each sample is an important factor for rebalancing our datasets. Since the number of samples annotated with intensity greater than zero for each AU is unbalanced (e.g., AU 1 and AU 25 in Fig. 4.1b) and so does the number of samples at

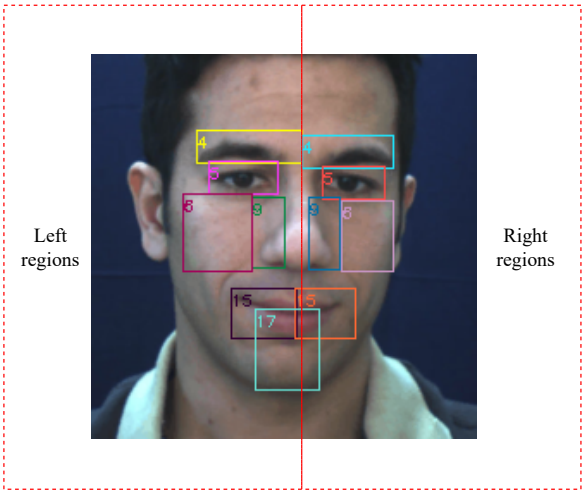
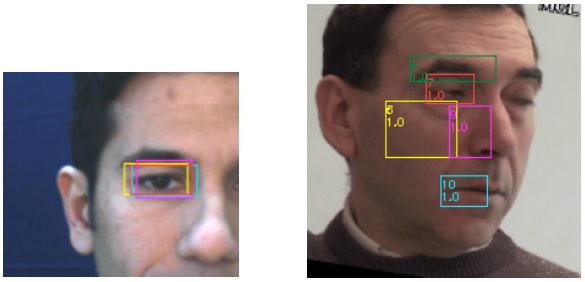


Figure 4.3: Facial regions bounding boxes extracted from face image using facial landmarks.



(a) Multi bounding boxes in the same region

(b) No bounding box in the face right side

Figure 4.4: Bounding boxes in training phase of our Facial region localisation module. There could be multiple bounding boxes proposed by RPN module for the same region (a) and there could also be no bounding box for some regions (b).

each intensity levels, we try to rebalance for both of them. Let U_i^j be the number of times the AU i with intensity j appears in our dataset. The duplicating weight of AU i is denoted as:

$$W_{U_i} = \frac{1}{\sum_{j=1}^{5} U_i^j} \times \max_{\alpha=0}^{T} \sum_{j=1}^{5} U_\alpha^j \tag{4.1}$$

With T denotes the total number of AUs that exists in our dataset. From Equation 4.1 we can see that the lower amount of samples annotated with AU i and intensity $j > 0$, the higher weight it gets. Therefore, this W_{U_i} ensures the balance of each AU $i \in \{0 \,..\, T\}$ inside the dataset.

Next, the duplicating weight of AU i with intensity j is denoted as:

$$W_{U_i^j} = \frac{1}{U_i^j} \times \max_{\beta=0}^{5} U_i^\beta \tag{4.2}$$

Again, we can see that the higher amount of samples annotated with intensity $j \leq 5$, the lower weight it gets. While the Equation 4.1 tries to rebalance the dataset at AU

level, Equation 4.2 tries to rebalance the dataset at intensity level. Finally, let x^i be the intensity of a sample x with its annotated AU i, the final weight of x is denoted as:

$$W_x = \frac{1}{T}\sum_{i=0}^{T} W_{U_i} \times \left(W_{U_i^{x^i}} \times \frac{1}{f} + 1 \right) \tag{4.3}$$

With f denotes the hyper-parameter weight-factor that we will have to select to determine how much we want to penalise the weight for the minority categories. Since the difference in the amount of data between AU intensities is huge (see Figure 4.1), we need this term to prevent oversampling too much the minority samples. The last term (plus one) in the equation is used to ensure the term $\frac{1}{f}$ rescaling the weight of the minority categories effectively.

In this work, we have set $f = 5$ based on our preliminary testing. Results of the rebalancing for the DISFA dataset are shown in Figure 4.2b. It can be seen that the balancing of the dataset has been improved compare to the original dataset (see Figure 4.1). Eventhough we can further reduce the value of f to get a nicer distribution of AU intensity. However, as we have mentioned earlier, duplicating too much will lead into the problem of overfitting, therefore we keep this data balancing configuration for all of our experiments.

4.4 Faster-RCNN for facial action unit intensity estimation approach

4.4.1 Facial region bounding boxes definition

In order to generate ground truth for training our network to localise the active appearance locations for each AU in our databases, we have defined the facial region bounding boxes based on the provided facial landmarks as can be seen in Figure 4.3. Each facial region contains one facial structure, which is the main active appearance of one or more facial AUs. One can see these facial regions as a more generic version of the heatmap

Table 4.1: Region definition of each facial AUs. A region contains one facial structure and contains one or more facial AUs.

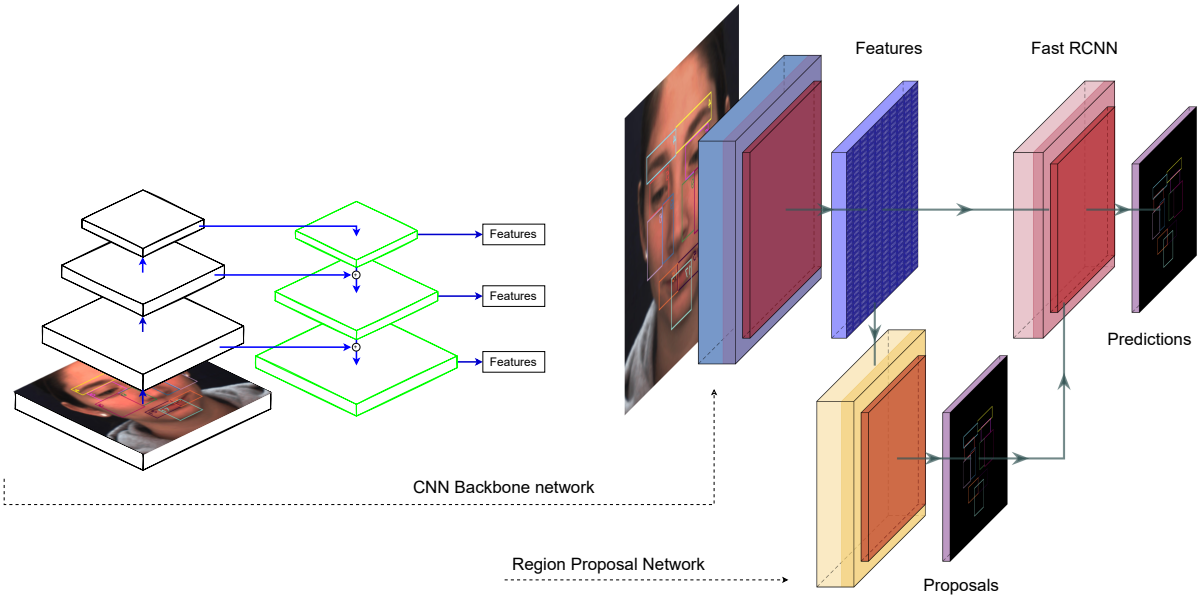
Region ID	Position	Definition	AU(s) included
4	Left & right	Browns area	AU 1, 2, 4
5	Left & right	Eyes area	AU 5, 7, 43
6	Left & right	Cheeks area	AU 6
9	Left & right	Nose wings area	AU 9
15	Left & right	Mouth area	AU 10, 12, 15, 20, 25, 26
17	Center	Chin area	AU 17

ground-truth defined in [SLTV18, FLL20, VBADE21]. While their heatmap ground-truth highly emphasises the central of the region, our facial region approach treats each pixel in the region equally and leaves the decision of which place to emphasise to the higher layer of neural network. Since the human facial structures in the left and right of the face are balanced in general, we have defined the left and right regions for each of the facial AUs accordingly. Table 4.1 shows the predefined facial regions and their containing facial AUs.

4.4.2 FFAU network architecture

4.4.2.1 Facial region localisation module

As we have mentioned earlier, the face region localisation module of our network takes the responsibility of localising each of the face regions that we defined in Table 4.1. In this work, we have chosen Faster RCNN network [RHGS15] architecture for this part of the network since it is one of the SOTA neural networks for objects detection and it also is an unified end-to-end trainable neural network. Faster RCNN network (see Figure 4.5b) consists of three main modules: the CNN backbone network for extracting features, the Region Proposal Network (RPN) for generating proposal regions and the Fast RCNN module for detecting objects in the proposed regions. In the next paragraphs, we explain step-by-step the way we configured each module inside this network.



(a) Feature Pyramid Network architecture (b) Faster RCNN Network architecture

Figure 4.5: Facial region localisation module visualisation. The module consists of a Faster RCNN network (b) built on top of a Feature Pyramid Network (a).

CNN backbone module: For the CNN backbone, we adopt the Feature Pyramid Network (FPN) [LDG+16] architecture as the backbone for our FFAU network since it has shown to be effectively improved the performance of the object detection tasks [HGDG17, LDG+16]. This network augments a standard convolutional network by adding a top-down pathway and lateral connections into the original network, enabling the network to construct a rich and multi-scale feature pyramid from a single resolution input image. The illustration of this network can be seen in Figure 4.5a.

Similar to [HGDG17, LDG+16], we have built a FPN on top of a ResNet 50 architecture [HZRS15]. Following [LDG+16], we have constructed 4 levels of FPN feature pyramid using the last 4 blocks of the ResNet architecture. The number of channels for each FPN pyramid layer is set to 256 channels.

RPN module: For the RPN module, we configured this module to generate anchor sizes of $\{16^2, 32^2, 64^2, 96^2, 128^2\}$ pixels and their aspect ratios (height:width) of {1:4,

1:2, 1:1, 3:2, 2:1} to cover a variety of potential shapes and sizes. Each of these anchor is assigned to a one-hot vector of classification targets with length γ, where γ is the total number of facial regions in our dataset. From Table 4.1, we can see that there are 5 pair regions with two positions left and right, one region with only one center position and one final region is reserved for the background. Totally, the number of regions are $\gamma = 12$ regions. To assign an anchor to a classification target, we use a similar assignment rule as in [RHGS15]. Specifically, anchors are assigned to target object boxes if the Intersection Over Union (IoU) between them is no less than 0.7 and to background if it is no greater than 0.3. The Non-Maximum Suppression (NMS) is set to 0.8 to suppress the anchors that overlaps too much. During training, we also have modified the positive anchors selective algorithm in RCNN network to ensure the presentation of all types of regions for each of the training image.

Table 4.2: Per region RoI layer and its corresponding Conv layer configuration to ensure the same output of 5×5 for each facial region.

Region ID	RoI output size	Conv layer	
		kernel size	stride
4	11×17	3×5	2×3
5	11×17	3×5	2×3
6	13×13	5×5	2×2
9	13×13	5×5	2×2
15	11×17	3×5	2×3
17	13×13	5×5	2×2

Fast RCNN module: The Fast RCNN module [Gir15, RHGS15] consists of three parts: a Region-of-Interest (RoI) pooling layer for extracting the interesting part of the backbone features using the given RPN anchor boxes; a classification network for classifying the extracted features and a regression network for regressing the offset from each anchor box to its nearby target object. For the RoI pooling layer, we have selected RoIAlign [HGDG17] instead of the original RoI pooling layer as in [Gir15, RHGS15] since it has shown to be more precise than the original one [HGDG17]. For the classification net-

work, we use a stack of Linear and ReLU layers with output of k classes. Similarly, we use the same architecture but with output of $k \times 4$ for the regression network. In inference phase, we only keep the predicted boxes which have its classification confidence scores equal or higher than 0.7.

4.4.2.2 AU intensity estimation module

The AU intensity estimation module is the module that utilises the predicted regions from the *Facial region localisation module* (section 4.4.2.1) and the shared backbone features to estimate the intensity for each of our facial AUs. This module consists of three sub-modules, including a *Per region RoI pooling layer* for extracting region features at different sizes from backbone features, a set of *Region feature extractor* sub-modules for exploiting features from each of the facial regions and a set of *AU intensity estimator* sub-modules for estimating the intensity for each of our facial AUs. Figure 4.6 shows the overview of this module.

Per region RoI pooling layer The main role of this layer is about extracting regional features from our shared backbone features by using the predicted region bounding boxes of the *Facial region localisation module*, the core of this layer is the RoIAlign pooling layer [HGDG17]. Since the aspect ratio of each of our region bounding boxes are quite different, e.g., the appearance of region 6 is most often a square shape, while it is a long rectangle for region 4 (see Figure 4.3). Therefore, if we apply the same square RoI pooling size for all of these regions, some information of the long rectangle shape region will be lost due to the RoI max pooling operation applying on a large receptive field. To avoid this problem, we propose a *Per region RoI pooling layer*, which is basically a mapping between each region with its corresponding RoI layer as can be seen in Table 4.2. From this table, we can see that region with long rectangle shape (i.e region 4) is mapped with a RoIAlign pooling layer with long rectangle output size (11×17), therefore reducing the lost of information.

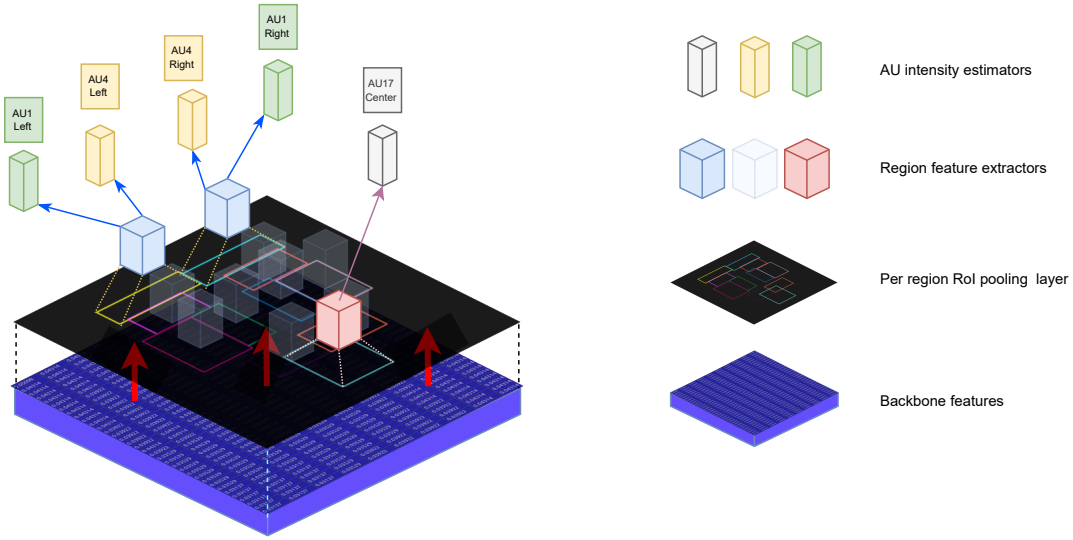


Figure 4.6: AU intensity estimation module. The *Per region RoI pooling layer* extracts regional features from our shared backbone features and routes it to pass though the corresponding *Region feature extractor* model based on the type of each region. Finally, the *AU intensity estimation* model extracts features and estimates the AU intensity for each of the given regional features.

Another responsibility of this layer is about resolving the uncertainty of the region proposals. Since the bounding box proposed by the *Facial region localisation* module could be including multiple or zero bounding boxes for a region (see Figure 4.4), in training, we randomly pick one bounding box per region to extract features. In evaluating, we pick the highest confidence bounding box to evaluate. If there is no bounding box proposed, we just return an empty feature vector.

Region feature extractor Since the output of our *Per region RoI pooling layer* is a map of regional features with different size, according to the type of region (see Table 4.2), we need to have a CNN network to extract features and also reduce the dimension for each of these regional features to the same size. To fulfill this requirement, we have constructed a set of m different *Region feature extractor* neural networks for each of our facial regions. Each of these networks consists of a Reduction block and an Inception-ResNet block as can be seen in Figure 4.7. These two blocks are parts of the Inception-ResNet architecture [SIVA16] with some modifications in number of channels

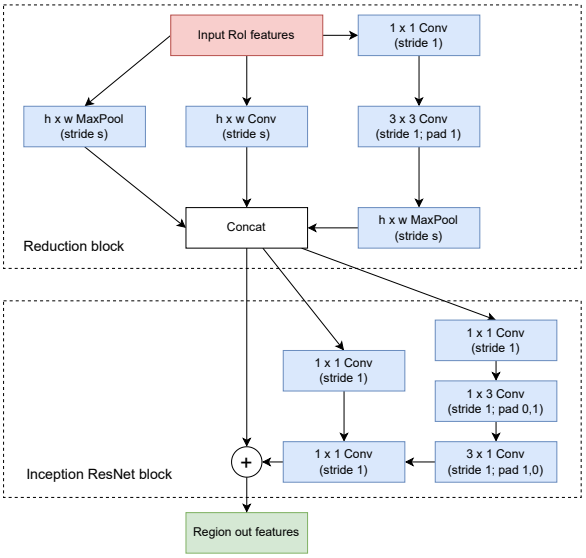


Figure 4.7: Network architecture visualisation of a *Region feature extractor* sub-module. The h, w and s parameters are corresponding to the Conv kernel size (h, w) and stride (s) that are defined in Table 4.2.

