Deep regression models for spatio-temporal expression recognition in videos

by

Gnana Praveen RAJASEKHAR

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, JUNE 2, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Eric Granger, Thesis supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Patrick Cardinal, Thesis Co-Supervisor
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Rafael Menelau Cruz, Chair, Board of Examiners
Department of Software and Information Technology Engineering, École de technologie supérieure

Mr. Matthew Toews, Member of the Jury
Department of Systems Engineering, École de technologie supérieure

Mr. François Brémond, External Examiner
INRIA Sophia-Antipolis, France

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MAY 22, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I like to express my deep sense of gratitude and appreciation to my supervisors Prof. Eric Granger and Prof. Patrick Cardinal for their invaluable support and encouragement, without whom this work certainly would not have been successful. I am greatly indebted to their constant guidance and meticulous editing, which contributed significantly to this thesis. I thank them for investing their valuable time in several discussions and enabling me to publish several research articles throughout this thesis. I am also grateful for their financial support for my thesis, as well as to attend international conferences to present my research work.

I would like to thank all my lab mates, especially Faniya, Akhil, and Shakeeb, who have helped me at crucial times during my Ph.D. On a personal note, I would like to thank my friend Madhu for many thought-provoking discussions personally and professionally, which made my Ph.D. experience memorable in my life.

I also would like to thank my Master's advisor Prof. Kannan Karthik, who instilled in me a passion for research, and Prof. R Venkatesh Babu, who introduced me to Machine Learning and enabled me to improve my research skills.

I am greatly indebted to my parents for their endless love, sacrifice, care, and supporting me in every way they can during all these years. Especially, I am grateful to my wife Priyanka for her support to pursue this research work. It would have been difficult for me to complete this thesis without her understanding and sacrifices, which enabled me to spend several nights in the lab. I also thank my son Phineas for the joy and wonderful time with him, which gave me the strength to endure tough times during the course of this thesis. Last, but not least, I would like to thank all my friends in Montreal, especially Harish, Naresh, Appa Rao, Anusha, Nikhita, Grace, Tony, and Jonathan, who have been a great support for me at critical times during the course of this journey.

Modèles de régression profonde pour la reconnaissance d'expressions spatio-temporelles d'expression dans les vidéos

Gnana Praveen RAJASEKHAR

RÉSUMÉ

La reconnaissance des expressions (RE) est un problème difficile dans le domaine de l'informatique affective, qui joue un rôle important dans la compréhension automatique des expressions et des émotions humaines. La reconnaissance des expressions peut être formulée autant comme un problème de classification que de régression. Bien que l'identification automatique des expressions exprimée comme un problème de régression joue un rôle crucial dans de nombreuses applications de santé, telles que l'estimation des niveaux de douleur et de fatigue, elle reste relativement peu explorée par rapport à la classification des expressions. La détection du niveau de fatigue est largement utilisée dans un certain nombre d'applications telles que la conduite autonome, les soins de santé et l'engagement des employés. Dans le même ordre d'idée, l'évaluation automatique du niveau de douleur a une valeur diagnostique potentielle importante pour les personnes telles que les nourrissons, les jeunes enfants et les personnes souffrant de troubles de la communication ou de troubles neurologiques. On a constaté que la fatigue est synchrone avec la douleur, une fatigue élevée étant associée à une douleur élevée, ce qui peut être constaté par la corrélation des scores analogiques visuels (VAS) de la fatigue et de la douleur. Cependant, l'expression de la douleur se produit sur une période plus courte, alors que la fatigue se manifeste sur une période plus longue.

Dans cette thèse, nous nous sommes principalement concentrés sur le développement de modèles profonds (DL) pour la reconnaissance des expressions basée sur la régression en tirant parti des relations spatio-temporelles ainsi que des modalités audio et visuelles disponibles dans les enregistrements vidéos. Les problèmes de régressions posent certains défis, tels que la capture de subtiles variations relatives à l'intensité des expressions entre deux images contiguës, les variations entre les individus et les conditions de capture, l'entraînement des modèles DL avec des vidéos faiblement étiquetées, la fusion efficace des modalités audio et visuelles, etc. Afin de d'amoindrir les effets de ces défis, nous nous concentrons sur le développement de modèles DL pour deux problèmes : (1) l'adaptation au domaine dans un contexte d'entrainement faiblement supervisée (WSDA) dans un problème d'estimation de l'intensité de la douleur, et (2) la fusion audio-visuelle (A-V) pour la reconnaissance dimensionnelle des émotions appliquée à la reconnaissance dimensionnelle des émotions.

Dans un premier temps, nous avons présenté une revue détaillée des approches d'apprentissage faiblement supervisé (WSL) dans le contexte de l'analyse du comportement facial. Afin de fournir une revue complète du domaine, nous avons également inclus l'utilisation des unités d'action (UA) en plus des expressions pour la classification et l'analyse du comportement facial. Nous avons également ajouté des unités d'action (UA) aux expressions pour les problèmes de régression. En particulier, nous avons fourni une taxonomie des méthodes existantes basées sur des scénarios WSL ainsi que leurs forces et limites respectives. Un examen des ensembles

de données largement utilisés, des protocoles expérimentaux et des résultats expérimentaux sont également présentés et discutés. Enfin, notre analyse critique de ces méthodes permet de mieux comprendre les directions de recherche potentielles pour exploiter les données faiblement étiquetées dans le cadre de l'analyse du comportement facial. Cette revue conclut que les méthodes WSL sont prometteuses pour gérer les données faiblement étiquettés dans les bases de données contenant des expressions faciales obtenus à partir de scénarios réels mais qu'elles ne sont pas suffisament explorées dans la littérature. Il y a par conséquent beaucoup de place pour améliorer la performance de l'ER faciale à partir de données faiblement annotées.

La deuxième contribution propose un nouveau modèle de DL pour l'adaptation de domaine, avec régression ordinale (WSDA-OR) afin d'estimer l'intensité de la douleur et de la fatigue dans des enregistrements vidéos, le tout dans un problème de régression. L'adaptation au domaine a été largement exploré afin d'atténuer les problèmes dus aux changements de domaines, principalement causés par des conditions de captures différentes entre les données utilisées pour l'entrainement (en laboratoire) et en production. Dans ce travail, l'adaptation au domaine est exploitée pour adapter un modèle DL à différentes personnes et conditions d'enregistrement dans le contexte où les vidéos sont faiblement annotées. Contrairement aux modèles WSL de pointe utilisés pour l'estimation de l'intensité de la douleur dans les vidéos, le modèle proposé renforce la relation ordinale entre les niveaux d'intensité de la douleur des séquences cibles en même temps que la cohérence temporelle sur plusieurs images consécutives. En particulier, il apprend des représentations de caractéristiques qui sont à la fois discriminantes et invariantes par rapport au domaine en intégrant l'apprentissage d'instances multiples avec l'apprentissage contradictoire, où des étiquettes gaussiennes sont utilisées pour représenter efficacement les étiquettes ordinales faibles au niveau des séquences du domaine cible. Les résultats expérimentaux sur les ensembles de données UNBC-McMaster, BIOVID et Fatigue (private) indiquent que l'approche proposée peut améliorer significativement les performances lorsque comparée aux modèles de pointe, ce qui permet d'atteindre une plus grande précision dans la localisation de la douleur.

En troisième lieu, un modèle d'attention croisée est proposé pour la fusion A-V pour la reconnaissance dimensionnelle des émotions basée sur les modalités faciales et vocales. La plupart des méthodes de pointe pour la fusion A-V reposent sur des réseaux récurrents ou des mécanismes d'attention conventionnels qui n'exploitent pas efficacement la nature complémentaire des modalités A-V. Dans ce travail, la relation complémentaire entre les modalités A-V est explorée afin d'extraire les caractéristiques saillantes, ce qui permet une prédiction précise des valeurs continues de valence et d'excitation. Les résultats expérimentaux sur RECOLA et Affwild2 indiquent que notre modèle de fusion A-V inter-attentionnel fournit une solution rentable qui peut surpasser les approches les plus récentes.

Les travaux décrits dans cette thèse indiquent clairement que l'adaptation efficace des modèles DL avec des vidéos faiblement étiquetées montre une amélioration significative par rapport aux méthodes de pointe précédentes dans le contexte de l'estimation des niveaux de douleur et de fatigue. En outre, ce travail a montré que l'exploitation de la relation complémentaire entre les modalités A-V joue un rôle crucial dans la fusion efficace des modalités dans le domaine de la reconnaissance dimensionnelle des émotions. Ce travail montre en outre qu'exploiter la

complémentarité entre les modalités A et V est un axe de recherche prometteur. L'approche inter-attentionnelle conjointe proposée pourrait également être améliorée en utilisant des mécanisme de porte ("gating") pour efficacement modéliser les relations intra et intermodales. De plus, l'approche proposée est plus résiliente lorsqu'une modalité n'est pas disponible.

Mots-clés: apprentissage en profondeur, apprentissage à instances multiples, adaptation domaine, évaluation de la douleur, fusion audiovisuelle, reconnaissance dimensionnelle des émotions, modèles d'attention

Deep regression models for spatio-temporal expression recognition in videos

Gnana Praveen RAJASEKHAR

ABSTRACT

Automatic expression recognition (ER) is a challenging problem in the field of affective computing, playing an important role in human behavior understanding in, e.g., human-computer interaction, sociable robots, and driver assistance. ER can be formulated as the problem of classification or regression of expressions. Though regression of expressions plays a crucial role in many healthcare applications, such as estimating pain and fatigue levels, it remains relatively less explored compared to the classification of expressions. Fatigue detection is widely used in applications such as autonomous driving and employee engagement. Similarly, automatic pain assessment has an important potential diagnostic value for infants, young children, and people with communicative or neurological impairments. Fatigue is synchronous with pain, where high fatigue is associated with high pain, which can be found with the correlation of Visual Analog Scores (VASs) of fatigue and pain. Often pain expressions happen over a shorter period of time, while fatigue happens over a longer duration.

Some of the major challenges in dealing with regression of expressions are subtle variations across individuals, ambiguity across the contiguous frames pertinent to the intensities of expressions, identity bias, and sensor capture conditions. Moreover, most deep learning (DL) models demand a huge amount of data with annotations, which requires a lot of human support with domain expertise. Therefore, leveraging DL models for the regression of expressions with limited annotations remains to be a major bottleneck. Although audio-visual fusion is expected to outperform the unimodal performance, failing to efficiently leverage the complementary relationship across the audio and visual modalities often results in poor performance. This Thesis focus on the development of DL models for two problems: (1) weakly supervised domain adaptation (WSDA) for estimating the levels of pain and fatigue and (2) audio-visual (A-V) fusion for dimensional emotion recognition.

As a first contribution, a detailed review of weakly supervised learning (WSL) approaches is presented for facial behavior analysis. To provide a comprehensive review, action units (AUs), which is defined by the fundamental actions of individual facial movements or a group of facial movements, are also included along with expressions for both classification and regression. In particular, a taxonomy of methods in the literature for different WSL scenarios has been provided, along with their respective strengths and limitations. A review of widely used public datasets, experimental protocols, and experimental results is also provided for the evaluation of these state-of-the-art methods. Finally, our critical analysis of these methods provides insight into the potential research directions to leverage weakly-labeled data for facial behavior analysis. This review concludes that although WSL methods are promising in handling the weak labels of facial expressions in real-world scenarios, they are not effectively explored in the literature, and there is much room for advancing the state-of-the-art facial ER performance given data with weak annotations.

As a second contribution, a novel DL model for WSDA with ordinal regression (WSDA-OR) is proposed to estimate the levels of pain and fatigue from videos. DA has been widely explored to alleviate the problem of domain shifts that typically occur between video data captured across various source (laboratory) and target (operational) domains. In this work, WSDA is leveraged to adapt a DL model to different persons and capture conditions when the videos are weakly annotated. Contrary to prior state-of-the-art WSL models for estimating pain intensity in videos, the proposed model enforces the ordinal relationship among the pain intensity levels of the target sequences along with the temporal coherence of multiple consecutive frames. In particular, it learns discriminant and domain-invariant feature representations by integrating multiple-instance learning with deep adversarial DA, where soft Gaussian labels are used to efficiently represent weak ordinal sequence-level labels from the target domain. Experimental results on UNBC-McMaster, BIOVID, and Fatigue (private) datasets indicate that our proposed approach can significantly improve performance over state-of-the-art models, allowing us to achieve a greater pain localization accuracy.

As a third contribution, a joint cross-attention model is proposed for A-V fusion in dimensional ER based on facial and vocal modalities. Most state-of-the-art methods for A-V fusion rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of A-V modalities. In this work, the complementary relationship across A-V modalities is effectively explored to extract the salient features, allowing for accurate prediction of continuous values of valence and arousal. Experimental results on RECOLA and Affwild2 indicate that our joint cross-attentional A-V fusion model provides a cost-effective solution that can outperform state-of-the-art approaches.

The work described in this Thesis indicates that efficiently adapting DL models with weakly labeled videos shows significant improvement over prior state-of-the-art methods for estimating pain and fatigue levels. This work shows that there is much room to further improve the proposed WSDA model to leverage the potential of DL models for unsupervised domain adaptation for the regression of expressions. This work has further shown that leveraging the complementary relationship across A and V modalities is a promising research direction for effective AV fusion. The proposed joint cross-attentional approach can also be further improved using gating mechanisms for effective modeling of intra and intermodal relationships as well as to handle corrupted modalities.

Keywords: deep learning, multiple instance learning, domain adaptation, pain assessment, audio-visual fusion, dimensional emotion recognition, attention models

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Audio
ABAW	Affective Behaviour Analysis in the Wild
AD	Action Descriptor
AMILP	Adaptive MIL Pooling
AU	Action Unit
A-V	Audio-Visual
BoW	Bag of Words
BoVW	Bag of Visual Words
BN	Bayesian Network
BLSTM	Bidirectional LSTM
BMS	Between Mean Squares
CA	Cross Attention
CNN	Convolutional Neural Network
CRNN	Convolutional Recurrent Neural Network
CE	Cross Entropy
CCC	Concordance Correlation Coefficient
DA	Domain Adaptation
DFT	Discrete Fourier Transform
DL	Deep Learning

ECG	Electro Cardio Gram
EDA	Electro Dermal Activity
ER	Expression Recognition
EMS	Error Mean Squares
FACS	Facial Action Coding System
FER	Facial Expression Recognition
GAN	Generative Adversarial Network
GCN	Graph Convolutional Network
GM	Gaussian Modeling
GRL	Gradient Reversal Layer
GT	Ground Truth
HMM	Hidden Markov Model
ICU	Intensive Care Unit
I3D	Inflated 3D CNN
ICC	Intraclass Correlation Coefficient
JCA	Joint Cross Attention
JBLSTM	Joint-BLSTM
LOSO	Leave-One-Subject-Out
LSTM	Long Short-Term Memory Network
LBP	Local Binary Pattern

LLD	Low-Level Descriptors
ML	Machine Learning
MSE	Mean Square Error
MAE	Mean Absolute Error
MFCC	Mel Frequency Cepstral Coefficients
MIR	Multiple Instance Regression
MPCNN	Multichannel Pose aware CNN
MIL	Multiple Instance Learning
MLE	Maximum Likelihood Estimation
NLP	Natural Language Processing
OR	Ordinal Regression
OPI	Observer Pain Intensity
PSPI	Prkachin and Solomon Pain Intensity
PCC	Pearson Correlation Coefficient
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
RBM	Restricted Boltzmann Machine
RCNN	Recurrent Convolutional Neural Network
SSL	Semi-Supervised Learning
SIFT	Scale Invariant Feature Transform

SEM	Structural Expectation Maximization
SGD	Stochastic Gradient Descent
STFT	Short-Term Fourier Transform
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
UDA	Unsupervised Domain Adaptation
UBLSTM	Unimodal-BLSTM
V	Visual
VAS	Visual Analog Scale
WSL	Weakly Supervised Learning
WSDA	Weakly Supervised Domain Adaptation

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

C_a	Joint correlation matrix across X_a and J
C_v	Joint correlation matrix across X_v and J
D	Dataset
d_a	Dimension of A feature representations
d_v	Dimension of V feature representations
d	Dimension of concatenated features
G_f	Feature mapping function
G_l	Mapping of feature vectors to labels of the source domain
$G_w l$	Mapping of feature vectors to weak sequence level labels of the target domain
G_d	Mapping of feature vectors to domain label
H	Predictions of hidden instances (frames) of the target domain
H_i	Hidden space of i^{th} clip
H_a	Attention maps of A modality
H_v	Attention maps of V modality
J	Deep feature vectors of joint A-V representations
K	Number of ordinal intensity levels
L_s	Source domain loss
L_T	Target domain loss
L_d	Domain classification loss

m_i	Number of instances (frames) in i^{th} bag (video clip)
N	Number of training samples
N_s	Number of training samples in the source domain
N_T	Number of training samples in the target domain
S	Source domain
T	Target domain
W_{ja}	Learnable weight matrices across A and J
W_{jv}	Learnable weight matrices across V and J
W_{ca}	Learnable weight matrices for C_a
W_a	Learnable weight matrices for X_a
W_{cv}	Learnable weight matrices for X_{cv}
W_v	Learnable weight matrices for X_v
W_{ha}	Learnable weight matrices for H_a
W_{hv}	Learnable weight matrices for H_v
X	Training data
X_i	i^{th} Bag (video clip) of training data
X_a	Deep feature vectors of A modality
X_v	Deep feature vectors of V modality
X_a^j	j^{th} Feature vector of A modality
X_v^j	j^{th} Feature vector of V modality

$X_{att,a}$	Attended features of A modality
$X_{att,v}$	Attended features of V modality
x_i^t	t^{th} instance (frame) of i^{th} bag (video clip)
$\widehat{X}$	Concatenated AV representations of $X_{att,a}$ and $X_{att,v}$
Y	Labels
Y_i	Label of i^{th} bag
θ_f	Parameters of the function G_f
θ_l	Parameters of function G_l
θ_{wl}	Parameters of function G_{wl}
θ_d	Parameters of G_d
μ_x	Mean of predictions
μ_y	Mean of ground truth
σ_y^2	Variance of ground truth
σ_x^2	Variance of predictions
σ_{xy}^2	Covariance of ground truth and predictions
ρ_c	Concordance correlation coefficient between ground truth and predictions

INTRODUCTION

Affective computing is an emerging research area, which deals with the study and development of systems that can recognize, interpret, and simulate human emotions. This area is interdisciplinary, spanning computer science, psychology, and cognitive science to understand the emotional state of an individual (Calvo, D'Mello, Gratch & Kappas, 2015). Human emotions can often be conveyed through various modalities such as face, voice, text, physiology, etc. Of all the modalities, facial and vocal expressions are the predominant contact-free channels, which carry a complementary relationship with each other (Shivappa, Trivedi & Rao, 2010). Expressions indicate the emotions being felt i.e., expressions display a wide range of modulations across face and voice but human emotions are limited (Matsumoto & Hwang, 2011). Emotions represent high-level information about the mood of the person, while expressions convey low-level information about the emotions being expressed. Expression Recognition (ER) plays a crucial

Figure 0.1 Examples of primary universal emotions. From left to right: neutral, happy, sad, fear, anger, surprise, disgust
Adapted from Compound facial expressions of emotion database Du *et al.* (2014)

role in the automatic understanding of human emotions. It is used to assess the affective health or emotional state of individuals such as anger, fatigue, depression, pain, motivation, and stress in health care, e-learning, security, etc. Recognizing expressions is a challenging problem in real-world scenarios as human expressions are often diverse in nature across individuals (Green & Guo, 2018), cultures (Chen & Jack, 2017), and sensor capture conditions (Kong, Suresh, Soh & Ong, 2021). Ekman and Fries conducted a cross-cultural study on facial expressions,

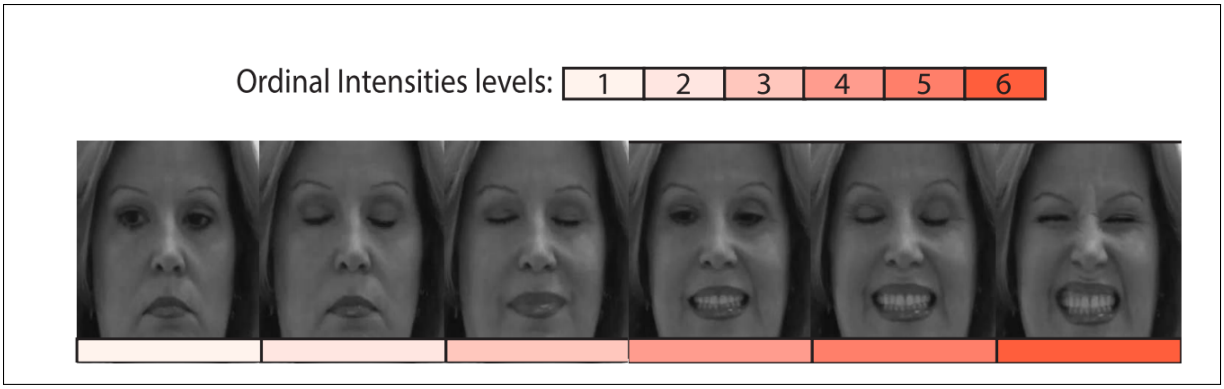

Figure 0.2 Ordinal pain intensity levels
Adapted from UNBC-McMaster database Lucey *et al.* (2011)

showing that there are six basic universal emotions across human ethnicity and cultures – Anger, Disgust, Fear, Happiness, Sadness, and Surprise (Ekman & Friesen, 1976) as shown in Figure 0.1. Subsequently, Contempt has been added to these basic emotions (Matsumoto, 1992). Given the simplicity of discrete or categorical representation, these seven prototypical emotions are the most widely used categorical model for the classification of emotions.

Though emotion classification was widely explored, ER has also been formulated as a regression problem to model the wide range of human expressions. In the case of regression, ER can be further categorized as the problem of ordinal regression or dimensional regression. Ordinal regression deals with the estimation of discrete ordinal or intensity levels of expressions such as pain intensity levels, depression levels, etc. as shown in Figure 0.2. Dimensional regression is the task of estimating the wide range of expressions on a continuous scale of valence and arousal as shown in Figure 0.3. Valence reflects the wide range of emotions in the dimension of pleasantness, from being negative (sad) to positive (happy). In contrast, arousal spans a range of intensities from passive (sleepiness) to active (high excitement). Dimensional modeling of emotions is more challenging than the categorical or ordinal case since it is difficult to obtain a continuous scale of annotations compared to discrete emotions. Given the continuous range

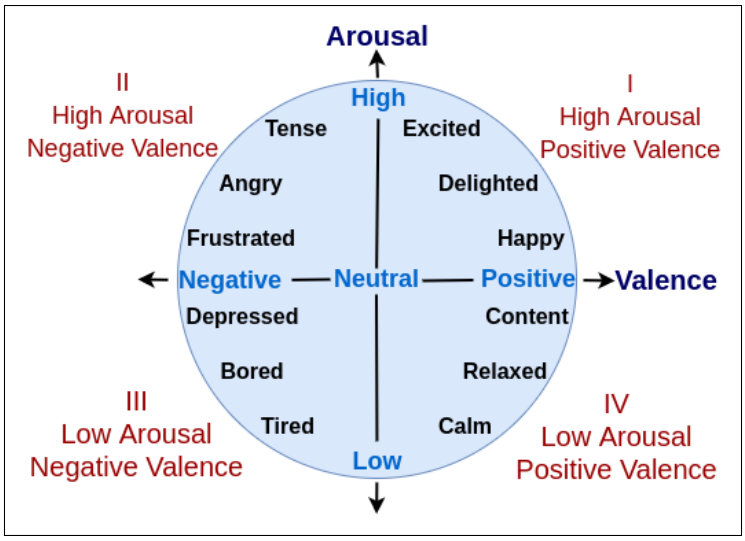

Figure 0.3 Dimensional emotion recognition in valence-arousal space
Taken from Praveen *et al.* (2023a)

of emotions, the annotations tend to be more noisy and ambiguous compared to the tasks of ordinal regression and classification.

ER systems can be divided into two main categories: image-based ER and video-based ER. In recent years, the development of ER systems has evolved from systems that perform image analysis under controlled laboratory conditions to video-based recognition under more challenging real-world scenarios. In image-based ER, only the spatial information is used to encode the image, whereas video-based ER exploits both the spatial and temporal relationship across the contiguous frames to obtain the feature representations. Leveraging the temporal dynamics of the evolution of expressions plays a pivotal role in developing a robust ER system for videos.

With the advancement of deep learning (DL) architectures, there has been significant progress in the performance of audio (A) and visual (V) recognition systems. One of the major advantages of DL models is end-to-end training, where the features and the classifier/regressor can be trained

together in an end-to-end fashion. This helps to execute the feature learning by itself without the need to explicitly hand-engineer the features. Since features are learned automatically depending on the task at hand, it helps to obtain more robust features tailor-made for the specific task at hand and thereby results in a high level of accuracy. Inspired by their performance, several approaches have been proposed in recent years for video-based ER using CNNs (Gavade, Bhat & Pujari, 2021; Li, Wen & Qiu, 2023) and Vision Transformers (ViTs) (Chaudhari, Bhatt, Krishna & Mazzeo, 2022; Ma, Sun & Li, 2021). They have shown significant improvement over the classical ML methods, which rely on hand-crafted feature extraction and classifier/regressor (Noor *et al.*, 2020; Abdulrahman & Eleyan, 2015). However, the performance of these DL models is constrained by the quality and quantity of representative annotated data. The need for a large amount of data acquisition demands the requirement of annotations. The labeling process of such training data demands much human support with strong domain expertise for the expressions. Moreover, the labeling process is highly vulnerable to the ambiguity of expressions due to the bias induced by domain experts. This Thesis focuses on developing robust and accurate DL models for video-based ER that perform regression based on weakly-labeled or unlabeled video data.

0.1 Motivation

ER has been widely used in many applications, such as estimating customer or student motivation and engagement levels in business or education settings respectively (Yang, Wang, Peng & Qiao, 2018), detecting fatigue and stress levels for driver assistance applications (Qiang Ji, Zhiwei Zhu & Lan, 2004), and assessing the level of depression or pain in healthcare (Tavakolian, Bordallo Lopez & Liu, 2020). ER systems for automatic estimation of fatigue or pain are relevant in the healthcare domain. Fatigue is a subjective feeling of tiredness reported by the patient rather than an objective one, which can be observed externally. It can be caused due to mental or physical stress that prevents a person from being able to function normally

Figure 0.4 Facial expressions of pain (left) and no pain (right)
Taken from Bellantonio *et al.* (2017)

(Adão Martins, Annaheim, Spengler & Rossi, 2021). Even though fatigue is quite common in many cases and alleviated by periods of rest, it is the most prevalent and primary symptom for most diseases, especially neurological disorders. Similarly, pain is also subjective in nature, usually self-reported by patients, either through clinical inspection, on a linear scale from 0 (no pain) to 10 (severe pain), or using the Visual Analog Scale (VAS) (Martinez, Rudovic & Picard, 2017). However, self-reported pain assessment is vulnerable to bias induced by the individual's perception of pain as shown in Figure 0.4. In general, pain and fatigue are highly related to each other, and coexist as primary symptoms of many diseases, which helps to diagnose and alleviate the intensity of major health problems in advance. Hence, in clinical applications, automatic detection of fatigue or pain has an immense need as it will be tedious to monitor the patients manually for a longer duration. Unlike images, videos carry a lot more information pertinent

to subtle expressions across the spatial and temporal dimensions. Though frames pertinent to fatigue exist for a longer duration, expressions related to pain occur only for fewer frames in the video. Many frames will be neutral in videos of pain expressions, which results in imbalanced data and poses a major challenge in leveraging the performance of DL models. Given the limited amount of relevant data, designing an ER system remains a challenging task as it will be difficult to capture the wide range of variations among subjects as well as sensor capture conditions. In addition to that, it requires experts for labeling, and the obtained labels are highly vulnerable to label ambiguity due to the bias induced by domain experts. Often only partial or weak labeling information is provided for the data e.g., video tags, as it involves costly manual intervention and high complexity to obtain labels for entire data.

Another challenging problem is to effectively fuse multiple modalities in order to achieve robust performance for ER. A-V fusion is one of the promising research directions, which can outperform unimodal performances by leveraging the complementary relationship across each other. For instance, the A modality can be leveraged to estimate the emotional state when the facial modality is missing due to pose, blur, low illumination, etc. Similarly, during silent regions in the A modality, the rich information in the V modality can be leveraged. Though A-V fusion has been widely explored in literature (Tzirakis, Trigeorgis, Nicolaou, Schuller & Zafeiriou, 2017; Schoneveld, Othmani & Abdelkawy, 2021), efficiently capturing the complementary relationship across the A and V modalities for ER remains to be a challenging problem, which is crucial for effective A-V fusion to develop a robust ER system that outperforms unimodal performances. Therefore, this Thesis explores DL models for video-based ER to leverage the rich spatiotemporal information based on A and V modalities captured in videos for estimating pain and fatigue levels. Specifically, this Thesis focuses on the development of DL models for two problems in video-based ER.

Firstly, the problem of expression intensity estimation is explored in the framework of weakly supervised learning (WSL) to address the complex process of labeling data, while still leveraging the potential of state-of-the-art DL architectures. Specifically, video-based pain intensity estimation has been addressed in the context of multiple instance learning (MIL) (Sikka, Dhall & Bartlett, 2014). Obtaining annotations for all the frames in the videos requires much human support with strong domain expertise, which is difficult to obtain in real-time environments. Moreover, the labeling process is highly vulnerable to the ambiguity of expressions, especially when labels are intensities, due to the bias induced by the domain experts. Therefore, there is a growing demand for the development of automatic pain estimation systems to ensure effective treatment and ongoing care. Given the cost and challenges of annotating data, techniques for WSL are very appealing as they allow exploiting of weak labels to train DL models. WSL can be applied in scenarios involving incomplete supervision, inexact supervision, and ambiguous or inaccurate supervision (Zhou, 2018). The inexact supervision scenario is relevant to our application, where training data sets only require global annotations for an entire video, or periodically for video sequences. MIL is one of the widely used approaches for inexact supervision (Carbonneau, Cheplygina, Granger & Gagnon, 2018). However, existing MIL-based approaches for automatic pain intensity estimation are based on traditional ML approaches due to the lack of sufficient data as facial expressions pertinent to pain are sparse in nature. They fail to leverage the potential of DL models to improve the performance of pain intensity estimation with weak annotations. This Thesis explores the prospect of training DL models with limited annotations using MIL as well as adapting DL models to different operational capture conditions through weakly supervised domain adaptation. Though MIL is used for both bag-level and instance-level prediction, our primary focus is on instance-level prediction for pain localization in videos.

Second, the problem of fusing A and V modalities is explored for dimensional ER, where human emotions are estimated in the valence-arousal space. Multiple modalities often provide diverse

and comprehensive information, which is not available in individual modalities. Therefore, effectively fusing the A and V modalities is expected to outperform the performance of individual modalities. Most of the existing approaches (Tzirakis *et al.*, 2017; Schoneveld *et al.*, 2021; Ortega, Cardinal & Koerich, 2019) in the literature that combines facial and vocal channels for dimensional ER focus on intra-modal relationships for multi-modal feature representations. Although inter-modal relationships play a crucial role in capturing the complementary relationship across the modalities, it is not effectively explored in the literature. In recent years, few approaches (Tzirakis, Chen, Zafeiriou & Schuller, 2021; Parthasarathy & Sundaram, 2021) have explored to capture the inter-modal relationships based on cross-modal attention of A and V modalities using transformers. By leveraging the cross-modal interactions across the modalities, transformers based on cross-modal attention has made significant improvement over the prior approaches, that focus only on intra-modal relationships. However, they are limited by their ability to effectively capture both inter-modal complementary relationship as well as intra-modal relationships among A and V modalities to improve the fusion performance over that of uni-modal approaches. This Thesis seeks to develop DL models that can leverage complementary relationships across A and V modalities, while still retaining the intra-modal relationships to improve the performance of the system.