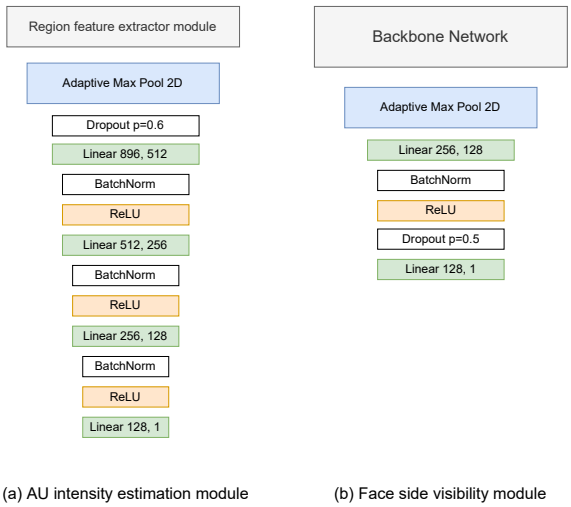


Figure 4.8: Network architecture of the AU intensity estimator module (a) and Face side visibility module (b).

and kernel size of its 2D Convolutional layers (Conv). Specifically, the kernel size (h, w) and stride (s) of the Conv layers in the Reduction block are configured differently according to the type of region to ensure that it produces the same output size as can be seen in Table 4.2. From this table and Table 4.1, we can see that there are 6 different facial regions (excluding background region) and 5 of them are pair regions (left and right). In this approach, we design our network to use a single *Region feature extractor* model for a pair regions. Therefore, there are totally $m = 6$ number of *Region feature extractor* models that are constructed in our approach.

Once we have all the regional features from our *Region feature extractor* models, it's time to estimate the intensity for each of our facial AUs.

AU intensity estimator The *AU intensity estimator* network consists of a sequence of Linear, BatchNorm, Dropout and ReLU layers as can be seen in Figure 4.8a. Features from the previous *Region feature extractor* module are shrinked to the size of 1×1 by Max Pooling operation before going through this module for estimating AU intensity.

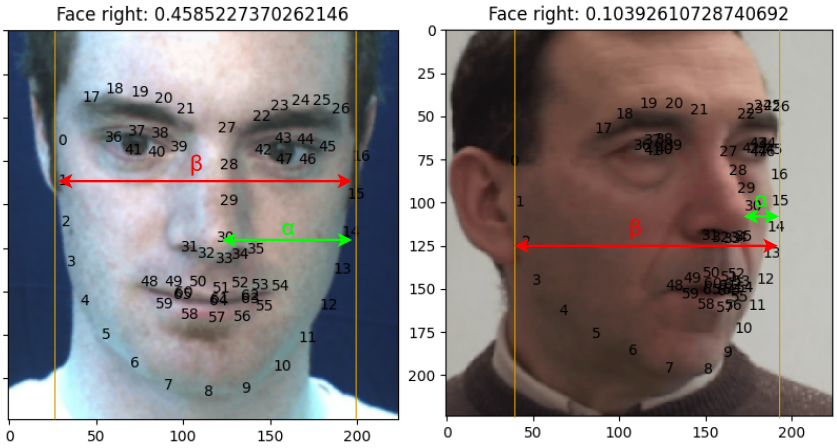


Figure 4.9: Face side visibility ground truth generation using the provided facial landmarks.

One can note that for one facial region, there could be containing multiple *AU intensity estimator* modules (see Figure 4.6). As we have mentioned earlier, considering the cases like AU 12 (oblique lip raising) and AU 25 (lips parting), it makes sense to have them sharing the same mouth region, since the most active appearance of these two AUs is the mouth region.

4.4.2.3 Face side visibility module

Handling head pose is a crucial step in many face-related domains [BMP09, BM08]. Since the patient not always looks directly to the camera when recording, it is important to be able to handle the visible, half-visible and obscured parts of the face correctly. An example of these cases can be seen in Figure 4.9b, the right part of the face is obscured and therefore it is incorrect to treat them in the same way as the left part of the face, we need to tell our network to focus more in the left side instead.

In order to solve this problem, we propose the *Face side visibility* module network, which is a network for estimating the percentage of the visibility of the right face side of the given face image. The network consists of a sequence of Linear, BatchNorm, Dropout and ReLU layers (similar to the *AU intensity estimator module*) that we have

put it on top of our CNN backbone network (see Figure 4.8b). In training, we generate ground truth to train this network by relying to the position of the nose point in the face. Specifically, for a given aligned face image sample i with its facial landmarks matrix M_i, let α_i be the horizontal distance between the nose point and M_i's farthest point to the right. Let β_i be the maximum horizontal distance of M_i, the ground truth of face right side percentage p_i^r of sample i is defined as:

$$p_i^r = \frac{\alpha_i}{\beta_i} \tag{4.4}$$

From Equation 4.4, we can see that the ground truth p_i^r emphasises the percentage of visibility of the right side of the face. Since the face image is aligned, the percentage of visibility of the face left side can be calculated as $p_i^l = 1 - p_i^r$. Because of having this relationship between p_i^r and p_i^l, we only need to train our *Face side visibility* module to estimate the value of p_i^r, then wen can calculate the value of p_i^l accordingly.

4.4.2.4 Movement exploitation module

In order to capture the temporal dynamics between the consecutive frame images, we designed the *Movement exploitation* module, as can be seen in Figure 4.10. For each facial AU, features from the first ReLU layer of the *AU intensity estimator* module (see Figure 4.8a) are extracted and fed into a bidirectional LSTM network. LSTM is a variant of RNN which has a capability of keeping long-term information from previous inputs. By learning the changing of the facial expression over time through the sequence of frames, we expect that this network will be able to capture the facial movements by using both the past and future information (bidirection) and combining it with the information from the current frame to make a better decision. As there are two positions for each AU (except AU 17, see Table 4.1), we take their output predictions and aggregate them with the percentage of face side visibility using the same function as described in Equation 4.6.

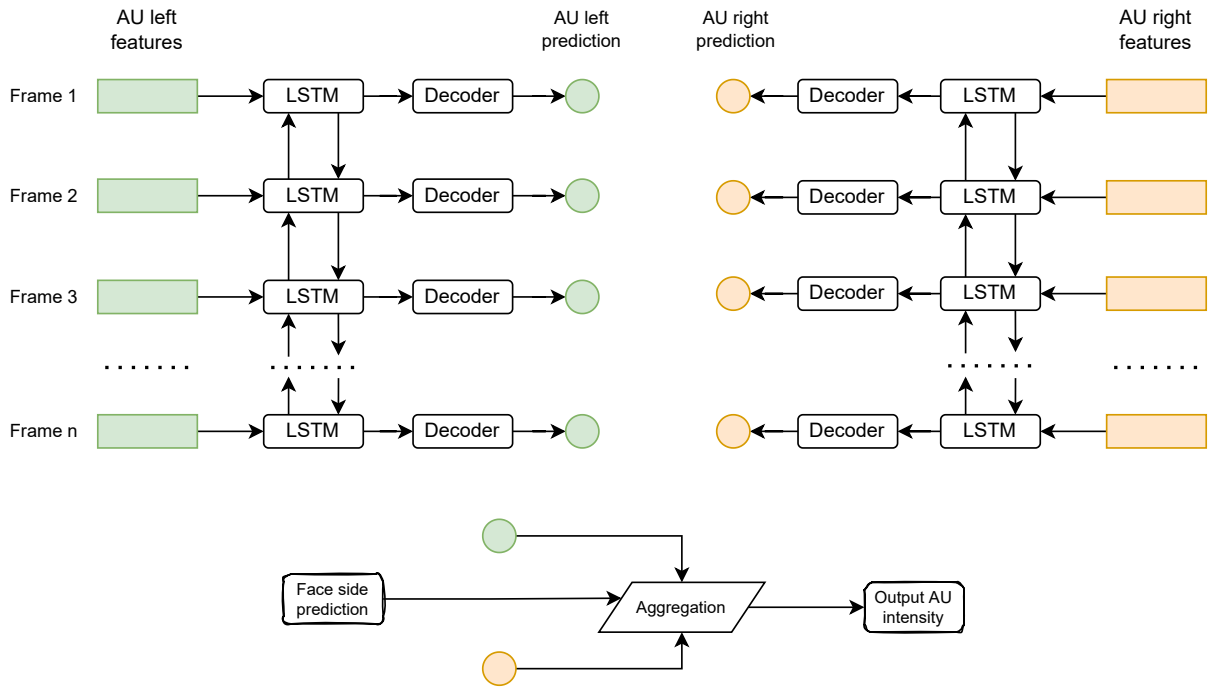


Figure 4.10: Network architecture of our *Movement exploitation module*. For each AU, features from each face side are fed into a bidirectional LSTM for exploiting the temporal dynamics between the consecutive frame images. Then, these features are decoded and aggregated to obtain a final estimate of the AU intensity using the predicted face side visibility percentage of the given frame image.

4.4.3 Loss functions

4.4.3.1 Face side visibility module

As we have mentioned earlier, in our approach, we try to automatically predict the percentage of the right face side visibility for better handling the patient's head pose cases. The percentage of the left face side visibility will be easily calculated according to the predicted of the right face side visibility percentage. In order to train the network right face side visibility prediction, for each sample i in our dataset, we try to minimise the L_1 distance between the predicted y_i^r and the ground truth p_i^r face right side percentage that we have defined in section 4.4.2.3. The objective loss function for this module is denoted as:

$$L_i^r = \|y_i^r - p_i^r\| \tag{4.5}$$

By minimising L_i^r, we expect that our network will be able to capture the important features of head posing in face images and then measure the visibility of the right face side correctly.

4.4.3.2 AU intensity estimator module

To optimise the parameters for this module, firstly we need to estimate the intensity for each of our facial AUs. For a sample i in our dataset, let $r\theta_i^j$ be the intensity estimation of our module for AU j for the right side of the face and $l\theta_i^j$ be the intensity estimation for the left side of the face. The final intensity estimation for AU j of sample i is aggregated as follows:

$$\theta_i^j = (1 - p_i^r) \times l\theta_i^j + p_i^r \times r\theta_i^j \tag{4.6}$$

Where the parameter p_i^r denotes the face side visibility percentage. Since our network is being trained using two stages strategy, we use the generated ground truth p_i^r to aggregate rθ_i^j and lθ_i^j in training phase. For the testing phase, we use the predicted y_i^r of our *Face side visibility* module instead. From this equation, we can see that our network tries to weight the importance of the left compare to the right face side prediction by using the face side visibility factor p_i^r. This way, we are able to take into account both side of the face, as well as eliminate the problem of the over turning head pose cases like in Figure 4.9b, since the invisible or hard to see parts of the face will have a low weight compared to the visible parts.

Regarding the objective loss function for this module, we utilise the average L_1 distance between the prediction θ_i^j and ground truth $\hat{y}_i^j$ for each of our AUs as follows:

$$L_i^{aus} = \frac{1}{T} \sum_{j=0}^{T} \|\theta_i^j - \hat{y}_i^j\| \tag{4.7}$$

Where T denotes the total number AUs in our dataset.

4.4.3.3 Facial region localisation module

For training our *Facial region localisation* module, we adopt the same two-stage training strategy as [RHGS15, HGDG17]. For each training sample i, we minimise the RPN loss L_i^{RPN} at first stage and both the classification loss L_i^{cls} and bounding-box loss L_i^{box} in parallel at the second stage:

$$L_i^{detection} = L_i^{RPN} + L_i^{cls} + L_i^{box} \tag{4.8}$$

Where the definition of L_i^{RPN}, L_i^{cls} and L_i^{box} loss functions are identical to [RHGS15, HGDG17].

4.4.3.4 Final objective loss function

Finally, the objective loss function for optimising the whole of our network is defined as the sum of the objective loss functions of all of our modules, as follows:

$$L = \frac{1}{N}\sum_{i=0}^{N}\left(L_i^{detection} + L_i^{aus} + L_i^{r}\right) \tag{4.9}$$

Where N denotes the total number of samples.

4.5 Experiments and Results

4.5.1 Implementation details

The whole network system is implemented using PyTorch framework [PGM+19]. During the training phase, Adam optimiser [KB14] was employed with the initial learning rate set to $4e^{-5}$ for the DISFA database and $3e^{-4}$ for the UNBC database. The training batch size set to 42 images per batch based on our preliminary testing. The training and validating processes were performed on an Intel Workstation machine with a NVIDIA Gerforce RTX 2080 Ti 11G GPU.

Regarding model parameters initialisation, we have initialised our model with a ResNet model pretrained [VBA21] on a database called Aff-Wild2 [KSHZ, KSZ21, KZ21, KZ19b, KSZ19, KTN+19, ZKN+17, KKHZ21], which is a large in-the-wild database for emotion recognition.

For data augmentation, we also have used Image Translation, Horizontal Flip and ColorJitter, as we have mentioned in our previous approach, which is described in Section 3.5.2.

In order to train our *Movement exploitation* module, we have applied two steps training strategy for training our network. In the first step, we train only the CNN part of the network using independent images. For the second step, we freeze the CNN part that has been trained in the first step and train only the *Movement exploitation* module using sequence frame images. The reason for that is because the likeness nature of sequence images. Since the images in a sequence are mostly look-alike each other, feeding them directly into a CNN layer will make the network easily to be overfitted. Although previously we have mentioned about under-sampling and over-sampling techniques for rebalancing the dataset. However, it is mainly to improve the performance of the CNN part of the network when training with independent images. For the RNN part, we still have to rely on the mentioned two steps training to avoid the problem of overfitting.

To evaluate our approach and compare with other articles' results, for the DISFA database, we report the 3-fold subject independent cross-validation results. For the UNBC McMaster database, we report the leave-one-subject-out cross-validation results.

4.5.2 Evaluation metrics

To compare the performance of our method within author scheme and with the SOTA approaches, we use ICC and MAE for both the two DISFA and UNBC McMaster databases. We exclusively compute MSE and PCC for the UNBC McMaster database to be inline with other pain-related works. Since our approach is including the object detection task for detecting the facial regions, therefore we also report

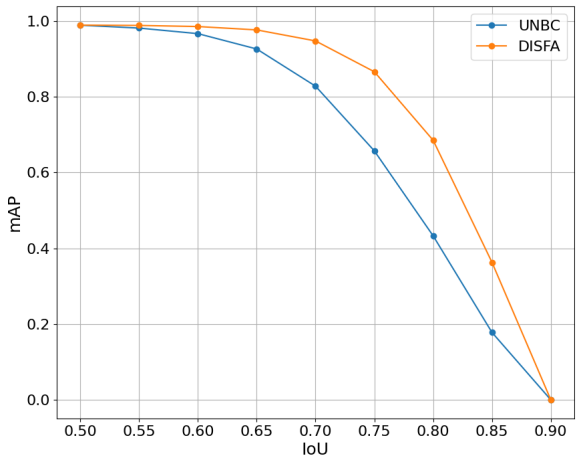


Figure 4.11: mAP vs. IoU overlap ratio on the UNBC and DISFA databases

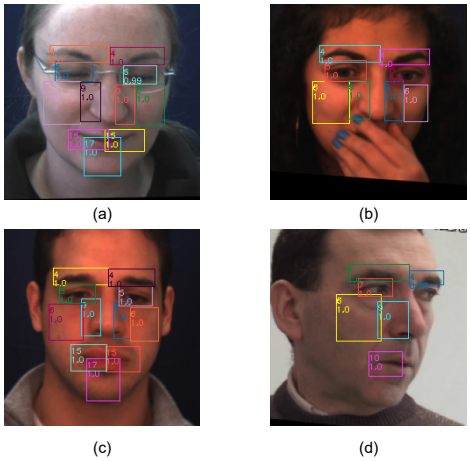


Figure 4.12: Prediction samples of our *Face region localisation* module. It can be seen that our network is able to predicted correctly both normal cases (a, c) and special cases (b, d).

the mean Average Precison (mAP) for different IoU including mAP@.5, mAP@.75 and mAP@[.5, .95], which are standard COCO evaluation metrics for object detection task [LMB+14, RHGS15, HGDG17].

4.5.3 Evaluation results

In this section, we provide the experimental results for evaluating the performance of our proposing approach within the author's scheme. We show the effectiveness and also the impact of each module to the final network.

4.5.3.1 Facial region localisation results

As we have mentioned earlier, our approach follows the *divide-and-conquer* paradigm, therefore it is crucial to verify the correctness of the dividing part. To evaluate the performance of this module, we report the average mAP of our approach on both the DISFA and UNBC databases, as shows in Table 4.3. From this table, we can see that at $IoU = 0.5$, our model have reached $\approx 99\%$ of mAP on the DISFA database and $\approx 98\%$ of mAP on the UNBC database, which basically means that our model have predicted

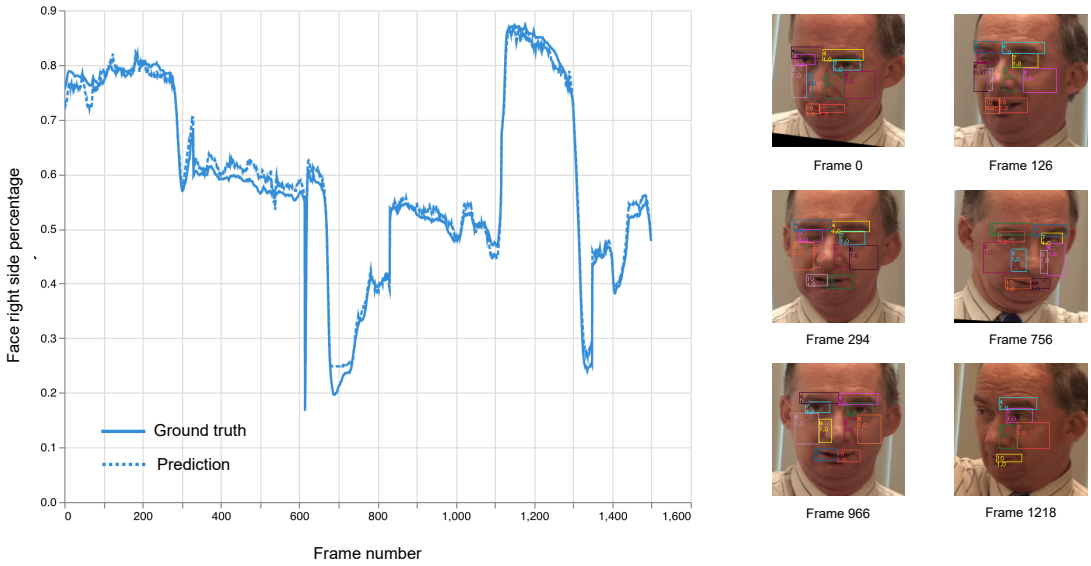


Figure 4.13: Face right side percentage prediction and its corresponding ground-truth of a subject in UNBC McMaster dataset.

almost perfectively at this IoU level. On top of that, our network is still be able to reach $mAP > 80\%$ at $IoU = 0.70$ as can be seen in Figure 4.11. This is an important information which shows that our network is capable of localising facial regions and is reliable for the next phase of our network to work on the predicted regions of this module.