0.2 Research Objectives and Contributions

Following the challenges and limitations highlighted above, the objective of this research is to develop DL models for expression behavior analysis in videos. Specifically, two research directions have been explored pertinent to pain localization through weakly labeled domain adaptation, and attention-based A-V fusion for dimensional ER. The main contributions of this Thesis are summarized as follows:

1) A comprehensive survey of weakly supervised learning models for facial behavior analysis: First, an exhaustive survey on WSL models for facial behavior analysis has been

provided including facial expressions and action units. This work investigates and highlights the research gaps in the literature along with potential research directions to develop robust models for facial behavior analysis. Effectively leveraging the deep models by capturing the relevant data with weak annotations was shown to significantly improve the performance of the system. Further details can be found in Chapter 2. This resulted in our first contribution along with the detailed findings of our comprehensive survey on existing methods in the following paper:

- **R Gnana Praveen**, Patrick Cardinal, Eric Granger "Weakly Supervised Learning for Facial Behavior Analysis: A Review" IEEE Transactions on Affective Computing (TAC), 2022 (Under Review).

2) Automatic pain localization using weakly labeled videos with domain adaptation: Most of the approaches in the literature fails to leverage the efficacy of DL models due to the lack of relevant training data, limited annotations, etc. Though few approaches have exploited DL models with fully supervised learning for pain localization in videos, developing prediction models with weak labels using DL models was found to be a challenging problem. The second contribution of this Thesis is to investigate the prospect of using DL models with domain adaptation to build robust prediction models for weakly labeled videos of video-level annotations. To achieve this goal, weakly supervised domain adaptation (WSDA) has been presented, which also helps to deal with variations in operational capture conditions. The proposed framework of WSDA has been applied for pain localization using weakly labeled videos. Experimental results on UNBC-McMaster, Biovid, and Fatigue data sets significantly outperforms the state-of-the-art models, resulting in greater pain localization accuracy. Further details can be found in Chapter 3. The second contribution resulted in the following papers:

- **R Gnana Praveen**, Eric Granger, Patrick Cardinal "Deep Weakly Supervised Domain Adaptation for Pain Localization in Videos" IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020.

- **R Gnana Praveen**, Eric Granger, Patrick Cardinal "Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos" Image and Vision Computing, vol 110, pp 104167, 2021.

3) Audiovisual fusion for dimensional emotion recognition: Conventional approaches to A-V fusion rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of A-V modalities. Even though transformer models are explored to capture the inter-modal relationships using cross-modal attention, they are limited by their ability to leverage the intra-modal relationships. Most of these approaches fail to effectively capture both the inter-modal and intra-modal relationships across the A and V modalities. This led to the third contribution to investigate the prospect of effectively exploiting the complementary relationship across A and V modalities to improve the performance of the system. It has been found that efficiently capturing the complementary relationship across A and V modalities significantly improves the performance of the system. To capture the A-V relationships effectively, a joint cross-attentional fusion model is proposed for dimensional ER. Experimental results indicate that the joint cross-attentional A-V fusion model provides a cost-effective solution that can outperform state-of-the-art approaches, even when the modalities are noisy or absent. Further details can be found in Chapter 4. The third contribution resulted in the following papers:

- **R Gnana Praveen**, Eric Granger, Patrick Cardinal "Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition" IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2021.
- **R Gnana Praveen**, Patrick Cardinal, Eric Granger "Audio-Visual Fusion for Dimensional Emotion Recognition Using Joint Cross-Attention" IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM), 2023.
- **R Gnana Praveen**, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, Patrick Cardinal,

Eric Granger; "A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition" IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022.

- **R. Gnana Praveen**, Patrick Cardinal, Eric Granger "Recursive Joint Attention for audio-visual fusion in regression-based emotion recognition," 48th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

The full list of publications that resulted from this research can be found in Appendix II.

0.3 Thesis Outline

Figure 0.5 presents an overview of the organization of this Thesis. **Chapter 1** provides the background on ER from videos, related to facial behavior analysis in videos as well as A-V fusion for dimensional ER. A detailed review of relevant works on facial behavior analysis in videos and A-V fusion for dimensional ER has been presented, followed by limitations of the literature. **Chapter 2** provides a comprehensive review of WSL models for facial behavior analysis. This chapter briefly introduces the framework of WSL and its relevance for facial behavior analysis. The literature on facial behavior analysis in videos as well as images is rigorously analyzed with weak annotations. Finally, the research gaps in the literature are highlighted with new potential research directions. This work corresponds to the paper "Weakly Supervised Learning for Facial Behavior Analysis: A Review" which was submitted to IEEE Transactions on Affective Computing (TAC). **Chapter 3** then introduces a new method for WSDA to adapt DL models for automatic pain localization in videos using weakly labeled videos. This work corresponds to the paper "Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos" published in the Image and Vision Computing (IVC) journal. **Chapter 4** introduces our framework of A-V fusion for dimensional ER based on videos. This work seeks to develop DL models that leverage the inter-modal relationships across

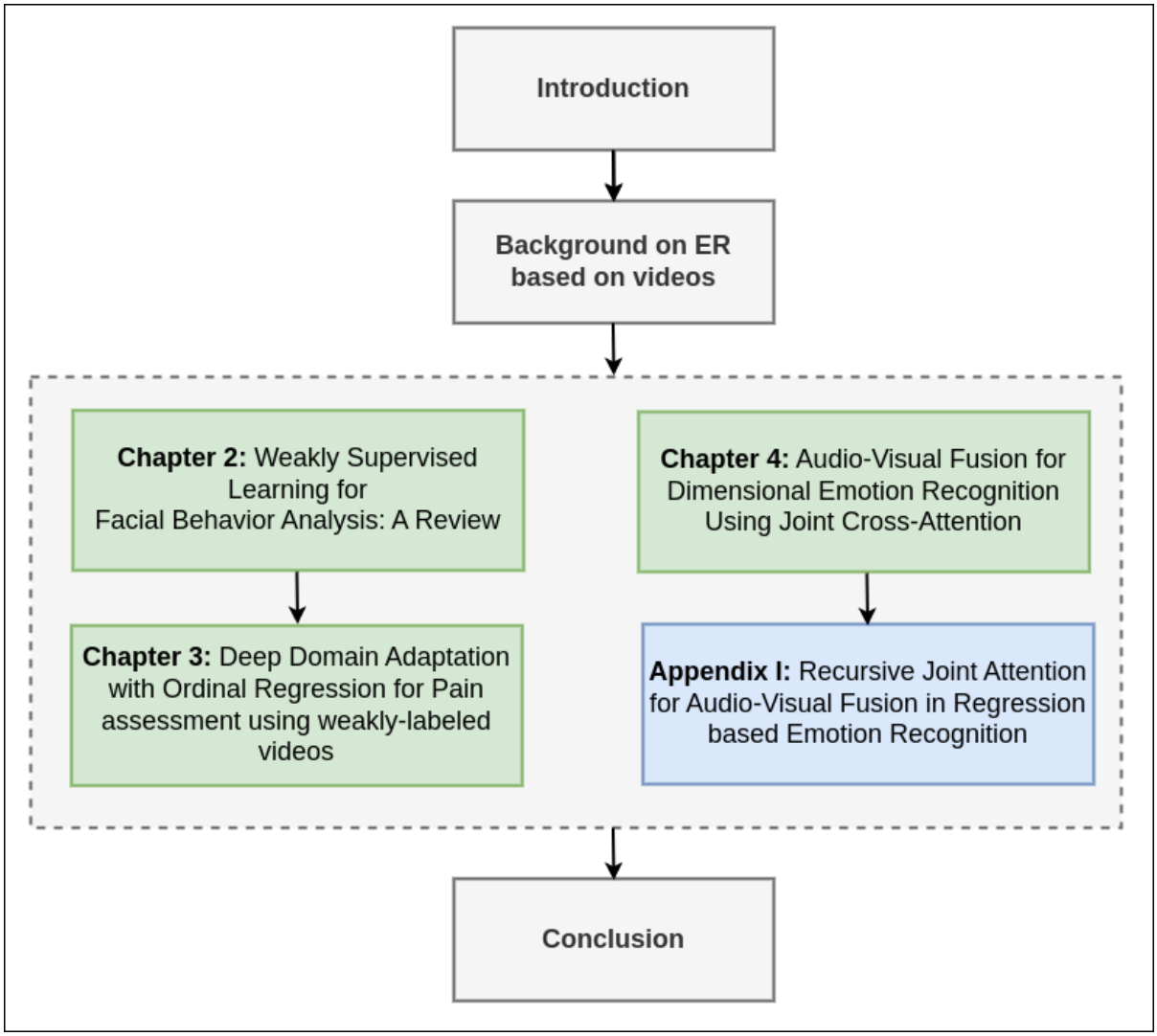

Figure 0.5 Organization of this Thesis. The arrows indicate the dependencies between the chapters and appendices

the A and V modalities and showed significant improvement over the existing approaches. This work corresponds to the paper "Audio-Visual Fusion for Dimensional Emotion Recognition Using Joint Cross-Attention" which has been published in the IEEE Transactions on Biometrics, Behaviour, and Identity Science (BIOM) journal. Appendix I presents recurrent joint attention for A-V fusion in regression-based ER, which relies on recursive fusion and LSTMs to further improve the joint cross-attentional model for A-V fusion. **Chapter 5** summarizes the major contributions of this dissertation and discusses its limitations along with potential future research

directions. The computational complexity of the proposed Joint Cross Attentional (JCA) A-V fusion model is provided in Appendix II. Finally, Appendix III provides the list of publications, obtained during this Ph.D. study.

CHAPTER 1

BACKGROUND ON EXPRESSION RECOGNITION

This chapter introduces the basic concepts of video-based ER. In this Thesis, we focus on DL models for ER using facial expressions as the visual modality, and vocal expressions as the audio modality.

1.1 Machine Learning Models

Machine Learning (ML) is the automation of the learning process of machines without human intervention. In simple words, it is nothing but imitating human intelligence in machines so that they can be able to perform tasks without being explicitly programmed for any specific task. In recent days, ML has been found to be so pervasive that we use it in most of our daily activities even without our knowledge. The process of automation is achieved using data, through which a model can be trained and used to make predictions on new data. Therefore, the performance of the ML system depends on the training model, which in turn relies on the quality of the data. The advancement of computing capabilities in handling huge data has fostered the deployment of ML systems in real-time applications. Depending on the availability of labels, ML systems are broadly classified into three categories: Supervised Learning, Unsupervised Learning, and Weakly Supervised Learning.

- **Supervised Learning:** The data is provided with their desired labels, where the training model is developed by reducing the error between the actual targets with the predicted ones. Some of the most widely used supervised learning approaches are Support Vector Machines (SVM), Linear regression, etc.
- **Unsupervised learning:** The data is not provided with corresponding labels. So it finds the commonalities among the unlabeled input data and estimates the hidden pattern in the structure of the data. Since it is difficult to obtain labeled data in many cases, unsupervised learning was found to be highly valuable in analyzing the hidden patterns of the data and for efficient data clustering and representation. Some of the commonly used unsupervised techniques are k-means clustering, Hierarchical Cluster Analysis, etc.

- **Weakly Supervised Learning:** It refers to the class of ML algorithms, where the data is provided with partial or noisy labels. The training model is developed using partial or noisy labels without the actual labels. Some of the widely used WSL approaches are Multiple Instance Learning (MIL), Active Learning, etc. The category of WSL models is discussed in detail in Section 1.5.

The core idea of ML models is to learn the patterns underlying the data. ML algorithms can be broadly classified into two major categories based on the mode of learning from data: Traditional ML approaches and Deep ML approaches as shown in Figure 1.1.

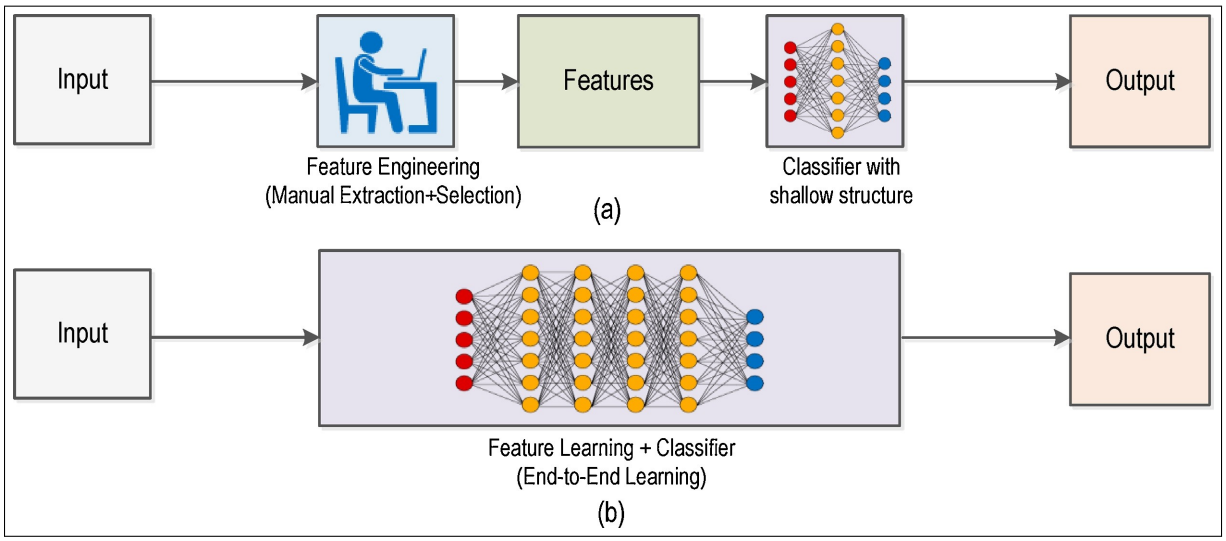

Figure 1.1 Demonstration of (a) Traditional Machine Learning and (b) Deep Learning Taken from Wang *et al.* (2018)

1.1.1 Traditional ML Approaches

Traditional ML belongs to the class of approaches, where the features are hand-engineered and the prediction models of classification or regression are trained separately. In traditional ML approaches, the features are chosen by subject experts, and the underlying patterns of the features are learned by the chosen prediction model. Traditional ML models often work better with small data captured under constrained environments. Since the features are directly

hand-engineered and the prediction models are decoupled from the features, the traditional ML models are deterministic and easier to interpret. Training the prediction models of traditional ML algorithms is not very expensive and takes less training time as they are decoupled from feature learning. Some of the widely used traditional ML algorithms are Support Vector Machine (SVM) (Cortes & Vapnik, 1995), Adaboost, Logistic Regression, and Random Forests. SVM is one of the most widely used supervised learning algorithms before the DL era. The basic idea of SVM is to estimate the hyperplane, which can separate the data points of different classes. The parameters of the hyperplane are optimized and learned by maximizing the distance of the hyperplane to the nearest data points of different classes. These nearest data points are termed "support vectors" as they support determining the decision boundary. Due to their robust performance and ability to be generalized in high-dimensional spaces, several variants have been subsequently proposed and successfully explored for several applications in computer vision, speech processing, and NLP. SVMs have also been explored along with deep features and found to be promising for several applications (Tang, 2013).

Another class of traditional ML models widely explored in the literature is ensemble learning, where the idea is to combine the predictions of multiple models to produce an optimal estimation. Of all the ensemble learning models, two major classes of ensemble learning models are bagging and boosting. Bagging refers to bootstrap aggregating, which involves a diverse group of individual learners by varying the amount of training data based on bootstrap sampling (sampling with replacement). Decision trees belong to the class of bagging, out of which random forests (Breiman, 2001) is the most widely used technique in the literature due to the low correlation and diverse nature of the base learners (trees). Random Forests is a variant of the tree-based model, where the basic idea is to leverage a large number of relatively uncorrelated and diverse models (trees) to produce ensemble predictions that are more accurate than individual predictions. The idea of bagging is also explored along with SVMs, where multiple SVMs are trained independently and the outputs are combined via majority voting (Kim, Pang, Je, Kim & Bang, 2002).

On the other hand, boosting relies on a set of weak learners, where the weak learners are combined progressively in a sequential fashion to form a strong learner. Some of the most popular boosting techniques are Adaboost (Freund & Schapire, 1996) and Gradient Boosting (Friedman, 2001), which has been successfully used across different domains (Viola & Jones, 2001; Nguyen, Ng & Nguyen, 2012). Adaboost explores a greedy approach by iteratively adjusting the weights based on the misclassified points to minimize the training error. Gradient boosting extends this idea by generalizing this framework using arbitrary differential loss function while Adaboost minimizes exponential loss function.

Apart from the above-mentioned approaches, there are also other successfully used traditional ML algorithms such as Naive Bayes Classifier, Logistic Regression, etc. Even though traditional ML algorithms are found to be successful in many domains, they often fail to capture the complex patterns underlying the distribution of data in real-world environments. Due to the shallow learning of prediction models, they tend to get saturated in limited performance, despite the availability of huge amounts of data. Moreover, the features need to be chosen by domain experts to reduce the complexity of data and make patterns visible for the learning algorithms to work. Due to the shallow learning of features, the performance of the prediction models is constrained by the limited learning capability of the feature representations.

1.1.2 Deep ML Approaches

In recent days, with the advancement of DL models, several DL-based techniques were found to be quite promising in handling complex real-world problems. The advancement of computing power and availability of massive amounts of data have revolutionized the advancement of neural networks over the past few years, drastically outperforming the traditional ML models in several domains. One of the major advantages of DL models is the ability to automatically learn feature representations, tailor-made for the specific task at hand. By learning the features in a cascaded fashion, DL models are able to obtain deeper feature representations, which can effectively capture complex data patterns in real-world conditions, resulting in drastic improvement in system performance. Another major advantage of DL models is the adaptability

and generalization capability of the features to novel tasks using transfer learning, where the pre-trained models can be fine-tuned and adapted to novel datasets or tasks. Moreover, DL models are highly scalable i.e., they can leverage the massive amounts of available data, yielding better feature representations to improve the performance of the system.

The major breakthrough in the field of DL has been achieved by AlexNet (Krizhevsky, Sutskever & Hinton, 2012), which has drastically improved the performance of the system on ImageNet ILSVRC challenge (Russakovsky *et al.*, 2015). Since then, a lot of researchers have explored the potential of DL models, significantly improving the performance of DL systems, by advancing the DL architectures, optimization techniques as well as loss functions. Inspired by the performance of AlexNet (Krizhevsky *et al.*, 2012), several variants of CNN architectures such as VGG (Simonyan & Zisserman, 2015), ResNet (He *et al.*, 2016), Inception (Szegedy *et al.*, 2015) are proposed in computer vision and achieved great results surpassing the performance of humans. The 2D CNN architectures have also been further extended to 3D CNNs (Tran, Bourdev, Fergus, Torresani & Paluri, 2015) to learn the spatiotemporal patterns and showed improvement in video-based applications.

Another widely used architecture to capture the temporal patterns in DL is the LSTM (Hochreiter & Schmidhuber, 1997) (a variant of RNNs), which learns the temporal patterns based on recurrent connections. Although 3D CNNs are efficient in capturing short-term temporal dynamics, they fail to capture long-term temporal patterns due to the computational complexity of 3D CNNs. On the other hand, LSTMs are effective in capturing long-term patterns using recurrent connections. Recently, transformers (Vaswani *et al.*, 2017) have replaced LSTMs to capture the temporal context of sequence patterns, which is a major breakthrough in the field of NLP. Following the success of transformers in NLP, it has gradually evolved achieving good results in other fields also. For instance, vision transformers (Dosovitskiy *et al.*, 2021) have been widely explored in computer vision applications, achieving state-of-the-art results. Despite the rapid advancement of DL systems in various domains, it is often constrained by the availability of massive amounts of data and computational resources.

1.2 Expression Recognition Systems

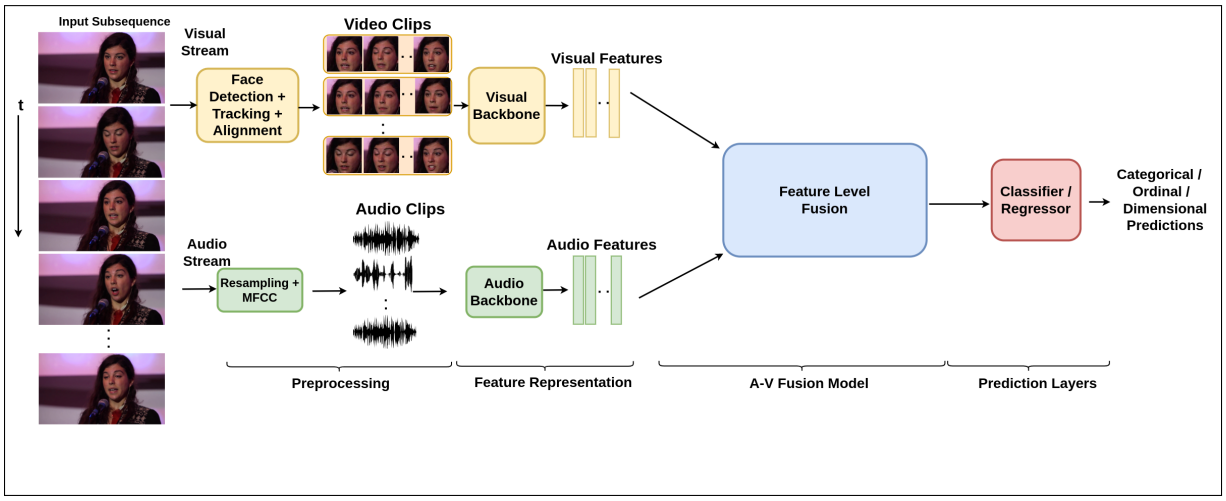

Figure 1.2 Block diagram of an Audio-Visual (A-V) Expression Recognition (ER) system. In this case, the system relies on facial expressions (visual modality) and vocal expressions (audio modality), and feature-level A-V fusion

Automatic recognition of emotions is an important task that facilitates natural interaction between humans and machines. Depending on the type of labels, emotion recognition can be formulated as a discrete classification problem (e.g., a person eliciting happy or sad emotions), or as a continuous regression problem (e.g. continuous values of valence and arousal). Though classification conveys the type of emotion being expressed, it fails to capture the wide range of emotions on a finer granularity. Valence and arousal are widely used for estimating emotion intensities in the continuous domain, where valence spans a wide range of emotions from sad to happy, and arousal reflects the energy or intensity of the emotions. Although human emotions can be expressed through various modalities such as the face, text, voice, and physiological signals, vocal and facial modalities are the predominant contact-free channels through which they can be efficiently expressed. There are four major building blocks to develop an automatic recognition system for ER, as shown in Figure 1.2.

1.2.1 Preprocessing

Visual modality: Face detection is the first and foremost step of an end-to-end ER system. The objective is to localize and extract the face region in the video frames. One of the most widely used approaches for face detection prior to the DL era is the Viola-Jones face detector (Viola & Jones, 2001), which used Ada-boost classifiers with Haar features to build a cascaded structure. With the advancement of DL architectures, several approaches have been proposed for face detection and significantly improved the performance of the system. (Li, Lin, Shen, Brandt & Hua, 2015) introduced a CNN-based calibration stage to improve the localization accuracy and also operates at multiple resolutions. This approach has been further extended by (Zhang, Zhang, Li & Qiao, 2016) by formulating face detection as well as alignment in a joint learning framework. They further improved the system by leveraging carefully designed cascaded architecture and an online hard-sampling mining strategy. Inspired by the performance of R-CNN models (Ren, He, Girshick & Sun, 2015) for object detection, (Jiang & Learned-Miller, 2017) explored faster-RCNN (Ren *et al.*, 2015) for face detection and achieved state-of-the-art results. Following the success of Faster-RCNN for face detection, several subsequent works (Wu, Yin, Wang & Xu, 2019; Sun, Wu & Hoi, 2018b) have been proposed using R-CNNs and showed significant improvement. Given the detected face, face alignment also plays a key role in extracting robust features to reduce the scale and in-plane rotation. One of the widely used approaches for face alignment is based on facial landmarks.

Most of the works based on facial landmarks employ cascaded regression (Lv, Shao, Xing, Cheng & Zhou, 2017; Sun, Wang & Tang, 2013; Zhang, Shan, Kan & Chen, 2014a) or RNNs (Trigeorgis, Snape, Nicolaou, Antonakos & Zafeiriou, 2016; Xiao *et al.*, 2016) to progressively refine the predictions of landmark coordinates. In contrast to cascaded regression, (Newell, Yang & Deng, 2016) designed a stacked hourglass network to estimate the landmarks based on heat maps and showed great success. This idea has been further explored by many subsequent works (Huang, Deng, Shen, Zhang & Ye, 2020; Yang, Liu & Zhang, 2017; Wang, Bo & Fuxin, 2019) and greatly improved the performance of the system. Contrary to landmark-based approaches, some of the works (Hayat, Khan, Werghi & Goecke, 2017; Zhong, Chen & Huang, 2017)

have explored spatial transformer networks (Jaderberg, Simonyan, Zisserman & kavukcuoglu, 2015) to optimize the face alignment along with face representation in a joint framework. Illumination and contrast variations are other challenging factors that can result in large intra-class variances, especially in unconstrained environments. Histogram Equalization is widely used for normalizing the variations in illumination and contrast by increasing the global contrast of the images. Various filtering-based approaches such as homomorphic filtering, Difference-of-Gaussian (DoG) filtering are also explored for effective face normalization. Finally, all the detected faces have to be normalized to a standard image size for effective feature representation.

Audio modality: Prepossessing the speech signal is an initial step for the development of any speech-based application. It suppresses the silence regions and noise in the signal such as background noise, cross-talk, etc, and amplifies the significant regions of the speech signal i.e., voiced segments for efficient feature extraction (Deb & Dandapat, 2019a). The first step is sampling, which is the process of converting the analog signal to a digital signal. Typically, sampling is performed by taking the discrete samples of the signal corresponding to the frequencies which is more than twice the maximum frequency of the signal, which is called the Nyquist rate. Since most of the sounds produced by the speech signal correspond to the range of 100 Hz - 4KHz, the sampling rate of the speech signal is considered to be 8KHz, which is twice the maximum frequency of the signal (4 KHz).

Pre-Emphasis is the process of enhancing the high-frequency components and decreasing the amplitude of low-frequency components. It is observed that high-frequency components carry significant information about the speech signal which corresponds to the voiced regions compared to the low-frequency components. It improves the overall signal-to-noise ratio of the signal and balances the magnitude of the frequency spectrum as low-frequency components have higher magnitudes compared to high-frequency components. Typically, pre-emphasis is carried out using a first-order filter for filtering operation. In recent days, pre-emphasis does not have significant importance in modern systems with the advancement of computing capabilities and can be replaced with simple normalization.

Normalization is primarily used to control the variations in the speech signal. In general, normalization is carried out using the energy of the signal, which is useful for phoneme discrimination. The objective is to discard unwanted energy variations of the speech signal due to background noise, different loudness levels in multiple speakers, etc. The most prevalent approach is to normalize the cepstral coefficients of the speech signal, which carries the energy of the signal. This is known as 'Cepstral Mean Normalization', which simply subtracts the mean of each coefficient from all the frames. In addition to cepstral mean normalization, there are also several other efficient approaches for the normalization of speech signals. Since speech signals are highly stochastic in nature, it will be difficult to analyze the entire signal as a whole, which led to the requirement of the segmentation of the speech signal as it is assumed to be stationary within a short duration of the speech signal.

The short segment of the speech signal is also called as 'frame'. Typically the frame length of the speech signal is considered to be 20 - 40 msec. If the frame length is much shorter, then we may fail to get enough spectral estimates for efficient feature extraction. On the other hand, longer segments of the speech signals will become stochastic, which will be difficult to analyze the signal. Many windowing techniques have been proposed for framing speech signals. In practice, these windowing operations will have a transient response at the borders of the windows. As a consequence to handle the toning down of the edges, overlapping of the frames is generally used for continuity of the speech signal during reconstruction. Otherwise, the reconstructed or unframed signal will be distorted. Conventionally the given speech signal is first segmented for frequency domain representation (normally for 60 ms). Then the low-frequency components are set to zero and the time domain samples of the signal are obtained by inverting the frequency domain representation. Now windowing mechanism is used to focus only on the central part of the speech segment and the borders of the segment are faded away. Generally, the Hamming window is used, which has a nice property of summation of the magnitude to unity when 50% of overlapping among the frames is considered. This phenomenon of considering overlapping windows and reconstruction of the speech signal by summing the overlapped windows is called the Overlap-add method.

1.2.2 Feature Extraction

Visual modality: Facial expressions in videos involve both appearance and temporal dynamics of video sequences. Efficient modeling of the spatial and temporal dynamics of the video sequences plays a crucial role in extracting robust features, which in turn improves the overall system performance. The spatiotemporal features capture the dynamic information of the appearance of faces across the frames of the video. Some of the conventional approaches for dynamic representation are based on local spatiotemporal features such as LBP-TOP (Zhao & Pietikainen, 2007), HOG-TOP (Chen, Chen, Chi & Fu, 2014), etc. Conventional hand-crafted features tend to be very local in nature and thereby dominated by the patterns of the nearest neighbor. They offer promising results in a constrained environment but fail to perform well in uncontrolled real-time environments. On the other hand, features learned from DL architectures learn non-locally by a series of nonlinear transformations and capture the high-level features of the image in a hierarchical fashion. Since the features are data-driven and learned automatically without any prior knowledge, they are quite robust in handling the challenges in uncontrolled environments. The feature extraction process is fully automatic and data-driven without any human intervention, which is a major breakthrough for its intensive use in many applications. Since most of the conventional features are found in local descriptors, they can be visualized as the initial layers of the DL architecture as they represent local features.

A simple way to obtain spatiotemporal features is by aggregating frame-level features such as the average or maximum of frame-level of features (Bargal, Barsoum, Ferrer & Zhang, 2016). Some of the works also explored matrix-based models such as eigenvector and covariance matrices for aggregation (Liu *et al.*, 2014). Recurrent Neural Networks (RNN) are the most widely used approach for capturing temporal information in various fields such as computer vision, natural language processing, speech processing, etc. Long Short-Term Memory (LSTM) is a special type of RNN explicitly designed to solve issues with long-term dependency problems such as gradient vanishing and exploding problems using short-term memory. The classic backpropagation through time (BPTT) is used to train the LSTM network. LSTM is extensively used for modeling sequential images for two major advantages. First, LSTM models are more

flexible in fine-tuning end-to-end when integrated with other deep networks such as CNN. In many approaches, LSTM has been used in combination with CNN to capture the effective latent appearance representation along with temporal dynamics (Kim, Baddar, Jang & Ro, 2019). Second, LSTM does not depend on the length of the inputs as it can support both fixed-length and variable-length inputs or outputs.