Table 4.3: Performance of our *Facial region localisation* module on the two DISFA and UNBC McMaster databases.

Database	mAP@.5	mAP@.75	mAP@[.5, .95]
DISFA	98.7	84.7	67.6
UNBC	97.8	62.0	57.3

Moving on to the mAP results at $IoU = 0.75$ we can see that our model is still quite confident with $mAP \approx 85\%$ on the DISFA database and drop slightly to 62% of mAP on the UNBC McMaster database. The reason for this dropping could be because the overtunning head pose case happens quite a lot in the latter compared to the former database. The overall mAP@[.5, .95] results of the DISFA and UNBC database shows once again that our *Facial region localisation* module is working effectively. Figure 4.12

shows some visualisation examples of the predictions of our network. It can be seen that our network has correctly localised the facial regions in both normal cases when the left and right parts of the face are roughly balanced and hard cases when the patient cover a part of their face (b) or when they overturn their head pose (d).

4.5.3.2 Face side visibility results

Table 4.4: Performance of our *Face side visibility* module on the two DISFA and UNBC McMaster databases.

Database	MAE	PCC
DISFA	0.04	0.62
UNBC McMaster	0.04	0.94

For evaluating the correctness of our network in estimating face right side percentage, we report the cross-validation MAE and PCC of both the UNBC McMaster and DISFA databases as shows in Table 4.4. From this table we can see that our network has reached $MAE = 0.04$ for both the two databases which indicates that on average the difference between our prediction and ground truth face right side visibility percentage is 0.04. Since the ground truth and prediction are both percentages, we can see it as $\approx 4\%$ error in prediction, which is an acceptable error threshold and shows the reliability of the module.

Regarding the PCC results, we can see that our network has reached the correlation of 0.94 on the UNBC McMaster database and 0.62 on the DISFA database. The reason for this difference in correlation results could be due to the nature of each of these two databases. As the participants in the DISFA database were watching a video to elicit spontaneous AUs, their faces tended not to move too much compared to the UNBC McMaster database when the patients were performing different movements on their arms to elicit the pain emotion. The difference in variations and frequencies of changing in the face side visibility in each of these databases appears to be the reason for the difference in the ICC correlation results. All in all, as the average difference in term of

Table 4.5: Comparison of the rebalancing techniques on the DISFA database. The final balancing is about applying both under-sampling and over-sampling techniques that we have proposed.

	AU	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC	No balancing	**.63**	.59	.69	.36	.51	.47	.83	.00	.29	.00	**.93**	.67	.50
	Only under-sampling	.61	.69	.71	.36	.42	**.57**	.80	.00	.46	.00	.92	**.68**	.52
	Final rebalancing	.60	**.70**	**.74**	**.72**	**.53**	.53	**.86**	**.27**	**.56**	**.34**	**.93**	.64	**.62**
MAE	No balancing	**.19**	**.14**	**.34**	**.04**	.30	.17	.25	**.09**	.21	**.07**	**.23**	.26	**.19**
	Only under-sampling	**.19**	**.14**	.38	**.04**	.33	**.13**	.25	**.09**	.22	**.07**	.27	**.22**	**.19**
	Final rebalancing	.20	.15	**.34**	**.04**	**.29**	.15	**.22**	.10	**.18**	.08	.25	.26	**.19**

MAE is only 4%, the changing in term of ICC correlation between these two databases seems acceptable. Figure 4.13 shows a visualisation of our network for a subject in the UNBC McMaster database. From this figure, we can see that our network has correctly predicted the face right side visibility percentage, even in hard cases like overturning left or overturning right.

4.5.3.3 AU intensity estimator results

Re-balancing dataset techniques evaluation: For evaluating the performance of our network in AU intensity estimation as well as evaluating the use of different dataset re-balancing techniques that we have mentioned earlier, we report the ICC and MAE results for each of the AUs as shows in Table 4.5. From this table, we can see that in term of MAE, the changing in performance when applying different data balancing techniques is not significant. However, in term of ICC correlation results, we can see that there are some different results when appling different data re-balancing techniques. Specifically, we have achieved the correlation of 0.5 when not applying any data re-balancing techniques. This correlation has improved to 0.52 when applying our under-sampling data that we explained in section 4.3, which is about 2% of improvement compared to the model trained without data re-balancing. This result suggests that the under-sampling technique that we applied reduces the imbalancing of the dataset. Yet, since the dataset is ways too imbalanced (see Figure 4.2), the improvement when appling our under-sampling technique is not significant. However, when we applied

Table 4.6: Performance comparison of our network on the DISFA database when training with and without the *Per region RoI pooling* layer.

	AU	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC	RoIAlign pooling	.53	.60	.73	.66	**.56**	**.56**	.85	**.30**	.50	.29	**.93**	**.66**	.60
	Per-reg. RoI pooling	**.60**	**.70**	**.74**	**.72**	.53	.53	**.86**	.27	**.56**	**.34**	**.93**	.64	**.62**
MAE	RoIAlign pooling	.24	.21	.35	**.04**	**.27**	**.15**	.24	.11	.25	**.08**	.27	.27	.21
	Per-reg. RoI pooling	**.20**	**.15**	**.34**	**.04**	.29	**.15**	**.22**	**.10**	**.18**	**.08**	**.25**	**.26**	**.19**

both under-sampling and over-sampling techniques, we have reached the correlation of 0.62, which is about 12% of improvement compared to the model trained without any data re-balancing. This result confirms the effectiveness of our proposed data-rebalancing technique in improving performance of the model on a highly imbalanced dataset.

Per region RoI pooling layer evaluation: For evaluating the effectiveness of the *Per region RoI pooling layer*, we report the performance of our network when training with and without this layer as in Table 4.6. From this table, we can see that the layer improved the performance of our network in term of both ICC and MAE evaluations, especially for the case of average MAE result with about 10% of improvement. This result clearly shows the effectiveness of using the right RoI pooling size for each of the facial regions compared to the one which uses the same pooling size for all regions. The drawback of this approach is the fact that we need to know the average shape of the region in advance, in order to choose the RoI pooling size accordingly. It works in the case of face analysis because we know the average shape of each of our facial regions when they are visible. In the case of invisible or incomplete facial regions due to head pose, we reduce the weight of these cases when optimising network parameters using the *Face side visibility module*.

Movement exploitation module: To evaluate the effectiveness of the module in exploiting temporal dynamic information, we report the performance of our FFAU network when training with and without this module, as well as at different sequence length con-

Table 4.7: Performance of our *Movement exploitation* module on the DISFA database.

Model	Seq. lengh	Avg MAE	Avg ICC(3,1)
FFAU	-	**.19**	**.62**
FFAU-LSTM	8	.20	.61
FFAU-LSTM	16	.20	**.62**
FFAU-LSTM	32	.20	**.62**
FFAU-LSTM	64	.20	**.62**

figurations as shown in Table 4.7. From this table, we can see that there are not much difference in performance when integrating the *Movement exploitation* module into our FFAU network. Specifically, there is no improvement in term of average ICC and also there is a slightly decrease in performance in term of MAE (5%). The average ICC of our model when trained with the sequence length of 8 frame images also seems to decrease slightly compared to the other configurations, which is probably due to the initialisation of the model parameters. From these results, we can conclude that our *movement exploitation* module could not effectively exploit the temporal dynamic information from sequence of frame images. The main reason for that seems to be related to the design of our FFAU neural network. As the AU regions proposing by our *facial region localisation* module can be of different sizes and shapes, features extracted from them consist quite a lot of spatial-feature variations. While having spatial-feature variations are great for CNN network to improve its generalisability, it is hard for LSTM network to model the changing of a certain feature(s) over time [BR19]. One way to solve this problem is to eliminate these spatial-feature variations out of the sequence features before feeding into the LSTM network. Another approach is to improve the LSTM module to learn to ignore these spatial-feature variations. However, due to the limited duration of the thesis, this will be an open research direction to further improve the overall performance of the network.

Table 4.8: Comparison to the SOTA AU intensity estimation methods on the DISFA database using 3-fold cross validation. Numbers in bold denote the best performance.

	AU	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC(3,1)	BORMIR[ZZD+18]	.20	.25	.30	.17	.39	.18	.58	.16	.23	.09	.71	.15	.28
	CCNN-IT[WOR+17]	.20	.12	.46	.08	.48	.44	.73	.29	.45	.21	.60	.46	.38
	KJRE[ZWD+19]	.27	.35	.25	.33	.51	.31	.67	.14	.17	.20	.74	.25	.35
	KBSS[ZDHJ18]	.23	.11	.48	.25	.50	.25	.71	.22	.25	.06	.83	.41	.36
	CFLF[ZJW+19]	.26	.19	.46	.35	.52	.36	.71	.18	.34	.21	.81	.51	.41
	2DC[LTWE+17]	.70	.55	.69	.05	.59	**.57**	**.88**	**.32**	.10	.08	.90	.50	.50
	SCC[FLL20]	**.73**	.44	.74	.06	.27	.51	.71	.04	.37	.04	.94	**.78**	.47
	DPG[SCW+21]	.46	.46	**.75**	.63	**.61**	.48	.84	.29	.44	.18	**.95**	.63	.56
	FFAU (ours)	.60	**.70**	.74	**.72**	.53	.53	.86	.27	**.56**	**.34**	.93	.64	**.62**
MAE	BORMIR[ZZD+18]	.88	.78	1.24	.59	.77	.78	.76	.56	.72	.63	.90	.88	.79
	CCNN-IT[WOR+17]	.73	.72	1.03	.21	.72	.51	.72	.43	.50	.44	1.16	.79	.66
	KJRE[ZWD+19]	1.02	.92	1.86	.70	.79	.87	.77	.60	.80	.72	.96	.94	.91
	KBSS[ZDHJ18]	.48	.49	.57	.08	.26	.22	.33	.15	.44	.22	.43	.36	.33
	CFLF[ZJW+19]	.33	.28	.61	.13	.35	.28	.42	.18	.29	.16	.53	.40	.33
	2DC[LTWE+17]	.32	.39	.53	.26	.43	.30	.25	.27	.61	.18	.37	.55	.37
	SCC[FLL20]	**.16**	.16	**.27**	**.03**	**.25**	**.13**	.32	.15	.20	.09	.30	.32	.20
	DPG[SCW+21]	.29	.26	.39	**.03**	.27	.14	.27	**.10**	.25	.11	**.24**	.34	.22
	FFAU (ours)	.20	**.15**	.34	.04	.29	.15	**.22**	**.10**	**.18**	**.08**	.25	**.26**	**.19**

4.6 Comparison with State of the art

In this section, we compare the performance of our approach with SOTA approaches on the two domains including facial AU intensity estimation and pain intensity estimation.

4.6.1 Facial action unit intensity estimation

We compared our proposed FFAU model with other works that related to facial AU intensity estimation on the DISFA dataset. CCNN-IT [WOR+17], KBSS [ZDHJ18] are deep networks that leverage structural or dynamic information. KJRE [ZWD+19], BORMIR [ZZD+18] and KBSS [ZDHJ18] combine prior knowledge or semantic information for facial AU intensity estimation. 2DC [LTWE+17] is another model that combines the deep model and probabilistic model. CFLF [ZJW+19] is an approach that tries to utilise spatial relationships among AUs. SCC [FLL20] and DPG [SCW+21] are deep methods

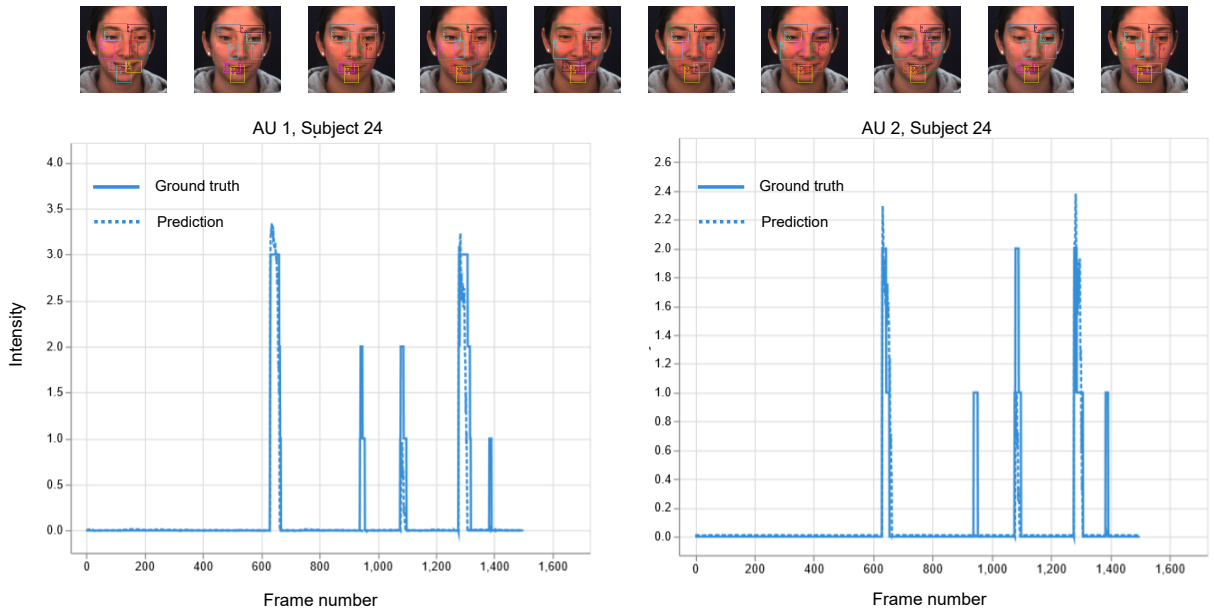


Figure 4.14: An example of the predicted results for AU1 and AU2 of the same subject compared to the corresponding ground-truth from the DISFA dataset.

that utilising graph network for AU intensity estimation. Table 4.8 shows the comparative results for the above mentioned methods evaluated on the DISFA dataset. From this table, we observe that the proposed FFAU network outperforms all other approaches on average with higher ICC and lower MAE. Specifically, the average MAE of our approach shows a decrease of $5\% \sim 79\%$ compared to the other mentioned approaches. On the other hand, in term of ICC correlation, the average ICC of our approach shows an increase of $6\% \sim 34\%$ compared to the other mentioned SOTA approaches. These results suggest the advantages of our approach in dividing the facial regions and conquering their intensity estimation. An example of the prediction of our method can be seen in Figure 4.14. We can see that for both two AUs (AU 1 and AU 2) in the figure, our network have predicted the AU intensity quite close to the ground truth of the dataset. To the best of our knowledge, the proposed FFAU network have achieved the best performance with the highest average ICC, as well as the lowest average MAE, for the DISFA database.

4.6.2 Pain intensity estimation

Next, for evaluating our approach and comparing with other works in the pain estimation domain, we train our network to predict 6 facial action units intensity that related to pain, which are including: AU4, AU6, AU7, AU9, AU10 and AU43. Afterwards, according to Prkachin and Solomon [PS08], we aggregate these AUs to compute PSPI pain intensity level using the Equation 1.1 (see Chapter 1.5.2). As our network only works with facial AU intensity estimation, we can turn our network to a pain intensity estimation network by applying Equation 1.1. Hence, we are able to compare our work with other SOTA methods in pain estimation domain.

4.6.2.1 The MSE scale issue in pain domain literature

One problem that we found in reviewing the literature in the domain of PSPI estimation is the misalignment of the PSPI intensity scales. The works [KRP12, FFV14, ZHSZ16, TH18a, VBADE21, HQX+21] use 16 discrete pain intensity levels $[0-16]$ (group A) while the works [RPP13, ZGWJ16, RCG+17, WXL+17] use the aggregated 6 discrete pain intensity levels $[0-5]$ (group B). Therefore, putting these two groups into a single SOTA leaderboard as in [HQX+21, RCG+17, VBADE21, TH18a] could cause a misleading to the readers since the two ranges are not the same. As the MSE metric that we have been used to evaluate these approaches is extremely sensitive to the outlier by definition, which penalises too much the evaluation results of group A than group B, leading to a better results in group A compared to group B. To solve this issue, we propose to use two different SOTA leaderboards for the two PSPI intensity scales. We also report the results of our approach on both of these two leaderboards for comparison.

Note that we exclude the works that neither follow the two mentioned PSPI scales nor use the same training or evaluating data, i.e. Bargshady et al. [BZD+20] and Xin et al. [XLY+21] since they are using 4-level PSPI scale, Hoang et al [HXMF20] since they have selected data from 19 of 25 subjects in the dataset (the evaluating set when applying leave-one-subject-out cross-validating is no longer the same as other works).

Table 4.9: PSPI 16-level comparison against Leave-One-Subject-Out method with MSE, MAE, PCC, and ICC on the UNBC McMaster database. The best results are shown in bold.

Model	MSE	MAE	PCC	ICC
Kaltwang et al. [KRP12]	1.39	-	0.59	0.50
Florea et al. [FFV14]	1.21	-	0.53	-
Zhou et al. [ZHSZ16]	1.54	-	0.64	-
Huang et al. [HQX+21]	0.76	0.40	**0.82**	-
Tavakolian et al. [TH18a]	0.69	-	0.81	-
3Stages (ours)	0.60	**0.35**	**0.82**	0.80
FFAU (ours)	**0.56**	0.38	**0.82**	**0.81**

4.6.2.2 State of the art 16-level PSPI estimation

We compare our proposed FFAU network to other deep networks that related to the domain of estimating PSPI intensity with 16 levels $[0-16]$, in which including Kaltwang et al [KRP12] with their shape and appearance features fusion network, Florea et al. [FFV14] with HoT and SVM classifier, Zhou et al. [ZHSZ16] with Recurrent Convolution Neural Network, Hoang et al. [HQX+21] with a hybrid network, Tavakolian et al. [TH18a] with Deep Binary Representation network, and our 3Stages training model that we have mentioned in the previous chapter (see Chapter 3). Table 4.9 shows the evaluation results of leave-one-subject-out cross-validation method for the above mentioned approaches evaluated on the UNBC McMaster database. At the first sign, it can be seen that our two approaches (FFAU and 3Stages) outperformed all the other SOTA approaches in all of the four evaluation metrics (MSE, MAE, PCC, ICC). Since we have already done the comparison between our 3Stages model and the previous SOTA approaches (see Section 3.5.6), we will compare the evaluation results between our 3Stages and our FFAU approaches in the next paragraph.

From table 4.9, we that our FFAU approach is 7% lower in term of MSE but 9% higher in term of MAE compared to our 3Stages model. Since the MSE score is much more sensitive to the outliers compared to MAE by definition, these results show that this FFAU approach predicts the PSPI with high intensity levels better than the other method. In term of ICC and PCC correlation, the predictions of our approach are slightly

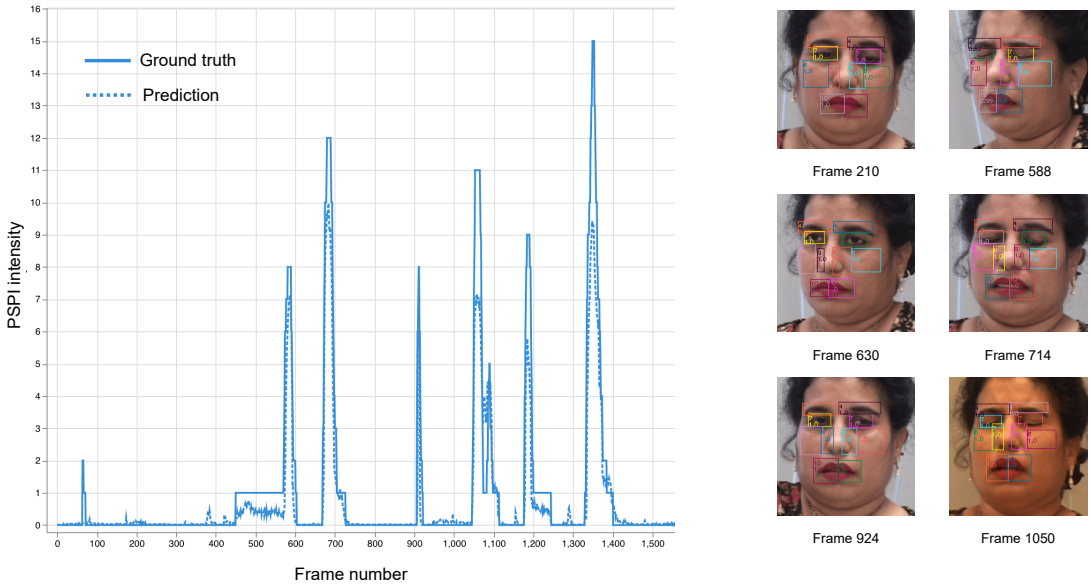


Figure 4.15: An example of the PSPI intensity prediction compared to the corresponding ground-truth from the UNBC McMaster dataset.

Table 4.10: PSPI 6-level comparison against Leave-One-Subject-Out method with MSE, MAE, PCC, and ICC on the UNBC McMaster database. The best results are shown in bold.

Model	MSE	MAE	PCC	ICC
Rudovic et al. [RPP13]	-	0.80	-	0.70
Zhao et al. [ZGWJ16]	-	0.81	0.60	0.56
Wang et al. [WXL+17]	0.80	0.46	0.65	-
Rodriguez et al. [RCG+17]	0.74	0.50	**0.78**	0.45
FFAU (ours)	**0.43**	**0.35**	**0.78**	**0.78**

better than those of our 3Stages model, but not significant. All in all, we can argue that our FFAU approach is slightly better than our 3Stages approach in term of 16 levels PSPI intensity estimation. Figure 4.15 shows an example of our network prediction versus ground truth of a subject in the UNBC McMaster database. It can be seen that our network has predicted quite correctly for both low and high intensity levels, which demonstrates the capability in pain intensity estimation of our FFAU network.

4.6.2.3 State of the art 6-level PSPI estimation

We compare the performance of our approach with other SOTA approaches that use PSPI 6 levels, which are including Rudovic et al. [RPP13] with their Conditional Ordinal Random Field (CORF) model, Zhao et al. [ZGWJ16] with OSVR regression model, Wang et al. [WXL+17] with a regularized deep neural network and Rodriguez et al. [RCG+17] with their VGG + LSTM hybrid network. In order to be inline with these works, we aggregate the results of our PSPI prediction using the same logic as in [RPP13, ZGWJ16, RCG+17, WXL+17]. Specifically, pain intensity levels 0, 1, 2 and 3 are kept the same. Pain levels 4, 5 are merged and pain levels $6+$ become $5th$ level. Table 4.10 shows the comparison results of our approach compared to the mentioned SOTA approaches. From this table, we can see that there are a large difference between our 16-level PSPI results (Table 4.9) and our 6-level PSPI results, despite being derived from the same set of prediction. Specifically, the 6-level PSPI is 23% lower in term of MSE and 8% lower in term of MAE compare to our 16-level PSPI results. These results again confirm the problem we have raised earlier, which is the fact that we can't compare the evaluation results of the two different prediction scales together, and there is a need of using different SOTA leaderboard for each scale.

Regarding the comparison within SOTA approaches for 6-levels PSPI evaluation, from Table 4.10 we can see that our network has outperformed all other approaches in almost all the evaluation metrics. Specifically, in term of MSE, our method has achieved 42% of improvement compared to previous SOTA approach [RCG+17] and in term of MAE, our method has reached 24% of improvement compared to previous SOTA approach [WXL+17]. In term of correlation, our approach has the same PCC with SOTA approach [RCG+17] but in term of ICC, our approach is 8% higher than previous SOTA approach [RPP13]. To the best of our knowledge, the proposed FFAU network have achieved the best performance with the highest average ICC and average PCC, as well as lowest MSE and MAE, for the UNBC McMaster database using PSPI 6-level estimation protocol.

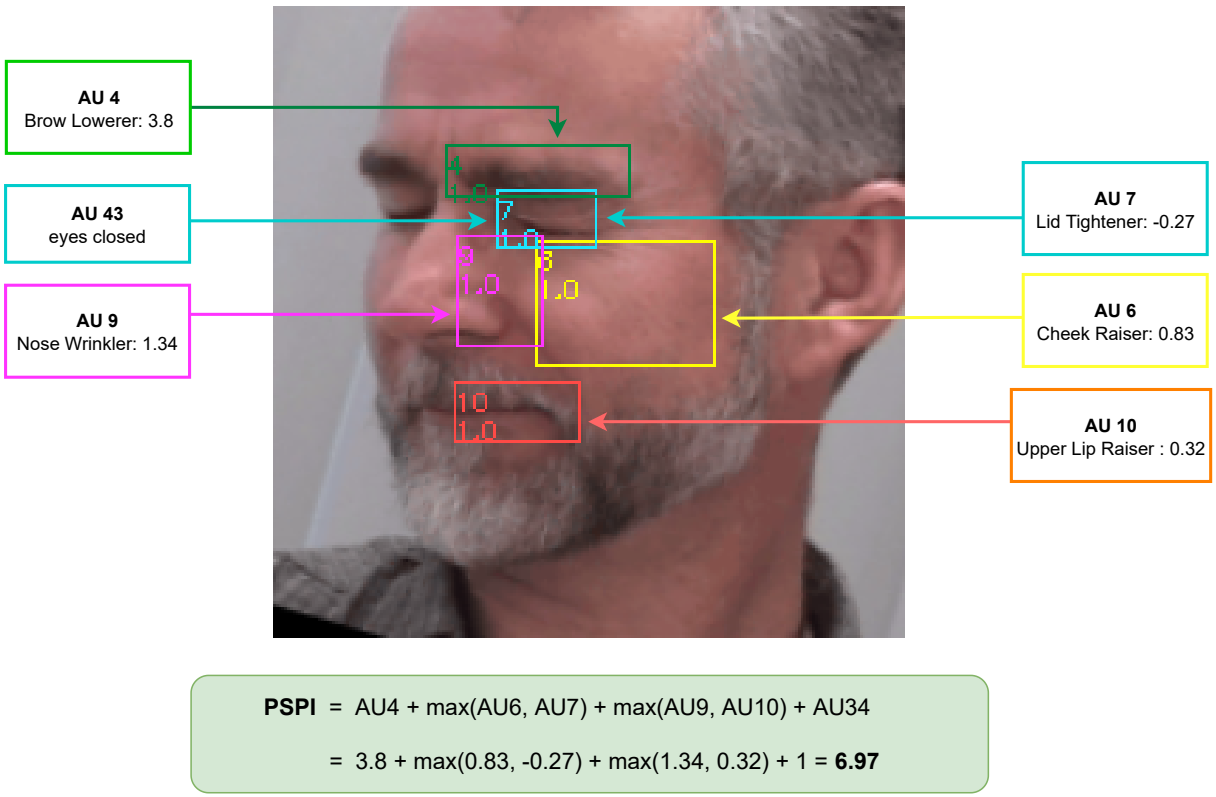


Figure 4.16: The visualisation of the PSPI intensity prediction by our network. We can see that our network is able to explain why it gives a pain level for an image by saying where it got the pain-related AU from and what score it gave to the AU.

4.7 Towards explainable PSPI pain assessment

One of the main drawbacks of deep learning methods is the lack of ability to explain why the network makes a particular decision, which is due to the black-box nature of deep learning algorithms. The end-to-end learning paradigm hides the entire decision process behind the complicated inner-workings of deep learning models, making its decisions less understandable to humans and prohibits their use in safety-critical applications [CL18]. In the domain of health care and medical fields, automatic system to complement medical professionals such as pain measurement system should have a certain amount of explainability and allow the human expert to retrace the decisions and use their judgment. However, there is still a huge gap between explainability and accuracy in the domain of PSPI pain intensity estimation, as SOTA approaches in this

domain only try to improve the PSPI score evaluation results, without any effort of telling why their model makes such a decision.

In this work, we take a step towards better explainability of deep neural network in PSPI pain intensity estimation. As our approach tries to isolate each pain-related AU before estimate its intensities, we can tell explicitly where are the pain-related AUs in the face images and which score we have estimated for each of these AUs. Moreover, as we have integrated the *Face side visibility* module to our FFAU network, we can also tell which face side that we have focused the most from the face image. Figure 4.16 shows an example of the prediction of our network. For each face image, our network is able to effectively localise the region of each facial AU and estimate its intensity. From these prediction results, professionals can see if the AU regions are located correctly or not. They can also be able to evaluate the intensity for each of these AU to ensure the correctness of our pain-related AU intensity prediction. Compared to SOTA approaches that only return PSPI score, our network provides much richer information to interpret, making it easier for practitioner or medical doctor to see and evaluate the correctness of our prediction.

While having many improvements in explainability of the model behaviour compared to SOTA approaches in PSPI pain intensity estimation, there are still some limitations that need to be addressed in future works. i.e., the ability to explain why the model assigns an AU to a particular region in the face image or how does it predict an intensity score for a region of the AU. These are some important questions which help to better understanding model's behaviour and prove its reliability. All in all, despite having these limitations, our approach still provides a great explainability in PSPI pain intensity estimation compared to SOTA approaches.

4.8 Conclusion

The breakthrough success of deep learning is mainly due to the availability of large-scale labelled datasets. However, large-scale labelled datasets are not always available

in some domains. Facial action unit and pain intensity estimation are among those domains that suffer from lacking of labelled training data. For these face-related domains, it requires a costly and time-consuming labeling effort by trained human annotators to be able to construct a dataset. Moreover, the work of collecting images and videos of faces is also difficult due to many reasons such as ethical, privacy, cultural variations, etc. Therefore, it is crucial to design learning techniques that can learn to extract correct feature representations from face image with a limited amount of labelled data and that are able to generalise for predicting on the newly unseen data.

The main reason why deep learning requires to have massive amout of training data is to learn to extract important features from images through the gradient descent algorithm. Then, one way to help the network to learn better feature representations from small dataset is to integrate the information regarding where to find these important features from the image. In this work, we have introduced a new approach called learning to isolate regions-of-interest for better extracting feature representations. Based on the concept of *divide and conquer* paradigm, our approach firstly tries to localise and isolate the regions-of-interest (*divide*) by utilising object detection network, then we estimate the AU intensity for each of the isolated regions (*conquer*), accordingly. This way, we not only tell the network where to find the regions that contains important information, but also isolate these regions for better feature extraction. Besides extracting features, we also introduced a module to evaluate the face side visibility, allowing our network to take into account the correct face side in case of head posing. Experiments on the two widely known databases UNBC McMaster and DISFA show that our approach outperforms other SOTA apporaches by a large margin on both the two databases. Furthermore, in term of explainability, our proposed approach provides much better view of the prediction results, especially in term of PSPI pain intensity estimation, compared to SOTA approaches. This result demonstrates the effectiveness of our approach in learning to extract correct feature representations from face image. On top of that, as our approach is about measuring facial AU intensity, which is the most basic building block to describe facial expressions. Hence, our approach can be adopted to measure

any higher order facial expression representation.

Chapter 5

Discussion and conclusion

Contents

In this work, we gave an overview on measuring facial expressions by utilising facial action units, with an application on automatic PSPI pain intensity estimation. As any human facial expression can be decomposed into a set of facial action units and their intensities, automatic measuring facial action unit intensity seems to be the key step towards better understanding human facial expression and assessment. In this chapter, we summarise all the findings and contributions that we have proposed. In addition, we discuss the obstacles to automatic facial expressions assessment and present future research challenges.

5.1 Contributions of the thesis

Facial action units are the most basic building blocks for facial expression assessment since they describe human facial muscle movements precisely. In this thesis, we addressed the problem of facial action unit intensity estimation, by proposing learning methods to focus and isolate regions-of-interest for better extracting feature representations. Additionally, we adopted the proposing approaches for the application of PSPI pain intensity estimation. The main contribution of this thesis is two fold, i.e., (1) learning to focus on regions-of-interest, and (2) learning to isolate regions-of-interest.

5.1.1 Learning to focus on regions-of-interest

Deep learning methods have achieved great success in learning visual representations thanks to the availability of large-scale labelled datasets. However, large-scale labelled dataset is not always available in some domains, especially in the domains of facial AUs and PSPI intensity estimation due to costly and time-consuming labeling effort by trained annotator. Hence, there is a need of developing a learning approach which is capable of learning to exploit correct feature representations from a limited amount of data. In order to tackle this problem, we have proposed an approach of learning to focus on regions-of-interest for better extracting feature representations. As deep learning

in general requires massive amount of training data to learn to extract correct features from images, therefore if we can tell the neural network where are the important places to focus in the image, it will ease the training process and improve the generalisability of the network. Based on that idea, our approach first tries to combine multi-database together and then trains a CNN network as heatmap regression for estimating facial AU intensity. This heatmap regression plays the role of guiding our network to focus on our predefined pain-related AU regions (regions-of-interest), which helps the network to learn to extract feature representations from the correct regions in the face image. Next, we utilise the knowledge that has been learned on the multi-database combination for feature extraction and PSPI pain intensity estimation, showing some great improvements compared to the SOTA approaches on the same domain. From the experimental results, we emphasised the importance of learning to focus on regions-of-interest for better extracting feature representations and reducing the effect of overfitting when training on a limited amount of data.

5.1.2 Learning to isolate regions-of-interest

Lacking of large-scale labelled training data seems to be the major issue in development of machine learning approaches in many domains, including facial AUs and PSPI pain intensity estimation. To effectively learn to extract correct feature representations from limited amount of data, we have proposed an approach of learning to focus on regions-of-interest, which have significantly improved the performance of the network compared to SOTA approaches. However, this approach still have some limitations as it does not take into account the head pose issue and the predefined heatmap emphasises too much the central location of AU region. Hence, we have extended the previous work and proposed a new approach called learning to isolate regions-of-interest. With this approach, we are not just focusing but isolating the regions-of-interest for better extracting feature representations. Besides extracting features, our network also measures the percentage of face side visibility and incorporates this factor into the equation

of AU intensity estimation, thus solving the head pose problem. Experiments on the two widely known DISFA and UNBC McMaster databases show that our proposing approach outperforms SOTA approaches by a large margin on both the two databases, which once again confirms the effectiveness of our learning to isolate regions-of-interest approach when training on a limited amount of data. Besides demonstrating the improvements in performance, we also have shown that our approach also provides a great level of explainability, especially in term of PSPI intensity estimation, compared to the SOTA approaches on the same domain. As more and more deep learning techniques are involved in human life, the ability to explain the outcome of the model's prediction is attracting more and more attention as a way to better understand the behavior of the model and prove its reliability.

5.2 Opening challenges

Despite having many promising advancements in automatic facial AUs and PSPI pain intensity estimation, there are still a number of challenges to be addressed for developing reliable and applicable methods for measuring human facial expression. These challenges can be categorised to three groups, including dataset (section 5.2.1), method (section 5.2.2) and computational efficiency (section 5.2.3). In the following, we list the existing challenges and discuss the potential solutions as the future work.