Another widely used approach for capturing the facial dynamics of videos are C3D network (Tran *et al.*, 2015), which simultaneously captures the spatial and temporal features. C3D-based approaches (Fan, Lu, Li & Liu, 2016; Ouyang *et al.*, 2017) are robust in capturing the short-term temporal dynamics while CNN in combination with LSTM is efficient in capturing long-term dynamics. Facial landmarks have also been explored to capture the dynamic variations of facial expressions in consecutive frames using trajectories of facial landmarks (Yan *et al.*, 2016; Kim, Lee, Choi & Song, 2017). Another line of research for capturing temporal dynamics in videos is based on cascaded networks or network ensembles. (Baccouche, Mamalet, Wolf, Garcia & Baskurt, 2012) employed a convolutional sparse autoencoder to capture sparse and shift-invariant features, followed by an LSTM classifier for temporal evolution. (Sun, Li, Huan, Liu & Han, 2019a) proposed a multichannel network using CNN features and optical flow-based features to capture temporal information and investigated three feature-fusion strategies: score-average fusion, Support Vector Machine (SVM)-based fusion, and neural-network-based fusion.

Audio modality: The aim of the feature extraction of the vocal signal is to obtain a compact and efficient representation of the vocal signal. Typically, acoustic features can be broadly classified into segmental features and suprasegmental features. (Schuller & Rigoll, 2009) provided a comparative review of feature-wise comparison between segmental features and suprasegmental features. Segmental features are the ones, which are estimated over a short duration of speech signal using windowing techniques (typically for a duration of 10 - 30 ms). Most of the features under this category are related to spectral characteristics and their derivatives such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), etc. Some of the important features widely used in the literature for emotion recognition are

MFCC (Korkmaz & Atasoy, 2015), LPCC (Chamoli, Semwal & Saikia, 2017), and perceptual linear predictive coefficients (LPC) (Glodek *et al.*, 2011). However, these conventional features perform poorly in the case of speaker-independent emotion recognition.

Recently, new spectral features are introduced for speech emotion recognition that significantly outperforms the conventional features (MFCC, LPCC, and PLP). (Tao, Liang, Zha, Zhang & Zhao, 2016) proposed spectral features based on local Hu moments of the Gabor-spectrogram and showed that Hu moments provide an excellent measure to discriminate emotion. The spectral features are promising to discriminate emotions and are highly correlated to the valence dimension. On the other hand, suprasegmental features are derived from the segmental features and are estimated over a large duration of the speech signal, typically 30-100 ms. Most of the emotion-related features are computed at the suprasegmental level as it conveys paralinguistic information better than segmental features. The features under this category are related to prosodic features such as fundamental pitch frequency, energy, shimmer, speech rate, spectral balance, spectral tilt, jitter, and normalized amplitude quotient. (Luengo, Navas & Hernáez, 2010) investigated the impact of prosody and spectral features and showed that spectral features outperformed prosody features. However, combining prosody and spectral features is found to be complementary to each other and improves the accuracy. The prosody features discriminate well between the inter-valence and inter-arousal, whereas spectral features distinguish between the intra-valence and intra-arousal. (Wu, Falk & Chan, 2011) proposed long-term spectro-temporal representation based on auditory and modulation filter-banks, that capture both acoustic frequency and temporal modulation frequency components.

The Short-Term Fourier Transform (STFT) overcomes the limitations of the conventional Fourier transform, which is widely used for DL techniques for speech emotion recognition. However, STFT fails to capture both time and frequency components as well as the spectro-temporal representation of the entire speech (Shah Fahad, Ranjan, Yadav & Deepak, 2021). Therefore, wavelet transform has been explored to extract the features as it highlights the instantaneous changes in the spectral evolution for emotions. (Deb & Dandapat, 2019b) proposed multi-scale amplitude features using wavelet decomposition and showed significant improvement in emotion

recognition. Though wavelet-based features are found to be promising, selecting a suitable wavelet plays a crucial role in effective emotion classification. Inspired by the production of an emotional speech, where non-linear pressure is exerted at the vocal cords, nonlinear features are explored. (Tamulevičius, Karbauskaitė & Dzemyda, 2017) introduced new features based on fractal dimensions and showed the superiority of these features over traditional features. Although emotion recognition using voice has been widely explored using conventional handcrafted features such as MFCC and global features (Sethu, Epps & Ambikairajah, 2015), there has been a significant improvement over the recent years with the introduction of DL models.

Low-level descriptor (LLD) features are widely used as input to CNNs or LSTMs, or a combination of both. However, it has been shown that spectrograms are found to carry significant paralingual information about the affective state of a person (Ma *et al.*, 2018b; Satt, Rozenberg & Hoory, 2017). (Chen, He, Yang & Zhang, 2018) proposed a CNN-LSTM model based on spectrograms and showed that delta and delta-delta log spectrograms preserve the emotionally relevant information and reduce the impact of emotionally irrelevant factors such as speaker identity, speaking style, and environments. (Sun, Chen, Xie & Gu, 2018a) explored the fusion of both shallow and deep features and showed improvement over that of individual features. (Ghosh, Laksana, Morency & Scherer, 2016) investigated the potential of transfer learning from dimensional attributes (valence and arousal) to categorical emotions and stacked autoencoder with RNN for emotion recognition. They have shown that a glottal spectrogram performs better than a conventional spectrogram, which is encoded with a stacked autoencoder and fed to RNN for classification.

1.2.3 Classification / Regression

After obtaining the refined feature vectors from the fusion model, A and V feature vectors are concatenated or further fed to a fully connected layer to obtain joint representation. This joint representation provides A-V feature representation, which is finally fed to the classifier or regressor to obtain the final predictions of the task at hand. Some of the widely used traditional

classifiers for A-V fusion of ER are SVM (Pérez Rosas, Mihalcea & Morency, 2013), HMM (Zeng *et al.*, 2007), etc. (Valstar *et al.*, 2013) used a support vector regressor (SVR) to generate the predictions of valence and arousal using the concatenated A-V feature representation. With DL models, fully connected layers are used to obtain the predictions from the A-V feature representation (Kuhnke, Rumberg & Ostermann, 2020). However, they do not effectively capture semantic information relevant to emotion recognition. LSTM-based models are found to be promising in efficiently capturing the relevant information, which in turn improves the performance of the system. One of the primitive approaches for A-V fusion-based emotion recognition was proposed by (Tzirakis *et al.*, 2017), which used a 2-layer LSTM network to obtain the predictions of valence and arousal from concatenated deep A-V features. (Schoneveld *et al.*, 2021) also explored LSTM model-based fusion and showed significant improvement in system performance. Several works have also explored ER using RNN models for generating the predictions of classifier or regressor (Caridakis *et al.*, 2006; Karpouzis *et al.*, 2007).

1.2.4 Audio-Visual Fusion

Typically, A-V fusion for emotion recognition can be achieved by three major strategies: decision-, feature-, and model-level fusion (Wu, Lin & Wei, 2014) as shown in Figure 1.3. In decision-level fusion (late fusion), multiple modalities are trained end-to-end independently, and then the predictions obtained from the individual modalities are fused to obtain the final predictions. Although decision-level fusion is easy to implement and requires less training, it neglects the interactions across the individual modalities, thereby resulting in limited improvement over uni-modal approaches. Conventionally, feature-level fusion (early fusion) is achieved by concatenating the features of A and V modalities immediately after they are extracted, which is further used for predicting the final outputs. Though feature-level fusion allows interaction between the modalities at the low-level features, it fails to leverage the interactions (inter-modal relationships) across the individual A and V modalities, thereby resulting in limited improvement in performance (Wu *et al.*, 2014). Model-level fusion is the most effective way to leverage the complementary nature of the modalities to obtain comprehensive feature representations. A

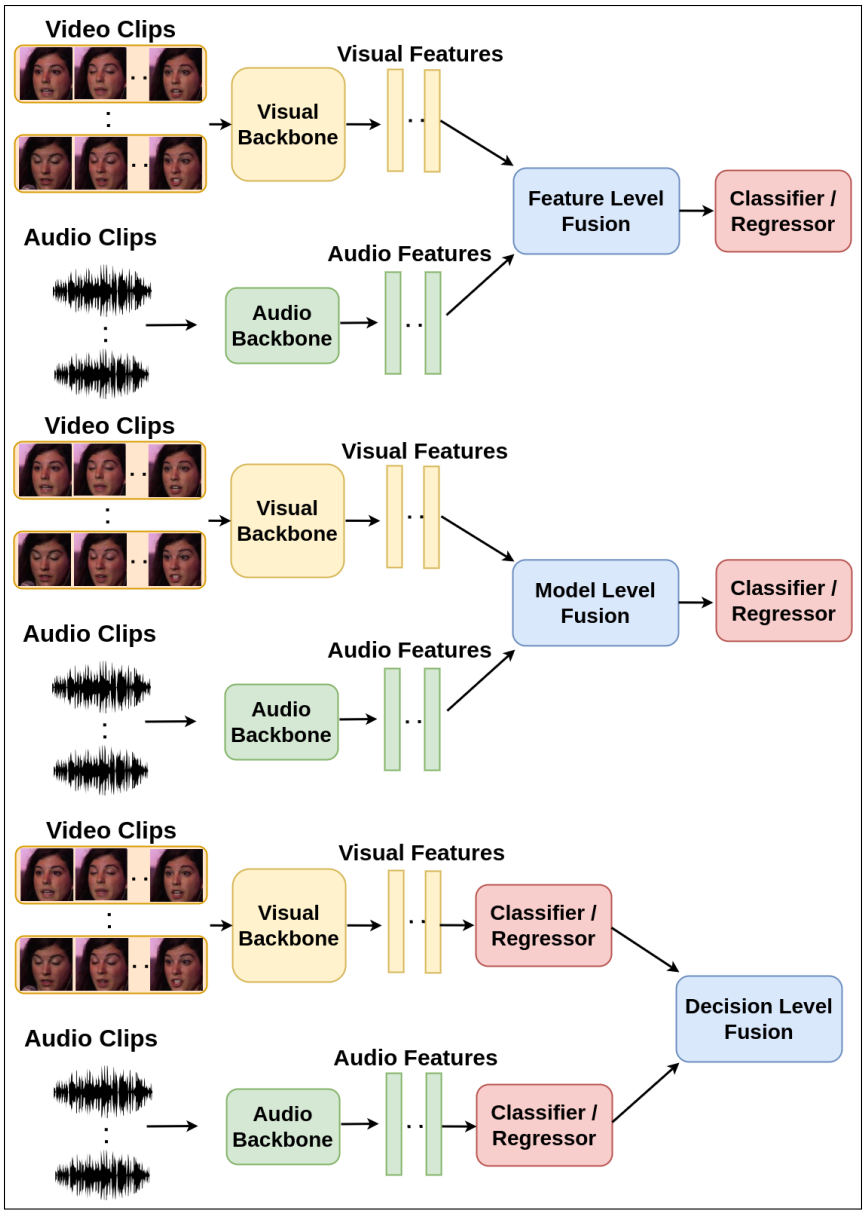

Figure 1.3 Fusion strategies of Audio-Visual (A-V) fusion model

summary of existing approaches based on A-V fusion models for ER is reviewed on 1.6.2 and 1.8.2.

1.3 Deep Learning Models for Expression Recognition

This section presents state-of-the-art DL approaches for ER that have been proposed in the framework of supervised learning.

1.3.1 Facial Expressions

Inspired by the performance of DL models, several approaches have been proposed for the analysis of facial expressions, where recognition of facial expressions provides semantic information over a short sequence of images. Since the temporal dynamics of facial expressions convey significant information, several DL models have been leveraged to capture the temporal dynamics of facial expressions in videos. This section provides DL-based approaches for the estimation of expression intensity as well as pain and fatigue levels.

1.3.1.1 Expression Intensity Recognition

(Ren, Hu & Deng, 2017) explored CNN features in combination with the RankBoost algorithm for expression intensity estimation. They generated sequential image pairs for the expression sequence by considering the neutral and apex frames, where relative expression intensity is estimated instead of the absolute intensity for every frame. They have used LeNet architecture to extract CNN features and modified the RankBoost algorithm to obtain the relative expression intensity. (Shiomi, Nomiya & Hochin, 2022) suggested two strategies for estimating the expression class as well as expression intensity in an implicit and explicit fashion. They extracted feature vectors based on facial movements and used a neural network structure to estimate the expression class, followed by expression intensity. (Chen *et al.*, 2022a) explored a framework to estimate both the expression class as well as its intensity in a unified framework. They have used label distribution learning to encode the expression intensity of each frame using linear interpolation and Gaussian function, where a Siamese network is used to learn the expression model using the label distribution as supervised information.

(Zheng, Yang, Liu & Cui, 2020) explored facial expression intensity using CNNs and attention mechanism, where a migration network is used to solve the problem of overfitting with small data, and an attention mechanism is deployed to obtain robust facial features sensitive to expression intensities. (Lee & Xu, 2003) explored cascaded neural networks and support vector machines to model the relationship between the trajectories of facial feature points and expression intensity levels. (Jaiswal, Egede & Valstar, 2018) used cumulative attributes with a DL model as a two-stage cascaded network. In the first stage, original labels are converted to cumulative attributes, and the CNN model is trained to output a cumulative attribute vector. Next, a regression layer is used to convert the cumulative attribute vectors to real-valued output. They have also evaluated the system with Euclidean loss and log loss and found that the latter outperforms the former. (Kawashima, Nomiya & Hochin, 2021) investigated the prospect of leveraging DL models for estimating facial expression intensity with small data and showed that fine-tuning CNNs on a small dataset was found to perform better than that of hand-crafted facial features. (Sabri & Kurita, 2018) investigated the ability of siamese and triplet networks for expression intensity estimation in videos. Sequential image pairs capturing the temporal dynamics of the sequence are obtained similarly to (Ren *et al.*, 2017), where two consecutive pairs are taken for Siamese networks and three images are considered for triplet networks. They have shown that the internal representations of triplet networks are efficient in capturing accurate localization of discriminative features, which improves the generalization capability of the network.

(Walecki, Rudovic, Pavlovic, Schuller & Pantic, 2017) proposed a novel copula CNN approach by exploiting the AU dependencies using conditional random fields (CRF) and estimating complex efficient feature representations simultaneously using CNN architectures. They have used multiple data sets for training the CNN model of 3 convolutional layers while the CNN features are fed to the CRF graph to capture AU dependencies. All the parameters of CNN and CRFs are jointly optimized by ensuring iterative batches that are most representative of the target structure during optimization. (Wei, Bozkurt, Morency & Sun, 2019) introduced a feature set that combines saliency map-based hand-crafted features and low-level CNN features

for spontaneous smile intensity estimation. They further exploited the mutual complementary relationship between the opponent-color characteristic of saliency maps and CNN features at multiple levels, and showed improvement in the performance of the system.

(Kollias, Psaroudakis, Arsenos & Theofilou, 2023) proposed a network for expression intensity estimation in videos to estimate, expressions, action units as well as valence and arousal. They have introduced a representation extractor with an RNN network, followed by a mask layer, which handles the varying input lengths by dynamically selecting the RNN outputs. (Lu & Zhang, 2019) explored happiness intensity estimation for a group of people, by estimating the happiness intensity level of each individual in the group using a CNN model pretrained on the VGGFace model. The group happiness intensity level is estimated using a weighted average of the intensity level of individual faces. (Thuseethan, Rajasegarar & Yearwood, 2019) proposed a metric-based mechanism for defining the primary emotions using the relation between action unit intensities and primary emotions, followed by a deep CNN-based approach to estimate the intensities of the primary emotions from posed and spontaneous video sequences.

1.3.1.2 Pain Estimation

Facial expressions of pain are often conveyed through the gold standard of facial expression research, FACS (Ekman & Friesen, 1978). Using FACS, facial expressions can be composed of a small subset of facial activities, namely lowering the brows (AU4), cheek raise (AU6)/lid tightening (AU7), etc. Nevertheless, the facial expressions of pain may not always be observed with a specific facial expression. (Wu *et al.*, 2022) investigated the prospect of leveraging DL-based models for pain classifiers based on facial expressions of critically-ill patients. (Monwar & Rezaei, 2006) explored DL models to recognize pain as a binary classification problem from facial expressions based on the location and shape features of the detected faces. (Wang *et al.*, 2017) addressed the problem of a limited dataset of facial expressions by fine-tuning the face recognition network using a regularized regression loss and additional data with expression labels. They have shown that their approach achieves state-of-the-art

performance benefiting from the rich feature representations trained on a huge amount of data for face verification.

(Xiang, Ye, Gregory & Trac, 2018) proposed temporal convolutional networks, where 1 -D convolution is performed over the frame-based feature vectors along the temporal dimension. They can detect pain levels both for frames as well as videos. (Egede, Valstar & Martinez, 2017) used handcrafted features based on shape and appearance in combination with deep-learned features and showed that their approach performs well even in small sample settings. They have also included temporal information by considering the previous and successive frames. Finally, the pain intensities are estimated using a relevance vector regressor (RVR). (Zhou, Hong, Su & Zhao, 2016a) proposed Recurrent Convolutional Neural Network (RCNN) by adding recurrent connections to the convolutional layers of the CNN architecture for estimating the pain intensities. However, choosing fixed temporal kernel depth fails to capture varying levels of temporal ranges as the duration of facial expressions may vary from short to long temporal ranges. To address the problem of fixed temporal depth, (Tavakolian & Hadid, 2018) designed a novel 3D CNN-based architecture using a stack of convolutional modules with varying kernel depths for efficient dynamic spatiotemporal representation of faces in videos. (Tavakolian *et al.*, 2020) proposed a self-supervised learning framework for pain intensity estimation from facial videos with a minimal amount of labeled data. They introduced a similarity function to learn generalized representations using a Siamese network, which is further fine-tuned for pain intensity estimation.

(Rodriguez *et al.*, 2018) used VGG Face pre-trained CNN network (Parkhi, Vedaldi & Zisserman, 2015) for capturing the facial features and further fed to the LSTM network to exploit the temporal relation between the frames. They have further shown that the performance of the pain recognition system can be enhanced by relying on the entire face images instead of only facial features. (Wang *et al.*, 2017) addressed the problem of a limited dataset of facial expressions by fine-tuning the face recognition network using a regularized regression loss. (Rodriguez *et al.*, 2018) used VGG Face pre-trained CNN network (Parkhi *et al.*, 2015) for capturing the facial features and further fed to the LSTM network to exploit the temporal relation between

the frames. Apart from 2D models with LSTM, researchers also explored optical flow and 3D CNN-based approaches for the temporal modeling of facial expressions. Compared with 2D models, 3D-CNNs are found to be quite promising in capturing the temporal dynamics of video sequences. The limitation of these approaches is that they require frame-level intensity labels, which is a major bottleneck in real-time scenarios.

1.3.1.3 Fatigue Estimation

(Sikander & Anwar, 2019) provided an in-depth review of existing technologies for fatigue detection in the context of driver assistance systems. (Ghazal, Abu Haeyeh, Abed & Ghazal, 2018) proposed a simple CNN with two convolutional layers followed by max-pooling layers and two fully connected layers, which can be deployed in a low-cost and real-time embedded system for fatigue detection. (Reddy, Kim, Yun, Seo & Jang, 2017) also explored a real-time drowsiness detection method based on DL models, which can be deployed in low-cost embedded devices while still retaining high accuracy. Specifically, they have proposed a compressed version of the heavy baseline model by exploiting the idea of "distillation" of neural network (Hinton, Vinyals & Dean, 2015) for transferring the knowledge from a huge model to a small model. (Long, Guojiang, Yuling & Junwei, 2021) explored facial key points to extract the feature vectors of each frame using DLIB (open-source software library). The feature vectors are further sliced into short temporal feature sequences and fed to the LSTM network to obtain the fatigue levels. (Zuopeng *et al.*, 2020) proposed a CNN model, named EM-CNN, to detect the fatigue levels based on eyes and mouth from the detected facial regions. They have considered the percentage of eyelid closure over time and mouth opening degree as major factors to determine driver fatigue levels using driving images. However, detecting eye states can be affected by wearing sunglasses. To solve this problem, (Zhang, Su, Geng & Xiao, 2017) explored infrared videos for detecting eye states based on the CNN model, which is further used to estimate eyelid closure over time and eye blink frequency to monitor driver fatigue levels.

(Li, Xia, Cao, Zhang & Feng, 2021) proposed a face detection network named Little Face to detect and classify faces to small yaw angle and large yaw angle. Then, the supervised descent

method (SDM) is optimized only using the normal low yawn states for accurate face alignment, which is further used to determine the driver fatigue level of distract state faces. (Dwivedi, Biswaranjan & Sethi, 2014) utilized CNN to learn features from the localized face regions obtained using the Viola-jones method (Viola & Jones, 2001) and the extracted CNN features are fed to the softmax layer for drowsiness detection. (Zhang, Murphey, Wang & Xu, 2015) explored yawning detection as a cue for driver fatigue monitoring, where nose tracking is preferred over mouth tracking as it offers more accurate tracking and deploying a neural network for yawning detection based on features extracted from nose tracking, gradient features around the mouth and facial movement. Static Bayesian networks are also explored for fatigue detection (Qiang Ji *et al.*, 2004), however, they fail to capture the temporal dynamics of the facial expressions. To incorporate the temporal dynamics, Dynamic Bayesian Networks (Li & Ji, 2005) were considered to improve the performance of fatigue detection by modeling the temporal dynamics of the driver's affective state (Qiang Ji, Lan & Looney, 2006). (Yang, Lin & Bhattacharya, 2010) explored physiological features such as ECG and EEG along with facial features for driver fatigue detection based on dynamic Bayesian networks.

1.3.2 Vocal Expressions

Vocal ER is widely explored in the context of emotion recognition, which is carried out by relying heavily on the acoustic model, however, paralinguistics is used for emotion recognition as it conveys the emotional state of the person. (Badshah, Ahmad, Rahim & Baik, 2017) explored spectrograms for speech emotion recognition, where the spectrograms are fed to a deep CNN with 3 convolutional layers and 3 fully connected layers. They further investigated the impact of transfer learning and showed that a freshly trained model performs better than a finetuned model. (Mao, Dong, Huang & Zhan, 2014) investigated the prospect of obtaining affect-salient features using CNN features in a two-stage framework. Firstly, they extracted local invariant features of unlabeled samples using a variant of the sparse autoencoder, which is further processed to obtain salient discriminative features using a novel objective function. (Yenigalla *et al.*, 2018) explored phoneme sequences with spectrograms and conducted various experiments with different kinds

of DL models. They have shown that the phoneme sequence carries significant emotional content when combined with spectrograms and fed to the CNN model improving the performance of the system. (Issa, Fatih Demirci & Yazici, 2020) introduced a new architecture, which extracts multiple spectral features and used them as input to the 1-D CNN model for emotion detection.

(Niu, Zou, Niu, He & Tan, 2017) proposed a novel approach for emotion recognition by designing a data augmentation algorithm inspired by the imaging principle of the retina for handling the problem of a limited dataset. They have designed Deep Retinal Convolutional Neural Networks (DRCNN) for obtaining robust high-level features with the help of existing DL architectures. (Han, Yu & Tashev, 2014) proposed a system for speech emotion recognition using a deep neural network. The probability distribution of the emotional states of each segment of the speech signal is obtained using DNNs, using which utterance level features are estimated. These features are further fed to a simple and efficient single-hidden layer neural network called an extreme learning machine (ELM) for the emotion classification of utterances. (Zhou, Guo & Bie, 2016b) proposed a two-stage approach of two affective models for emotion recognition: one based on a stacked autoencoder network for feature extraction and the other based on a deep belief network for the classification of emotions.

(Satt *et al.*, 2017) proposed a deep network which is a combination of CNN and RNN, which is applied to the spectrograms of the speech signal. The spectrogram features are found to be effective in suppressing the background non-speech such as music and crowd noise. (Zhao, Mao & Chen, 2019) explored the potential of both 1D CNN and 2D CNNs in combination with LSTMs to capture the local and global emotion-related features using speech and log Melspectrograms respectively. They also show that the combination of 2D CNN with LSTM outperforms traditional networks such as Deep Belief Networks (DBN) and CNNs. (Lee & Tashev, 2015) explored the impact of long-term contextual effect to capture the temporal dynamics pertinent to the high-level representation of emotional states using BLSTM models. To tackle the uncertainty of emotional labels, the label of each frame is considered as a sequence of random variables. (Fayek, Lech & Cavedon, 2017) evaluated the potential of DL models including feed-forward and recurrent neural network architectures and their variants for vocal

emotion recognition. They have shown that convNets have better discriminative performance than other architectures for vocal emotion recognition. (Zheng, Yu & Zou, 2015) proposed a systematic approach using log-spectrogram and principal component analysis (PCA), where the latter is used to suppress the interferences. The PCA-whitened spectrogram is split into nonoverlapping segments and fed to deep CNN to learn efficient features and detect emotions. (Neumann & Vu, 2017) explored attention mechanism for emotion recognition using attentive convolutional neural network in combination with multi-view learning objective function. They have conducted various experiments and investigated the impact of variations in the length of the input signal, input acoustic features, and emotions.

1.3.2.1 Fatigue Estimation

Speech is one of the significant cues and has an inevitable dependency on fatigue, which has opened a wide range of applications in the medical domain. (Chen *et al.*, 2022c) explored the potential of DL models and showed that the phonetic features pertinent to fatigue can be effectively captured using DL models, especially BLSTM network. (Krajewski *et al.*, 2010) explored features of nonlinear dynamics (NLD) as it provides additional information related to dynamics and structure of fatigue speech than conventional features such as cepstral coefficients, formats, etc. They have shown that NLD features are highly correlated with fatigue, resulting in improved accuracy. (Greeley *et al.*, 2006) showed that the attentiveness of the participants and fatigue level is influenced by the formant frequencies and MFCC coefficients. They have further shown that voice-based fatigue detection is related to precise phonetic identification and alignment and developed techniques for fatigue detection based on phonetic alignments. (Krajewski, Wieland & Batliner, 2008) proposed a novel framework for fatigue detection based on speech characteristics such as prosody articulation and speech quality.

(Nicholas, Vidhyasaharan, Julien, Sebastian & Jarek, 2015) investigated that a speaker's level of depression is associated with the acoustic properties of the speech such as spectral features and energy and presented a novel method for estimating the acoustic volume using Monte Carlo Sampling of the feature distribution. They have also provided a review of the current assessment

methods based on speech characteristics. They proposed a novel method 'relevance Vector Machines' for predicting depression levels in clinical settings based on paralinguistic properties. They have further provided state-of-the-art for speech analysis in health-care applications as well as the impact of DL models (Nicholas, Alice & Björn, 2018). (Shen & Wei, 2021) addressed the problem of over-fitting with a limited fatigue dataset as it is very time-consuming to obtain labeled fatigue data. They proposed a novel DL framework by integrating Active Learning (AL) with complex speech features extracted using stacked sparse autoencoder networks, followed by a densely connected convolutional autoencoder from spectrograms. (Wu & Sun, 2022) also proposed an approach for Air traffic controller fatigue detection using ensemble learning based on a self-adaption quantum genetic algorithm (SQGA). To cover a wide range of diversity among the learners, traditional classifiers such as K-Nearest Neighbor (k-NN), and SVMs are used along with neural networks. Finally, a weighted summation of the individual predictions is computed to obtain final fatigue-level predictions. (Bayerl, Wagner, Baumann, Bocklet & Riedhammer, 2023) investigated the impact of neural embeddings such as x-vectors, ECAPA-TDNN, and wav2vec 2.0 for fatigue detection based on vocal expressions and showed that all the neural embeddings can effectively capture vocal fatigue when temporal smoothing and normalization are applied to the extracted embeddings.

1.4 General Challenges of Expression Recognition

This section presents generic challenges related to video-based ER.

Variability Across Subjects: Subjective Bias is one of the major factors induced by the subjective nature of the subjects i.e., different subjects can express the same level of pain at different intensity levels. In addition to that, different subjects exhibit different facial appearances though they have the same facial structure, which may result in hard samples in discriminating the expressions at various intensity levels. Therefore, fine-tuning pre-trained models on face recognition still retains the traces of subjective appearances i.e., face-dominated information pertinent to subject identity, which will deteriorate the performance of recognizing expression levels as they rely only on the facial dynamics. To address this problem, (Ding, Zhou & Chellappa,

2017) proposed a novel architecture named FaceNet2ExpNet. They have shown that the fully connected layers, succeeding the convolutional layers play a crucial role in capturing the task-relevant features. Therefore, only the convolutional layers are fine-tuned with the target labels and fully connected layers are trained from scratch.

Variability in Capture Conditions: Though there has been a shift in data capture from laboratory-controlled conditions to in-the-wild uncontrolled environments, the datasets are generally captured in a specific environment, which may vary across datasets and thereby results in different data distribution of various datasets. Typically, state-of-the-art approaches are evaluated on limited datasets and show superior performance. However, when deployed on different datasets, these algorithms may fail to retain their superior performance due to the differences in the distribution of datasets, often termed data-set bias, which is a prevalent problem in the field of machine learning. To address the problem of data-set bias, a few approaches (Benitez-Quiroz, Srinivasan & Martinez, 2016) have used multiple datasets for training by merging the datasets and evaluated on different datasets. Even though merging multiple data sets may increase the training data and thereby achieve better generalization, it may suffer from label subjectivity. A few more approaches conducted cross-database experiments to validate the generalizability of the algorithm by evaluating the algorithm on a dataset different from training data (Ruiz, d. Weijer & Binefa, 2015; Wang, Peng & Ji, 2018c).

Limited Relevant Data and Annotations: With the advent of DL models, there has been a significant boost in the performance of prediction models for various applications in computer vision, natural language processing, speech processing, etc. However, the performance of these predictive models is highly constrained by the quantity and quality of data. An inferior method trained with abundant data may often result in better performance than a superior method with limited data. In the case of ER, most of the frames correspond to neutral frames, and expressions are portrayed only in a few frames though humans can exhibit a wide range of expressions. Therefore, deploying DL models for facial expression intensity estimation poses a major challenge, especially with limited annotations. Since the V appearance of the face varies from person to person due to age, civilization, ethnicity, cosmetics, eyeglasses, etc.,

the detection of facial expression intensities is a challenging task. In addition to the personal attributes, variations due to pose, occlusion, and illumination are prevalent in unconstrained scenarios of facial expressions, which leads to high intra-class variability. Therefore, there is an immense need for large-scale data-set with a wide range of intra-class variation. Though humans are capable of exhibiting a wide range of facial expressions, most of the existing data sets are developed based on basic universal expressions as they are more frequent in our everyday life.

Label Ambiguity: Compared to other problems of computer vision, labeling expressions is a highly complex process as it is subjective in nature. Manual annotation of expression intensity levels is even more challenging compared to the prototypical categorical expressions due to the increased range of facial behavior. Moreover, the annotation of expression intensity levels requires domain expertise certified by FACS coding system, which is a time-consuming and laborious task and thereby highly prone to errors induced by annotators. The process of annotating expression intensity levels is highly complex due to the subtle differences between different intensity levels, which is very challenging even for expert annotators. The continuous dimensional model further complicates the process of labeling especially when annotators are asked to label every frame of the video sequences as a continuous range of values for the intensities will be more sensitive than discrete values, which will result in differences in the labels for the same intensity of facial expression. Another major factor in obtaining annotations for the continuous dimensional models is the reaction time of annotators. To alleviate the impact of label subjectivity, the dataset is typically labeled by the strategy of crowd-sourcing, where labels are refined from several annotators (Li, Deng & Du, 2017). In addition to label subjectivity, there can be variations in the facial appearance due to heterogeneity of subjects termed as identity bias i.e., ambiguity induced by the subjective nature of humans. For instance, the expression of sadness is often misinterpreted as a neutral expression as the V appearance of sadness is very close to that of a neutral expression.

Efficient Feature Representation: With the ubiquity of DL-based approaches for various computer vision problems, there has been a shift from handcrafted features to learned features, where pre-trained models of face verification are used for finetuning with the facial expressions or

AU data due to the limited data pertinent to facial expressions or AU. However, as mentioned in (Ding *et al.*, 2017), the abstract feature representation of higher layers still retains identity-relevant information even after fine-tuning the face verification net with FER data. To overcome the problem of limited data and exploit the success of DL for FER, (Egede *et al.*, 2017) have explored the fusion of handcrafted and learned features for automatic pain intensity estimation and showed superior performance over state-of-the-art approaches. RNN-based approaches are found to be promising in capturing the temporal dynamics which has shown robust performance in various computer vision, speech, and NLP applications. (Kim *et al.*, 2019) explored efficient feature representation robust to expression intensity variations by encoding the facial expressions in two stages. First, spatial features are obtained through CNN using five objective terms to enhance the expression class separability. Second, the obtained spatial features are fed to LSTM to learn the temporal features. (Wang, Wang & Ji, 2013) studied the contribution of spatiotemporal relationship among facial muscles for efficient FER. They have modeled the facial expression as a complex activity of temporally overlapping facial events, where they proposed an Interval Temporal Bayesian Network to capture the temporal relations of facial events for FER.

1.5 Weakly Supervised Learning

The category of machine learning approaches that deal with weakly annotated data is termed Weakly Supervised Learning (WSL). Unlike supervised learning, accurate labeling will not be provided for entire data in most real-world applications due to the tedious process of obtaining annotations. Since the basic idea of machine learning approaches is to learn from data, labeling of data plays a crucial role in controlling the performance of the prediction model. Therefore, WSL is gaining attention in recent years as it has immense potential in improving the bottleneck of the labeling mechanism, which will result in efficient learning from the data. Depending on the mode of availability of labels (annotations), WSL can be classified into three categories: Incomplete Supervision, Inexact Supervision, and Inaccurate Supervision. The details regarding each of these categories are discussed elaborately in (Zhou, 2018). The pictorial representation of the classification is shown in Figure 1.4.

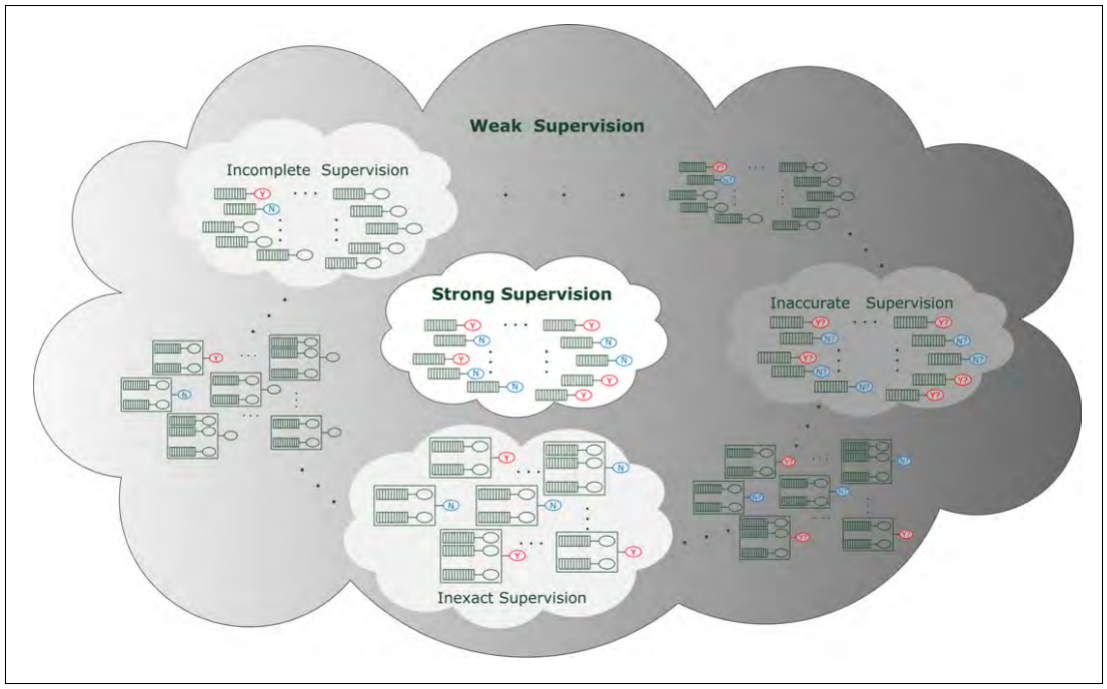

Figure 1.4 Pictorial illustration of WSL. Bars denote feature vectors; red/blue marks labels; "?" implies inaccurate labels, intermediate subgraphs depict in-between situations with mixed types of weak supervision
Taken from Zhou (2018)

Incomplete Supervision : It refers to the family of ML algorithms which deals with the situation where only a small amount of labeled data is provided, which is not sufficient to achieve a good learner despite the availability of abundant unlabeled data. For instance, it will be easy to collect a large number of videos pertinent to various expressions, however, labeling the entire data with the corresponding expressions remains to be a tedious task that demands much human labor. One of the major solutions to handle this problem is semi-supervised learning (SSL), where data distribution is assumed to occur in clusters. Semi-supervised learning relies on the assumption of local smoothness ie., samples that lie close to each other are assumed to have similar labels, based on which it makes use of unlabeled data by modeling the distribution of the data, where the labels of the unlabeled data are obtained using data distribution.

Inexact Supervision: In this category, coarsely-grained labeling is provided for the entire data instead of the exact labeling of data. The goal is to predict the accurate labels of unknown

data using coarsely labeled data. One of the major approaches to tackle this problem is based on MIL. It has a wide range of applications in the field of machine learning, where fully labeled information is difficult to attain due to the high cost of the labeling process. Due to the ubiquity of problems that are naturally formulated as the setting of MIL such as image and video classification, document classification, sound classification, etc., it emerged as a highly useful tool in many real-world applications.

Inaccurate Supervision: It refers to the scenario, where labeling information provided is not always correct, unlike inexact and incomplete supervision. Though labeling information is provided for entire data similar to supervised learning, labeling of data tends to be noisy. Since inaccurate labeling degrades the performance of the prediction model, the goal is to identify the potentially mislabeled samples from the given labeled data. Since labeling the frames of the video with the corresponding expressions is a laborious task, the annotators are more vulnerable to mislabeling the frames of the videos. Crowd-sourcing is a simple and effective way of compensating the mislabeled samples, where the same labeling task is provided to a group of individuals, and the correct labels of the samples are computed by taking the aggregate of the labels provided by different individuals using the majority voting strategy.

1.5.1 Multiple Instance Learning

MIL has drawn much attention in the past few years as it has shown robust performance in dealing with weakly annotated data. Mathematically, the problem can be formulated as the task of predicting the learning function $f : \mathbf{X} \rightarrow \mathbf{Y}$ from the training data-set $\mathbf{D} = \{(\mathbf{X_1}, \mathbf{Y_1}), (\mathbf{X_2}, \mathbf{Y_2}), ..., (\mathbf{X_n}, \mathbf{Y_n})\}$ where $\mathbf{X_i}$ denotes a set of instances of input data, $\mathbf{Y_i}$ represents the label information and N is the number of training data. These set of instances of each input data $\mathbf{X_i} = \{x_{i1}, x_{i2}, ..., x_{im_i}\}$ is called a bag, $x_{ij,}$ represents an instance of the bag where $j \in \{1, 2, ..., m_i\}$ and m_i is the number of instances in bag $X_i \subseteq X$. A bag X_i is said to be positive only if there exists at least one instance $x_{ij,}$ of X_i belongs to Y_i where $y_{ij} = Y_i$ i.e., at least one of the instances of X_i should have the label Y_i and y_{ij} denotes the label of instance j of bag X_i. Otherwise, the bag is said to be negative, where all the instances of bag X_i do not

have the label Y_i. The goal of the algorithms is to predict the labels of the unseen bags using the learning function f obtained from the training data D. This phenomenon is known as Multiple Instance Learning. MIL can also be extended to regression problems, where the major difference from the classification framework is that each bag is associated with a real-valued label but not a class.

Even though the primary goal is to predict the bag-level prediction, many techniques have been proposed to perform instance-level prediction, where each instance of the bag can be predicted. For example, in the case of the classification task, the instances of the bags can be classified along with the bag-level classification. These algorithms can be broadly classified as instance-level predictions and bag-level predictions.

- **Instance Level Algorithms:** These algorithms tend to predict the label of each instance of the bag, which is in turn used to perform bag level classification.
- **Bag Level Algorithms:** These methods represents the instances of the entire bag as a single feature, thereby transforming the problem into supervised learning. To predict the label of a new bag, distance metrics are used for discriminating the bags.

Several factors influence the performance of MIL algorithms such as Bag Composition, Data Distribution, and Label Ambiguity, which is discussed in detail by (Carbonneau *et al.*, 2018). In recent days, many MIL algorithms have been proposed using DL architectures to address the problems with classical approaches. (Xinggang, Yongluan, Peng, Xiang & Wenyu, 2018) proposed a novel approach based on DL architectures for bag representations with a focus on estimating the instance labels and showed that DL-based MIL approaches outperform the classical approaches. (Ilse, Tomczak & Welling, 2018) addressed the problem of conventionally used max-pooling (or a differentiable approximation) operation for combining the responses of instance level classifier by proposing attention mechanism to give high priority to significant instances (high weightage for witness) using neural-network-based permutation-invariant aggregation operator.

1.5.2 Applications in Facial Expression Recognition

(Ruiz, Van de Weijer & Binefa, 2014) proposed a multi-concept MIL framework based on multi-concept assumption i.e., multiple expressions in a video for estimating high-level (bag-level) semantic labels of videos, which is influenced by multiple discriminative expressions. A set of k hyper-planes are modeled to discriminate k concepts (facial expressions) in the instance space and the bag-level representation is obtained using the probability of bag for each concept, which is further classified using a linear classifier. (Wang *et al.*, 2020) proposed an automatic depression detection system using landmarks of facial expressions through the framework of MIL. LSTM is used to model the relationship between the instances (sub-sequences) of a bag (video) and global max pooling is deployed to identify depression-related instances and to generate the depression label of a test sequence. (Liu, Xu, Wang, Rao & Burnett, 2016) also used Bag of Visual Words (BoVW) at the pixel level with probabilistic latent semantic analysis (pLSA) for feature extraction and formulated the problem of pixel-level emotion detection using MIL framework, where the image is assumed to be a bag and local patches of the image as instances. The emotional content is detected for each patch using the Bayes rule, which is in turn used to predict the emotion of the test image.

(Xie, Tao & Wei, 2019) proposed an online MIL framework for early expression detection in videos i.e., detecting an expression as soon as it starts and before it ends by extending a max-margin early event detector (MMED) with a nonlinear kernel and further accelerated the training process by reformulating MIL based EED (MIED) in an online setting. For each training sequence, sub-sequence pairs along with ranking relationships are generated and a nonlinear instance-level classifier is trained by treating bag as a sub-sequence and some of its candidate subsets as instances for detecting expression of test sequences. Instead of relating bags and instances in the model generation, (Fang & Chang, 2015) explored the significance of sparse representation for effective feature learning in FER as a MIL problem, where a bag is treated as a set of images. The sparse feature representation is derived by considering binary labels for features and two strategies are investigated, one with mean subtraction and the other with subtracting neutral face to predict the expression of test sequences.

(Ruiz, Rudovic, Binefa & Pantic, 2016) proposed multi-instance dynamic ordinal random fields (MI-DORF) for estimating ordinal intensity levels of frames, where the ordinal variables are modeled as normal distribution and the relationship between the given observation (frame) and latent ordinal value is obtained by projecting the given observation (frame) onto the ordinal line, which is divided by the consecutive overlapping cutoff points of the normal distributions. Next, the temporal information is modeled across the consecutive latent ordinal variables to ensure the smoothness of the latent ordinal states. (Gnana Praveen, Granger & Cardinal, 2020) further improved the performance using deep 3D CNN model (I3D (Carreira & Zisserman, 2017)) by integrating MIL into adversarial deep domain adaptation (Ganin & Lempitsky, 2015) framework for pain intensity estimation, where source domain is assumed to have fully annotated videos and target domain has periodically annotated weak labels. (Yang *et al.*, 2018) proposed an approach for student engagement prediction in the wild using multiple-instance regression, where the input video (bag) was divided into segments (instances) and spatiotemporal features of each segment are fed to an LSTM network followed by 3 fully connected layers to obtain the regressed value of engagement intensity.

1.6 Attention Models

Attention models are specific types of models, that focus selectively on a particular aspect of information more relevant to the downstream tasks. Inspired by the human visual processing system, attention models have been explored in the field of machine learning for various applications in computer vision (Guo *et al.*, 2022), Natural Language Processing (NLP) (Vaswani *et al.*, 2017) and speech (Karmakar, Teng & Lu, 2021). With the advent of transformers (Vaswani *et al.*, 2017), attention models have become extremely popular in the field of DL. They deploy the notion of relevance by allowing the model to pay specific attention to certain parts of the input data, which is more relevant to perform the specific task at hand. In addition to improving the performance of the models, they are also useful in analyzing the interpretability of the DL models. Attention models can be categorized as soft attention, hard attention, and self-attention based on the type of attention weights computed for various components of the

input data (de Santana Correia & Colombini, 2022). In soft attention, attention weights are computed for each input element from 0 to 1. It used a softmax function to estimate the weights, which is deterministic and differentiable. In the case of hard attention, the attention weights are computed as 0 or 1 for each input element i.e., only the elements of input pertinent to the task at hand will be retained by discarding the rest. Hard attention is not differentiable and reinforcement learning techniques are used to train the models. Both soft and hard attention weights are estimated based on the relevance of input data and targets. However, self-attention focused on relevance among the input elements irrespective of the target data. It allows the input data to interact with each other and determines the salient regions to be emphasized.

1.6.1 Audio-Visual Attention for Video-Based Applications

(Tian, Shi, Li, Duan & Xu, 2018) investigated temporal localization tasks in a supervised and weakly supervised setting along with cross-modality localization. They proposed a dual multimodal residual network for A-V fusion, where A-V correlations are explored for A-guided V attention, and introduced an A-V distance learning network to deal with cross-modality localization. (Xue, Zhong, Cai, Chen & Wang, 2023) proposed a co-attention model that leverages the spatial and semantic correlations across A and V features, which helps to obtain more discriminative features for better localization of events in videos. (Hu *et al.*, 2021) introduced a novel framework of the deep multimodal attention network for sound localization as well as event localization in videos, where a multi-modal separator and multi-modal matching classifier module are deployed to address sound separation and modal synchronization problems. Recently, joint co-attention has been explored by (Duan *et al.*, 2021) in a recursive fashion for A-V event localization. They have shown that recursive training of joint co-attention yields more discriminant and robust feature representations for multimodal fusion. (Lee, Jain, Park & Yun, 2021) proposed multi-stage cross-attention, where A-V fusion is performed collaboratively to fuse A and V features for localizing and classifying actions in videos. (Lee, Yun & Jain, 2022) further improved the cross-attention model by introducing a leaky gated mechanism, where the gates are used to adaptively choose the cross-attended features based on the semantic

relevance across the A and V features. (Nagrani, Yang, Arnab, Schmid & Sun, 2021) proposed a transformer-based architecture to fuse the A and V modalities at multiple layers using fusion bottlenecks, which helps to collate and condense more relevant information in each modality, resulting in improved fusion performance at a reduced computational cost.

(Wang, Gao, Zhao & Wu, 2020a) addressed the problem of multi-modal feature fusion along with frame alignment issues between A and V modalities using cross-attention for speech recognition. (Hu, Wang, Nie & Li, 2019) proposed dense multi-modal fusion by densely integrating the representation at multiple shared layers to capture hierarchical correlations across the modalities and evaluated on cross-modal retrieval and speech recognition. (Vukotić, Raymond & Gravier, 2016) proposed a cross-modal deep network architecture, where the weights of two deep networks are enforced to be symmetry, yielding joint representation in a common feature space. They have further evaluated the proposed approach to multimodal retrieval tasks.

1.6.2 Audio-Visual Attention for Expression Recognition

(Zhao *et al.*, 2020) proposed an end-to-end architecture for emotion classification by integrating spatial, channel-wise, and temporal attention into V network and temporal attention into A network. (Ghaleb, Niehues & Asteriadis, 2020) explored attention to weigh the time windows of a video sequence to efficiently exploit the temporal interactions between the A-V modalities. They used transformer (Vaswani *et al.*, 2017) based encoders to obtain the attention weights through self-attention for emotion classification. (Lee, Kim, Kim & Sohn, 2018) proposed spatiotemporal attention for the V modality to focus on emotional salient parts using Convolutional LSTM (ConvLSTM) modules and a temporal attention network using deep networks for A modality. Then the attended features are concatenated and fed to the regression network for the prediction of valence and arousal. However, these approaches focused on modeling the intra-modal relationships and failed to effectively exploit the inter-modal relationship of the A-V modalities. (Wang *et al.*, 2020) investigated the prospect of exploiting the implicit contextual information along with the A and V modalities. They have proposed an end-to-end architecture using cross-attention based on transformers for A-V group ER. (Zhang *et al.*, 2021b) investigated

the prospect of improving the fusion performance over individual modalities and proposed leader-follower attentive fusion for dimensional ER. They have leveraged the audio modality to boost the performance of visual modality base on the attention weights obtained from the cross-modal interactions of A and V modalities. (Zhang, Huang, Zeng & Shan, 2020) proposed an attentive fusion mechanism, where the obtained A and V features are further re-weighted using weights, obtained from scoring functions based on the relevant information in the individual modalities for dimensional ER. (Luo, Zou & Huang, 2018) investigated the potential of joint representation learning using Convolutional Recurrent Neural Networks (CRNN) for vocal ER. They have shown that the impact of time intervals significantly impacts the performance of the system.

(Parthasarathy & Sundaram, 2021) explored transformers with cross-modal attention for dimensional ER, where cross-attention is leveraged using cross-modal interactions across A and V modalities. (Tzirakis *et al.*, 2021) investigated various fusion strategies along with attention mechanisms for A-V fusion-based dimensional ER. They have further explored self-attention as well as cross-attention fusion based on transformers to enable the extracted features of different modalities to attend to each other. Although these approaches have explored cross-modal attention with transformers and showed significant improvement, they fail to leverage semantic relevance among the A-V features as well as to simultaneously capture the intra-modal relationships. Typically, transformers are explored with cross-modal attention, where the query comes from one modality and keys and values come from another modality. However, we have explored joint cross-modal attention, where the query can be considered as a joint feature representation, which helps to simultaneously obtain a semantic measure of intra and inter-modal relevance among A and V modalities.

1.7 Domain Adaptation

With the advancement of CNN architectures, there has been significant progress in the performance of various applications with the availability of huge high-quality training data. However, the performance of these advanced CNN architectures may degrade when this is a lack of

sufficient manually annotated datasets. One of the approaches to handle this problem is to adapt the pretrained models trained on a huge dataset on the available unlabeled data, termed "transfer learning". The data used for the pretrained model is referred to as source data and data on which pretrained model is adapted is termed target data. Domain Adaptation (DA) is a special case of transfer learning that utilizes the knowledge of labeled data in the source domain to enhance the performance of the system on target data with different domain distributions but with the same task. DA approaches are further categorized into supervised, semi-supervised, and unsupervised

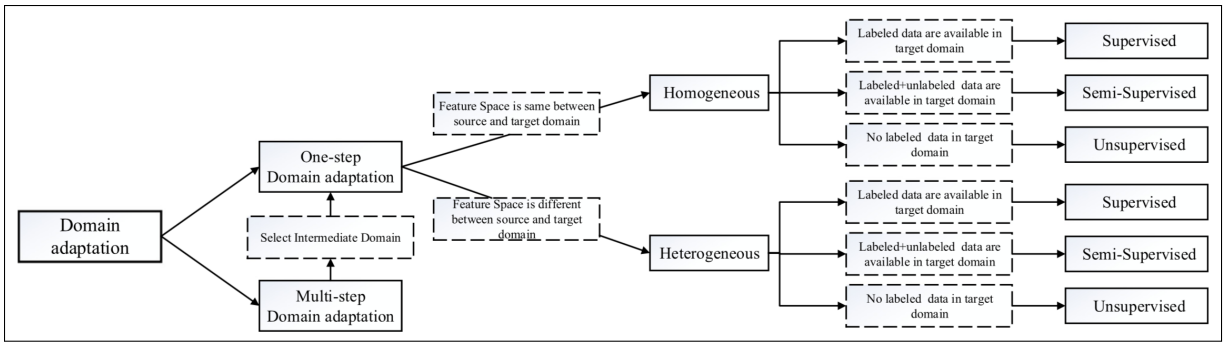

Figure 1.5 Overview of different DA approaches
Taken from Wang & Deng (2018)

DA approaches. In supervised DA approaches, labeled data is available for the target domain, which is used for DA. Semi-supervised DA deals with the case where the target data is partially labeled and finally, unsupervised DA refers to the scenario, where unlabeled data is available in the target domain. The overview of various categories in DA is shown in Figure 1.5.

Deep DA is the class of algorithms that leverage deep network architectures to enhance the performance of deep DA. DL architectures designed for DA were observed to achieve better results than shallow methods. (Wang & Deng, 2018) provided a survey on the DA approaches based on DL architecture with some of the applications related to visual categorization. With the advancement of neural-network-based DL approaches for various visual categorization applications such as image classification, face recognition, and object detection. It was observed that the deep features of the DL models converge from generic to specific features, where the transferability of the representation sharply decreases with the higher layers.

1.7.1 Domain Adaptation for Video-Based Applications

(Ganin & Lempitsky, 2015) proposed a novel approach of adversarial DA using DL models with partial or no target data labels using a simple gradient reversal layer. Inspired by their adversarial DA framework, (Jamal, Namboodiri, Deodhare & Venkatesh, 2018) explored deep DA in action space, where the distribution gaps across the source and target domains of action videos are minimized in an adversarial fashion using domain confusion loss. (Chen, Gao & Ma, 2022b) proposed a novel framework of Multi-level attentive adversarial learning with temporal dilation, where the distribution gap across the domains is minimized at multi-level temporal features using multiple domain discriminators in an adversarial fashion. The temporal features are dilated and further aggregated using the attention mechanism determined by individual domain discriminators. (Song *et al.*, 2021) proposed spatiotemporal DA to jointly learn the clip-level and video-level representation alignment in a self-supervised contrastive framework. They further introduced a domain metric scheme (video-based contrastive alignment) to optimize the category-aware video-level alignment and domain-invariance across source and target domains.

In recent years, several approaches have explored DA with multiple modalities. (Zhang, Doughty, Shao & Snoek, 2022) investigated the prospect of leveraging activity sounds for DA as they have less variance across domains. They proposed an audio-adaptive encoder and audio-infused recognizer that can effectively model the cross-modal interactions across domains and generate domain-invariant features for activity recognition. (Yang, Huang, Sugano & Sato, 2022) have shown that cross-modal interaction allows exploiting the complementary relationship across the modalities to effectively achieve cross-domain alignment. They have further leveraged the consensus of multiple modalities to obtain more relevant transferable information to achieve domain-invariance for action recognition. (Kim *et al.*, 2021) explored the multi-modal information in videos with cross-domain adaptation setting, where each modality of a domain is considered as a view and contrastive learning technique is leveraged to simultaneously regularize the cross-modal and cross-domain feature representations. (Munro & Damen, 2019) explored a multi-modal approach by deploying late-fusion of the two modalities, where the

alignment of modalities and action recognition tasks are jointly optimized using multiple domain discriminators.

1.7.2 Domain Adaptation for Facial Expression Recognition

DA has been widely used for many applications related to facial analysis such as face recognition, FER, smile detection, etc. The performance of the recognition system significantly degrades when there is a domain shift in the test images which can be due to variations in pose, illumination, resolution, expressions, and modality. (Sangineto, Zen, Ricci & Sebe, 2014) proposed a regression framework for personalized FER, where classifiers are generated for the individuals of the source data rather than a generic model for the entire source data. Further, parameter transfer is done to the target individual using the learned regression models from the source data without the need for labeled target data. (Wang, Wang & Ni, 2018) proposed an unsupervised DA approach for a small target dataset using GANs, where GAN-generated samples are used to fine-tune the model pretrained on the source dataset. (Zhu, Sang & Zhao, 2016) explored an unsupervised domain adaptation approach in the feature space, where the mismatch between the feature distributions of the source and target domains are minimized still retaining the discrimination among the face images related to facial expressions.

(Liu, Wu, Lu & Zhang, 2019) proposed a data augmentation method as a DA task to handle the problem of limited relevant data on facial expressions using a similarity-preserving generative adversarial network (SPGAN). (Ji, Hu, Yang & Shen, 2023) proposed Region Attention eNhanced Domain Adaptation (RANDA), where pseudo labels are iteratively assigned to the target domain, followed by adversarial learning to minimize the distribution gaps across the feature representations of source and target domains. They further deployed region attention learning guided by facial landmarks to obtain robust features. (Wang, Ding, Yan & Shen, 2022) explored a two-stage training pipeline for cross-domain FER, where the source domain is first pretrained to obtain semantic features, followed by learning domain-invariant features by minimizing the distance between samples of both domains with their prototype and maximizing the distance across the prototypes using adversarial loss function. (Kalischek, Thiam, Bellmann & Schwenker,

2019) studied the applicability of two DA frameworks, one for frame-level expression analysis and the other for sequence-level expressions based on the self-ensembling method. They further showed that DA is mostly applicable to person-specific FER.

1.7.3 Challenges

In this section, the challenges relevant to developing WSL models are presented for the estimation of pain intensity, followed by relevant work in the literature.

Level Imbalance: In the case of expression intensity estimation, expressions can be expressed at various intensity levels, however, they are generally sparse in nature as they are expressed in only a few frames, resulting in a huge imbalance among the various expression intensity levels, which is also reflected in UNBC-McMaster Pain database (Lucey *et al.*, 2011). This imbalance in expressions at various intensity levels is termed a level Imbalance. For instance, subtle expressions of pain can be more frequent compared to expressions of intense pain, thereby resulting in a huge imbalance among the samples of various expression levels. Moreover, neutral frames are highly dominant compared to the frames eliciting expressions, which can be seen in Figure 1.6. (He & Garcia, 2009) investigated and provided a comprehensive review of state-of-the-art approaches for learning from imbalanced data along with metrics used for evaluating the performance of the systems. (Jaiswal *et al.*, 2018) used cumulative attributes with a DL model as a two-stage cascaded network. In the first stage, original labels are converted to cumulative attributes, and the CNN model is trained to output a cumulative attribute vector. Next, a regression layer is used to convert the cumulative attribute vectors to real-valued output. They have also evaluated the system with Euclidean loss and log loss and found that the latter outperforms the former.

Limited Annotations: Due to the ambiguity and complex process of obtaining annotations, it was found that manual annotations are extremely challenging and even impossible to obtain for large scale data-sets. Therefore, WSL-based approaches have been explored to reduce the need for exact and complete annotations, which has motivated many researchers to deal with weak annotations in real-time applications. Another approach to counteract the problem of limited

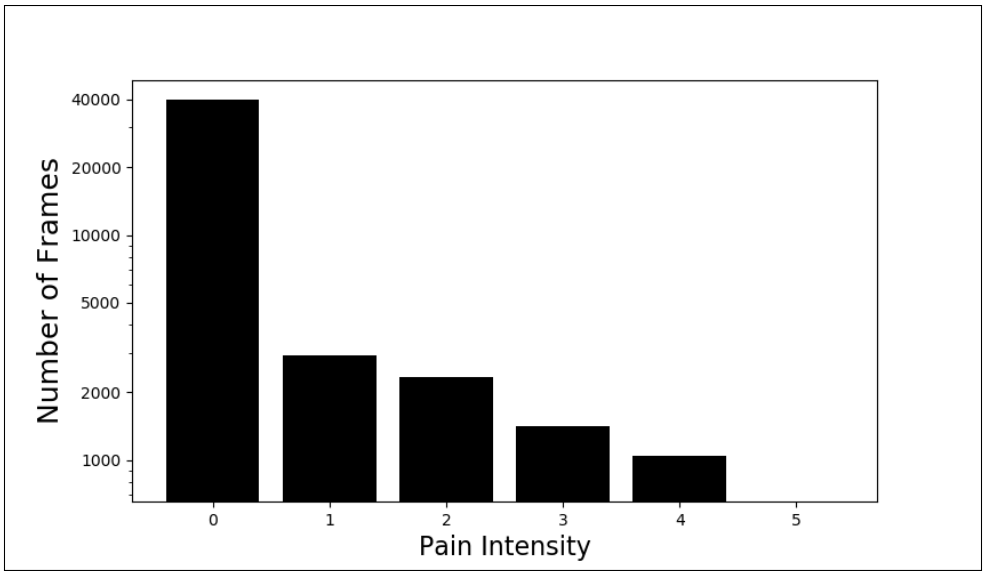

Figure 1.6 Frequency distribution of pain intensity levels in PSPI scale (0-5)

annotations is to develop a tool for semi-automatic annotation. Many researchers have explored the prospect of automating the process of annotations, which can be further refined by expert annotators to minimize the burden on human labor (Dhall, Goecke, Lucey & Gedeon, 2012). Since a fully automatic tool for obtaining annotations of expressions or AUs is not feasible and reliable, a semi-automatic tool seems to be a more plausible approach to obtain reliable annotations for large scale data-sets, especially for the case of continuous affect model which is more vulnerable to noise and highly complex process to discriminate the subtle variations across the frames.

1.8 Audio-Visual Fusion

In this section, the existing approaches of A-V fusion for video-based applications and ER are provided, followed by challenges pertinent to developing AV fusion models for dimensional ER.

1.8.1 Audio-Visual Fusion for Video Based Applications

A-V fusion has been mostly explored for action localization and event localization applications along with ER. (Kazakos, Nagrani, Zisserman & Damen, 2019) proposed a novel architecture for fusing A and V modalities within a range of temporal off-sets, where A, V, and flow modalities are fused at the mid-level before temporal aggregation, with shared modality and fusion weights over time. (Brousmiche, Rouat & Dupont, 2019) investigated the performance of several fusion strategies for the A-V recognition task of event localization in videos. They further introduce feature-wise linear modulation layers to exploit the semantic relationship across A and V modalities and showed that A-V fusion performs better than unimodal performance. (Liu, Quan, Liu & Yan, 2022) proposed a novel bi-directional modality fusion, which not only fused A and V features but also enhanced the fused features to obtain more robust A-V feature representations, where two forward-backward fusion modules are deployed in both directions for event localization. (Zhao, Gong & Li, 2021) developed an A-V recurrent network for video summarization, which is composed of three modules: LSTM networks to model the temporal dependency of individual A and V modalities, A-V fusion LSTM to fuse the A and V features based on latent consistency between them, and self-attention video encoder to capture the global dependency in the video. (Rhevanth, Ahmed, Shah & Mohan, 2022) proposed a video summarization technique by extracting the keyframes based on the structural similarity index, where MFCC features are obtained for the corresponding keyframes and further fused with V features of keyframes. The obtained A-V features of the keyframes are further refined using a CNN model for the final summarization of videos.