5.2.1 Facial action unit intensity dataset

Well-labelled large-scale dataset is crucial for developing automatic facial AU intensity estimation systems and proving their usefulness. For the field of deep learning and machine learning, it is even more critical to have more high-quality data due to the data-hungry nature of these learning algorithms. This is one of the core challenges in the domains of facial AUs and PSPI pain intensity estimation, since the amount of well-labelled, publicly available data in these domains are still limited. The main reason is

because it requires a costly and time-consuming labeling effort by trained human annotators. For instance, it may take more than an hour for an expert annotator to code the intensity of AUs in one second of a face video [LTWE+17]. Furthermore, AUs intensity coding requires profound knowledge of the FACS and additional training by FACS experts to be able to correctly label data. Approximately, it requires about 100 hours of time involved in this FACS training [Prk09] for a single FACS coder. In constructing a FACS based dataset, it requires to have at least two (or more) FACS coders to ensure the correctness and consistency of the dataset. Therefore, it is challenging to obtain a large-scale high-quality FACS annotated dataset. Besides labeling data, the work of collecting images and videos of faces is also challenging due to many reasons such as ethical, privacy, cultural variations, etc. These face images and videos, when collected, must also include people from different countries and cultures, of different ages and genders to ensure the coverage of a wide range of different facial traits in the dataset. Lacking of any of these facial traits when training a learning model could lead to the issue of unable to predict well in the real-life cases, e.g. models trained only on young faces do not generalise well to older faces due to the textural differences caused by ageing and variations in facial muscle elasticity and facial dynamics [Has17]. Hence, these facial traits play a vital role in providing a large amount of facial expression variations to improve the generalisability of learning methods. In addition to the covering of human facial traits, the environmental variations are also one another important factor which could impact the performance of automatic facial AU measurement [WRGEP20]. As existing facial AU intensity datasets were created under controlled conditions, they do not sufficiently cover environmental variations such as lighting conditions, backgrounds, pose changes, and occlusion. Hence, it is important to consider such environmental variations during the data collection process to increase the robustness of automatic pain assessment methods. All in all, it is a great challenge in constructing a well-labelled large-scale dataset which covers a wide range of ethnicities, ages, genders, and environmental variations. In the end, the release of such dataset would be of great help in improving the robustness of automatic facial expression assessment methods.

5.2.2 Multi-modal expressions assessment

Since any human facial expression can be decomposed into a set of facial action units and their intensities, automatic facial AUs intensity estimation appears to be a great way to measure facial expressions. The intensity estimation of these facial AUs can also be used to measure any higher order facial expression representations, e.g. Pain expression. Therefore, constructing deep learning model for estimating facial AUs automatically is a key step towards better understanding human expression and assessment. However, facial AU is only one source of information and if we rely entirely on this source alone, there will certainly be a time when something goes wrong, e.g. the person cover entirely their face due to experiencing a high level of emotion. Therefore, a learning approach that consists of multiple models for exploiting different sources of emotion seems to be an optimal choice for a better and stable automatic human expression measurement. Regarding the sources of emotion that we can exploit, as we have mentioned in Chapter 1, according to Rosenthal non-verbal communication model [Ros05], the subject experiences an internal state and expresses through his external features. These external features could be facial expressions, body gestures, non-verbal vocalisations, speech or different physiological signals. While the measurement procedure of the physiological signals is complicated and intrusive, body gestures and vocalisations seems to be a great source of emotion, as an alternative to facial expressions. Multi-modal including the fusion estimations of facial expression, body gesture and vocalisation models is a great way to ensure the accuracy and stability of the network. Previous works [KCC10, ZLCJ18, KRO20] have shown that using multi different modalities in combination greatly increases performance over unimodal emotion recognition systems. Since humans use more than one modality to recognise emotions and process signals in a complementary manner, it is expected that an automatic system demonstrate similar behavior. Overall, the development of such a multi-modal system capable of complementing each other in the case that some modality feature values are missing or unreliable is a great reward but also is a great challenge.

5.2.3 Real-time expressions assessment

In the fields of healthcare and medicine, the ability to continuously monitor a patient in real-time is very important to ensure the well-being of the patient, especially for in-patients with cognitive disorders or serious illnesses. Hence, when constructing a deep neural network for human expressions assessment, besides improving the generalisability of the model for more accurate predictions, the ability to keep up in real-time configuration is another important aspect that need to be taken into account. The general trend in deep learning is towards deeper, wider, and more complicated networks in order to achieve higher accuracy [HZC+17]. However, this approach makes deep networks heavier and slower, which is not suitable for the requirements of real-time applications. On the other hand, shallow network is much faster than deep network and seems to be perfectively fit for a real-time application. However, empirical work shows that it is difficult to train shallow nets to be as accurate as deep nets [BC14]. Another promising direction is relying on the Knowledge Distillation (KD) technique [HVD+15] (see Appendix A for a learning approach that utilising this technique), whose idea is to train a shallow student network to mimic the ability of a deep teacher model. This way, we can still have a light-weight shallow network that is fast enough for running in real-time configuration, and at the same time, as accurate as deep network thanks to the KD technique.

One another problem with facial expressions assessment systems is the fact that they do not work directly with original camera images but only with face cropped images. Traditional approach for inference these expressions assessment systems consists of several steps, including detection of facial landmarks, alignment and cropping of the face, and finally facial expressions assessment (e.g., [VBADE21, WXL+17]). Each of these steps consumes quite a bit of time and this is one of the main reasons that slow down the whole application. As features of the image have been extracted twice, once inside the facial landmark detection network and once inside the facial expressions assessment network, future work can try to reuse the features from the first step for the second step,

e.g., by using RoI pooling layer [HGDG17] to extract features from face regions. This way, we can reduce half of the time to extract features from images, improving the speed of the system for better real-time predictions.

5.3 Conclusion

In this thesis, we presented different approaches for efficient learning from limited amount of data, for automatic facial expression assessment with application to the pain emotion. We investigated the importance of integrating information about the location of regions-of-interest in the face image into the training process for better extracting feature representations. In our first approach, we have tried to train our network to focus on regions-of-interest by utilising AU heatmaps regression on a combination of multi-database, reaching a great level of performance compared to SOTA approaches on the same domain. Then, we have expanded the idea from focusing to isolating the regions-of-interest by proposing an approach that relies on object detection network, i.e., the FFAU neural network, which further improves the performance of deep neural network on both facial AUs and PSPI pain intensity estimation. Besides the improvements in performance, our FFAU network has also reached a great level of explainability in term of PSPI pain intensity estimation. While other SOTA methods only return the PSPI score without any explanation, our FFAU network can indicate the intensity and also the regions for each of the AUs that construct the PSPI score, which improve the reliability of our network's predictions. Despite having these improvements, one of the drawbacks of our FFAU network is that it has not be able to exploit the temporal dynamics between consecutive frame images of a video. This drawback, however, is also an opportunity to further improve the performance of our network in the future work. Considering the challenges that we encountered in this research and those mentioned in Section 5.2, in our future work, we would like to develop a multi-modal computational model which is capable of exploiting both spatial and temporal information from multi different modalities such as facial expressions, vocalisations, and body gestures. With

the Knowledge Distillation technique, we expect such model to be both generalisation to newly unseen samples and fast enough for real-time running configuration.

Bibliography

[AA21] Sharmeen M Saleem Abdullah and Adnan Mohsin Abdulazeez. Facial expression recognition based on deep learning convolution neural network: A review. *Journal of Soft Computing and Data Mining*, 2(1):53–65, 2021.

[ABRD15] Grigory Antipov, Sid-Ahmed Berrani, Natacha Ruchaud, and Jean-Luc Dugelay. Learned vs. hand-crafted features for pedestrian gender recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1263–1266, 2015.

[AEAKAS20] Rasha Al-Eidan, Hend Al-Khalifa, and AbdulMalik Al-Salman. Deep-learning-based models for pain recognition: A systematic review. *Applied Sciences*, 10:5984, 08 2020.

[Aga18] Abien Fred Agarap. Deep learning using rectified linear units (relu). *ArXiv*, abs/1803.08375, 2018.

[AHRH00] Marc Archinard, Véronique Haynal-Reymond, and Michel Heller. Doctor's and patients' facial expressions and suicide reattempt risk assessment. 2000.

[AMBD18] Benjamin Allaert, José Mennesson, Ioan Marius Bilasco, and Chabane Djeraba. Impact of the face registration techniques on facial expressions recognition. *Signal Processing: Image Communication*, 61:44–53, 2018.

[ASC05] Zara Ambadar, Jonathan W Schooler, and Jeffrey F Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological science*, 16(5):403–410, 2005.

[ASH+22] Mouath Aouayeb, Catherine Soladie, Wassim Hamidouche, Kidiyo Kpalma, and Renaud Seguier. Spatiotemporal features fusion from local facial regions for micro-expressions recognition. *Frontiers in Signal Processing*, page 23, 2022.

[Bal15] Davide Ballabio. A matlab toolbox for principal component analysis and unsupervised exploration of data structure. *Chemometrics and intelligent laboratory systems*, 149:1–9, 2015.

[BAPD13] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013.

[BC14] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.

[Bel66] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[BGV92] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[BLF+06] Marian Stewart Bartlett, Gwen Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, Javier R Movellan, et al. Automatic recognition of facial actions in spontaneous expressions. *J. Multim.*, 1(6):22–35, 2006.

[BM08] Kevin Bailly and Maurice Milgram. Head pose determination using synthetic images. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 1071–1080. Springer, 2008.

[BMP09] Kevin Bailly, Maurice Milgram, and Philippe Phothisane. Head pose estimation by a stepwise nonlinear regression. In *International conference on computer analysis of images and patterns*, pages 25–32. Springer, 2009.

[BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[BPL10] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118, 2010.

[BR19] Wissam J Baddar and Yong Man Ro. Mode variational lstm robust to unseen modes of variation: Application to facial expression recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3215–3223, 2019.

[Bur98] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[BW06] Lisa Feldman Barrett and Tor D. Wager. The structure of emotion: Evidence from neuroimaging studies. *Current Directions in Psychological Science*, 15(2):79–83, 2006.

[BWL20] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[BZD+20] Ghazal Bargshady, Xujuan Zhou, Ravinesh C. Deo, Jeffrey Soar, Frank Whittaker, and Hua Wang. Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Systems with Applications*, 149:113305, Jul 2020.

[CAE07] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007.

[CBC+06] Linda A Camras, Roger Bakeman, Yinghe Chen, Katherine Norris, and Thomas R Cain. Culture, ethnicity, and children's facial expressions: a study of european american, mainland chinese, chinese american, and adopted chinese girls. *Emotion*, 6(1):103, 2006.

[CBL+00] John T Cacioppo, Gary G Berntson, Jeff T Larsen, Kirsten M Poehlmann, Tiffany A Ito, et al. The psychophysiology of emotion. *Handbook of emotions*, 2(01):2000, 2000.

[CE05] Jeffrey F Cohn and Paul Ekman. Measuring facial action. 2005.

[Cho21] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[CHP+21] Weijun Chen, Hongbo Huang, Shuai Peng, Changsheng Zhou, and Cuiping Zhang. Yolo-face: a real-time face detector. *The Visual Computer*, 37(4):805–813, 2021.

[CHWS16] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *European Conference on Computer Vision*, pages 122–138. Springer, 2016.

[CHZH07] Deng Cai, Xiaofei He, Wei Vivian Zhang, and Jiawei Han. Regularized locality preserving indexing via spectral regression. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 741–750, 2007.

[CILS12] Tim F Cootes, Mircea C Ionita, Claudia Lindner, and Patrick Sauer. Robust and accurate shape model fitting using random forest regression

voting. In *European conference on computer vision*, pages 278–291. Springer, 2012.

[CJK04] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[CKM+09] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[CL18] Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018.

[COG19] Ozan Cakiroglu, Caner Ozer, and Bilge Gunsel. Design of a deep face detector by mask r-cnn. In *2019 27th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2019.

[Coh88] J Cohen. Statistical power analysis for the behavioral sciences lawrence earlbaum associates. 20th–, 1988.

[CPG11] Kenneth D. Craig, Kenneth M. Prkachin, and Ruth E. Grunau. *The facial expression of pain.*, page 117–133. The Guilford Press, New York, NY, US, 2011.

[Cra92] Kenneth D Craig. The facial expression of pain better than a thousand words? *APS Journal*, 1(3):153–162, 1992.

[Cra09] Kenneth D Craig. The social communication model of pain. *Canadian Psychology/Psychologie canadienne*, 50(1):22, 2009.

[CSX+18] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[CTC17] Wen-Sheng Chu, Fernando Torre, and Jeffrey Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. pages 25–32, 05 2017.

[CvMG+14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[CWD+18] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[CWS+18] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn, 2018.

[DBK+96] Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

[DCS20] Didan Deng, Zhaokang Chen, and Bertram E. Shi. Multitask emotion recognition with incomplete labels, 2020.

[DKMJ18] Kartik Dutta, Praveen Krishnan, Minesh Mathew, and CV Jawahar. Improving cnn-rnn hybrid networks for handwriting recognition. In

2018 16th international conference on frontiers in handwriting recognition (ICFHR), pages 80–85. IEEE, 2018.

[DP98] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[DRGP+11] Lies De Ruddere, Liesbet Goubert, Ken Martin Prkachin, Michael André Louis Stevens, Dimitri Marcel Leon Van Ryckeghem, and Geert Crombez. When you dislike patients, pain is taken less seriously. *PAIN®*, 152(10):2342–2347, 2011.

[DSI+01] Ying Dai, Yoshitaka Shibata, Tomoyuki Ishii, Koji Hashimoto, K Katamachi, K Noguchi, N Kakizaki, and Dawei Cai. An associate memory model of facial expressions and its application in facial expression recognition of patients on bed. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 151–151. IEEE Computer Society, 2001.

[dSP22] Claudio Filipi Gonç alves dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys*, jan 2022.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[DT18] Suresh Dara and Priyanka Tumma. Feature extraction by using deep learning: A survey. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1795–1801. IEEE, 2018.

[DTM14] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

[DV16] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning, 2016.

[DVGT17] Andrei Dobrescu, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Leveraging multiple datasets for deep leaf counting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[DZR12] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[EA12] Reda Elbarougy and Masato Akagi. Speech emotion recognition system based on a dimensional approach using a three-layered model. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9. IEEE, 2012.

[EDF90] Paul Ekman, Richard J Davidson, and Wallace V Friesen. The duchenne smile: Emotional expression and brain physiology: Ii. *Journal of personality and social psychology*, 58(2):342, 1990.

[EF71] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2), 1971.

[EF78] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[EF82] Paul Ekman and Wallace V Friesen. Felt, false, and miserable smiles. *Journal of nonverbal behavior*, 6(4):238–252, 1982.

[EFH02] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. *Facial Action Coding System: The Manual on CD ROM*. A Human Face, 2002.

[ESH19] Konstantin Eckle and Johannes Schmidt-Hieber. A comparison of deep networks with relu activation function and linear spline-type methods. *Neural Networks*, 110:232–242, 2019.

[EV18] R. Ezhilarasi and P. Varalakshmi. Tumor detection in the brain using faster r-cnn. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pages 388–392, 2018.

[FCAL04] Erika E Forbes, Jeffrey F Cohn, Nicholas B Allen, and Peter M Lewinsohn. Infant affect during parent–infant interaction at 3 and 6 months: Differences between mothers and fathers and influence of parent history of depression. *Infancy*, 5(1):61–84, 2004.

[FCNL12] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[FFV14] Corneliu Florea, Laura Florea, and Constantin Vertan. Learning pain from emotion: Transferred hot data representation for pain intensity estimation. page 778–790, Sep 2014.

[Fis38] Ronald A Fisher. The statistical utilization of multiple measurements. *Annals of eugenics*, 8(4):376–386, 1938.

[Fit13] Maria Fitzgerald. Central nociceptive pathways and descending modulation. *Oxford Textbook of Paediatric Pain*, pages 74–81, 2013.

[FL00] Beat Fasel and Juergen Luettin. Recognition of asymmetric facial action unit activities and intensities. In *Proceedings of the International*

Conference on Pattern Recognition (ICPR 2000), volume 1, pages 1100–1103, 2000.

[FLL20] Yingruo Fan, Jacqueline C. K. Lam, and Victor O. K. Li. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. *arXiv:2004.09681 [cs]*, Apr 2020.

[FLR+17] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. Dssd : Deconvolutional single shot detector, 2017.

[Fri98] Jerome H Friedman. Data mining and statistics: What's the connection? *Computing science and statistics*, 29(1):3–9, 1998.

[FS97] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[FVFP17] Diego R Faria, Mario Vieira, Fernanda CC Faria, and Cristiano Premebida. Affective facial expressions recognition for human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 805–810. IEEE, 2017.

[Gab46] Dennis Gabor. Theory of communication. part 3: Frequency compression and expansion. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):445–457, 1946.

[GB20] Darshan Gera and S Balasubramanian. Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information, 2020.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GCJC17] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 581–588. IEEE, 2017.

[GDDM13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013.

[GHR13] Sarah D Gunnery, Judith A Hall, and Mollie A Ruben. The deliberate duchenne smile: Individual differences in expressive control. *Journal of Nonverbal Behavior*, 37(1):29–41, 2013.

[GHZ+18] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.

[Gir15] Ross Girshick. Fast r-cnn, 2015.

[GMS10] Vicente García, Ramon A Mollineda, and J Salvador Sánchez. Theoretical analysis of a performance measure for imbalanced data. In *2010 20th International Conference on Pattern Recognition*, pages 617–620. IEEE, 2010.

[Gow75] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[GPAM+20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[GRvdVB14] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251, 2014.

[GS08] Kasey M Griffin and Michael A Sayette. Facial reactions to smoking cues relate to ambivalence about smoking. *Psychology of Addictive Behaviors*, 22(4):551, 2008.

[GSC00] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[GWK+18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.

[Haa11] Alfred Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 71(1):38–53, 1911.

[HAG+17] Maxwell L Hutchinson, Erin Antono, Brenna M Gibbons, Sean Paradiso, Julia Ling, and Bryce Meredig. Overcoming data scarcity with transfer learning. *arXiv preprint arXiv:1711.05099*, 2017.

[Has17] Teena Chakkalayil Hassan. *Recognizing emotions conveyed through facial expressions*. 2017.

[HBK95] Ursula Hess, Rainer Banse, and Arvid Kappas. The intensity of facial expression is determined by underlying affective state and social situation. *Journal of personality and social psychology*, 69(2):280, 1995.

[HBK97] Ursula Hess, Sylvie Blairy, and Robert E Kleck. The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4):241–257, 1997.

[HC12] Zakia Hammal and Jeffrey F Cohn. Automatic detection of pain intensity. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 47–52, 2012.