1.8.2 Audio-Visual Fusion for Expression Recognition

(Tzirakis *et al.*, 2017) proposed A-V fusion-based dimensional emotion recognition using DL models, where A and V features are obtained using ResNet50 and 1D CNN respectively. The obtained features are then concatenated and fed to a Long short-term memory (LSTM) for the prediction of valence and arousal. (Ortega *et al.*, 2019) investigated an empirical study of fine-tuning pretrained CNN models by freezing various convolutional layers. (Kuhnke *et al.*,

2020) proposed two stream A-V network, where V features are extracted from R(2plus1)D model (Tran *et al.*, 2018) and A features are obtained from Resnet18 model (He *et al.*, 2016). The obtained features are further concatenated for the final prediction of valence and arousal. (Wang, Wang, Qi & Suzuki, 2021) further improved their approach (Kuhnke *et al.*, 2020) by introducing teacher-student model in a semi-supervised learning framework. The teacher model is trained on the available labels, which are further used to obtain pseudo labels for unlabeled data. (Schoneveld *et al.*, 2021) explored knowledge distillation using a student-teacher model for the V modality and a CNN model for the A modality using spectrograms. The deep feature representations are combined using an RNN-based fusion strategy. Inspired by the deep auto-encoders, (Nguyen *et al.*, 2021) investigated the prospect of how to simultaneously learn compact representative features from A and V modalities using deep auto-encoders. They have proposed a deep model of two-stream auto-encoders and LSTM for efficiently integrating V and A streams for dimensional emotion recognition. (Rasipuram, Bhat & Maitra, 2020) also explored the fusion of A and V features along with head pose, eye gaze, and action unit intensities, where the temporal modeling is performed on the individual modalities using GRUs. Though the above-mentioned approaches have shown significant improvement in dimensional emotion recognition, they fail to capture the inter-modal relationships and relevant salient features specific to the task.

1.8.3 Challenges

Differences in Learning Dynamics: Multiple modalities are often exploited to capture diverse and comprehensive information among multiple modalities to obtain superior performance than uni-modal approaches. Leveraging multiple modalities allows us to retain comprehensive information which is often missing in some of the modalities. So, multi-modal approaches are expected to perform better than uni-modal approaches. However, it was shown that multi-modal approaches do not always outperform uni-modal approaches (Wang, Tran & Feiszli, 2020b). This has been attributed to the fact that A and V channels exhibit different learning dynamics while training the system in an end-to-end manner. Therefore, A and V modalities generalize

at different learning rates, which results in poor performance of joint training than uni-modal performance. Coping with the variations of learning dynamics of A and V modalities seems to be crucial to develop a robust system that outperforms uni-modal systems.

Synchronization Issues: Exhibiting emotions are temporally dynamic events, which can be inferred from both A and V modalities. A and V channels often exhibit different frame rates, which results in the mismatch of temporal alignment of A and V features. However, proper alignment of A and V modalities is a fundamental step to effectively capture the correlation across the A and V features for emotion recognition. Therefore, improper alignment of A and V features will result in poor performance of the A-V system for emotion recognition.

Audio-Visual Representation: A-V representation refers to the task of representing the data from A and V modalities in the form of a joint representation. Since A and V channels often contain complementary and redundant information, it is very important to obtain a robust A-V feature representation in an efficient and meaningful way. For instance, in a specific video clip, A modalities might be exhibiting significant modulations in the vocal expression, while facial modality might be exhibiting poor semantics pertinent to emotions. So, some of the challenges pertinent to A-V feature representation include different noise levels, missing data in one of the modalities and effectively capturing both intra and inter-modal relationships from A and V modalities.

Effective Fusion: Effectively fusing the A and V modalities allows us to capture the complementary information across the A and V modalities i.e., the ability to retain the relevant information pertinent to emotions even if one of the modalities is missing. With the advent of DL, multimodal representation and fusion have been intertwined since the representation is learned along with the task of classification or regression layers. Moreover, the heterogeneous nature of the A and V modalities poses a major challenge in fusing the modalities. Typically, A-V fusion can be broadly categorized as model-agnostic and model-based approaches. Model-agnostic approaches refer to the class of approaches, which does not depend on any specific machine learning model. Model-based approaches explicitly rely on specific models for fusion like

neural networks. Though DL models show great progress in their performance, it often lacks interpretability, which is a major challenge to analyze the fusion performance.

1.9 Conclusion

Based on the above-mentioned challenges and reviewing the current literature, two potential research directions have been investigated to leverage DL models for weakly labeled videos with minimum annotation and effectively capture the complementary relationship across A and V modalities. Both of these research directions remain at a rudimentary level and are found to be very promising in improving the performance of the system to build a robust ER system for videos. A detailed review of our contributions along with a comprehensive review of existing approaches to WSL for facial behavior analysis is presented in the following chapters.

CHAPTER 2

WEAKLY SUPERVISED LEARNING FOR FACIAL BEHAVIOUR ANALYSIS: A REVIEW

Gnana Praveen Rajasekhar[a] , Eric Granger[a] , Patrick Cardinal[b]

[a]Department of Systems Engineering, École de technologie supérieure,
[b]Department of Software and IT Engineering, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper submitted for publication, IEEE Transactions on Affective Computing, September 2022

Abstract

Given the recent advances in deep learning (DL), and in sensor and computing technologies, there has been considerable progress in the development of systems that analyze facial behavior, evolving from systems that perform image analysis under controlled laboratory conditions, to those for video analysis under more challenging real-world conditions. However, some key challenges in real-world applications include the significant variations over time of facial expressions for different people and capture conditions and the limited amount of data to train predictive models. DL models typically require supervised training with large-scale datasets to provide a high level of performance, and the collection and annotation of such datasets is a costly undertaking that relies on domain experts. Moreover, the annotation process is highly vulnerable to the ambiguity of expressions or action units due to the bias induced by the domain experts. Therefore, there is an imperative need to address the problem of facial behavior analysis with weak annotations. This paper provides a comprehensive review of weakly supervised learning (WSL) approaches that are suitable for facial behavior analysis, either using weak categorical or dimensional labels. First, a taxonomy of scenarios for WSL is introduced, along with challenges related to each scenario. For both classification and regression applications (i.e., prediction of categorical and intensity levels, respectively), we provide a systematic review of state-of-art ML/DL models for each scenario, along with their respective strengths and limitations. A review of the widely-used public datasets, experimental protocols, and experimental results is also provided for these state-of-art ML/DL models. Finally, we present a critical analysis of

Figure 2.1 Examples of primary universal emotions. From left to right: neutral, happy, sad, fear, anger, surprise, disgust
Adapted from Compound facial expressions of emotion database Du *et al.* (2014)

models for different applications and scenarios, the key challenges, and opportunities, along with the potential research directions to leverage weakly-labeled data to address real-world facial behavior analysis problems.

2.1 Introduction

Facial Behavior Analysis is an emerging area of interest in computer vision and affective computing, where it has great potential for many applications in human-computer interaction, sociable robots, autonomous-driving cars, etc. It was shown that only one-third of human communication is conveyed through verbal components and two-thirds of communication occurs through non-verbal components (Mehrabian, 2017a). Although several nonverbal components are available, facial behavior plays a major role in conveying the mental state of a person, which can be reflected by the movements of the facial muscles of a person. Ekman and Fries conducted a cross-cultural study on facial expressions, showing that there are six basic universal facial emotions across human ethnicity and cultures – Anger, Disgust, Fear, Happy, Sad, and Surprise (Ekman & Friesen, 1976) as shown in Figure 2.1. Subsequently, Contempt has been added to these basic emotions (Matsumoto, 1992). Given the simplicity of discrete representation, these seven prototypical emotions are the most widely used categorical model for the classification of facial emotions. Though the terms "expression" and "emotion" are alternatively used in many research papers, the primary difference is facial emotion conveys the mental state of a person,

Figure 2.2 Examples of action units
Taken from Martinez & Valstar (2016)

whereas facial expression is the indicator of the emotions being felt, i.e., facial expressions display a wide range of facial modulations but facial emotions are limited.

Ekman developed the Facial Action Coding System (FACS) (Ekman, 2002), a taxonomy of facial expressions that defines all observable facial movements for every emotion. It is comprised of 32 action units (AUs), and 14 additional action descriptors (ADs). Action units are described by the fundamental actions of individual muscles or groups of muscles to form a specific movement as shown in Figure 2.2. Action descriptors are unitary movements that account for the head pose, gaze direction, and miscellaneous actions such as jaw thrust, blow, and bite, etc. It has been used as a standard for manually annotating facial expressions as it defines a set of rules to express any possible facial expression in terms of specific AUs and also measures the intensity of facial expressions at five discrete levels ($A < B < C < D < E$) where A being the minimum intensity level and E being the maximum intensity level. To further enhance the range of facial expressions, continuous models over affect dimensions are proposed (Gunes & Schuller, 2013). Due to the immense potential of AUs in interpreting the expressions for deriving high-level information, we have focused on the analysis of both expressions and AUs in the framework of categorical as well as regression labels. In the setting of categorical labels, expressions or action units are analyzed as the problem of classifying or detecting expressions or AUs. In the

case of regression labels, they can be formulated as the problem of either ordinal regression or continuous (dimensional) regression. Ordinal regression deals with the estimation of discrete ordinal or intensity levels of expressions or AUs, whereas continuous regression is the task of estimating the wide range of emotions on a continuous scale of valence and arousal as a dimensional problem.

Most of the conventional approaches for facial expression recognition (FER) rely on hand-crafted features, such as Scale Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), LBP on three orthogonal planes (LBP-TOP) descriptors, which are deterministic and shallow in nature. Therefore, the performance declines in real-time scenarios as it fails to capture the wide range of intra-class variations of expressions within the same class due to factors such as age, gender, race, cultural background, and other person-specific characteristics in uncontrolled environments. With the advancement of deep learning architectures and computing capability, there has been a breakthrough in the field of machine learning, which has made it possible to cope with a wide range of variations to develop intelligent systems in uncontrolled real-time environments and perform at par with human ability. With the increase in the training data, labeling the data remains a tedious task that demands a lot of human support and is time-consuming, thereby not feasible in real-time applications. To achieve minimal competency as a FACS coder, it takes over 100 hours of training, and each minute of video takes approximately one hour to score (Ekman & Friesen, 1978). Moreover, labels of intensity levels provided by the annotators are subjective in nature, resulting in the ambiguity of the annotations due to the bias induced by the annotators. Annotating dimensional labels i.e., valence and arousal on a continuous scale becomes even more challenging as it increases the level of ambiguities due to the wide range of emotions compared to discrete intensity levels. Therefore, there is an immense need to deal with weak annotations pertinent to facial analysis in real-time environments to fully leverage the potential of deep learning models.

In recent years, though exhaustive surveys including deep learning approaches are published on facial behavior analysis (Martinez & Valstar, 2016; Sariyanidi, Gunes & Cavallaro, 2015; Samal & Iyengar, 1992; Li & Deng, 2020), weakly supervised learning (WSL) based approaches

for analyzing facial behavior is not yet done to the best of our knowledge, despite the immense need of WSL approaches in real-life situations. In this survey, we have conducted a comprehensive review of facial behavior analysis with weak annotations, consolidated and provided a taxonomy of existing work over the last decade, primarily focusing on the classification or regression of facial expressions or action units.The contribution of this review is as follows.

- We present the relevance of facial behavior analysis in the context of various types of weak annotations and the corresponding problem formulations associated with it in the context of classification and regression.
- State-of-the-art approaches for expression and AU analysis in the framework of WSL are extensively reviewed, consolidated, and discussed insights along with advantages and limitations.
- Widely used datasets in the context of weak annotations are provided along with comparative results and the corresponding evaluation strategies.
- Prospective challenges and opportunities associated with the development of a robust FER system pertinent to WSL scenarios in real-life situations along with an insight into potential research directions are discussed.

The rest of the paper is organized as follows. The overview of facial behavior analysis in the context of weakly annotated data is presented in Section 2.2. A comprehensive survey of the existing approaches for facial behavior analysis with weak annotations related to classification and regression is considered in Section 2.3 and Section 2.4 respectively. The databases widely considered in the framework of WSL for evaluating their approaches are mentioned in Section 2.3.4.1 and Section 2.4.3.1. The challenges with the existing state-of-the-art approaches and opportunities along with potential research directions are discussed in Section 2.5.

2.2 Weakly Supervised Learning for Facial Behavior Analysis

The category of machine learning approaches that deal with weakly annotated data is termed "Weakly Supervised Learning (WSL)". Unlike supervised learning, accurate labeling will not be provided for entire data in most real-world applications due to the tedious process of obtaining

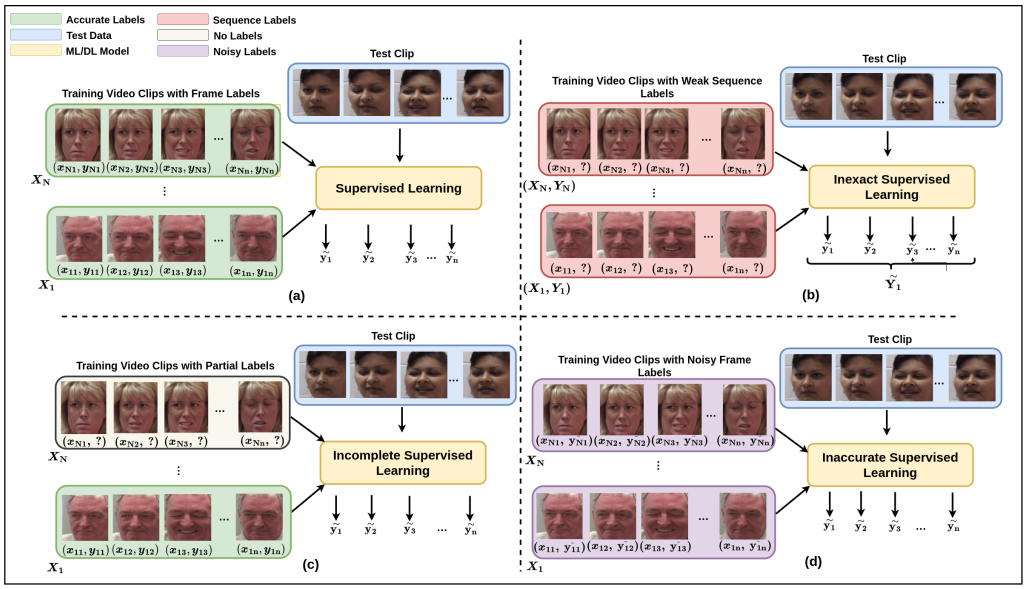

Figure 2.3 Illustration of WSL scenarios for expression recognition in videos. (a) Supervised Learning with accurate frame-level labels. (b) Multiple Instance Learning with Sequence Level labels. (c) Semi-supervised Learning with partial labels (d) Inaccurate Supervised Learning with noisy labels
Taken from Gnana Praveen *et al.* (2021)

annotations. Therefore, WSL is gaining attention in recent years as it has immense potential in a lot of vision applications such as object detection, image categorization, etc. Depending on the mode of availability of labels (annotations), WSL can be classified into three categories: Inexact Supervision, Incomplete Supervision, and Inaccurate Supervision. The details regarding each of these categories are discussed elaborately in (Zhou, 2018). In this section, we will briefly introduce these categories and their relevance to the recognition of facial expressions and action units, which is depicted in Figure 2.3 and Figure 2.4 respectively. We have primarily focused on four specific problems pertinent to facial behavior analysis in the context of various categories of weakly supervised learning: Expression detection, Expression intensity estimation, AU detection, and AU intensity estimation. The objective of expression or AU detection is to classify various expressions or action units in a given image or video, whereas expression or AU intensity estimation refers to estimating the intensities of expressions or AUs.

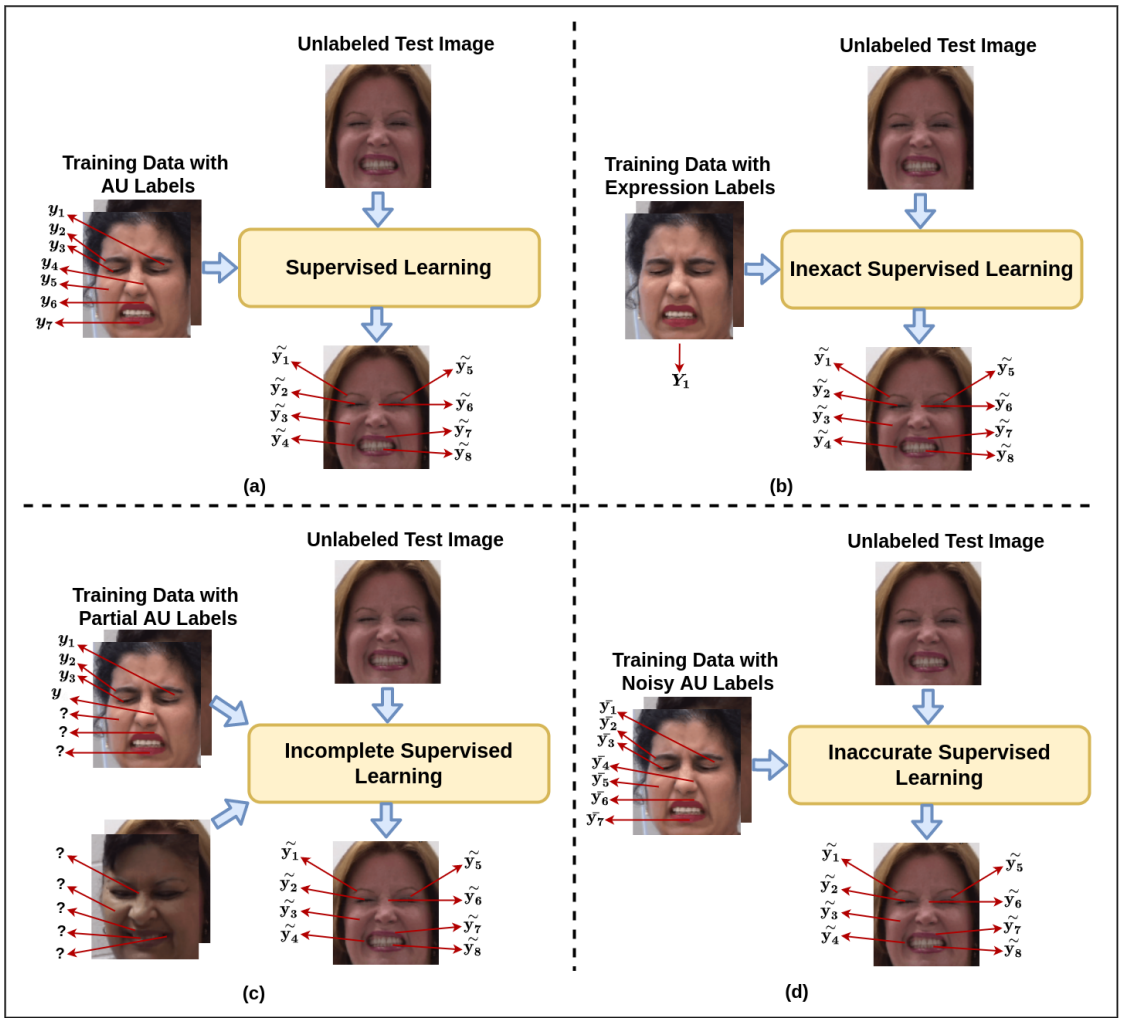

Figure 2.4 Pictorial illustration of WSL scenarios for AU recognition in images. (a) Supervised Learning with accurate AU annotations. (b) Inexact Supervised Learning with Image-Level expression annotations. (c) Incomplete Learning with partial AU annotations (d) Inaccurate Supervised Learning with noisy AU annotations Taken from Gnana Praveen *et al.* (2021)

2.2.1 Inexact Supervision

In this category, coarsely-grained labeling is provided for the data samples instead of the exact labeling of data. The goal is to predict the accurate labels of unknown test data using the coarsely labeled training data. One of the major approaches to tackle this problem is based on Multiple Instance Learning (MIL) (Foulds & Frank, 2010). It has a lot of potential in a wide range of

applications, where fully labeled information is difficult to attain due to the high cost of the labeling process. Due to the ubiquity of problems that are naturally formulated in the setting of MIL such as image and video classification (Chen & Wang, 2004), document classification (Settles, Craven & Ray, 2007), object detection (Viola, Platt & Zhang, 2006), etc., it has emerged as a highly useful tool in many real-world applications.

The coarsely labeled data is considered a "bag" and the samples within the bag contributing to the coarse annotation are referred to as "instances". In most computer vision applications, the bag is considered to be an image or video, and the instances of the bag are considered as patches in images and segments of frames or frames in videos. Even though the primary goal is to predict the bag-level prediction, many techniques have been proposed to perform instance-level prediction also, where each instance of the bag can be predicted. For example, in the case of the classification task, the instances of the bags can be classified along with the bag-level classification. Several factors influence the performance of MIL algorithms such as bag composition, data distribution, and label ambiguity, which is discussed in detail by (Carbonneau *et al.*, 2018). These MIL algorithms can be broadly classified as instance-level predictions and bag-level predictions:

- **Instance Level Algorithms:** These algorithms tend to predict the label of each instance of the bag, which is in turn used to perform bag level classification.
- **Bag Level Algorithms:** These methods represents the instances of the entire bag as a single feature, thereby transforming the problem into supervised learning.

In the case of expression detection, the task is to localize and predict the expressions of short-video clips or frames of the videos using training data with sequence-level labels. Even though data acquisition of videos of various expressions is not very difficult to attain, the major bottleneck is to get the exact label information of all the frames in the training videos. This problem can be circumvented using MIL algorithms, where sequence level predictions can be estimated along with frame level predictions as shown in Figure 2.3b. In the context of multiple instance regression (MIR) for expression intensity estimation, the objective is to estimate the intensities of frames or sequences using training data of videos with sequence level intensities of

expressions, where the sequence level label is given by the maximum or average of labels of individual frames of a given sequence (Ray & Page, 2001). The state-of-the-art approaches for detecting facial expressions and expression intensities from sequence level labels are discussed in Section 2.3.1 and Section 2.4.1 respectively.

The task of AU detection in the context of inexact annotations is formulated as the problem of AU label estimation from the expression labels of training data without AU labels, where expression labels act as weak supervision for AU labels as shown in Figure 2.4b. The relevant approaches for the problem of AU detection from the training data of expression labels are elaborated in detail in Section 2.3.1. Similar to MIR for expression intensity estimation, AU intensity estimation can also be formulated to estimate the AU intensities of frames (instances) or sequences (bags) using training data of sequence-level AU intensities.

2.2.2 Incomplete Supervision

It refers to the family of ML algorithms which deals with the situation where only a small amount of labeled data is provided, despite the availability of abundant unlabeled data. For instance, it will be easy to get a huge number of videos or images of facial expressions. However, labeling the entire dataset with the corresponding expressions or action units remains to be a tedious task that demands a lot of human labor. One of the major solutions to handle this problem is semi-supervised learning (SSL) (Chapelle, Schlkopf & Zien, 2010), where data distribution is assumed to occur in clusters. Semi-supervised learning relies on the assumption of local smoothness ie., samples that lie close to each other are assumed to have similar labels, based on which it makes use of unlabeled data by modeling the distribution of the data. Therefore, the labels of unlabeled data can be predicted using generative methods, which assume that the labeled and unlabeled data are generated from the same distribution. The model assumption for the data distribution plays a crucial role in improving the performance of the system as it influences the label assignment of the unlabeled data.

In the context of incomplete supervision for expression detection in videos (or images), annotations will be provided for a subset of videos (or images), which can drastically reduce the time required to annotate the data-set. The objective is to predict the labels of all the frames in a test sequence (or test image) using partial labels of videos (or images) of training data. For expression intensity estimation, the intensity levels will be provided only for a subset of videos (or images) as shown in Figure 2.3c and the relevant approaches are discussed in Section 2.4.2.

For AU detection, the problem of semi-supervised learning can be formulated in two ways: Missing Labels and Incomplete Labels as shown in Figure 2.4c. In the first case, each sample of training data is assumed to be provided with multiple labels of Action Units but with missing labels. The task is to train the AU classifier with the training samples of missing labels to predict the complete set of labels for the test sample. On the other hand, for incomplete labels, the entire label set of multiple AUs is provided to the training data but only for a subset of images in the training data, where the rest of the training samples do not have AU labels (but may have expression labels). The problem of incomplete annotation can be extended further to AU intensity estimation, where AU intensity levels are provided only for the key frames within a video sequence of the training data or only for a subset of data-set of images and the relevant approaches for AU detection and AU intensity estimation are discussed in Section 2.3.2 and Section 2.4.2.

2.2.3 Inaccurate Supervision

It refers to the scenario where labeling information is provided for the entire dataset similar to supervised learning, however, the labels tend to be highly noisy. Since inaccurate labels degrade the performance of the prediction model, the goal is to overcome the challenges imposed by noisy labels and generate a robust predictive model to estimate the accurate labels of the test data. Some of the widely used approaches to handle this problem are data-editing (Muhlenbach, Lallich & Zighed, 2004) and crowd-sourcing (Brabham, 2008). In data editing, the training samples are considered to be nodes, which are connected to each other based on labels, forming a graph-like structure. The edges connecting the samples of different labels are termed cut-edges.

A weighted statistic is estimated for cut-edges, with the intuition that a sample is considered a potentially mislabeled sample if it is associated with many cut-edges. Crowd-sourcing is a simple and effective way of compensating mislabeled samples, where the same labeling task is provided to a group of independent non-expert annotators, and the correct labels of the samples are computed by taking the aggregate of the labels provided by different individuals using a majority voting strategy.

In the case of expression detection, labeling frames of the video with the corresponding expressions is a laborious task and annotators are more vulnerable in mislabeling frames of the videos. For instance, the expression of "sadness" may look similar to the expression of "neutral". Therefore, there is an imperative need to handle the problem of noisy expression labels and predict the accurate expressions of test data. The ambiguity will be even more pronounced for compound expressions, which is a combination of more than one expression. The problem of inaccurate annotations is more prevalent in the case of regression i.e., estimating the level of the expressions. Regression can be done in two ways: ordinal regression, where the expression levels are assumed to be discrete and continuous regression, where the expression levels are continuous. Since more variation is possible for continuous regression compared to ordinal regression, annotations of continuous regression tend to be noisier. The scenario of inaccurate annotations for expression recognition in videos is shown in Figure 2.3d and the relevant approaches for expression detection are discussed in Section 2.3.3. The framework of inaccurate annotations for expression recognition can also be further extended to AU recognition, where AU labels are considered to be noisy. The framework of inaccurate annotations for AU recognition is shown in Figure 2.4d and the relevant approaches for AU detection are discussed in Section 2.3.3.

2.3 Weakly Supervised Learning for Classification

The classification of facial expressions is a well-explored research problem in the field of affective computing. In this section, we categorize the existing approaches on WSL for facial behavior analysis based on the type of weak annotation: Inexact, Incomplete and Inaccurate annotations,

which are further sub-classified to expression detection and AU detection. Moreover, we provide a summarization of the results of these approaches along with the evaluation strategies.

2.3.1 Inexact Annotations

This section deals with the review of existing approaches related to WSL of inexact annotations, where annotations are provided at a global coarse level instead of finer low-level annotations.

Detection of Expressions

In the case of inexact annotations, expression labels are provided at the sequence level and the goal of the task is to localize and predict the expressions of test sequences.

Instance Level Approaches: The detection of expressions with MIL was initiated by (Sikka *et al.*, 2014), where automatic pain localization was achieved by considering each video sequence as a bag comprising of multiple segments (sub-sequences or instances) and segment-level features are obtained by temporally pooling Bag of Words (BoW) representation of frames. An instance-level classifier is developed using MILBOOST (Viola *et al.*, 2006) and bag-level labels are predicted based on the maximum probabilities of the instances of the corresponding bag. (Wu, Wang & Ji, 2015b) further enhanced this approach by incorporating a discriminative Hidden Markov Model (HMM) based instance level classifier with MIL instead of MILBOOST to efficiently capture the temporal dynamics. Moreover, segment-level representation (instance) is obtained based on the displacement of facial landmarks between the current frame and the last one, which is further extended to expression localization. (Chen, Ansari & Wilkie, 2022d) explored pain classification with MIL using the relationship between AUs and pain expression under two strategies of feature representation: compact and clustered representation. In compact representation, each entry of the feature vector denotes the probability estimate of the corresponding AUs whereas clustered representation is obtained by clustering the co-occurrence of AUs. They have shown improvement using a clustered representation of AUs while an instance-level (sub-sequence) classifier is modeled using MILBOOST.

Bag Level Approaches: Unlike the aforementioned approaches that rely on a single concept assumption, (Ruiz *et al.*, 2014) proposed a multi-concept MIL framework based on multi-concept assumption i.e., multiple expressions in a video for estimating high-level (bag-level) semantic labels of videos, which is influenced by multiple discriminative expressions. A set of k hyper-planes are modeled to discriminate k concepts (facial expressions) in the instance space and the bag-level representation is obtained using the probability of bag for each concept, which is further classified using a linear classifier. (Sikka, Sharma & Bartlett, 2016) also investigated the temporal dynamics of facial expressions in videos, where the ordering of the discriminative templates (neutral, onset, apex, or offset) in a video sequence is associated with a cost function, which captures the likelihood of the occurrence of different temporal orders. Each video is assigned a score based on the ordering of the templates using the model, which is learned with Stochastic Gradient Descent (SGD) by minimizing the regularized max-margin hinge loss function.

Unlike conventional MIL methods, (Huang, Ngai, Hua, Chan & Leong, 2016) developed a novel framework of Personal Affect Detection with Minimal Annotation (PADMA) for handling user-specific differences based on the association between key facial gestures and affect labels i.e., if an instance occurs frequently in bags of particular class but not in others, the instance has a strong association with the label. The facial feature vectors in a video are clustered as key facial gestures (expressions) and the video is represented as a sequence of corresponding affect labels. The sequence level affects are predicted by affect frequency analysis using the association between facial gestures and facial affects. (Xu & Mordohai, 2010) represented a sequence with a set of 20 dominating motion fields and a dictionary of motion words is obtained from the motion fields of training sequences, which are labeled at the sequence level. Each frame (motion field) is then represented as a histogram of motion words and labels are obtained using the nearest neighbor classifier. Then the sequence level label is predicted using a majority vote of its frame labels. (Wang *et al.*, 2020) proposed an automatic depression detection system using landmarks of facial expressions through the framework of multiple instance learning. LSTM is used to model the relationship between the instances (sub-sequences) of a bag (video) and global max

pooling is deployed to identify depression-related instances and to generate the depression label of a test sequence.

Detection of Action Units

Since any expression can be characterized as a combination of action units, many psychological studies have shown that there exists strong relation between expressions and AUs (Lewis, Haviland-Jones & Barrett, 2010). Due to the expensive and laborious task of obtaining AU annotations compared to expressions and the close relation between expressions and AUs as shown in Table 2.1, many researchers have explored the problem of AU detection in the framework of weakly supervised learning, where expression labels are considered as weak coarse labels for AUs.