[HCLW19] Yunxin Huang, Fei Chen, Shaohe Lv, and Xiaodong Wang. Facial expression recognition: A survey. *Symmetry*, 11(10):1189, 2019.

[HCR92] Elaine Hatfield, John T Cacioppo, and Richard L Rapson. Primitive emotional contagion. 1992.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2017.

[Hjo69] C.H. Hjortsjö. *Man's Face and Mimic Language*. Studentlitteratur, 1969.

[HLHM00] Thomas Hadjistavropoulos, Diane LaChapelle, Carla Hale, and Farley K MacLeod. Age-and appearance-related stereotypes about patients undergoing a painful medical procedure. *The Pain Clinic*, 12(1):25–33, 2000.

[HMD15] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[HO00] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[HQX+21] Yibo Huang, Linbo Qing, Shengyu Xu, Lu Wang, and Yonghong Peng. Hybnet: a hybrid network structure for pain intensity estimation. *The Visual Computer*, Feb 2021.

[HRVB90] Heather D Hadjistavropoulos, Michael A Ross, and Carl L Von Baeyer. Are physicians' ratings of pain affected by patients' physical attractiveness? *Social Science & Medicine*, 31(1):69–72, 1990.

[HT10] Dong Huang and Fernando De la Torre. Bilinear kernel reduced rank regression for facial expression synthesis. In *European conference on computer vision*, pages 364–377. Springer, 2010.

[HVD+15] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[HXMF20] Dong Huang, Zhaoqiang Xia, Joshua Mwesigye, and Xiaoyi Feng. Pain-attentive network: a deep spatio-temporal attention model for pain estimation. *Multimedia Tools and Applications*, 79(37–38):28329–28354, Oct 2020.

[HZC+17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[HZRS14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision – ECCV 2014*, pages 346–361. Springer International Publishing, 2014.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[HZZ+16] Xiaopeng Hong, Guoying Zhao, Stefanos Zafeiriou, Maja Pantic, and Matti Pietikäinen. Capturing correlations of local features for image representation. *Neurocomputing*, 184:99–106, 2016.

[HZZ+20] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*, 2020.

[JCDLT13] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 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.

[JGCDLT13] László A Jeni, Jeffrey M Girard, Jeffrey F Cohn, and Fernando De La Torre. 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.

[JGY+12] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[JL16] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016.

[JLM17] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 650–657, 2017.

[JNK20] Mira Jeong, Jaeyeal Nam, and Byoung Chul Ko. Lightweight multilayer random forests for monitoring driver emotional status. *Ieee Access*, 8:60344–60354, 2020.

[Jol05] Ian Jolliffe. Principal component analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[JXYY12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[JZGX21] Yue Jin, Tianqing Zheng, Chao Gao, and Guoqiang Xu. A multi-modal and multi-task learning method for action unit and expression recognition, 2021.

[JZH21] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[KB17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[KCC10] Loic Kessous, Ginevra Castellano, and George Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1):33–48, 2010.

[KFL12] Miriam Kunz, Nicole Faltermeier, and Stefan Lautenbacher. Impact of visual learning on facial expressions of physical distress: A study on voluntary and evoked expressions of pain in congenitally blind and sighted individuals. *Biological psychology*, 89(2):467–476, 2012.

[KGAS15] Heysem Kaya, Furkan Gürpinar, Sadaf Afshar, and Albert Ali Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 459–466, 2015.

[Kim10] Tai-hoon Kim. Pattern recognition using artificial neural network: A review. In Samir Kumar Bandyopadhyay, Wael Adi, Tai-hoon Kim, and Yang Xiao, editors, *Information Security and Assurance*, pages 138–148, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[Kin80] Ross Kindermann. Markov random fields and their applications. *American mathematical society*, 1980.

[KKHZ21] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.

[KL14] Miriam Kunz and Stefan Lautenbacher. The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain. *European Journal of Pain*, 18(6):813–823, 2014.

[KMK06] Shinobu Kitayama, Batja Mesquita, and Mayumi Karasawa. Cultural affordances and emotional experience: socially engaging and disengaging emotions in japan and the united states. *Journal of personality and social psychology*, 91(5):890, 2006.

[KMKB13] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, and Saida Bouakaz. Pain detection through shape and appearance features. In

2013 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2013.

[KML19] Miriam Kunz, Doris Meixner, and Stefan Lautenbacher. Facial muscle movements encoding pain—a systematic review. *Pain*, 160(3):535–549, 2019.

[KMSL95] Dacher Keltner, TERRIE E MOFfiTT, and Magda Stouthamer-Loeber. Facial expressions of emotion and psychopathology in adolescent boys. *Journal of Abnormal Psychology*, 104(4):644, 1995.

[KR19] Sudhakar Kumawat and Shanmuganathan Raman. Lp-3dcnn: Unveiling local phase in 3d convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4903–4912, 2019.

[KRO20] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020.

[KRP12] Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing*, pages 368–377. Springer, 2012.

[KS14] George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[KSHZ] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800.

[KSZ19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[KSZ21] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[KSZQ20] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, Apr 2020.

[KTN+19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019.

[KTP15] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Latent trees for estimating intensity of facial action units. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 296–304, 2015.

[KTTP17] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[KZ18] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition, 2018.

[KZ19a] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition, 2019.

[KZ19b] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[KZ20] Dimitrios Kollias and Stefanos Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 12(3):595–606, 2020.

[KZ21] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[L+89] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19(143-155):18, 1989.

[LAE+16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

[LAZ17] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[LBE15] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LBL+12] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[LCK+10] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[LCP+12] Patrick Lucey, Jeffrey Cohn, Kenneth Prkachin, Patricia Solomon, Sien Chew, and Iain Matthews. Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 30:197–205, 03 2012.

[LD20] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020.

[LDG+16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2016.

[LDH19] Harshit Kumar Lohani, S Dhanalakshmi, and V Hemalatha. Performance analysis of extreme learning machine variants with varying intermediate nodes and different activation functions. In *Cognitive Informatics and Soft Computing*, pages 613–623. Springer, 2019.

[LEF90] Robert W Levenson, Paul Ekman, and Wallace V Friesen. Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, 27(4):363–384, 1990.

[LGTB97] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[Lie98] Jenn-Jier James Lien. *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity*. University of Pittsburgh, 1998.

[LKG19] Zeju Li, Konstantinos Kamnitsas, and Ben Glocker. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 402–410. Springer, 2019.

[LLS+15] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[LRVL10] Iulia Lefter, Léon Rothkrantz, and David Van Leeuwen. Emotion recognition from speech by combining databases and fusion of classifiers. volume 6231, pages 353–360, 09 2010.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[LTWE+17] Dieu Linh Tran, Robert Walecki, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3190–3199, 2017.

[LZH+19] Liang Li, Xinge Zhu, Yiming Hao, Shuhui Wang, Xingyu Gao, and Qingming Huang. A hierarchical cnn-rnn approach for visual emotion classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–17, 2019.

[LZZ+17] Jiaxing Li, Dexiang Zhang, Jingjing Zhang, Jun Zhang, Teng Li, Yi Xia, Qing Yan, and Lina Xun. Facial expression recognition with faster r-cnn. *Procedia Computer Science*, 107:135–140, 12 2017.

[Man84] George Mandler. *Mind and body: Psychology of emotion and stress*. WW Norton & Company Incorporated, 1984.

[Mar08] Serge Marchand. The physiology of pain mechanisms: from the periphery to the brain. *Rheumatic disease clinics of North America*, 34(2):285–309, 2008.

[MBPG14] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European conference on computer vision*, pages 720–735. Springer, 2014.

[MCB13] Pedro Martins, Rui Caseiro, and Jorge Batista. Generative face alignment through 2.5 d active appearance models. *Computer Vision and Image Understanding*, 117(3):250–268, 2013.

[MCMC09] Mohammad H Mahoor, Steven Cadavid, Daniel S Messinger, and Jeffrey F Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 74–80. IEEE, 2009.

[MCT+89] Carol Z Malatesta, Clayton Culver, Johanna Rich Tesman, Beth Shepard, Alan Fogel, Mark Reimers, and Gail Zivin. The development of emotion expression during the first two years of life. *Monographs of the society for research in child development*, pages i–136, 1989.

[Mer79] HAFD Merskey. Pain terms: a list with definitions and notes on usage. recommended by the iasp subcommittee on taxonomy. *Pain*, 6:249–252, 1979.

[MF67] Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3):248, 1967.

[MG14] Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2014.

[MHM19] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, jan 2019.

[MLLB21] Shervin Minaee, Ping Luo, Zhe Lin, and Kevin Bowyer. Going deeper into face detection: A survey. *arXiv preprint arXiv:2103.14983*, 2021.

[MM14] S Mohammad Mavadati and Mohammad H Mahoor. Temporal facial expression modeling for automated action unit intensity measurement. In *2014 22nd International Conference on Pattern Recognition*, pages 4648–4653. IEEE, 2014.

[MMB+13] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[MR73] Albert Mehrabian and James A Russell. A measure of arousal seeking tendency. *Environment and Behavior*, 5(3):315, 1973.

[MSTA18] Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. In *Intelligent Engineering Informatics*, pages 623–632. Springer, 2018.

[MVBP12] Brais Martinez, Michel F Valstar, Xavier Binefa, and Maja Pantic. Local evidence aggregation for regression-based facial point detection. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1149–1163, 2012.

[MVC+11] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 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, 2011.

[NSCD17] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017.

[NSD19] Mahyar Najibi, Bharat Singh, and Larry S Davis. Fa-rpn: Floating region proposals for face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2019.

[NYD16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016.

[NZY21] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[OPM02] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[Pal11] Mahesh Pal. Support vector machines/relevance vector machine for remote sensing classification: A review. *arXiv preprint arXiv:1101.2987*, 2011.

[PC92] Gary D Poole and Kenneth D Craig. Judgments of genuine, suppressed, and faked facial expressions of pain. *Journal of personality and social psychology*, 63(5):797, 1992.

[PC95] Kenneth M Prkachin and Kenneth D Craig. Expressing pain: The communication and interpretation of facial pain signals. *Journal of nonverbal behavior*, 19(4):191–205, 1995.

[PCBH17] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[PFA20] Simone Porcu, Alessandro Floris, and Luigi Atzori. Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics*, 9(11), 2020.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[PKB15] Kenneth M Prkachin, Kimberley A Kaseweter, and M Erin Browne. Understanding the suffering of others: the sources and consequences of third-person pain. *Pain, emotion and cognition*, pages 53–72, 2015.

[Prk09] Kenneth M Prkachin. Assessing pain by facial expression: facial expression as nexus. *Pain Research and Management*, 14(1):53–58, 2009.

[PRP05] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[PS08] Kenneth Prkachin and Patricia Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139:267–74, 06 2008.

[PYC+19] Xianzhang Pan, Guoliang Ying, Guodong Chen, Hongming Li, and Wenshu Li. A deep spatial and temporal aggregation framework

for video-based facial expression recognition. *IEEE Access*, 7:48807–48815, 2019.

[QJM+19] Rong Qi, Rui-Sheng Jia, Qi-Chao Mao, Hong-Mei Sun, and Ling-Qun Zuo. Face detection method based on cascaded convolutional networks. *IEEE Access*, 7:110740–110748, 2019.

[RCG+17] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M. Gonfaus, Kamal Nasrollahi, Thomas B. Moeslund, and F. Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE Transactions on Cybernetics*, page 1–11, 2017.

[RDGF15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015.

[RE05] Erika L Rosenberg and Paul Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2005.

[RE20] Erika L Rosenberg and Paul Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 2020.

[RF17] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[RFD97] James A Russell and José Miguel Fernández-Dols. *The psychology of facial expression*. Cambridge university press Cambridge, UK, 1997.

[RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[RM77] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

[RM19] Aprameyo Roy and Deepak Mishra. Ecnn: Activity recognition using ensembled convolutional neural networks. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 757–760. IEEE, 2019.

[RMN09] Rajat Raina, Anand Madhavan, and Andrew Y Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880, 2009.

[Ros05] Robert Rosenthal. Conducting judgment studies: Some methodological issues. *The new handbook of methods in nonverbal behavior research*, pages 199–234, 2005.

[RPC17] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):121–135, 2017.

[RPP13] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. *Automatic Pain Intensity Estimation with Heteroscedastic Conditional Ordinal Random Fields*, volume 8034 of *Lecture Notes in Computer Science*, page 234–243. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[RRBP16] Adria Ruiz, Ognjen Rudovic, Xavier Binefa, and Maja Pantic. Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation, 2016.

[RSCC17] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face

analysis. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 17–24. IEEE, 2017.

[Rus80] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[Rus95] James A Russell. Facial expressions of emotion: What lies beyond minimal universality? *Psychological bulletin*, 118(3):379, 1995.

[RVBS17] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise, 2017.

[RW04] Michael E Robinson and Emily A Wise. Prior pain experience: influence on the observation of experimental pain in men and women. *The Journal of Pain*, 5(5):264–269, 2004.

[SA19] Kanwar Bharat Singh and Mustafa Ali Arat. Deep learning in the automotive industry: Recent advances and application examples. *arXiv preprint arXiv:1906.08834*, 2019.

[SB09] Klaus R Scherer and Tobias Brosch. Culture-specific appraisal biases contribute to emotion dispositions. *European Journal of Personality: Published for the European Association of Personality Psychology*, 23(3):265–288, 2009.

[SBAO18] Sanun Srisuk, Apiwat Boonkong, Damrongsak Arunyagool, and Surachai Ongkittikul. Handcraft and learned feature extraction techniques for robust face recognition: A review. In *2018 International Electrical Engineering Congress (iEECON)*, pages 1–4. IEEE, 2018.

[SBB+03] Stefan R Schweinberger, Lyndsay M Baird, Margarethe Blümler, Jürgen M Kaufmann, and Bettina Mohr. Interhemispheric cooperation for face recognition but not for affective facial expressions. *Neuropsychologia*, 41(4):407–414, 2003.

[SBR10] Klaus R Scherer, Tanja Bänziger, and Etienne Roesch. *A Blueprint for Affective Computing: A sourcebook and manual*. Oxford University Press, 2010.

[SC10] Martin Schiavenato and Kenneth D Craig. Pain assessment as a social transaction: beyond the "gold standard". *The Clinical journal of pain*, 26(8):667–676, 2010.

[SCGH05] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, volume 5670, pages 56–67. International Society for Optics and Photonics, 2005.

[Sch99] Emery Schubert. Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, 51(3):154–165, 1999.

[SCW+21] Tengfei Song, Zijun Cui, Yuru Wang, Wenming Zheng, and Qiang Ji. Dynamic probabilistic graph convolution for facial action unit intensity estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4843–4852, Nashville, TN, USA, Jun 2021. IEEE.

[SDF+21] Teresa Souto, Ana Rita Dias, Maria Ferreira, Cristina Queirós, and Vanessa Figueiredo. How to assess emotional recognition in individuals with a diagnosis of schizophrenia and other psychotic disorders: a pilot study. *Current Psychology*, pages 1–10, 03 2021.

[SEZ+13] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013.

[SGC14] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[SHW+07] Miiamaaria V Saarela, Yevhen Hlushchuk, Amanda C de C Williams, Martin Schürmann, Eija Kalso, and Riitta Hari. The compassionate brain: humans detect intensity of pain from another's face. *Cerebral cortex*, 17(1):230–237, 2007.

[SIVA16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[SKG20] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[SLBW13] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Detecting and aligning faces by image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3467, 2013.

[SLJ+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[SLTV18] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, and Michel Valstar. Joint action unit localisation and intensity estimation through heatmap regression. *arXiv preprint arXiv:1805.03487*, 2018.

[SMB10] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[SS18] Hanan Salam and Renaud Seguier. A survey on face modeling: building a bridge between face analysis and synthesis. *The Visual Computer*, 34(2):289–319, 2018.

[SSB12] Arman Savran, Bulent Sankur, and M Taha Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, 2012.

[STE13] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. *Advances in neural information processing systems*, 26, 2013.

[SZP13] Georgia Sandbach, Stefanos Zafeiriou, and Maja Pantic. Markov random field structures for facial action unit intensity estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 738–745, 2013.

[SZWR11] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll. Using multiple databases for training in emotion recognition: To unite or to vote? In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolu-

tional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[TBLH17] Jérôme Thevenot, Miguel Bordallo Lopez, and Abdenour Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 10 2017.

[TDHL18] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European conference on computer vision (ECCV)*, pages 797–813, 2018.

[TFSK19] Iuliana Tabian, Hailing Fu, and Zahra Sharif Khodaei. A convolutional neural network for impact detection and characterization of complex composite structures. *Sensors*, 19(22):4933, 2019.

[TH18a] Mohammad Tavakolian and Abdenour Hadid. Deep binary representation of facial expressions: A novel framework for automatic pain intensity recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1952–1956. IEEE, 2018.

[TH18b] Mohammad Tavakolian and Abdenour Hadid. Deep spatiotemporal representation of the face for automatic pain intensity estimation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 350–354. IEEE, 2018.

[TH19] Mohammad Tavakolian and Abdenour Hadid. A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics. *International Journal of Computer Vision*, 127(10):1413–1425, 2019.

[Tha90] Robert E Thayer. *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[Tip01] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[TMD+05] Fabien Trémeau, Dolores Malaspina, Fabrice Duval, Humberto Corrêa, Michaela Hager-Budny, Laura Coin-Bariou, Jean-Paul Macher, and Jack M Gorman. Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects. *American Journal of Psychiatry*, 162(1):92–101, 2005.

[TSV20] Gaurav Tripathi, Kuldeep Singh, and Dinesh Kumar Vishwakarma. Violence recognition using convolutional neural network: A survey. *Journal of Intelligent & Fuzzy Systems*, 39(5):7931–7952, 2020.

[TWC99] Auke Tellegen, David Watson, and Lee Anna Clark. On the dimensional and hierarchical structure of affect. *Psychological science*, 10(4):297–303, 1999.

[TWH+18] Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018.