Table 2.1 List of AUs observed in expressions

Expression	**AUs**
Anger	4, 5, 7, 10, 17, 22-26
Disgust	9, 10, 16, 17, 25, 26
Fear	1, 2, 4, 5, 20, 25, 26, 27
Happiness	6, 12, 25
Sadness	1, 4, 6, 11, 15, 17
Surprise	1, 2, 5, 26, 27
Pain	4, 6, 7, 9, 10, 12, 20, 25, 26, 27, 43

(Ruiz *et al.*, 2015) investigated the prospect of learning AU classifiers using expression labels by exploiting the relationship between expressions and AUs. Each input sample is mapped to an Action unit, which is in turn mapped to an expression, and action unit classification is considered a hidden task due to the lack of AU labels. Each expression classifier is learned before training using an empirical study of the relationship between expressions and AUs (Miyato, Maeda, Ishii & Koyama, 2018), which is in turn used to learn AU classifiers using gradient descent. Instead of using only the relationship between basic expressions and the corresponding AU probabilities, (Wang *et al.*, 2018c) exploited both expression-dependent and

expression-independent AU probabilities without using any extra large-scale expression-labeled facial images. First, the domain knowledge of expressions and AUs are summarized, based on which pseudo AU labels are generated for each expression. Then a Restricted Boltzmann Machine (RBM) is used to model prior joint AU distribution from the pseudo AU labels. Finally, AU classifiers are assumed to be linear functions with a sigmoid output layer and learned using Maximum Likelihood estimation (MLE) with regard to the learned AU label prior. Using a similar approach, (Peng & Wang, 2018) explored adversarial training for AU recognition instead of maximizing the log-likelihood of the AU classifier with regard learned AU label prior. Inspired by generative adversarial networks (GAN), AU classifiers are learned by minimizing the differences between AU output distribution from AU classifiers and pseudo AU label distribution derived from the summarized domain knowledge.

Unlike the prior approaches, (Wang, Peng, Chen & Ji, 2018b) modeled the relation between expressions and AU probabilities as inequalities instead of exact probabilities i.e., higher probabilities of occurrence have higher rankings than those with lower probabilities. Then expression-dependent ranking order among AUs is exploited to train the AU classifiers as a multi-label ranking problem by minimizing rank loss. Similarly, (Zhang, Dong, Hu & Ji, 2018a) comprehensively summarized the domain knowledge and exploited both expression-dependent and expression-independent AU probabilities to model the relationship among AU probabilities. Then multiple AU classifiers are jointly learned by leveraging the prior probabilities on AUs i.e., the derived relationships among AUs are represented in terms of inequality constraints among AU probabilities, which is incorporated into the objective function instead of the multi-label ranking framework as in (Wang *et al.*, 2018b).

Inspired by the idea of dual learning, (Wang & Peng, 2019) integrated the task of face synthesis along with AU recognition, where the latter is considered as the main task and the former as an auxiliary task. Specifically, AU labels are predicted using AU classifiers learned from the domain knowledge, which is formulated as the first objective term and a synthetic face is generated using predicted AU labels, which are optimized using the second objective term indicating the difference between the original and generated faces. Two conditional distributions

are learned, one for each task which is jointly optimized using the stochastic gradient descent method. All of the above-mentioned approaches for AU detection with weak expression labels have been extended to the problem of incomplete AU annotations by deploying additional loss components for the available partial AU annotations.

2.3.2 Incomplete Annotations

In this subsection, we have reviewed the existing algorithms, which fall under the category of WSL with incomplete annotations.

Detection of Expressions

(Cohen, Sebe, Cozman & Huang, 2003) investigated the prospect of exploiting unlabeled data for improving the classification performance of the system and proposed a structure learning algorithm of the Bayesian network. The facial features are extracted as motion features of nonrigid regions of the face, which capture the deformation of the face and are fed to the proposed Bayesian classifier for recognizing expressions. (Happy, Dantcheva & Bremond, 2019) addressed the problem of limited data with incomplete annotations for classifying facial expressions even with low intensity, as in real-life data. Label smoothing is used to prevent the model from obtaining high confidence scores, thereby retaining expressions with low intensities. Initially, they train a CNN model with limited labeled data in a supervised manner until an adequate performance is achieved. Subsequently, model parameters are further updated by finetuning using a portion of unlabelled data with high-confidence predictions, obtained by the current model in every epoch. (Florea, Badea, Florea, Racoviteanu & Vertan, 2020) improved the idea of center loss (Wen, Zhang, Li & Qiao, 2016) by maximizing the distance between the centroids of the different classes in the loss function. Pseudo labels are estimated based on the distances to centroids of different classes and used along with mix-up augmentation (Hongyi Zhang,Moustapha Cisse & Lopez-Paz, 2018) to avoid over-fitting.

Detection of Action Units

Missing Labels: (Song, McDuff, Vasisht & Kapoor, 2015) developed a Bayesian framework for AU recognition by encoding the sparsity i.e., only very few AUs are active at any moment, and co-occurrence structure of AUs i.e., overlapping of AUs in multiple groups via compressed sensing and group-wise sparsity inducing priors. (Wu, Lyu, Hu & Ji, 2015a) proposed a multi-label learning framework with missing labels to learn multi-label classifiers by enforcing the constraints of consistency between predicted labels and provided labels (label consistency) as well as with label smoothness i.e., labels of similar features should be close to each other along with modeling the co-occurrence relationships among AUs. However, (Li *et al.*, 2016) found that the constraint of label smoothness with shared feature space among AUs is violated for the task of AU recognition due to the diverse nature of the occurrence of AUs i.e., different AUs occur in different face regions, thereby features selected for one AU classifier are not discriminative for other AU classifier. Therefore, discriminative features are learned for each AU class before deploying the constraint of label smoothness. (Li, Wu, Zhao, Yao & Ji, 2019) further extended the idea of (Li *et al.*, 2016) to address the problem of class imbalance in two aspects: the number of positive AUs being much smaller than negative AUs in each sample (image) and the rate of positive samples of different AUs being significantly different. They have explored class cardinality bounds, where the model is learned by imposing the lower and upper bounds, obtained using a histogram of positive AUs into the objective function.

Incomplete Labels: (Shangfei, Quan & Qiang, 2017) deal with incomplete AU labels but complete expression labels by modeling the dependencies among AUs and the relationship between expressions and AUs with Bayesian Network (BN) using Maximum Likelihood estimation. However, Structural Expectation Maximization (SEM) is used to learn the parameters of BN for data with no AU labels. They have further extended the approach to estimate AU intensities. (Peng & Wang, 2019) explored an adversarial GAN-based approach with dual learning by leveraging domain knowledge of expressions and AUs along with facial image synthesis from predicted AUs. Specifically, the probabilistic duality between tasks and the dependencies among facial features, AUs, and expressions are explored in an adversarial learning

framework. Next, reconstruction loss is deployed by considering the constraints of the dual task of AU predictions from facial images and facial image synthesis from predicted AUs in addition to the standard supervised loss pertinent to AU labels and facial features.

Unlike the above two approaches, (Wu, Wang, Pan & Ji, 2017) used only incomplete AU annotations without expression labels and models the prior relationships among AUs using Restricted Boltzmann Machine (RBM). Multiple SVMs are used to learn AU classifiers using deep features by minimizing the error between predicted labels and ground-truth labels while simultaneously maximizing the log-likelihood of the AU label distribution model to be consistent with learned AU label distributions. Inspired by the idea of co-training, (Niu, Han, Shan & Chen, 2019) further improved the performance using a novel approach of multi-label co-regularization for semi-supervised AU recognition without expression labels.Specifically, a multi-view loss is designed to ensure the features generated from the two views are conditionally independent by orthogonalizing weights of AU classifiers of two views. Next, a co-regularization loss is designed to enforce the AU classifiers from two views to have similar predictions by minimizing the distance between two predicted probability distributions from two views. Subsequently, a graph convolutional network (GCN) is also used to model the strong relationships among different AUs and fine-tune the network.

2.3.3 Inaccurate Annotations

Since the process of the annotation of expressions or AU labels is a complex process, the annotations are highly vulnerable to noise. Therefore, there is an immense need to refine the inaccurate annotations of expression or AU labels.

Detection of Expressions

(Mollahosseini *et al.*, 2016b) investigated the consistency of noisy annotations of images crawled from web and performance of deep networks in handling noisy labels. AlexNet (Krizhevsky *et al.*, 2012) and WACV-Net (Mollahosseini, Chan & Mahoor, 2016a) are trained using training

data of true labels only, true labels with noisy labels, and true labels with noisy labels along with noise modeling (Tong Xiao, Tian Xia, Yi Yang, Chang Huang & Xiaogang Wang, 2015). All three scenarios are evaluated using the test set with true labels. It was observed that Alexnet outperforms WACVNet in all cases with the scenario of training with true labels having the best performance followed by noisy labels with noise modeling and noisy labels. (Barsoum, Zhang, Ferrer & Zhang, 2016) trained a deep CNN with noisy labels obtained from 10 taggers, which is analyzed using four different strategies for effective label assignment: majority-voting, multi-label learning, probabilistic label drawing, and cross-entropy loss. It was observed that strategies of multi-label learning fully exploit the label distribution and outperform the single-label strategy. (Zeng, Shan & Chen, 2018) proposed a 3-step framework, Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) to discover the latent true labels of noisy data with multiple inconsistent annotations. First, predictive models are trained for individually labeled data-sets. Next, pseudo annotations are generated from the trained predictive models to obtain multiple labels for each image of the labeled data-sets as well as large-scale unlabelled data.Finally, LTNet is trained to estimate the latent true labels by maximizing the log-likelihood of observed multiple pseudo annotations. (Zhang, Xu & Xu, 2021a) proposed a pose-invariant model by leveraging the noisy data on the web to boost FER performance in the wild. They have used clean data for jointly modeling the pose and classification task in order to stabilize the network. Then the noisy data is further exploited to enhance the pose-invariant feature learning by jointly learning the pose-modeling, noise-modeling, and classification tasks.

Detection of Action Units

(Zhao, Chu & Martinez, 2018) explored weakly supervised clustering on large-scale images from the web with inaccurate annotations to derive a weakly-supervised spectral algorithm that learns an embedding space to couple image appearance and semantics. Next, the noisy annotations are refined using rank order clustering by identifying groups of visually and semantically similar images They have further enhanced the approach by invoking stochastic extension to deal with large-scale images. (Benitez-Quiroz, Wang & Martinez, 2017) proposed a global-local loss

function by combining the local loss function, which emphasizes accurate detection by focusing on salient regions, however, requires very accurate labels for better convergence, which is circumvented by combining the global loss function that captures the global structure of images yielding consistent results for AU recognition.

2.3.4 Experimental Results

In this section, we will present the datasets used for validating the WSL models for classification proposed in the framework of weakly supervised learning along with the critical analysis of results.

2.3.4.1 Datasets

UNBC-McMaster (Lucey *et al.*, 2011): The database contains 200 video sequences of 48398 frames captured from 25 participants, who self-identified with shoulder pain. Each frame of the video sequence is labeled with 5 discrete intensity levels of AUs pertinent to pain ($A < B < C < D < E$)obtained by three certified FACS coders. Only the action units related to pain are considered: brow-lowering (AU4), cheek-raising (AU6), eyelid tightening (AU7), nose wrinkling (AU9), upper-lip raising (AU10), oblique lip raising (AU12), horizontal lip stretch (AU20), lips parting (AU25), jaw-dropping (AU26), mouth stretching (AU27) and eye-closure (AU43), however AU43 is assigned only binary labels. In addition to the annotations based on FACS, they have also provided labels of discrete pain intensities both at sequence-level and frame-level. Prkachin and Solomon Pain Intensity Scale (PSPI) of pain intensities are labeled at frame level with 16 discrete levels from 0-15 and Observer Pain Intensity (OPI) ratings are provided at sequence-level on a scale of 0 - 5.

CK+ (Lucey *et al.*, 2010) : The database consists of 593 video sequences captured in controlled laboratory conditions, where the emotions are spontaneously performed by 123 participants. All the video sequences are considered to vary from neutral face to peak formation of the facial expressions and the duration of the sequences varies from 10 to 60 frames. The video sequences

are labeled based on FACS with seven basic expression labels (including Contempt) and each emotion is defined by a prototypical combination of specific action units (AUs). Out of 593 sequences, it was found that only 327 sequences satisfy the labeling strategy, where each video sequence is labeled with the corresponding emotion label. Since the emotion labels are provided at video-level, static approaches assign the emotion label of the video sequence to the last one to three frames that exhibit the peak formation of the expression, and the first frame is considered a neutral frame.

MMI (Pantic, Valstar, Rademaker & Maat, 2005; Valstar & Pantic, 2010): The database contains 326 video sequences spontaneously captured in laboratory-controlled conditions from 32 subjects though it includes challenging variations such as large interpersonal variations, pose, etc compared to the CK+ database. The sequences are captured as onset-apex-offset i.e., the sequence starts with a neutral expression (onset), reaches the peak (apex), and returns again to a neutral expression (offset). As per the standard of FACS, 213 sequences are labeled with six basic facial expressions (excluding contempt) at the video level, of which 205 sequences are captured in frontal view. They have also provided frame-level annotations for some of the sequences. For approaches based on static images, only the first frame (neutral expression) and peak frames (apex of expressions) are considered.

DISFA (Mavadati, Mahoor, Bartlett, Trinh & Cohn, 2013): Denver Intensity of Spontaneous Facial Action dataset (DISFA) is created using 9 short video clips from YouTube, where the participants of 27 adults are allowed to watch the short video clips pertaining to various emotions. The facial expressions of each of the participants are captured with a high-resolution video of 1024x768 pixels with a frame rate of 20fps resulting in 1,30,754 frames in total. Each of these frames is annotated with action units along with the discrete intensity levels by FACS expert raters. The action units related to the expressions in the database are AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU17, AU20, AU25, and AU26, whose intensities are provided on a six-point ordinal scale (neutral < A < B < C < D < E).

Table 2.2 Comparative evaluation of performance measures for classification of expressions under various modes of WSL setting on most widely evaluated datasets

WSL Problem	Dataset	Method	Task	Model-type	Learning Model	Validation	Performance
Inexact	UNBC - McMaster	(Sikka *et al.*, 2014)	Pain [2 classes]	Dynamic	Gradient Boosting	LOSO	83.70
		(Wu *et al.*, 2015b)	Pain [2 classes]	Dynamic	HMM	LOSO	85.23
		(Chen *et al.*, 2022d)	Pain [2 classes]	Dynamic	Gradient Boosting	10-fold	85.60
		(Ruiz *et al.*, 2014)	Pain [2 classes]	Static	Gradient Descent	LOSO	85.70
		(Sikka *et al.*, 2016)	Pain [2 classes]	Dynamic	SGD	LOSO	87.00
		(Huang *et al.*, 2016)	Pain [2 classes]	Static	RF-IAF	LOSO	84.40
	CK+	(Wu *et al.*, 2015b)	Expression [7 classes]	Dynamic	HMM	LOSO	98.54
		(Sikka *et al.*, 2016)	Expression [7 classes]	Dynamic	SGD	10-fold	95.19
	Oulu-CASIA VIS	(Sikka *et al.*, 2016)	Expression Detection	Dynamic	Classical	10-fold	74.0
		(Xie *et al.*, 2019)	Expression Detection	Dynamic	Classical	5-fold	87.71
	BU-4DFE	(Xu & Mordohai, 2010)	Expression Detection	Dynamic	Classical	10-fold	63.83
Incomplete	RAF-DB	(Florea *et al.*, 2020)	Expression [7 classes]	Static	Margin-Mix	Conventional	70.68
	CK+	(Happy *et al.*, 2019)	Expression [7 classes]	Static	CNN	Conventional	99.35
	FER+	(Florea *et al.*, 2020)	Expression [7 classes]	Static	Margin-Mix	Conventional	81.25
Inaccurate	RAF-DB	(Zeng *et al.*, 2018)	Expression [7 classes]	Static	IPA2LT	Conventional	86.77
		(Zhang *et al.*, 2021a)	Expression [7 classes]	Static	CNN	Conventional	88.89
	AffectNet	(Zeng *et al.*, 2018)	Expression [7 classes]	Static	IPA2LT	Conventional	55.11
		(Zhang *et al.*, 2021a)	Expression [7 classes]	Static	CNN	Conventional	60.04

BP4D (Zhang *et al.*, 2014b): The dataset is captured from 41 participants, where each subject is requested to exhibit 8 spontaneous expressions and thereby 2D and 3D videos are obtained for each task. A total of 328 video sequences are obtained, where the frames exhibiting a high density of facial expressions are annotated with facial AUs. Due to the intensive process of FACS coding, the most expressive temporal segments i.e., 20-second segments are encoded. A total of 27 AUs are coded for the expression sequences and AU intensities are also coded on an ordinal scale of 0-5 for AU 12 and AU 14. Similar to BU-4DFE, this dataset is also used for analyzing facial expressions in dynamic 3D space.

Table 2.3 Comparative evaluation of performance measures for classification of action units under various modes of WSL setting on widely evaluated data-sets

WSL Problem	Dataset	Reference	Task	Model-type	Learning Model	Validation	Performance
Inexact	UNBC - McMaster	(Ruiz *et al.*, 2015)	AU [14 AUs]	Static	Gradient Descent	5-fold	0.235
		(Wang *et al.*, 2018c)	AU [6 AUs]	Static	RBM	5-fold	0.351
		(Wang & Peng, 2019)	AU [6 AUs]	Static	RBM	5-fold	0.400
		(Peng & Wang, 2018)	AU [6 AUs]	Static	RAN	5-fold	0.376
		(Zhang *et al.*, 2018a)	AU [3 AUs]	Static	LBFGS(Moller, 1993)	5-fold	0.510
	CK+	(Ruiz *et al.*, 2015)	AU [12 AUs]	Static	Gradient Descent	5-fold	0.469
		(Wang *et al.*, 2018c)	AU [12 AUs]	Static	RBM	5-fold	0.705
		(Wang & Peng, 2019)	AU [12 AUs]	Static	RBM	5-fold	0.740
		(Peng & Wang, 2018)	AU [12 AUs]	Static	RAN	5-fold	0.715
		(Zhang *et al.*, 2018a)	AU [8 AUs]	Static	LBFGS(Moller, 1993)	5-fold	0.732
	MMI	(Ruiz *et al.*, 2015)	AU [14 AUs]	Static	Gradient Descent	5-fold	0.431
		(Wang *et al.*, 2018c)	AU [13 AUs]	Static	RBM	5-fold	0.516
		(Wang & Peng, 2019)	AU [13 AUs]	Static	RBM	5-fold	0.530
		(Peng & Wang, 2018)	AU [13 AUs]	Static	RAN	5-fold	0.520
		(Zhang *et al.*, 2018a)	AU [8 AUs]	Static	LBFGS(Moller, 1993)	5-fold	0.481
	DISFA	(Ruiz *et al.*, 2015)	AU [12 AUs]	Static	Gradient Descent	5-fold	0.371
		(Wang *et al.*, 2018c)	AU [12 AUs]	Static	RBM	5-fold	0.424
Incomplete	UNBC-McMaster	(Shangfei *et al.*, 2017)	AU [6 AUs]	Static	Bayesian Network	5-fold	0.183
		(Song *et al.*, 2015)	AU [6 AUs]	Static	Bayesian Model	5-fold	0.450
		(Wu *et al.*, 2015a)	AU [6 AUs]	Static	Gradient Descent	5-fold	0.146
		(Ruiz *et al.*, 2015)	AU [6 AUs]	Static	Gradient Descent	5-fold	0.292
		(Wang *et al.*, 2018c)	AU [6 AUs]	Static	RBM	5-fold	0.502
		(Wang & Peng, 2019)	AU [6 AUs]	Static	RBM	5-fold	0.514
		(Peng & Wang, 2018)	AU [6 AUs]	Static	RAN	5-fold	0.472
		(Wang *et al.*, 2018b)	AU [6 AUs]	Static	Classical	5-fold	0.521
		(Peng & Wang, 2019)	AU [6 AUs]	Static	DSGAN	5-fold	0.520
	CK+	(Shangfei *et al.*, 2017)	AU [13 AUs]	Static	Bayesian Network	5-fold	0.781
		(Song *et al.*, 2015)	AU [13 AUs]	Static	Bayesian Model	5-fold	0.696
		(Wu *et al.*, 2015a)	AU [13 AUs]	Static	Gradient Descent	5-fold	0.652
		(Ruiz *et al.*, 2015)	AU [14 AUs]	Static	Gradient Descent	5-fold	0.594
		(Wang *et al.*, 2018c)	AU [12 AUs]	Static	RBM	5-fold	0.787
		(Wang & Peng, 2019)	AU [12 AUs]	Static	RBM	5-fold	0.806
		(Peng & Wang, 2018)	AU [12 AUs]	Static	RAN	5-fold	0.792
		(Peng & Wang, 2019)	AU [12 AUs]	Static	DSGAN	5-fold	0.792
		(Zhang *et al.*, 2018a)	AU [8 AUs]	Static	LBFGS(Moller, 1993)	5-fold	0.754
	MMI	(Shangfei *et al.*, 2017)	AU [13 AUs]	Static	Bayesian Network	5-fold	0.438
		(Song *et al.*, 2015)	AU [13 AUs]	Static	Bayesian Model	5-fold	0.447
		(Wu *et al.*, 2015a)	AU [13 AUs]	Static	Gradient Descent	5-fold	0.432
		(Ruiz *et al.*, 2015)	AU [13 AUs]	Static	Gradient Descent	5-fold	0.530

Continued on next page

WSL Problem	Dataset	Reference	Task	Model-type	Learning Model	Validation	Performance
Incomplete	MMI	(Wang *et al.*, 2018c)	AU [13 AUs]	Static	RBM	5-fold	0.531
		(Wang & Peng, 2019)	AU [13 AUs]	Static	RBM	5-fold	0.561
		(Peng & Wang, 2018)	AU [13 AUs]	Static	RAN	5-fold	0.529
		(Peng & Wang, 2019)	AU [13 AUs]	Static	DSGAN	5-fold	0.537
	DISFA	(Shangfei *et al.*, 2017)	AU [12 AUs]	Static	Bayesian Network	5-fold	0.430
		(Song *et al.*, 2015)	AU [12 AUs]	Static	Bayesian Model	5-fold	0.426
		(Wu *et al.*, 2015a)	AU [12 AUs]	Static	Gradient Descent	5-fold	0.382
		(Ruiz *et al.*, 2015)	AU [12 AUs]	Static	Gradient Descent	5-fold	0.428
		(Wang *et al.*, 2018c)	AU [12 AUs]	Static	RBM	5-fold	0.522
	BP4D	(Song *et al.*, 2015)	AU [12 AUs]	Static	Bayesian Model	Training : 60%	0.400
		(Wu *et al.*, 2017)	AU [12 AUs]	Static	RBM	Validation : 20%	0.452

2.3.4.2 Experimental Protocol

Depending on the mode of annotation, the datasets are further modified to match the corresponding task at hand to validate the techniques. For the task of classification, the performance of expression and AUs are expressed in terms of percentage and F1-score respectively.

Inexact Annotations: Though UNBC-McMaster dataset is primarily used for both regression and classification, it has been explored for expression classification by converting the ordinal labels (OPI ratings) of the data-set to binary labels based on a threshold i.e., OPI > 3 is treated as pain and OPI=0 as no pain, which results in a total of 149 sequences with 57 positive bags and 92 negative bags. For AU classification, 7319 frames are chosen from 30 video sequences of 17 subjects that exhibit the expression of pain with PSPI > 4. Six AU labels are associated with the chosen frames i.e., AU4, AU6, AU7, AU9, AU10, and AU43, which have a dependency on expression labels of pain. The bag label of each pain sequence is considered as the maximum of frame labels. Out of 25 subjects, 15 are used for training, 9 for validation, and 1 for testing.

In the case of CK+ dataset, 327 sequences are considered for expression classification, whereas for AU classification, 309 sequences of 106 subjects are chosen from 593 sequences of 123 subjects based on the occurrence of AU labels i.e., AU labels, which are available for more than 10% of all frames are chosen to result in 12 AU labels. MMI dataset is used for AU classification by considering sequences, where AUs are available for more than 10% of all samples, resulting in 171 sequences from 27 subjects with 13 labels. For the classification of DISFA dataset, 482

apex frames are chosen based on AU intensity levels, for which expression labels are obtained by FACS. Similar to CK+ and MMI, 9 AUs are considered, whose occurrence is greater than 10% 5-fold cross-validation is deployed, where 20% of the whole database is used as validation set according to subjects. All the experiments are conducted as subject-independent protocol.

Incomplete Annotations: In the context of incomplete annotations, UNBC-McMaster, CK+, MMI, and DISFA datasets are used for AU detection, where training data is obtained with missing AU labels by dropping some of the AU labels. In the case of the BP4D dataset for AU classification, only partial AU annotations are considered without expression labels. Images of 60% of subjects are used for training, 20% for validation, and the last 20% for testing. For all the datasets, AU labels for 50% of the training samples are randomly removed to incorporate the setting of incomplete annotations. The experiment is repeated for times and the average score of the F1 measure is used for validation.

2.3.5 Critical Analysis

Although the classification of facial expressions or action units is well explored in the framework of supervised learning, it is still an under-researched problem in the setting of WSL. Recently, action unit detection has relatively drawn much attention compared to the problem of expression detection in the context of WSL. This could be due to the fact that facial AUs cover a wide range of facial expressions rather than a limited six basic universal expressions, which have huge potential in a lot of real-time applications. Facial expressions or AUs are dynamic processes, which evolve over time, and thereby temporal information of expression or AUs conveys significant information about facial behavior. However, temporal information is not well explored for the problem of AU detection though it has been investigated for pain detection in the framework of WSL (Sikka, 2014; Sikka *et al.*, 2016). Current works on expression detection in WSL have used max-pooling or displacement of facial landmarks for extracting the dynamic information of video sequences pertinent to expression. Displacement of facial landmarks was found to capture the temporal dynamic better than max-pooling, which has been reflected in the work of (Wu

et al., 2015b), where they have used displacement of facial landmarks in conjunction with HMM instead of max-pooling.

Most of the current research on facial behavior analysis pertinent to WSL is based on classical machine learning approaches (shallow networks) though deep-learning-based approaches are gaining attention in recent years. One of the major bottlenecks in using deep learning for FER is the lack of sufficient data for facial expressions or AUs as they are sparse in nature i.e., facial expressions or AUs occur only for a limited duration of time in a video sequence. Deep networks were proven to be robust in handling the wide range of intra-variations such as illumination, pose, identity, etc better than shallow networks when a large amount of data is provided (Wang *et al.*, 2017). Therefore, pretrained networks on face recognition (Parkhi *et al.*, 2015) are used for expression or AU recognition by retaining the lower layers and fine-tuning only the higher layers as they represent the task-relevant features to handle the problem of limited data-set of facial expressions (Kaya, Grpnar & Salah, 2017).

However, it was found that the abstract feature representation of higher layers still holds the expression-unrelated information pertinent to subject identity even after fine-tuning the face verification nets (Ding *et al.*, 2017). To tackle this problem, large-scale in-the-wild FER datasets (Benitez-Quiroz, Srinivasan, Feng, Wang & Martínez, 2017), (Mollahosseini, Hasani & Mahoor, 2019) have been captured and made available to the research community in recent years, where the noisy annotations are refined using WSL approaches related to inaccurate annotations (Zeng *et al.*, 2018), (Li *et al.*, 2017). Recently, few works (Wang *et al.*, 2018c), (Wang & Peng, 2019) have explored RBM for AU detection. To the best of our knowledge, no work has been done using deep learning architectures for expression detection in the framework of inexact annotations though it has been recently explored in WSL with incomplete annotations (Happy *et al.*, 2019).

Most of the current research on AU detection has been focused on exploiting the domain knowledge of dependencies among AUs and between basic expressions and AUs in the framework of WSL, where expression labels act as weak supervision for AU detection. Similarly,

expression detection is widely explored in the context of pain detection in the framework of WSL. Since the experimental strategy of (Ruiz *et al.*, 2015), (Shangfei *et al.*, 2017), (Song *et al.*, 2015) and (Wu *et al.*, 2015a) are different from that of (Wang *et al.*, 2018c), (Wang & Peng, 2019) and (Peng & Wang, 2018), (Wang *et al.*, 2018c) have re-conducted the experiments of (Ruiz *et al.*, 2015), (Shangfei *et al.*, 2017), (Song *et al.*, 2015) and (Wu *et al.*, 2015a) with the setup of (Wang *et al.*, 2018c) in order to have fair comparison. The detailed comparison of results of current state-of-the-art methods on widely used datasets is shown in Table 2.3. For expression detection, the performance measure reflects the performance of sequence-level classification for data with inexact annotations and frame-level classification for data with incomplete annotations. The performance metric of accuracy measure is used for expression detection and average F1-Score for AU detection.

2.4 Weakly Supervised Learning for Regression

Regression can be formulated as ordinal regression and continuous regression. Ordinal regression algorithms are the class of machine learning algorithms, which deals with the task of recognizing the patterns on a categorical scale that reflects the ordering between the labels. Continuous regression algorithms deal with estimating the intensities of continuous labels such as valence or arousal. Though the problem of regression is not well explored as in the case of the classification task, it has been recently gaining attention due to its immense potential in many real-world applications. In the context of facial behavior analysis, regression deals with the estimation of intensity levels of facial expressions or AUs. Similar to the classification of facial expressions, regression also follows the same methodology of major building blocks i.e., Preprocessing, Feature Extraction, and Model Generation. However, the model is trained to capture the intensities for the task of regression rather than classifying the behavior patterns. Even though few algorithms have been proposed for the task of estimating the intensities of facial expressions or AUs (Wang *et al.*, 2017; Tavakolian & Hadid, 2018) in a fully-supervised setting, the problem is still at the rudimentary stage in the framework of WSL. We have classified the existing approaches relevant to the regression of facial expressions or AUs in the framework of WSL as

inexact, incomplete, and inaccurate techniques based on the mode of WSL setting similar to that of classification as described in Section 2.3.

2.4.1 Inexact Annotations

In the case of regression with inexact annotations, the intensity levels of expressions or AUs are provided at the global level i.e., at the sequence level, the goal is to estimate the intensity level of individual frames or sub-sequences using sequence-level labels.

Expression Intensity Estimation

(Ruiz *et al.*, 2016) proposed multi-instance dynamic ordinal random fields (MI-DORF) for estimating ordinal intensity levels of frames, where the ordinal variables are modeled as normal distribution and the relationship between the given observation (frame) and latent ordinal value is obtained by projecting the given observation (frame) onto the ordinal line, which is divided by the consecutive overlapping cutoff points of the normal distributions. Next, the temporal information is modeled across the consecutive latent ordinal variables to ensure the smoothness of the latent ordinal states. (Gnana Praveen *et al.*, 2020) further improved the performance using deep 3D CNN model (I3D (Carreira & Zisserman, 2017)) by integrating multiple instance learning into adversarial deep DA (Ganin & Lempitsky, 2015) framework for pain intensity estimation, where source domain is assumed to have fully annotated videos and target domain has periodically annotated weak labels. (Yang *et al.*, 2018) proposed an approach for student engagement prediction in-the-wild using multiple-instance regression, where the input video (bag) was divided into segments (instances) and spatiotemporal features of each segment are fed to an LSTM network followed by 3 fully connected layers to obtain the regressed value of engagement intensity.

Action Units Intensity Estimation

(Ruiz, Rudovic, Binefa & Pantic, 2018) extended the idea of (Ruiz *et al.*, 2016) for AU intensity estimation by modeling the relationship between the weak sequence-level label and instance label using two strategies: maximum or relative values of instance labels. They have also further extended the approach to partially labeled data, where the sequence-level labels are provided along with partial instance-level labels. Unlike the conventional framework of MIL, (Zhang, Zhao, Dong, Hu & Ji, 2018b) explored domain knowledge of relevance using two labels of peak and valley frames. Specifically, they have considered three major factors: Ordinal relevance, intensity smoothness, and relevance smoothness based on the gradual evolving process of facial behavior Ordinal relevance ensures the relevant intensity levels based on the proximity of neighboring frames, whereas intensity and relevance smoothness constrains the smooth evolution of facial appearance and ordinal relevance respectively. Due to the relevance of local patches for AUs, (Zhang, Jiang, Wu, Fan & Ji, 2019b) further improved the performance by developing a patch-based deep model using attention mechanisms for feature fusion and label fusion to capture the spatial relationships among local patches and temporal dynamics pertinent to each AU respectively. Since the contribution of local patches and temporal dynamics vary for different AUs, the attention mechanism was further augmented by learnable task-related context.