[TYRW14] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[UAM+17] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE access*, 6:1155–1166, 2017.

[USGS13] Jasper Uijlings, K. Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 09 2013.

[VAG+15] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE, 2015.

[Val08] Michel François Valstar. Timing is everything: A spatio-temporal approach to the analysis of facial actions. 2008.

[Vap99] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[VBA21] Manh Tu Vu and Marie Beurton-Aimar. Multitask multi-database emotion recognition, 2021.

[VBADE21] Manh Tu Vu, Marie Beurton-Aimar, Pierre-yves Dezaunay, and Marine Cotty Eslous. Automated pain estimation based on facial action units from multi-databases. In *2021 Joint 10th International Conference on Informatics, Electronics Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision Pattern Recognition (icIVPR)*, pages 1–8, 2021.

[vBJM84] Carl L von Baeyer, Marianne E Johnson, and Marcia J McMillan. Consequences of nonverbal expression of pain: Patient distress and observer concern. *Social Science & Medicine*, 19(12):1319–1324, 1984.

[VDMPVdH+09] Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

[VJ01] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[Vod17] Damir Vodenicarevic. *Rhythms and oscillations: a vision for nanoelectronics*. PhD thesis, Université Paris Saclay (COmUE), 2017.

[WAHLE+16] Philipp Werner, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C Traue. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3):286–299, 2016.

[WAHLE+17] Philipp Werner, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C. Traue. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3):286–299, 2017.

[WGE+13] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C. Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O. Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE International Conference on Cybernetics (CYBCO)*, page 128–131, Jun 2013.

[WGT+18] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65, 2018.

[Wil02] Amanda C de C Williams. Facial expression of pain: an evolutionary account. *Behavioral and brain sciences*, 25(4):439–455, 2002.

[WJ16] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3400–3408, 2016.

[WJ19] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019.

[WLH+19] Fei Wang, Zhongheng Li, Fang He, Rong Wang, Weizhong Yu, and Feiping Nie. Feature learning viewpoint of adaboost and a new algorithm. *IEEE Access*, 7:149890–149899, 2019.

[WLMW+19] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, page 1–1, 2019.

[WLMW+22] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind W. Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 13(1):530–552, 2022.

[WLSR20] Yingying Wang, Yibin Li, Yong Song, and Xuewen Rong. The influence of the activation function in a convolution neural network model of facial expression recognition. *Applied Sciences*, 10(5):1897, 2020.

[WOR+17] Robert Walecki, Ognjen, Rudovic, Vladimir Pavlovic, Björn Schuller, and Maja Pantic. Deep structured learning for facial action unit intensity estimation. *arXiv:1704.04481 [cs]*, Apr 2017. arXiv: 1704.04481.

[WRGEP20] Nicola Webb, Ariel Ruiz-Garcia, Mark Elshaw, and Vasile Palade. Emotion recognition from face images in an unconstrained environment

for usage on social robots. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[WRPP16] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4902–4910, 2016.

[Wu17] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, 5(23):495, 2017.

[WW21] Lingfeng Wang and Shisen Wang. A multi-task mean teacher for semi-supervised facial affective behavior analysis, 2021.

[WXL+17] Feng Wang, Xiang Xiang, Chang Liu, Trac D. Tran, Austin Reiter, Gregory D. Hager, Harry Quon, Jian Cheng, and Alan L. Yuille. Regularizing face verification nets for pain intensity regression. *arXiv:1702.06925 [cs]*, 2017. arXiv: 1702.06925.

[WZ95] Ronald J Williams and David Zipser. Gradient-based learning algorithms for recurrent. *Backpropagation: Theory, architectures, and applications*, 433:17, 1995.

[XLW+16] Jun Xu, Xiaofei Luo, Guanhao Wang, Hannah Gilmore, and Anant Madabhushi. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, 191:214–223, 2016.

[XLY+21] Xuwu Xin, Xiaowu Li, Shengfu Yang, Xiaoyan Lin, and Xin Zheng. Pain expression assessment based on a locality and identity aware network. *IET Image Processing*, 15(12):2948–2958, 2021.

[XSH+18] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[YKA02] Ming-Hsuan Yang, David J Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 24(1):34–58, 2002.

[YLLT16] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[YYLL15] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Fine-grained evaluation on face detection in the wild. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.

[ZAA+21] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models, 2021.

[ZDHJ18] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 2314–2323, Salt Lake City, UT, Jun 2018. IEEE.

[ZDWG21] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Audio-visual attentive fusion for continuous emotion recognition, 2021.

[ZGC+21] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition, 2021.

[ZGW+16] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3438–3446, 2016.

[ZGWJ16] Rui Zhao, Quan Gan, Shangfei Wang, and Qiang Ji. Facial expression intensity estimation using ordinal information. page 3466–3474, Jun 2016.

[ZHSZ16] Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 1535–1543, Jun 2016.

[ZHZ+20] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, Shiguang Shan, and Xilin Chen. m^3t: Multi-modal continuous valence-arousal estimation in the wild, 2020.

[ZJW+19] Yong Zhang, Haiyong Jiang, Baoyuan Wu, Yanbo Fan, and Qiang Ji. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 733–742, Seoul, Korea (South), Oct 2019. IEEE.

[ZKN+17] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and

arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017.

[ZLCJ18] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pages 65–72, 2018.

[ZLL+16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[ZLLT18] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[ZLZ20] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, 36(5):1067–1093, 2020.

[ZNNS20] Junhui Zhao, Yiwen Nie, Shanjin Ni, and Xiaoke Sun. Traffic data imputation and prediction: An efficient realization of deep learning. *IEEE Access*, 8:46713–46722, 2020.

[ZPS17] Yuqian Zhou, Jimin Pi, and Bertram E. Shi. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 872–877, 2017.

[ZSCC13] Xiaowei Zhao, Shiguang Shan, Xiujuan Chai, and Xilin Chen. Cascaded shape space pruning for robust facial landmark detection. In

Proceedings of the IEEE International Conference on Computer Vision, pages 1033–1040, 2013.

[ZSS+22] Khalid Zaman, Zhaoyun Sun, Sayyed Mudassar Shah, Muhammad Shoaib, Lili Pei, and Altaf Hussain. Driver emotions recognition based on improved faster r-cnn and neural architectural search network. *Symmetry*, 14(4):687, Mar 2022.

[ZSZZ18] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 449–458, 2018.

[Zue90] Victor W Zue. Automatic speech recognition and understanding. In *AI in the 1980s and beyond*, pages 185–200. 1990.

[ZWD+19] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3452–3461, Long Beach, CA, USA, Jun 2019. IEEE.

[ZWHZ20] Jialiang Zhang, Xiongwei Wu, Steven CH Hoi, and Jianke Zhu. Feature agglomeration networks for single stage face detection. *Neurocomputing*, 380:180–189, 2020.

[ZXT18] Changzheng Zhang, Xiang Xu, and Dandan Tu. Face detection using improved faster rcnn, 2018.

[ZYC+14] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[ZZD+18] Yong Zhang, Rui Zhao, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 7034–7043, Salt Lake City, UT, Jun 2018. IEEE.

[ZZH15] Yang Zhang, Li Zhang, and M Alamgir Hossain. Adaptive 3d facial action intensity estimation and emotion recognition. *Expert systems with applications*, 42(3):1446–1464, 2015.

[ZZLS17] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep learning for biometrics*, pages 57–79. Springer, 2017.

[ZZXW18] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review, 2018.

Appendix A

Multitask Multi-database Emotion Recognition

Knowledge Distillation (KD) is a technique that transfers knowledge from a deep and complex model to a small student model. Hence, the main use of the technique is about compressing a deep learning teacher model to have a smaller and faster student model, while maintaining approximately the same level of performance as the teacher model. However, this technique can also be used to improve the performance of the student model over the teacher model in some specific cases. In this appendix, we summarise our findings in paper II, which utilise the KD technique to improve the performance of multi-task deep learning model in the context of Valence and Arousal estimation.

Paper II: M. T. Vu, M. Beurton-Aimar and S. Marchand, "Multitask Multi-database Emotion Recognition," 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 3630-3637, doi: 10.1109/ICCVW54120.2021.00406.

A.1 Introduction

Emotion recognition and analysis are the crucial parts of many applications and human-computer interactive systems, especially in health care and medical fields [TBLH17, AEAKAS20] since it is directly related to the health state of a patient. As results, more and more works have been conducted to try to analyse human emotions and behaviours [SCGH05, SKG20, WLMW+19]. In the same sense, the 2nd Affective Behavior Analysis in-the-wild (ABAW 2021) competition by Kollias *et al.*

[KSHZ, KSZ21, KZ21, KZ19b, KSZ19, KTN+19, ZKN+17, KKHZ21] provides a large-scale dataset Aff-Wild2 [KZ19a] for analysing human emotion in-the-wild settings. This dataset includes videos with annotations for three tasks including: valence-arousal estimation, action unit (AU) detection, and seven basic facial expression classification. Valence represents how positive or negative an emotional state is, whereas arousal describes how passive or active it is. The seven basic facial expressions include neutral, anger, disgust, fear, happiness, sadness, and surprise. AUs are the basic actions of individuals or groups of muscles for portraying emotions.

In this paper, we focus on two tasks: seven basic facial expressions classification and valence-arousal estimation. Inspired by the multitask training with incomplete label method from Deng *et al.* [DCS20] we propose a method to further exploit the inter-task correlations between these two tasks. Similar to Deng *et al.* [DCS20] we apply the distillation knowledge technique to train two multitask models: a teacher model and a student model. The student model will be trained using both ground truth labels and soft labels derived from the pretrained teacher model. However, instead of treating each task independently when training teacher model as in [DCS20], we add one more task to the training process, which is the combination of the two tasks above to train the network using data coming from AffectNet database [MHM19], in which contains labels for both of the two tasks. Since the data for this task has been annotated for both seven basic expressions and valence-arousal, this task will play the role of guiding the training, i.e. re-balancing the gradient backpropagation of the first two tasks and exploiting the inter-task correlations between the training tasks. Apart from that, taking into account that there are a huge number of videos that are annotated for both seven basic facial expressions and valence-arousal labels in the Affwild2 database, we integrate this information into the student model's training process for better exploiting inter-task correlations. With these improvements, our model has reached the performance on par with the state of the art on the test set of the official dataset Affwild2 of the competition.

A.2 Related Works

The challenges of human affect analysis have attracted lots of research efforts, especially in in-the-wild settings. In this section, we will briefly introduce some works related to this problem. Pan *et al.* [PYC+19] propose a framework to aggregate spatial and temporal convolutional features across the entire extent of a video. Deng *et al.* [DCS20] apply distillation knowledge technique to train their multitask model using data with incomplete labels. Kuhnke *et al.* [KRO20] propose a two stream aural-visual network for multi-task training. Gera *et al.* [GB20] propose a spatio-channel attention network, which is able to extract local and global attentive features for classifying facial expressions. Kollias *et al.* [KSZ21] proposed FaceBehaviorNet for large-scale face analysis, by jointly learning multiple facial affective behaviour tasks and a distribution matching approach. Wei Zhang *et al.* [ZGC+21] propose a heuristic that the three emotion representations including: categorical emotions, action units and valence-arousal are intrinsically associated with each other. They try to exploit these hierarchical relationships by developing a prior aided streaming network for multitask prediction. Wang *et al.* [WW21] extend the work of Kuhnke *et al.* [KRO20] by improving the preprocessing method of rendering mask and applying mean teacher model for utilising the unlabelled data. Su Zhang *et al.* [ZDWG21] propose an audio-visual spatial-temporal deep neural network with attention mechanism for valence-arousal estimation.

A.3 Methodology

In this section, we introduce our multitask multi-databases training method. Frame images are extracted from video and fed into a Convolution Neural Network (CNN) to train for analysing human's emotion in-the-wild. Then, features extracted from this network will go through a Recurrent Neural Network (RNN) to capture temporal information and finally, perform both the seven basic facial expressions classification and valence-arousal estimation. Because in our dataset, we do not always have all labels for

all of our tasks, we have applied the multitask training with missing labels method that is described in [DCS20] with some enhancements, which is described in the sections below.

A.3.1 Data Imbalancing

Similar to [DCS20], we also have used some external datasets to address the data imbalance problem in the Affwild2 dataset, e.g. most of the frames inside the Affwild2 dataset have their valence value in the range of $[0 - 0.4]$. The external datasets are including Expression in-the-Wild (ExpW) dataset [ZLLT18] for expression classification and AFEW-VA dataset [KTTP17] for valence-arousal estimation. After merging these datasets, we have applied the same dataset balancing protocol as [DCS20] to improve the balance of the dataset.

Different from [DCS20], as we have mentioned earlier, in this preliminary work we perform only two tasks: seven basic facial expressions prediction and valence-arousal estimation. Apart from that, we also want to include the AffectNet database [MHM19] into the training, since this database is annotated for both seven basic expressions and valence-arousal are available. After this step, for the training process, our dataset is including three parts:

Mixed EXPR The mixing set of the AffWild 2 (expressions part) and ExpW datasets for seven basic expressions. This dataset has no information about valence and arousal.

Mixed VA The mixing set of the AffWild 2 (valence-arousal part) and AFEW-VA datasets for valence and arousal. This dataset has no information about the seven basic expressions.

Affect EXPR_VA The AffectNet dataset, for both seven basic expressions and valence-arousal.

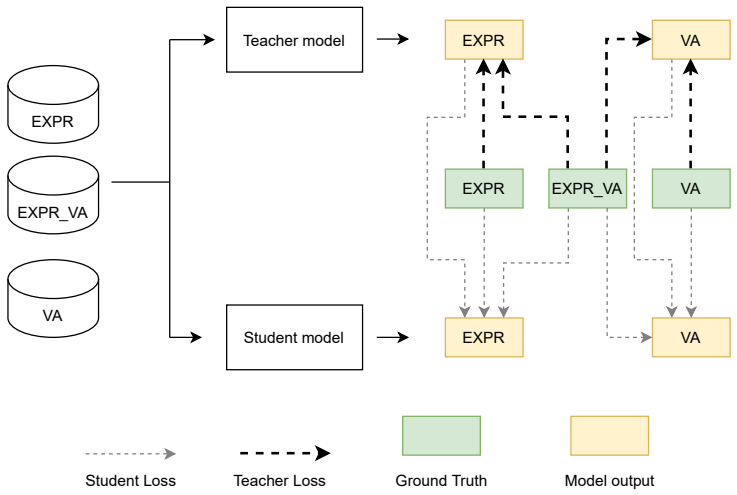


Figure A.1: The overview of our multitask training with missing labels.

Corresponding to these three dataset's parts are the three training tasks $\mathscr{T} \in \{1, 2, 3\}$, which are including: expression classification (EXPR), valence-arousal estimation (VA) and the mixing of these two tasks (EXPR_VA). One can note that even though we have three training tasks, our model has only two outputs, which are EXPR and VA, since the last training task reuses these two outputs for computing loss.

A.3.2 Multitask training with missing labels

Here we describe the formulars that are used to train our teacher and student models. Let (X, Y) be the training dataset, where X is a set of input vectors and Y is a set of ground truth training labels. Since our dataset contains three parts including: *Mixed EXPR*, *Mixed VA* and *Affect EXPR_VA*, therefore $(X, Y) = \{(X^{(i)}, Y^{(i)})\}_{i=1}^{3}$. For convenience of notation, we assume each subset i includes an equal number N of instances within a batch, i.e $(X^{(i)}, Y^{(j)}) = \{(x^{(i,n)}, y^{(i,n)})\}_{n=1}^{N}$ where n indexes the instance. Because the data from the last set *Affect EXPR_VA* is including both EXPR and VA annotations, we denote 3_{expr} and 3_{va} as the EXPR annotation and the VA annotation of this set, respectively. For example, instance $x^{(3,1)}$ belongs to *Affect EXPR_VA* dataset and has two annotations: $y^{(3_{expr},1)}$ and $y^{(3_{va},1)}$

The inputs for all instances have the same dimensionality, regardless of task. How-

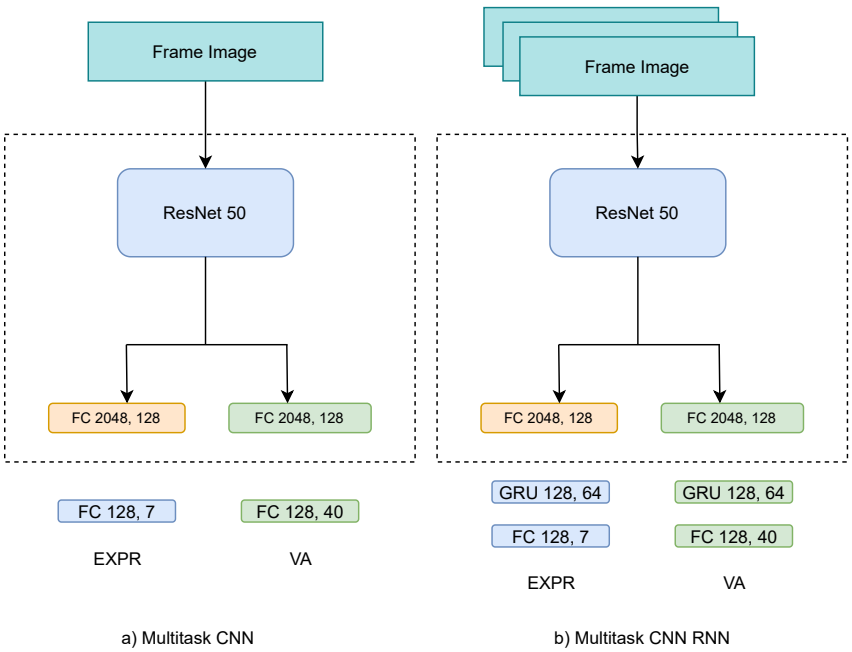


Figure A.2: The multitask CNN (a) and CNN-RNN (b) architectures, The two architectures share the same ResNet spatial feature extractor shown in the dashed box.

ever, the ground truth labels for different tasks have different dimensionality. The label for the first task (EXPR) is $y^{(1)} \in \{0,1\}^7$. The label for the second task (VA) is $y^{(2)} \in [-1,1]^2$. The label for the last task (EXPR_VA) is the mixed of the two tasks above.