2.4.2 Incomplete Annotations

In this framework, the intensities of the frames are provided only for a subset of the training data. The goal of the task is to generate a robust training model for predicting intensity values of test data at frame-level using partially labeled data along with unlabeled data.

Expression Intensity Estimation

To the best of our knowledge, only one work has been done related to this problem, where (Zhao, Gan, Wang & Ji, 2016) proposed a max-margin-based ordinal support vector regression using ordinal relationship, which is flexible and generic in handling the varying level of annotations

and a linear model is learned by solving the optimization problem using the Alternating Direction Method of Multipliers (ADMM) to predict the frame-level intensity of the test image.

Action Units Intensity Estimation

(Zhang, Dong, Hu & Ji, 2018) designed a deep convolutional neural network for intensity estimation of Action Units(AUs) using annotations of only peak and valley frames. The parameters of CNN are learned by encoding domain knowledge of facial symmetry, temporal intensity ordering, relative appearance similarity, and contrastive appearance difference. CNN is designed as 3 convolutional layers followed by 3 max-pooling layers and 1 fully connected layer. (Zhang, Fan, Dong, Hu & Ji, 2019a) further extended (Zhang *et al.*, 2018) to joint estimation of multiple AU intensities by introducing a task index to update the corresponding parameters of the fully connected layer. They have also used a lot of unlabelled frames in addition to labeled key frames for training to handle over-fitting under the framework of semi-supervised learning. (Wang, Pan, Wu & Ji, 2019) extended the idea of (Wu *et al.*, 2017) for AU intensity estimation, where RBM is used to model AU intensity distribution and regularize the prediction model of AU intensities. (Zhang *et al.*, 2019c) further improved the performance by jointly learning the representation and estimator using partially labeled frames by incorporating human knowledge of soft and hard constraints. Specifically, the sequences are segmented into monotonically increasing segments, and temporal label ranking and positive AU intensity levels are considered hard constraints, and temporal label smoothness and temporal feature smoothness are considered as soft constraints.

2.4.3 Experimental Results

In this section, the datasets used for validating the WSL models proposed in the framework of weakly supervised learning for ordinal regression along with the critical analysis of results pertinent to the WSL approaches are presented.

2.4.3.1 Datasets

With the advancement of state-of-the-art FER systems, regression-based approaches are gaining attention as humans make use of the wide range of intensity of facial expressions to convey their feelings. To cover this wide range of facial expressions, several databases have been developed with intensity levels of facial expressions such as pain, action units, etc.

FERA 2015 Challenge (Valstar *et al.*, 2015): The dataset is drawn from BP4D (Zhang *et al.*, 2014b) and SEMAINE (McKeown, Valstar, Cowie, Pantic & Schroder, 2012) databases for the task of AU occurrence and intensity estimation, where only five AUs from BP4D i.e., AU6, AU10, AU12, AU14, and AU17 are considered for AU intensity estimation and 14 AUs from both SEMAINE and BP4D for occurrence detection i.e., AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, AU25, AU28, and AU45. The original dataset of BP4D is used as the training set, where training data is drawn from 21 subjects, development set from 20 subjects, and the dataset is further extended for test-set captured from 20 subjects, resulting in 75,586 images in the training partition, 71,261 images in development partition and 75,726 in the testing partition. Similarly for the SEMAINE dataset, 48,000 images are used for training, 45,000 for development, and 37,695 for testing. The entire dataset is annotated frame-wise for AU occurrence and intensity level for the corresponding subset of AUs. For the BP4D-extended set, the onset and offsets are treated as B-level of intensity. In both datasets, most facially-expressive segments are coded for AUs and AU intensities. The intensity levels of AUs are coded on an ordinal scale of 0-5.

BU-4DFE (Yin, Chen, Sun, Worm & Reale, 2008): The database is an extended version of BU-3DFE, where the facial behavior of static 3D space is further extended to include dynamic space. The dataset contains 606 3D video sequences obtained from 101 subjects by allowing each subject to exhibit six prototypical facial expressions. Each video sequence has 100 frames with a resolution of 1040x1329, which results in a total of approximately 60,600 frames. This dataset is widely used for multi-view 3D facial expression analysis.

Table 2.4 Comparative evaluation of performance measures for regression of expressions or action units under various modes of WSL setting on most widely evaluated datasets

WSL Problem	Dataset	Reference	Task	Data Type	Learning Model	Validation	MAE	PCC	ICC
Inexact	UNBC-McMaster	(Ruiz *et al.*, 2018)	Pain [6 levels]	Static	DORF	LOSO	0.710	0.360	0.340
		(Gnana Praveen *et al.*, 2020)	Pain [6 levels]	Dynamic	WSDA	LOSO	0.714	0.630	0.567
		(Zhang *et al.*, 2018b)	Pain [6 levels]	Static	BORMIR	LOSO	0.821	0.605	0.531
	DISFA	(Ruiz *et al.*, 2018)	Action Unit [12 AUs]	Static	DORF	5-fold	1.130	0.400	0.260
		(Zhao *et al.*, 2016)	Action Unit [12 AUs]	Static	OSVR	5-fold	1.380	0.350	0.150
		(Zhang *et al.*, 2018b)	Action Unit [12 AUs]	Static	BORMIR	5-fold	0.789	0.353	0.283
		(Zhang *et al.*, 2019b)	Action Unit [12 AUs]	Dynamic	CFLF	3-fold	0.329	-	0.408
	FERA 2015	(Zhang *et al.*, 2018b)	Action Unit [5 AUs]	Static	BORMIR	(Valstar *et al.*, 2015)	0.852	0.635	0.620
		(Zhang *et al.*, 2019b)	Action Unit [5 AUs]	Dynamic	CFLF	(Valstar *et al.*, 2015)	0.741	-	0.661
Incomplete	UNBC-McMaster	(Ruiz *et al.*, 2018)	Pain [6 levels]	Static	DORF	LOSO	0.510	0.460	0.460
		(Zhao *et al.*, 2016)	Pain [6 levels]	Static	OSVR	LOSO	0.951	0.544	0.495
	DISFA	(Ruiz *et al.*, 2018)	Action Unit [12 AUs]	Static	DORF	5-fold	0.480	0.420	0.380
		(Zhao *et al.*, 2016)	Action Unit [12 AUs]	Static	OSVR	5-fold	0.800	0.370	0.290
		(Zhang *et al.*, 2018)	Action Unit [12 AUs]	Dynamic	CNN	3-fold	0.330	-	0.360
		(Zhang *et al.*, 2019a)	Action Unit [12 AUs]	Dynamic	CNN	3-fold	0.330	-	0.350
		(Wang *et al.*, 2019)	Action Unit [12 AUs]	Static	RBM	3-fold	0.431	0.592	0.549
		(Zhang *et al.*, 2019c)	Action Unit [12 AUs]	Dynamic	CNN	5-fold	0.910	0.370	0.350
	FERA 2015	(Wang *et al.*, 2019)	Action Unit [5 AUs]	Static	RBM	(Valstar *et al.*, 2015)	0.728	0.605	0.585
		(Shangfei *et al.*, 2017)	Action Unit [5 AUs]	Static	Bayesian Network	(Valstar *et al.*, 2015)	-	0.638	0.610
		(Zhao *et al.*, 2016)	Action Unit [5 AUs]	Static	OSVR	Valstar *et al.* (2015)	1.077	0.545	0.544
		(Zhang *et al.*, 2019a)	Action Unit [5 AUs]	Dynamic	CNN	(Valstar *et al.*, 2015)	0.640	-	0.670
		(Zhang *et al.*, 2018)	Action Unit [5 AUs]	Dynamic	CNN	(Valstar *et al.*, 2015)	0.660	-	0.670
		(Zhang *et al.*, 2019c)	Action Unit [5 AUs]	Dynamic	CNN	(Valstar *et al.*, 2015)	0.870	0.620	0.600
	CK+	(Zhao *et al.*, 2016)	Expression [11 levels]	Static	OSVR	10-fold	1.981	0.729	0.716
	BU-4DFE	(Zhao *et al.*, 2016)	Expression [11 levels]	Static	OSVR	LOSO	2.242	0.545	0.503

2.4.3.2 Experimental Protocol

In the case of regression, the performance of expression is expressed as Mean Square Error (MSE). The intensities of pain and AUs are evaluated in terms of Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and Intraclass Correlation Coefficient (ICC).

Inexact Annotations: In case of regression on the UNBC-McMaster dataset, PSPI labels of the frames are converted to 6 ordinal levels:0(0), 1(1), 2(2), 3(3), 4-5(4), 6-15(5). The bag label of each pain sequence is considered as the maximum of frame labels. Out of 25 subjects, 15 are used for training, 9 for validation, and 1 for testing. For the EmotiW dataset, the training and validation dataset provided by the organizers is used for training by manually splitting the data to compensate for the class imbalance. The performance measure is evaluated on the test dataset provided by the organizers.

Incomplete Annotations: For regression, only 8.8% of total annotations are considered for the UNBC-McMaster dataset for the task of pain regression. Similarly, CK+ and BU-4DFE datasets are used for expression regression, where 327 and 120 sequences with a total of 5876 and 2289 frames respectively are considered, out of which annotations are provided for onset and apex frames. For the task of AU regression in UNBC-McMaster and DISFA datasets, only 10% of annotated frames are considered in (Ruiz *et al.*, 2018), (Wang *et al.*, 2019) and (Zhao *et al.*, 2016) to incorporate the setting of incomplete annotations, whereas (Zhang *et al.*, 2018b) and (Zhang *et al.*, 2018) considered only the annotations of peak and valley frames. In the case of FERA 2015 dataset, official training and development sets provided by FERA 2015 challenge (Valstar *et al.*, 2015) are deployed. Similar to UNBC-McMaster and DISFA, (Wang *et al.*, 2019; Shangfei *et al.*, 2017) considers only 10% of annotated frames while (Zhao *et al.*, 2016; Zhang *et al.*, 2018b; Zhang *et al.*, 2018) considers annotations of peak and valley frames.

2.4.4 Critical Analysis

The intensity estimation of facial expressions or AUs is more challenging than the task of classification due to the complexity of capturing the subtle variation of facial appearance and obtaining the annotations of intensity levels as they are scarce and expensive. Therefore, in general, the task of intensity estimation is still an under-researched problem compared to the task of classification. A similar phenomenon has been observed in the framework of WSL for which the problem of intensity estimation is rarely explored. Since temporal dynamics plays a crucial role in conveying significant information for the task of intensity level estimation, (Zhang *et al.*,

2018b) and (Ruiz *et al.*, 2018) modeled the relevant ordinal relationship across temporal frames by incorporating intensity and relevance smoothness into the objective function while (Zhang *et al.*, 2018) and (Zhang *et al.*, 2019a) invoked facial symmetry and contrastive appearance difference in addition to the temporal relevance and regression in a deep learning framework, which was found to outperform the former approaches. (Kaur, Mustafa, Mehta & Dhall, 2018) and (Yang *et al.*, 2018) formulated the task of AU intensity estimation in the setting of deep multi-instance regression and captured the temporal dynamics of frames using LBP-TOP features and represented the video as maximum or average of the corresponding instance frames.

With the advent of deep learning architectures and their breakthrough performance in many applications under real-time uncontrolled conditions, few approaches have explored deep models for AU intensity estimation in recent days. However, one of the major challenges in using deep models for intensity estimation of facial expressions or AUs is the requirement of a large number of intensity annotations, which is very expensive and demands strong domain expertise for obtaining the annotations. Therefore, estimation of intensity levels using deep models in the framework of WSL still remains to be an open problem though few approaches have explored the problem in a fully supervised setting (Gudi, Tasli, den Uyl & Maroulis, 2015), (Walecki *et al.*, 2017). To the best of our knowledge, only two works have exploited deep models for AU intensity estimation i.e., (Zhang *et al.*, 2018) and (Wang *et al.*, 2019), and no work has been done for estimating expression intensity levels with deep models.

To compare the work of (Ruiz *et al.*, 2018) with the conventional approach of pain detection (Sikka *et al.*, 2014), MILBOOST is deployed, and the output probabilities of pain detection are normalized between 0 and 5 to have a fair comparison with that of (Ruiz *et al.*, 2018). The performance metric used for the validation of the state-of-the-art approaches for AU intensity estimation is Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and Intra-class correlation (ICC). PCC is normally used to measure the linear association between predicted values and ground truth and ICC is used for correlation among annotators. MAE measures the error between predictions and ground truth, which is typically used for applications pertinent to ordinal regression. The expression intensity values for student engagement level prediction in

(Yang *et al.*, 2018) and (Kaur *et al.*, 2018) are validated with the performance measure of Mean Square Error (MSE). The detailed comparison of results of current state-of-the-art methods on widely used datasets is shown in Table 2.4.

2.5 Challenges and Opportunities

In addition to the challenges related to data labeling, there are other challenges to developing a robust FER system. Though it has been discussed extensively in the existing literature (Gehrig & Ekenel, 2013), (Martinez & Valstar, 2016), we will briefly review some of the generic challenges along with the potential research directions pertinent to FER.

2.5.1 Challenges

In general, the task of facial analysis suffers from a lot of challenges, which is prevalent in most face-related applications such as identity recognition, attribute recognition, etc.

2.5.1.1 Dataset Bias

Though there has been a shift in data capture from laboratory-controlled conditions to in-the-wild uncontrolled environments, the datasets are generally captured in a specific environment, which may vary across datasets and thereby results in different data distribution of various datasets. Typically, state-of-the-art approaches are evaluated on limited datasets and show superior performance. However, when deployed on different datasets, these algorithms may fail to retain their superior performance due to the differences in the distribution of datasets, often termed data-set bias, which is a prevalent problem in the field of machine learning. To address the problem of data-set bias, a few approaches (Benitez-Quiroz *et al.*, 2016) have used multiple datasets for training by merging the datasets and evaluated on different datasets. Even though merging multiple datasets may increase the training data and thereby achieve better generalization, it may suffer from label subjectivity as discussed in 2.5.1.3. A few more approaches conducted

cross-database experiments to validate the generalizability of the algorithm by evaluating the algorithm on a dataset different from training data (Ruiz *et al.*, 2015), (Wang *et al.*, 2018c).

2.5.1.2 Data Sparsity and Class Imbalance

Since the V appearance of the face varies from person to person due to age, civilization, ethnicity, cosmetics, eyeglasses, etc., the detection of facial expressions is a challenging task. In addition to the personal attributes, variations due to pose, occlusion, and illumination are prevalent in unconstrained scenarios of facial expressions, which leads to high intra-class variability. Therefore, there is an immense need for large-scale data-set with a wide range of intra-class variation. In most machine-learning-based applications, the performance of the system is highly influenced by the quality and quantity of data, which has been reflected in the superior performance of deep learning architectures. Though humans are capable of exhibiting a wide range of facial expressions, most of the existing datasets are developed based on basic universal expressions and limited AUs as they are more frequent in our everyday life.

However, these facial expressions or AUs are generally sparse in nature as they are expressed only a few frames, resulting in a huge class imbalance, which is also reflected in the UNBC-McMaster Pain database (Lucey *et al.*, 2011). For instance, eliciting a smile is a frequently occurring expression, whereas expressions such as disgust, and anger are less common expressions, thereby resulting in limited data on those less frequent expressions. (He & Garcia, 2009) investigated and provided a comprehensive review of state-of-the-art approaches for learning from imbalanced data along with metrics used for evaluating the performance of the systems. (Jaiswal *et al.*, 2018) used cumulative attributes with a deep learning model as a two-stage cascaded network. In the first stage, original labels are converted to cumulative attributes, and the CNN model is trained to output a cumulative attribute vector. Next, a regression layer is used to convert the cumulative attribute vectors to real-valued output. They have also used evaluated the system with Euclidean loss and log-loss and found that the latter outperforms the former.

2.5.1.3 Label Subjectivity and Identity Bias

Label subjectivity and Identity Bias are two major factors induced by the subjective nature of the annotators and varied responses of expressions. Compared to other problems of computer vision, labeling facial expressions is a highly complex process as it is subjective in nature. Manual annotation of AUs is even more challenging compared to the prototypical categorical expressions due to the increased range of facial behavior. Moreover, the annotation of AUs requires domain expertise certified by FACS coding system, which is a time-consuming and laborious task and thereby highly prone to errors induced by annotators. The process of annotation becomes more complex for intensities of expressions or AUs as the difference between different intensity levels is very subtle, which is very challenging even for expert annotators. The continuous dimensional model further complicates the process of labeling especially when annotators are asked to label every frame of the video sequences as a continuous range of values for the intensities will be more sensitive than discrete values, which will result in differences in the labels for the same intensity of facial expression. Another major factor in obtaining annotations for the continuous dimensional model is the reaction time of annotators.

To alleviate the impact of label subjectivity, the dataset is typically labeled by the strategy of crowd-sourcing, where labels are refined from several annotators (Li *et al.*, 2017). In addition to label subjectivity, there can be variations in the facial appearance due to heterogeneity of subjects termed identity bias, i.e., ambiguity induced by the subjective nature of humans. For instance, the expression of sadness is often misinterpreted as a neutral expression as the V appearance of sadness is very close to that of a neutral expression.

2.5.1.4 Tool for Semi-Automatic Annotation

Due to the above-mentioned challenges in Section 2.5.1.3, it was found that manual annotations are extremely challenging and even impossible to obtain for large scale data-sets. Though WSL-based approaches reduce the need for exact and complete annotations, the need for minimal annotations for large scale data-sets has motivated many researchers to automate the process of

annotations, which can be further refined by expert annotators to minimize the burden on human labor (Dhall *et al.*, 2012). Since a fully automatic tool for obtaining annotations of expressions or AUs is not feasible and reliable, a semi-automatic tool seems to be a more plausible approach to obtain reliable annotations for large scale data-sets, especially for the case of continuous affect model which is more vulnerable to noise and highly complex process to discriminate the subtle variations across the frames.

2.5.1.5 Efficient Feature Representation

In the field of machine learning, one of the major characteristics of feature representation is to retain the relevant information on target labels while still minimizing the entropy of features. For the task of FER, the features are expected to be robust to face appearance variations such as pose, occlusion, illumination, blur, etc still retaining the relevancy of expressions. (Sariyanidi *et al.*, 2015) have provided a comprehensive analysis of local as well as global hand-crafted features such as Gabor, SIFT, LBP, etc by revealing its advantages and limitations to various key challenging factors. They have further analyzed the feature selection approaches to refine the feature representation and showed that fusion-based representations outperform individual feature representations. Another promising line of approach is to incorporate the temporal dynamics in the feature representation, which has outperformed the approaches based on static representation. Typically, Three Orthogonal Planes (TOP) features are extended with handcrafted features such as LBP-TOP (Zhao & Pietikainen, 2007), LGBP (Almaev & Valstar, 2013), etc to incorporate the dynamic information in the static feature representations.

With the ubiquity of deep learning-based approaches for various computer vision problems, there has been a shift from handcrafted features to learned features, where pretrained models of face verification are used for finetuning with the facial expressions or AU data due to the limited data pertinent to facial expressions or AU. However, as mentioned in (Ding *et al.*, 2017), the abstract feature representation of higher layers still retains identity-relevant information even after fine-tuning the face verification net with FER data. To overcome the problem of limited data and take advantage of the success of deep learning for FER, (Egede *et al.*, 2017) have explored

the fusion of hand-crafted and learned features for automatic estimation of pain intensity and showed superior performance over state-of-the-art approaches. RNN-based approaches are found to be promising in capturing the temporal dynamics which was shown robust performance in various computer vision, speech, and NLP applications. (Kim *et al.*, 2019) explored efficient feature representation robust to expression intensity variations by encoding facial expressions in two stages. First, spatial features are obtained through CNN using five objective terms to enhance the separability of the expression classes. Second, the obtained spatial features are fed to LSTM to learn the temporal features. (Wang *et al.*, 2013) studied the contribution of spatiotemporal relationship among facial muscles for efficient FER. They have modeled the facial expression as a complex activity of temporally overlapping facial events, where they proposed an Interval Temporal Bayesian Network to capture the temporal relations of facial events for FER.

2.5.2 Potential Research Directions

In this section, we will present some of the potential research directions for the advancement of facial behavior analysis in the framework of WSL.

2.5.2.1 Exploiting Deep Networks

With the massive success of deep learning architectures, many researchers have leveraged deep models for various applications in computer vision such as object detection, face recognition, etc, and showed significant improvement in performance over the traditional approaches in real-world conditions. However, the performance of deep models is not fully explored in the context of facial behavior analysis due to the limited training data and laborious task of annotations which demands human expertise. Despite the complex process of obtaining annotations for facial behavior, most of the existing approaches to FER based on deep learning have been focused on the fully supervised setting. (Li & Deng, 2020) provided a comprehensive survey on deep learning-based approaches for FER in the framework of supervised learning and gave insight into the advantages and limitations of deploying deep models for FER.

To the best of our knowledge, only five works have used deep networks with domain knowledge for AU recognition, and only one work on the prediction of engagement level in videos. (Wang *et al.*, 2018c), (Wang & Peng, 2019) and (Peng & Wang, 2018) have explored the domain knowledge of dependencies among AUs and the relationship between expressions and AUs using deep networks, where (Wang *et al.*, 2018c) and (Wang & Peng, 2019) used RBM for modeling the domain knowledge and (Peng & Wang, 2018) used Recognition Adversarial Network (RAN) to match the distribution of predicted labels with the pseudo AU labels obtained from domain knowledge. Only two works i.e., (Wang *et al.*, 2019) and (Zhang *et al.*, 2018) used unlabeled and partially labeled data for AU intensity estimation, where RBM is used for modeling global dependencies among AUs and CNN for modeling the ordinal relevance and regression respectively. (Kaur *et al.*, 2018) and (Dhall, Kaur, Goecke & Gedeon, 2018) used deep multi-instance learning for engagement level prediction of sequences, where (Kaur *et al.*, 2018) outperforms (Dhall *et al.*, 2018). To the best of our knowledge, no work has been done on expression detection in WSL framework using deep learning models.

2.5.2.2 Exploiting SpatioTemporal Dynamics

In most of the existing approaches for FER in WSL, only short-term dynamics across the temporal frames are exploited. (Kaur *et al.*, 2018) and (Dhall *et al.*, 2018) used LBP-TOP features (Zhao & Pietikainen, 2007) for capturing the temporal dynamics of the sub-sequences while (Chen *et al.*, 2022d), (Sikka *et al.*, 2014) and (Xie *et al.*, 2019) frame aggregation i.e., maximum of feature vectors of the frames are treated as spatiotemporal features. (Wu *et al.*, 2015b) explored displacement of facial landmarks with HMM for capturing temporal dynamics and showed that it outperforms simple max-based frame aggregation techniques. (Sikka *et al.*, 2016) capture the temporal order of the templates of the sequence, where temporal dynamics is obtained by appending frame-level features of the sequence while (Zhang *et al.*, 2018) captures temporal dynamics using contrastive appearance difference i.e., the difference between apex frame and neutral frame.

In recent days, LSTM was found to achieve superior performance in capturing both short-term and long-term temporal dynamics by exploiting the semantic connection across the frames. Another promising technique based on CNN i.e., C3D (Tran *et al.*, 2015) is also gaining much attention in the field of computer vision for modeling temporal information across the frames. Compared to RNN, C3D is efficient in capturing short-term temporal information. Though LSTM and C3D techniques are widely explored for FER in fully supervised settings (Kim *et al.*, 2019; Hasani & Mahoor, 2017), it is not yet explored for the framework of WSL for FER. Therefore, LSTM and C3D techniques when deployed in WSL setting for FER was expected to further enhance the performance of existing state-of-the-art approaches.

2.5.2.3 Dimensional Affect Model with Inaccurate Annotations

The problem of FER in the framework of inaccurate annotations is mostly explored in the context of classification as described in Section 2.3.3. Though the problem of noisy annotations is more pronounced in the case of the dimensional model as ordinal annotations have more subtle variations across the consecutive frames, not much work has been done on the problem of inaccurate annotations in the dimensional model. Due to the complexity of obtaining annotations in the dimensional model and lack of techniques to handle noisy dimensional annotations, most of the datasets are developed for the task of classification of facial expressions or AUs, which is mentioned in Section 2.3.4.1. Though few datasets have been explored for the task of ordinal regression as described in Section 2.4.3.1, the development of datasets for the continuous dimensional model is rarely explored. As far as we know, only two datasets i.e., (Kollias *et al.*, 2019) and (Mollahosseini *et al.*, 2019) have been developed for FER in a continuous dimensional model though a few multi-modal datasets are available (Ringeval, Sonderegger, Sauer & Lalanne, 2013), (Aung *et al.*, 2016).

The ordinal annotations of DISFA (Mavadati *et al.*, 2013) are obtained by two FACS-certified experts and noisy annotations are reduced by evaluating the correlation between the annotations provided by the two FACS-certifed experts. For UNBC-McMaster (Lucey *et al.*, 2011), the ordinal annotations are obtained from three FACS coders, which were then reviewed by a fourth

FACS coder and validated using Ekman-Friesen formulae (Ekman, Friesen & Hager, 2002). Similarly, (Mollahosseini *et al.*, 2019) hired 12 expert annotators and the annotations are further reviewed by two independent annotators. (Kollias *et al.*, 2019) obtained annotations from six trained experts and doubly reviewed by two more annotators. Finally, the final labels are considered as the mean of the annotations for each sample. They have also conducted statistical analysis to evaluate the inter-annotator correlations.

2.5.2.4 Continuous Affect Model

The continuous dimensional model conveys a wider range of expressions than ordinal regression and plays a crucial role in capturing the subtle changes and context sensitivity of emotions. Compared to the task of classification and ordinal regression, WSL-based approaches for facial expressions or AUs are hardly explored in the context of continuous dimensional space though few endeavors have been made in the context of fully supervised learning (Feng, Shu, Charless, Tao & Baiying, 2020; Kollias & Zafeiriou, 2018). (Gunes & Schuller, 2013) investigated the potential of continuous dimensional model and gave insights on existing state-of-the-art approaches and challenges associated with automatic continuous analysis and synthesis of emotional behavior. Due to the intricate and error-prone process of obtaining annotations, there is an imperative need to formulate the problem of the continuous dimensional model in the framework of WSL to handle noisy annotations and alleviate the negative impact of unreliable annotations. (Huang *et al.*, 2015) investigated the impact of annotation delay compensation and other post-processing operations for continuous emotion prediction of multi-modal data.

As far as we know, only one work (Pei, Jiang, Alioscha-Perez & Sahli, 2019) has been done based on WSL for the prediction of a continuous dimensional model i.e., valence and arousal in a multi-modal framework using audio and visual features. They have reduced the noise of unreliable labels by introducing temporal label, which incorporates contextual information by considering the labels within a temporal window for every time step. They have further used a robust loss function that ignores small errors between predictions and labels in order to further reduce the impact of noisy labels. Therefore, there is a lot of room for improvement to augment

the performance of the FER system in the continuous dimensional model using WSL-based approaches.

2.5.2.5 Domain Adaptation

DA is another promising line of research to handle data with limited annotations although it requires a source domain with accurate annotations as it exploits the knowledge of the source domain for modeling the target domain. It has been widely used for many applications related to facial analysis such as face recognition, facial expression recognition, smile detection, etc. (Wang *et al.*, 2018) proposed an unsupervised domain adaptation approach for a small target dataset using GANs, where GAN-generated samples are used to fine-tune the model pretrained on the source dataset. (Zhu *et al.*, 2016) explored an unsupervised domain adaptation approach in the feature space, where the mismatch between the feature distributions of the source and target domains are minimized still retaining the discrimination among the face images related to facial expressions. (Shao, Cai, Cham, Lu & Ma, 2019) exploited the collection of constrained images from the source domain with both AU and landmark labels and the unpaired collection of unconstrained wild images from the target domain with only landmark labels. The rich features of source and target images are disentangled to shape and text features, and the shape features (AU label information) of source images are fused with the texture information of the target feature. Therefore, the performance of FER can be enhanced using domain adaptation along with deep learning for handling data with limited annotations while still harnessing the potential of deep networks.

2.5.2.6 Localization of Action Unit patches

Attention mechanisms has been gaining attention for capturing the most relevant features and achieved great success for various computer vision problems such as object detection, semantic segmentation, etc. (Xie & Hu, 2019) proposed a deep-based framework for FER using aggregation of local and global features, where local features capture the expression-relevant details and global features models the high-level semantic information of the expression.

In recent days, a few approaches have been proposed in FER for localizing the AU regions based on facial landmarks (Li, Abtahi, Zhu & Yin, 2018; Li, Abtahi & Zhu, 2017; Shao, Liu, Cai & Ma, 2018). However, these approaches rely on prior knowledge of predefined AU attentions, which restrict the capacity to predefined AUs and fails to capture the wide range of non-rigid AUs. As far as we know, only two approaches (Shao, Liu, Cai, Wu & Ma, 2019; Liu *et al.*, 2016) have addressed the problem of AU localization in the WSL framework, which is discussed in Section 2.3.1. (Shao *et al.*, 2019) integrated the relations among AUs with attention mechanism in an end-to-end deep learning framework to capture more accurate attention while (Liu *et al.*, 2016) used BoVW to capture the pixel-wise attention for emotion detection using MIL framework. Due to the immense potential of expression-relevant local features and the tedious task of predefined attention, there is an imperative need to formulate the problem of capturing AU attention in the WSL framework.

2.5.2.7 Multimodal Affective Modeling

Humans exhibit emotions through a diverse range of modalities such as facial expressions, vocal expressions, physiological signals, etc. Multimodal analysis has drawn much attention over the past few years as it enhances the overall performance of the system over the isolated mono-modal approaches. The most effective way of using the multimodal framework is to use different modalities such as the face, speech, ECG, etc in a complimentary fashion to provide a comprehensive feature representation, resulting in higher accuracy of the system. Inspired by the performance of multimodal approaches, several multimodal datasets are developed for the advancement of the system to handle the problems of real-world challenging scenarios (Ringeval *et al.*, 2013; Busso *et al.*, 2008). Recently, deep learning architectures are found to outperform state-of-the-art techniques by capturing the complex non-linear interaction in multimodal data. (Rouast, Adam & Chiong, 2019) provided an exhaustive review of the role of deep architectures for affect recognition using audio, visual, and physiological signals.

Of all the modalities through which emotions can be expressed, facial images and vocal expressions play a crucial role in conveying emotions. In order to foster progress in multimodal

emotion recognition, several audio-visual challenges such as AVEC 2014 - AVEC 2017 (Ringeval *et al.*, 2015a; Ringeval, Schuller, Valstar, Cowie & Pantic, 2015b; Ringeval *et al.*, 2017), etc have been conducted. (Tzirakis *et al.*, 2017) extracted visual features and audio features using the deep residual network of 50 layers (ResNet-50) (Szegedy, Ioffe, Vanhoucke & Alemi, 2017) and CNN models respectively and deployed LSTM architectures for handling the outliers for better classification. However, most of these works have focused on the setting of supervised learning. As far as we know, only one work (Pei *et al.*, 2019) has been done on WSL based approach for multimodal affect recognition using audio and visual features.