Similar to [DCS20], we also apply the two steps training for capturing inter-task correlations. We train a single teacher model using only the ground truth labels in the first step. In the second step, we replace the missing labels with soft labels derived from the outputs of the teacher model. We then use the ground truth and soft labels to train a single student model. Different from [DCS20], we do not train multi student models for model ensemble because this approach is too costly in term of computation and the gain in performance is not significant. The overview of our network can be seen in Fig A.1 and the architecture of our model is in Fig A.2.

To be in the same line with [DCS20] in the sense of notation, we also denote the output of our multitask network by $f_\theta^{(i)}(\cdot)$ where θ contains the model parameters of either teacher model or student model, and $i \in \{1,2\}$ indicates the current task. For

example, $f_\theta^{(1)}(x^{(3)})$ indicates the output of the network for task 1 (EXPR) for an instance in the *Affect EXPR_VA* set. To avoid clutter, we will often refer to the output of the teacher network on task i by $t^{(i)}$ irrespective of what the input label is, i.e. $t^{(i)} = f_\theta^{(i)}(x^{(j)})$ for some $j \in \{1, 2\}$ and similarly to the output of the student network on task i by $s^{(i)}$.

Regarding the objective loss functions, similar to [DCS20], we also treat the problem of expression classification as a multiclass classification problem, and the problem of valence-arousal estimation as a combination of multiclass classification and regression problem. We will use the same Soft-max Function SF, the Cross Entropy function CE and the Concordance Correlation Coefficient function CCC, which have already been defined in [DCS20].

A.3.2.1 Supervision loss functions

Here we denote the loss functions that are used for optimizing our models parameters with the supervision of the ground truth labels for each of our training tasks.

EXPR task The supervision loss for the samples from the *Mixed EXPR* set is denoted as:

$$\mathscr{L}^{(1)}(y^{(1)}, t^{(1)}) = CE\left(y^{(1)}, SF(t^{(1)}, 1)\right) \tag{A.1}$$

VA task The supervision loss for the samples from the *Mixed VA* set is denoted as:

$$\mathscr{L}^{(2)}(y^{(2)}, t^{(2)}) = \sum_{i=1}^{2} \Bigg\{ CE\left(onehot(y_i^{(2)}), SF(t_i^{(2)}, 1)\right) + \frac{1}{B}\left(1 - CCC(y_i^{(2)}, t_i^{(2)})\right) \Bigg\} \tag{A.2}$$

EXPR_VA task For the samples from Affect EXPR_VA set, since the samples of this set are annotated for both VA and EXPR, the supervision loss for this task is denoted as:

$$\begin{aligned}\mathscr{L}^{(3)}(y^{(3)}, t^{(3)}) = CE\left(y^{(3_{expr})}, SF(f_{\theta_t}^{(1)}(x^{(3)}), 1)\right) \\ + \sum_{i=1}^{2}\Bigg\{CE\left(onehot(y_i^{(3_{va})}), SF(f_{\theta_t i}^{(2)}(x^{(3)}), 1)\right) \\ + \frac{1}{B}\left(1 - CCC(y_i^{(3_{va})}, f_{\theta_t i}^{(2)}(x^{(3)}))\right)\Bigg\} \quad \text{(A.3)}\end{aligned}$$

From this equation, we can see that for each sample of the dataset, we calculate the loss for both EXPR and VA tasks. Therefore, the gradient backpropagation derived from this task's loss is the most accurate one compared to the other two tasks. Because we can see that the loss of the EXPR task can be used to adjust the model's parameters for better EXPR prediction, but it has absolutely no idea of whether the VA estimation is correct or not, and the same goes for the loss of the VA task. Therefore, the EXPR_VA task plays the role of guiding the training process, i.e. re-balance the gradient backpropagation for the whole training process. In the same time, since this task compute the loss for both EXPR and VA tasks, it can exploit the inter-task correlations, which typically can help the network for better prediction.

A.3.2.2 Distillation loss functions

Here we denote the loss functions that are used to optimise our student model parameters with the supervision of both the ground truth labels (hard targets) and the pretrained teacher model's outputs (soft targets) for each of our training tasks. Similar to [DCS20], we use the KL divergence to measure the difference between two probability distributions (output of teacher model and student model). The KL divergence of two vectors p and q is denoted as: $KL(p, q) = \sum_i p_i log\left(\frac{p_i}{q_i}\right)$.

EXPR task Distillation loss for the samples from the *Mixed EXPR* set:

$$\mathscr{H}^{(1)}(t^{(1)}, s^{(1)}) = KL\left(SF(t^{(1)}, T), SF(s^{(1)}, T)\right) \quad \text{(A.4)}$$

VA task Distillation loss for the samples from the *Mixed VA* set:

$$\mathscr{H}^{(2)}(t^{(2)}, s^{(2)}) = \sum_{i=1}^{2} KL\left(SF(t_i^{(2)}, T), SF(s_i^{(2)}, T)\right) \tag{A.5}$$

EXPR_VA task Distillation loss for the samples from the *Affect EXPR_VA* set is the combination of the EXPR and VA distillation losses, which is denoted as:

$$\begin{aligned}\mathscr{H}^{(3)}(t^{(3)}, s^{(3)}) = KL\Bigg(SF\left(f_{\theta_t}^{(1)}(x^{(3)}), T\right), SF\left(f_{\theta_s}^{(1)}(x^{(3)}), T\right)\Bigg) \\ + \sum_{i=1}^{2} KL\Bigg(SF\left(f_{\theta_t i}^{(2)}(x^{(3)}), T\right), SF\left(f_{\theta_s i}^{(2)}(x^{(3)}), T\right)\Bigg)\end{aligned} \tag{A.6}$$

A.3.2.3 Batch-wise loss functions

Given a batch of data $(X, Y) = \{\{(x^{(i,n)}, y^{(i,n)})\}_{n=1}^{N}\}_{i=1}^{3}$, the parameters of teacher network and student networks are denoted as θ_t and θ_s, respectively. Since our last dataset *Affect EXPR_VA* contains annotation for both EXPR and VA, therefore, when $i = 3$ then $y^{(3,n)}$ contains both $y^{(3_{expr},n)}$ and $y^{(3_{va},n)}$.

The training teacher loss is denoted as:

$$\mathscr{F}_t(X, Y, \theta_t) = \sum_{i=1}^{3}\sum_{n=1}^{N} \mathscr{L}^{(i)}\left(y^{(i,n)}, f_{\theta_t}^{(i)}(x^{(i,n)})\right) \tag{A.7}$$

The student loss of a sample x with ground truth y from dataset i with $i \in \{1, 2, 3\}$ is denoted as:

$$\mathscr{G}_i(x, y, \theta_t, \theta_s) = \lambda \times \mathscr{L}^{(i)}\left(y, f_{\theta_s}^{(i)}(x)\right) + (1 - \lambda) \times \mathscr{H}^{(i)}\left(f_{\theta_t}^{(i)}(x), f_{\theta_s}^{(i)}(x)\right) \tag{A.8}$$

Similar to [DCS20], we also use the parameter λ to weight the supervision loss versus the distillation loss. The λ parameter is set to 0.6 to weight the ground truth slightly more than the soft labels.

The student loss is denoted as:

$$\begin{aligned}\mathscr{F}_t(X, Y, \theta_t, \theta_s) = \sum_{n=1}^{N} \mathscr{G}_3\left(x^{(3,n)}, y^{(3,n)}, \theta_t, \theta_s\right) + \sum_{i=1}^{2}\sum_{n=1}^{N}\Bigg\{\mathscr{G}_i\left(x^{(i,n)}, y^{(i,n)}, \theta_t, \theta_s\right) \\ + \sum_{j\neq i} \mathscr{H}^{(j)}\left(f_{\theta_t}^{(j)}(x^{(j,n)}), f_{\theta_s}^{(j)}(x^{(j,n)})\right)\Bigg\}\end{aligned} \tag{A.9}$$

As we have mentioned earlier, there are 164 videos that are annotated for both EXPR and VA in the Affwild2 database. Instead of treating all of these videos as if they are annotated with only one label like [DCS20], we check if the given video frame has been annotated with one or both EXPR and VA labels. Then, we compute the objective loss of the secondary task using the distillation loss alone or supervision loss plus distillation loss, respectively. Particularly, the student loss for taking into account this characteristic is denoted as:

$$\mathscr{F}_t(X, Y, \theta_t, \theta_s) = \sum_{n=1}^{N} \mathscr{G}_3\left(x^{(3,n)}, y^{(3,n)}, \theta_t, \theta_s\right) + \sum_{i=1}^{2}\sum_{n=1}^{N}\left\{\mathscr{G}_i\left(x^{(i,n)}, y^{(i,n)}, \theta_t, \theta_s\right) + \sum_{j\neq i}\left\{\begin{array}{ll}\mathscr{H}^{(j)}\left(f_{\theta_t}^{(j)}(x^{(j,n)}), f_{\theta_s}^{(j)}(x^{(j,n)})\right), & \text{if } y^{j,n} \text{ is NA} \\ \mathscr{G}_j\left(x^{(j,n)}, y^{(j,n)}, \theta_t, \theta_s\right), & \text{otherwise}\end{array}\right\}\right\} \tag{A.10}$$

A.3.3 Frame images analysis

For the video's frame images, face images with the size of 112×112 pixels are aligned and extracted from each frame. Then, we use these images to train a CNN model using the method mentioned in Section A.3.2. For this CNN model, we have selected the ResNet 50 [HZRS15] architecture as base network and added two head layers corresponding to the two outputs of the model: EXPR and VA (see Figure A.2). During training, we have applied some image-wise augmentation process with some filters to improve the performance of the model. These filters are including: random image translation [PFA20] and random image horizontal flip.

A.3.4 Temporal information exploitation

Once the CNN student model has been trained, we use this model to extract features from each video frame. Then, we group these features together to form a new dataset ds of feature's sequences with the sequence length of 32 frames per sequence. Finally,

we fed data from this new dataset ds into a bidirectional RNN network for exploiting temporal information, as well as predicting EXPR and VA. For this RNN network, we have selected the Gated Recurrent Units (GRU) architecture [CvMG+14] as it has been proven to be efficient in remembering long-term dependencies. Regarding this GRU model's parameters, we also use the training method in Section A.3.2 to train them. During the training, we have used the same augmentation process with filters that are mentioned in Section A.3.3 but in sequence level.

A.4 Experiments and Results

A.4.1 Implementation details

The whole network system is implemented using PyTorch framework [PGM+19]. During the training phase, Adam optimizer [KB17] was employed with the initial learning rate is set to $1e^{-4}$. The maximum number of epochs is 40 and the training process will stop when there is no improvement after five consecutive epochs. The number of batch size for the CNN part of the network is set to 64. For RNN network, the batch size is 16. The training and validating processes were performed on an Intel Workstation machine with a NVIDIA Gerforce RTX 2080 Ti 11G GPU.

A.4.2 Results

Here we report the results of different experiments to demonstrate the effectiveness of each of our changes comparing to the original method [DCS20]. For the evaluation metrics, we use the same criterion as outlined in [KSHZ]. Valence and Arousal estimation is based on the mean Concordance Correlation Coefficient (CCC). The seven basic expressions classification is measured by $0.67 \times F_1$ score $+ 0.33 \times$ total accuracy. For each of our experiments, we run it 10 times and report the mean of the evaluation results on the Validation set of the AffWild2 dataset.

Table A.1 shows the performance of the teacher network when training using Equation A.7 with only the first two tasks ($\mathscr{T} \in \{1, 2\}$) and with all the three tasks ($\mathscr{T} \in \{1, 2, 3\}$). From this table, we can see that when training with only two tasks, our model has already outperformed the baseline results of the competition. When we add the third task EXPR_VA into the training process

($\mathscr{T} \in \{1, 2, 3\}$), we can see that the performance of both EXPR and Valence have increased quite a lot, especially the later with 17% of improvement. Despite of having a slightly decreasing in term of Arousal (about 2%), the performance of the network has been improved in overall by a large margin, compared to the model trained without the EXPR_VA task.

Table A.1: Performance results of the teacher CNN models on the validation set of the Affwild2 database. The baseline results are provided by the ABAW 2021 competition organiser.

Method	EXPR	Valence	Arousal
Baseline	0.366	0.230	0.210
Multitask $\mathscr{T} \in \{1, 2\}$	0.498	0.374	**0.407**
Multitask $\mathscr{T} \in \{1, 2, 3\}$	**0.513**	**0.438**	0.398

After training the teacher model, we train student models with the supervision of both ground truth and the pretrained teacher model using Equation A.9 for the case of not using the shared annotations (No sharing), and using Equation A.10 for the case of using the shared annotations (With sharing). The results are shown in Table A.2. From this table, it can be seen that the performance of the model trained using the shared annotations (With sharing) is better than the one trained without using it (No sharing). This results indicate the importance of exploiting the sharing annotations in the database.

Once the student model is trained, we use this CNN model to extract features to train GRU network for exploiting temporal information. We train a teacher model using Equation A.7 and a student model using Equation A.10. Table A.3 shows the results of

Table A.2: Performance results of the student CNN models on the validation set of the Affwild2 database. The student models are trained using all three tasks $\mathscr{T} \in \{1, 2, 3\}$.

Method	EXPR	Valence	Arousal
No sharing	0.513	**0.472**	0.412
With sharing	**0.525**	0.471	0.421

Table A.3: Performance results of the CNN + GRU model. Both teacher and student models are trained using all three tasks $\mathscr{T} \in \{1, 2, 3\}$.

Method	EXPR	Valence	Arousal
Teacher model	0.555	0.523	0.543
Student model	**0.555**	**0.526**	**0.551**

these models. From this table, we can see that: the performance of the student model is equivalent to the performance of the teacher model in the task of EXPR prediction and better than the teacher model in all the other cases. When we compare the CNN + GRU model with the CNN model alone (in Table A.2), the former model outperformed the latter by a large margin.

A.4.3 Comparison with State of the art

Here we compare the performance of our model with the state of the art on the test set of Affwild2 dataset. In this 2nd challenge, the database has been updated by adding more videos and labels for the AU detection task, but since the data for EXPR recognition task and VA estimation task are almost unchanged, we are still able to compare the performance of our model with the works on the previous ABAW 2020 challenge [KSHZ].

Table A.4 shows the comparison results between the works on Affwild2 database. One can note that these results are the results of the test set of the database and have

Table A.4: Comparison with other works on the test set of the Affwild2 database

Method	Expression			CCC		
	F_1	Acc	Criterion	Valence	Arousal	Mean
Top entries to ABAW 2020:						
ICT-VIPL [ZHZ+20]	0.287	0.652	0.408	0.361	0.408	0.385
NISL2020 [DCS20]	0.270	0.680	0.405	0.440	**0.454**	**0.447**
TNT [KRO20]	**0.398**	**0.734**	**0.509**	**0.448**	0.417	0.433
Top entries to ABAW 2021:						
FlyingPigs [ZDWG21]	–	–	–	0.463	0.492	0.478
STAR [WW21]	0.476	0.732	0.560	0.478	**0.498**	0.488
Netease Fuxi Virtual Human [ZGC+21]	**0.763**	**0.807**	**0.778**	0.486	0.495	0.491
CPIC-DIR2021 [JZGX21]	0.683	0.771	0.712	–	–	–
NISL2021 (no publication)	0.431	0.654	0.505	**0.533**	0.454	**0.494**
Our model	0.351	0.668	0.456	0.505	0.475	0.490

been computed by the organiser of the competition for fair comparison. For the prior works on this dataset (ABAW 2020) we have: ICT-VIPL team [ZHZ+20] with their M^3T model, NISL2020 team [DCS20] with their multitask model trained on multiple datasets with incomplete labels and TNT team [KRO20] with their two streams aural-visual network. For the top entries to the challenge (ABAW 2021), we have: FlyingPigs team [ZDWG21] with their Audio-visual Attentive Fusion model, STAR team [WW21] with their multitask aural-visual model, Netease Fuxi Virtual Human team [ZGC+21] with their Prior Aided Streaming network, CPIC-DIR2021 team [JZGX21] with their multitask multimodal method for detecting AUs and classifying facial expressions. Apart from that, we also have the NISL2021 team, but without publicly available article. From this table, we can see that our model is significantly outperformed the original model (which we have adapted from) of the NISL2020 team [DCS20] in both of the two tasks and outperformed all other prior works in term of VA estimation. This results are clearly showing that our changes and improvements in the approach have improved the overall performance of the model significantly.

For the top entries to ABAW 2021 competition, our model has reached the third place in the VA estimation track leaderboard, over 40 teams that have participated in

this track of the challenge. Taking into account that the difference between the mean VA of our model and the best model (NISL2021) is only 0.82%, we can say that our model has reached the same performance level compared to the state of the art in term of VA estimation on the official Affwild2 database of the competition. In term of EXPR recognition, we are in the 8th place in the EXPR track leaderboard, over 55 teams that participated in this track of the challenge. The reason for this results could be because we are not using AUs annotations as the other competitors. e.g. the top two teams in this track: Netease Fuxi Virtual Human and CPIC-DIR2021, they are both trying to detect AUs beside recognise EXPR and have achieved a good performance compared to the others. There could be a strong link between action units and facial expressions that need to be identified in the future works.

A.5 Conclusion

In this paper, we have presented a method to optimise the multitask training with incomplete labels approach. On top of the original method based on teacher-student architecture, we have added a new task to train the deep neural network on a dataset that contains both seven basic expressions and valence-arousal values for better exploiting the inter-task correlations between the two tasks. In the same time, we have exploited the shared annotations inside the Affwild2 database during the training process of the student model. With these improvements, we have obtained a model that is on par with state of the art in term of valence and arousal estimation on the test set of the Affwild2 database. In future work, we will investigate about the link between action units and facial expressions, which could be the key to further improve the performance of both the facial expressions classification and valence-arousal estimation tasks.