2.5.2.8 Infrared and Thermal Images

Infrared and Thermal images are found to be efficient in capturing texture in images even under low illumination conditions, which has achieved success in applications such as Image-dehazing, low light imaging, etc. Inspired by the invariance of thermal images to illumination, a few approaches have been proposed to exploit thermal images with RGB images in a complementary fashion to augment the performance of FER system. (Wang, Pan, Chen & Ji, 2018a) explored thermal images for better feature representation by extracting features from visible and thermal images using two deep networks, which are further trained with two SVM models for expression recognition. The whole architecture is jointly refined using similarity constraints on the mapping of thermal and visible representations to expressions. (Pan & Wang, 2018) have further enhanced the performance of the approach by introducing a discriminator module to differentiate visible and thermal representation and enforcing the similarity between mapping functions of visible and thermal representation to expression labels through adversarial learning. Though the fusion of thermal images with RGB images was expected to be a promising line of research for FER, it still remains to be an under-researched problem.

2.5.2.9 3D and Depth Images

Despite the advancement of FER systems based on 2D images, pose-variance still remains to be a challenging problem. To overcome the problem of pose-variance and occlusion, 3D data has been

explored to obtain the comprehensive information displayed by the face and capture the subtle changes of facial AUs in detail using the depth of the facial surface. For instance, AU18(Lip pucker) is hard to differentiate from AU10+AU17+AU24 in a 2D frontal view. (Sandbach, Zafeiriou, Pantic & Yin, 2012) provided a comprehensive survey on the existing datasets and FER systems pertinent to 3D or 4D data. (Li, Sun, Xu & Chen, 2017) extracted six types of 2D facial attributes from textured 3D face scans and jointly fed them to the feature extraction and feature fusion subnets to learn the optimal 2D and 3D facial representations. The proposed approach is further enhanced by extracting deep features from different facial parts of texture and depth images and fused together with feedback mechanism (Jan, Ding, Meng, Chen & Li, 2018). (Hui Chen, Jiangdong Li, Fengjun Zhang, Yang Li & Hongan Wang, 2015) restored 3D facial models from 2D images and proposed a novel random forest-based algorithm to simultaneously estimate 3D facial tracking and continuous emotion intensities. Although few approaches have exploited 3D data in a fully supervised setting, we believe that 3D or depth images have not been explored for FER in WSL framework.

2.6 Conclusion

In this paper, we have introduced various categories of weakly supervised learning approaches and provided a taxonomy of approaches for facial behavior analysis based on various modes of annotations. A comprehensive review of state-of-the-art approaches pertinent to WSL is provided along with the comparative evaluation of the results. We have further provided insights into the limitations of the existing approaches and the challenges associated with them. Finally, we have presented potential research directions based on our analysis for the future development of facial behavior analysis in the framework of WSL.

CHAPTER 3

DEEP DOMAIN ADAPTATION WITH ORDINAL REGRESSION FOR PAIN ASSESSMENT USING WEAKLY-LABELED VIDEOS

Gnana Praveen Rajasekhar[a] , Eric Granger[a] , Patrick Cardinal[b]

[a]Department of Systems Engineering, École de technologie supérieure,
[b]Department of Software and IT Engineering, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in Image and Vision Computing, April 2021

Abstract

Estimation of pain intensity from facial expressions captured in videos has immense potential for healthcare applications. Given the challenges related to subjective variations of facial expressions, and to operational capture conditions, the accuracy of state-of-the-art deep learning (DL) models for recognizing facial expressions may decline. Domain adaptation (DA) has been widely explored to alleviate the problem of domain shifts that typically occur between video data captured across various source (laboratory) and target (operational) domains. Moreover, given the laborious task of collecting and annotating videos, and the subjective bias due to ambiguity among adjacent intensity levels, weakly-supervised learning (WSL) is gaining attention in such applications. State-of-the-art WSL models are typically formulated as regression problems and do not leverage the ordinal relationship among pain intensity levels, nor the temporal coherence of multiple consecutive frames. This paper introduces a new DL model for weakly-supervised DA with ordinal regression (WSDA-OR) that can be adapted using target domain videos with coarse labels provided periodically. The WSDA-OR model enforces ordinal relationships among the intensity levels assigned to target sequences and associates multiple relevant frames to sequence-level labels (instead of a single frame). In particular, it learns discriminant and domain-invariant feature representations by integrating multiple instance learning with deep adversarial DA, where soft Gaussian labels are used to efficiently represent the weak ordinal sequence-level labels from the target domain. The proposed approach was validated using the RECOLA video dataset as fully-labeled source domain data, and UNBC-McMaster shoulder pain video dataset as

Figure 3.1 Examples of video frames with pain (left) and without pain (right) pain Taken from Bellantonio *et al.* (2017)

weakly-labeled target domain data. We have also validated WSDA-OR on BIOVID and Fatigue (private) datasets for sequence-level estimation. Experimental results indicate that our proposed approach can significantly improve performance over the state-of-the-art models, allowing us to achieve a greater pain localization accuracy.

3.1 Introduction

Pain is a highly disturbing sensation caused by injury, illness, or mental distress. It is a primitive symptom of the malfunctioning of any system in our body (Zeng *et al.*, 2018). Pain is typically conveyed through a patient or an observer on a linear scale from 0 (no pain) to 10 (severe

pain). However, the assessment provided by a patient or an observer may not be reliable since it is subjected to bias induced by the individual's perception of pain as shown in Figure 3.1. Automatic estimation of pain is useful for people who lack verbal communication such as infants, patients suffering from neurological disorders, those in the intensive care unit (ICU) requiring assisted breathing, etc. Therefore, there is a growing demand for the development of automatic pain management systems to ensure effective treatment and ongoing care.

One of the primary channels through which pain can be effectively communicated is facial expressions. Over the years, there has been significant progress in the automatic estimation of pain intensities based on facial expressions in videos (Hassan *et al.*, 2019). In recent years, deep learning (DL) models have provided state-of-the-art performance in many visual recognition applications such as object detection, image classification, semantic segmentation, action recognition, etc (Zhao, Zheng, Xu & Wu, 2019). Compared to 2D-CNN models, 3D-CNNs are found to be efficient for encoding the spatiotemporal dynamics of facial expressions in videos (de Melo, Granger & Hadid, 2019). However, using DL models poses several challenges for real-world pain intensity estimation. An important challenge is the subjective variability of facial expressions across different individuals and the operational capture conditions of videos. Indeed, the performance of DL models for facial expression recognition may decline significantly when there is a considerable domain shift between data distributions of videos captured in the source (lab setting) and target (operational) domains (Wang & Deng, 2018).

Domain adaptation (DA) has been widely used to address the problem of domain differences in various visual recognition applications (Wang & Deng, 2018). In particular, unsupervised DA (UDA) is commonly used for applications related to facial analysis, such as smile detection, to learn robust domain-invariant CNN representations based on labeled source and unlabeled target domain data (Sangineto *et al.*, 2014; Wang *et al.*, 2018; Zhu *et al.*, 2016). The literature on UDA techniques focused on learning discriminant domain invariant embeddings by optimizing an adversarial loss to encourage domain confusion or a discrepancy loss between the two data distributions. Reconstruction-based approaches are another popular paradigm to learn the

mapping between source and target images such that images captured in different domains have similar appearances.

In contrast with the existing DA approaches for facial expression analysis, we explore the weakly-supervised DA (WSDA) case, where source data is fully labeled (at a frame level), and target data is weakly labeled. The authors have explored deep DA models to learn a common representation that diminishes the domain shift between source and target domains. A preliminary version of the approach proposed in this paper is presented in (Gnana Praveen *et al.*, 2020), where deep DA is explored for weakly-supervised pain localization in videos. In the present work, our approach is improved considerably by leveraging the ordinal relationship among intensity levels, and temporal coherence of multiple consecutive frames. We also provided a more detailed formulation and experimental validation of our method. Performing DA for pain intensity level estimation from videos of faces is a challenging problem, in particular when the reference video data is provided with a limited amount of annotations. Most of the existing DL models for pain intensity level estimation have been explored in the fully supervised setting, using frame-level labels (Tavakolian & Hadid, 2018; Zhou *et al.*, 2016a). However, annotating the pain intensity levels for large-scale datasets involves a costly and time-consuming process with domain experts. Moreover, the manual annotation process is vulnerable to subjective bias, resulting in ambiguous labels.

Recently, weakly supervised learning (WSL) has been gaining attention for its potential to train machine learning (ML) models using data with a limited amount of annotations (Zhou, 2018). Based on the availability of labels, WSL scenarios can be classified according to three categories: incomplete, inexact, and inaccurate supervision (Zhou, 2018). Incomplete supervision refers to the scenario where annotations are only provided for a subset of the training dataset. In scenarios involving inexact supervision, annotations are provided for the entire dataset, but at a global or coarse level compared to ones provided in a fully supervised scenario. Inaccurate supervision deals with scenarios where annotations are noisy and ambiguous. (Gnana Praveen *et al.*, 2021) provided a comprehensive review of WSL-based approaches for facial behavior analysis. In the context of pain assessment in videos, inexact supervision is a relevant scenario since it assumes

that pain intensities are annotated on a periodic basis or for entire videos (sequence-level), rather than at the frame level. In particular, multiple instance learning (MIL) methods have been widely used in applications such as object recognition (Miao, Tony, Ming-Chang & Khodayari-Rostamabad, 2016), text categorization (Andrews, Tsochantaridis & Hofmann, 2002), and context-based image retrieval (Zhang, Goldman, Yu & Fritts, 2002), to train ML models using data with coarse annotations. Therefore, we have formulated the problem of pain intensity level estimation from faces captured in videos in the framework of MIL, where sequences are considered to be bags and frames as instances. Most MIL methods proposed in the literature for pain intensity estimation rely on handcrafted features and conventional ML approaches (Sikka *et al.*, 2014; Wu *et al.*, 2015b; Ruiz *et al.*, 2018), due in part to the limited availability of training data with sequence-level annotations. In this paper, we investigate deep WSDA models of pain intensity levels using sequence-level labels.

Pain level assessment can be formulated as a classification or regression problem. In classification, pain estimation is often formulated as a binary problem, i.e., pain/no pain, whereas regression allows predicting a wider range of pain intensities. Recently, approaches based on the regression formulation have gained much attention in the literature because they provide more accurate localization of pain intensities (Tavakolian & Hadid, 2018), (Zhou *et al.*, 2016a). Regression-based approaches can in turn be classified into ordinal and continuous regression. Although continuous regression predicts a wider range of pain intensities, discrete pain intensity levels are often preferred in practical applications to ease video analysis and annotation. Ordinal relations among pain intensity levels convey a rich source of information, yet very few approaches have explored the ordinal relationship among pain intensity levels for automatic pain intensity estimation in videos (Ruiz *et al.*, 2018). The problem of ordinal regression has been widely explored for various applications, such as age estimation (Niu, Zhou, Wang, Gao & Hua, 2016), and image ranking (Liu, Liu, Zhong & Chan, 2011). However, the ordinal regression framework is less explored for pain intensity estimation in videos annotated at the sequence-level. This paper introduces a new deep WSDA model with ordinal regression (WSDA-OR). It learns discriminant and domain-invariant feature representations by integrating MIL with adversarial

DA, where soft Gaussian labels are used to represent the weak ordinal sequence-levels from target videos.

In most of the conventional MIL approaches for pain assessment, MIL pooling is performed using maximum operator, i.e., the sequence level label is associated with the frame corresponding to the maximum intensity level (Hsu, Lin & Chuang, 2014; Sikka, 2014). However, the maximum operator only relies on a single frame, failing to capture the relevant information available in multiple adjacent frames. (Ilse *et al.*, 2018) have shown that attention-based MIL pooling can significantly improve predictive accuracy. Inspired by their approach, we introduce adaptive MIL pooling, that relies on multiple relevant frames of the sequence (bag). It allows associating all the relevant frames of the corresponding sequence to the sequence-level label, and can significantly improve the accuracy of pain assessment. To the best of our knowledge, WSDA-OR is the first model to efficiently capture the ordinal relationship among pain intensity levels through Gaussian representation, in the context of multiple instance regression (MIR).

The main contributions of this paper are:

- a DL model for pain assessment that can adapt to diverse capture conditions and individuals using weakly-labeled target videos;
- Gaussian modeling through multiple instance regression (MIR) to efficiently capture the ordinal relationship among intensity levels;
- an adaptive MIL pooling to associate all the relevant frames of the corresponding sequence to the sequence-level label;
- an extensive set of experiments validating that our proposed WSDA-OR can outperform state-of-the-art models.

The rest of this paper is organized as follows. Section 3.2 provides some background on models for pain intensity estimation, deep DA, ordinal regression, and MIL. Our proposed WSDA-OR model is described in Section 3.3. Finally, Section 3.4 presents the experimental methodology (datasets, protocols, and performance metrics), and results for validation.

3.2 Related Work

3.2.1 Deep Models for Pain Intensity Estimation:

Though DL is explored extensively for fully supervised learning, it is still at a rudimentary level to deal with weakly-labeled data. (Zhou *et al.*, 2016a) proposed Recurrent Convolutional Neural Network (RCNN) using recurrent connections in the convolution layers to capture the temporal information without increasing the overload of parameters to avoid overfitting. In order to deal with the problem of limited data, (Wang *et al.*, 2017) used a pretrained face recognition network for fine-tuning using a regularized regression loss. (Rodriguez *et al.*, 2018) also used VGG Face pre-trained CNN network (Parkhi *et al.*, 2015) for capturing the facial features and LSTM network is used to exploit the temporal relation between the frames. Compared to 2D CNN models, 3D CNNs are found to be gaining attention in efficiently capturing the temporal dynamics of the video sequences. (Tavakolian & Hadid, 2018) propose a 3D-CNN based architecture using a stack of convolution modules with varying kernel depths for efficient dynamic spatiotemporal representation of faces in videos. A temporal pooling method to encode the spatiotemporal facial variations in video clips based on a two-stream model that performs a late fusion of appearance and dynamic information (Carneiro de Melo, Granger & Lopez, 2020). However, all these approaches have been proposed in the setting of fully supervised learning, thereby requiring frame-level labels. Inflated 3D-CNNs (I3D) have been employed for facial expression recognition, allowing to leverage pre-trained 2D-CNNs, yet benefit from the efficient modeling of temporal dynamics using 3D CNN models (Ayral, Pedersoli, Bacon & Granger, 2021; Carreira & Zisserman, 2017). Inspired by these benefits and their performance, we have relied on I3D for modeling the spatiotemporal dynamics of pain expressions for adversarial DA with weakly-labeled target videos.

3.2.2 Deep Domain Adaptation:

(Wang & Deng, 2018) provided a survey of the deep DA approaches, with applications in visual recognition. Deep DA can be primarily summarized into three categories: discrepancy-based,

adversarial-based, and reconstruction-based DA. Discrepancy-based approaches attempt to minimize the domain shift by fine-tuning the deep model with the labeled or unlabeled target data. Adversarial-based approaches deploy domain discriminators to classify whether a data sample is drawn from the source or target domain to diminish the domain shift. Finally, reconstruction-based approaches try to ensure feature invariance using data reconstruction of source or target samples to improve the performance of DA. (Sangineto *et al.*, 2014) proposed a regression framework for personalized facial expression recognition, where classifiers are generated for the individuals of the source data rather than a generic model for the entire source data. (Wang *et al.*, 2018) proposed an unsupervised DA approach for a small target dataset using Generative Adversarial Network (GAN), where GAN-generated samples are used to fine-tune the model pretrained on the source dataset. (Zhu *et al.*, 2016) explored the unsupervised DA approach in the feature space, where the mismatch between the feature distributions of the source and target domains are minimized still retaining the discriminative information among the face images related to facial expressions. (Bozorgtabar, Mahapatra & Thiran, 2020) investigated the use of adversarial DA to transform the visual appearances of simulated faces to real face images without losing the face details relevant to identity or expressions. By doing so, expression recognition models trained on labeled realistic face images with arbitrary head poses can be directly generalized on the unlabeled simulated images without the need for re-training.

Contrary to the existing DA approaches for facial expression analysis, we have explored DA in the context of adapting source domain data with full labels to target domain data with coarse labels. (Ganin & Lempitsky, 2015) proposed a novel approach of adversarial DA using deep models with partial or no target data labels using a simple gradient reversal layer. We have further extended their approach for the scenario of coarsely labeled target data for pain localization in videos.

3.2.3 Ordinal Regression:

(Zhao *et al.*, 2016) proposed a max-margin-based ordinal support vector regression using ordinal relationship, which is flexible and generic in handling varying levels of annotations. A linear

model is learned by solving the optimization problem using the Alternating Direction Method of Multipliers (ADMM) to predict the frame-level intensity of the test image. (Zhang *et al.*, 2018b) explored domain knowledge of Ordinal relevance, intensity smoothness, and relevance smoothness based on the gradually evolving process of facial behavior. (Zhang *et al.*, 2018) designed a CNN model for intensity estimation of Action Units(AUs) using annotations of only peak and valley frames, where the parameters of CNN are learned by encoding domain knowledge of facial symmetry, temporal intensity ordering, relative appearance similarity, and contrastive appearance difference. All of the above-mentioned approaches did not efficiently capture the ordinal relationship and are proposed for expression or action unit intensity estimation but not for pain intensity estimation.

3.2.4 Multiple Instance Learning:

Though MIL has been widely explored for many computer vision applications, relatively fewer techniques have been proposed for dynamic pain intensity estimation. (Sikka *et al.*, 2014) developed an automatic pain recognition system for pain localization in the framework of MIL, where video segments are represented as bags of multiple subsequences and MILBOOST (Viola *et al.*, 2006) is used for instance-level pain detection. (Wu *et al.*, 2015b) further enhanced the approach by incorporating a discriminative Hidden Markov Model (HMM) based instance level classifier in conjunction with MIL framework instead of MILBOOST to efficiently capture the temporal dynamics. (Chen *et al.*, 2022d) proposed a novel two-stage approach for pain detection by deploying a novel strategy to encode AU combinations using individual AU scores.

However, all of these approaches have been proposed for pain detection. (Ruiz *et al.*, 2018) proposed multi-instance dynamic ordinal random fields (MI-DORF) for modeling temporal sequences of ordinal instances, where bags are defined as temporal sequences labeled as ordinal variables. The instance labels are obtained by incorporating high-order cardinality potential relating bag and instance labels in the energy function. But they have not leveraged the superior performance of DL models. (Zhang *et al.*, 2018) designed a deep CNN based on weakly supervised learning for intensity estimation of Action Units(AUs) of facial expressions with

limited annotations of AUs, where only the annotations of peak and valley frames of the AUs are considered. Despite the advancement of MIL for various applications in computer vision, not much work has been explored for the estimation of pain intensity levels using state-of-the-art DL models. Unlike the above-mentioned approaches, our approach focused on pain intensity level estimation using DL models in conjunction with MIL framework for localization of pain intensity levels i.e., instance level prediction. We have further improved the pooling mechanism by introducing adaptive MIL pooling to efficiently leverage all the relevant frames in the sequence to associate with the sequence level label.

3.3 Proposed Approach

In this section, we elaborate on the proposed approach in detail. In the proposed framework, we have explored deep DA in the context of MIL for ordinal regression, where labels of intensity levels are provided for video sequences instead of individual frames. To efficiently model the ordinal relationship among the intensity levels, we have considered Gaussian modeling of the intensity levels (labels) instead of one-hot vectors. Unlike the conventional approaches of MIL (Gnana Praveen *et al.*, 2020; Ruiz *et al.*, 2018), (Sikka, 2014), where the sequence level label was associated with a single frame, we have exploited multiple frames, which are relevant to the sequence level label to enhance the performance of learning framework. Inspired by the performance of the I3D model (Carreira & Zisserman, 2017) with adversarial learning (Ganin & Lempitsky, 2015), we have used the framework of adversarial-based DA as it was shown to yield superior performance in the framework of DL models for videos (Jamal *et al.*, 2018). The overall block diagram of the proposed approach is shown in Fig 3.2.

Let $\mathbf{D} = \{(\mathbf{X_1}, \mathbf{Y_1}), (\mathbf{X_2}, \mathbf{Y_2}), ..., (\mathbf{X_N}, \mathbf{Y_N})\}$ represents the dataset of pain expressions of videos from source and target domains. $\mathbf{X_i}$ denotes a video sequence of the training data with a certain number of frames. In the case of the source domain, $\mathbf{Y_i}$ denotes a structured label vector with frame-level annotations of the corresponding video sequence $\mathbf{X_i}$, whereas, for the target domain, $\mathbf{Y_i}$ represents an ordinal intensity value i.e., sequence level ordinal intensity value of

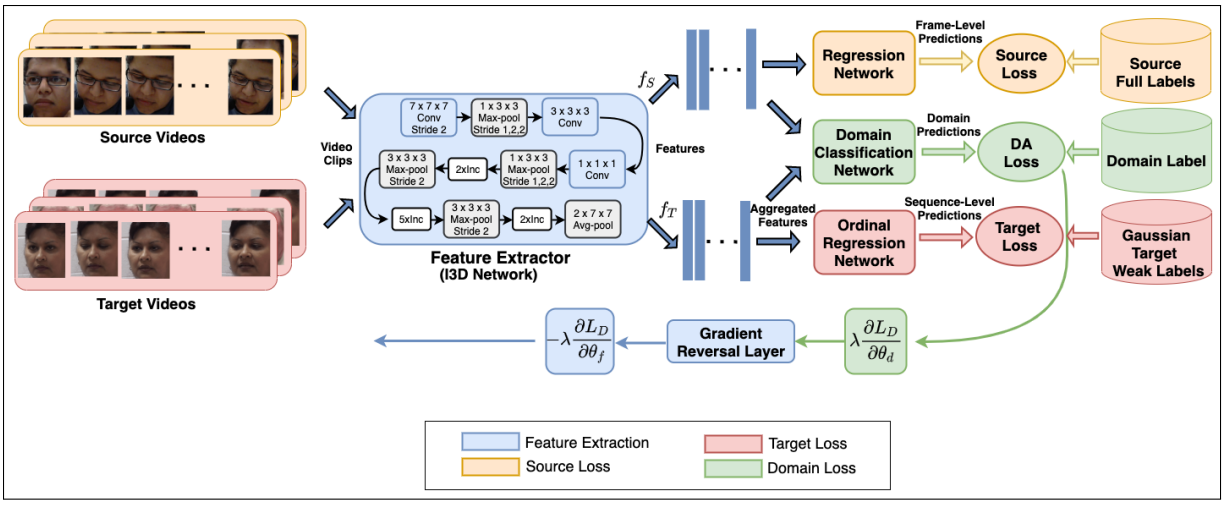

Figure 3.2 Overall architecture of the proposed approach (WSDA-OR). Inc denotes Inception module (Szegedy *et al.*, 2015). Different colors are used to discriminate data flow in different loss components. Best viewed in color
Taken from Rajasekhar *et al.* (2021b)

the corresponding video sequence, which is given by

$$\mathbf{Y_i} = \begin{cases} \{y_i^1,\ y_i^2,\ ....y_i^{n_i}\}\ if\ \mathbf{X_i} \in\ source\ domain \\ \qquad y_i \qquad\quad if\ \mathbf{X_i} \in target\ domain \end{cases} \tag{3.1}$$

where n_i denotes the number of frames in the corresponding sequence $\mathbf{X_i}$. N represents the number of training sequences. Specifically, $\mathbf{X_i} = \{x_i^1, x_i^2, ...x_i^{n_i}\}$ represents the temporal sequence of n_i observations (frames) and x_i^t denotes t^{th} frame in i^{th} sequence, where $t \in \{1, 2, ..n_i\}$.

The objective of the problem is to estimate a generic ordinal regression model $F : \mathbf{X} \rightarrow \mathbf{H}$ from the training data $\mathbf{D}$ to predict the pain intensity level of frames of unseen test sequences, where $\mathbf{X}$ denotes the video sequences of training data and $\mathbf{H}$ represents the hidden label space of frame-level annotations of the target domain. The estimated intensity levels of the individual frames of the sequences in the target domain are predicted as structured output $\mathbf{H_i} \in \mathbf{H}$, where $\mathbf{H_i} = \{h_i^1, h_i^2, ..., h_i^{n_i}\}$ and each frame x_i^t of the sequence is assigned a latent ordinal value h_i^t.

Let S represent the source dataset, which is fully labeled videos and T denote the target dataset, which is weakly labeled videos. Let G_f represent the feature mapping function, where the parameters of this mapping are denoted by θ_f. Similarly, the feature vectors of the source domain and target domain are mapped to the corresponding labels using G_l and G_{wl}, whose parameters are denoted by θ_l and θ_{wl} respectively. Finally, the mapping of the feature vector to the domain label is obtained by G_d with parameters θ_d.

3.3.1 Gaussian Modeling of Ordinal Intensity Levels

Due to the ordinal nature of pain intensity levels, we have formulated pain intensity estimation as an ordinal regression problem, which attempts to solve the classification problem while still retaining the ordinal relationship among the labels. Though ordinal regression can be formulated as a classification problem, it does not capture the relative ordering among the ordinal labels. For instance, if a particular sample has a pain intensity level of "4", a misclassification of "3" or "5" is more acceptable than a misclassification of "1". Although the objective is to predict the correct intensity level of "4", the system should, in the event of misclassification, logically predict an ordinal level as close as possible to the ground truth "4".

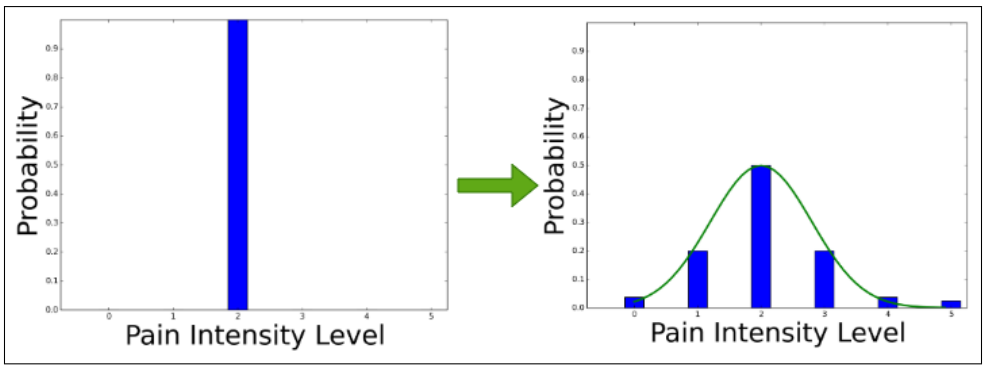

Figure 3.3 Gaussian representation of weak ordinal labels
Taken from Rajasekhar *et al.* (2021b)

In order to model the relative ordering among the ordinal labels, soft labels have been widely explored in the literature (Díaz & Marathe, 2019; Tan, Zhou, Wan, Lei & Li, 2017). Gaussian distribution was found to be promising in modeling the ordinal relationships, where the probability

(or contribution) of the nearby ordinal labels decreases exponentially as we move away from the ground truth on either side of the mean in a symmetric manner. This approach has been widely explored in the ordinal regression literature (Chu & Ghahramani, 2005), (Liu, Wang & Kong, 2019). Most of these approaches have imposed the constraints of Gaussian modeling on the predicted outputs. In this work, we propose a simple yet efficient approach of encoding the target labels as soft labels obtained from Gaussian distribution instead of one-hot vectors as shown in Fig 3.3. Specifically, the mean of the Gaussian model is considered as the corresponding ground truth label, and the variance controls the influence of neighboring ordinal levels. The intensity levels in close proximity to the corresponding label, therefore, have higher relevance compared to the intensity levels at far proximity. The soft Gaussian labels of the ordinal intensity levels are given by

$$q_i = e^{\frac{-(k-y_i)^2}{2\sigma^2}} \tag{3.2}$$

where σ denotes the Gaussian smoothing parameter (variance) of the Gaussian model and $k \in \{0, 1, 2, ..., K-1\}$ and K denotes the number of ordinal intensity levels.

The proposed approach of encoding the target labels using Gaussian distribution automatically learns the ordinal relationships without any explicit modification to the network architecture. Therefore, our method can also be used with any conventional classification networks with common categorical loss functions such as cross-entropy. Additionally, deploying a soft Gaussian version of the target labels also helps in counterfeiting the problem of limited data with deep networks. We show empirically that these soft representations obtained from Gaussian distribution efficiently capture the ordinal relationship among the pain intensity levels, and significantly improve the performance of the system.

3.3.2 Adaptive Multiple Instance Learning Pooling

In the framework of MIL, the choice of the pooling function plays a crucial role in associating the instance-level outputs to the bag label. Several pooling functions have been explored in the literature and a comparative study of various pooling techniques is discussed in (Wang,

Li & Metze, 2019). In the conventional setting of MIR (Hsu *et al.*, 2014), the sequence level label is associated with the frame corresponding to the highest intensity level in the framework of MIL (Gnana Praveen *et al.*, 2020; Ruiz *et al.*, 2018; Sikka, 2014). The relationship between the coarse bag-level label $\mathbf{Y_i}$ and latent instance-level labels $\mathbf{H_i}$ is modeled by assigning the maximum value of predicted instance-level outputs to the bag label, which is given by

$$\mathbf{Y_i} = \max_{h}(\mathbf{H_i}) \; \forall(\mathbf{X_i}, \mathbf{Y_i}) \in \mathbf{D} \tag{3.3}$$

If the label $\mathbf{Y_i}$ is 0, then all the frames in the sequence $\mathbf{X_i}$ will be assigned 0 i.e., neutral frame.

In the case of pain intensity levels, the sequence level label is associated with the frame corresponding to the highest intensity level in the framework of MIR (Gnana Praveen *et al.*, 2020), (Ruiz *et al.*, 2018; Sikka, 2014). The prediction of the weakly labeled sequence is given by

$$P(\mathbf{X_i}) = \max_{j \in (1,..n_i)}(G_{wl}(G_f(x_i^j))) \tag{3.4}$$

where $P(\mathbf{X_i})$ denotes the probabilities of the frame pertinent to the maximum intensity level among all the frames of the sequence, G_f and G_wl represent the feature extraction and weak ordinal regression layers respectively.

However, the maximum operator relies only on a single frame and does not efficiently exploit the information available in all the frames relevant to the sequence level label (Ilse *et al.*, 2018). To leverage the relevant information of multiple frames, several learnable pooling functions have been proposed with deep networks (Ilse *et al.*, 2018; Liu, Zhou, Sun, Zha & Zeng, 2017). Unlike prior approaches, we have proposed a simple pooling function, which adaptively chooses the relevant frames without the need to learn any additional parameters. We further show that the proposed pooling mechanism has significantly improved the performance of the system.

The frames relevant to the bag label (sequence level label) are selected based on the predicted instance-level outputs of the deep network. In the case of pain intensity estimation, there could be many frames predicted as having the maximum pain intensity level, which is relevant to the

sequence level label. The frames predicted as maximum intensity level are considered as frames relevant to sequence level labels. Next, MIL pooling is performed by averaging the output responses of the selected relevant frames, whose outputs represent the maximum intensity level within the sequence. Therefore, the frames that are irrelevant to the sequence-level (bag) label are discarded while the frames relevant to the sequence-level label are retained and deployed in the pooling mechanism.

Typically, pain expressions are sparse in nature, where most of the frames in the sequence are neutral along with few pain expression frames relevant to sequence-level labels. By deploying adaptive MIL pooling only on the frames pertinent to maximum predicted intensity levels, highly redundant neutral frames are discarded and the relevant multiple frames of higher intensity levels are effectively used in the training mechanism. For the sake of ordinal regression, the number of output units of the weak supervision layer G_{wl} (last fully connected soft-max layer) of the ordinal regression module of the target domain is equal to the number of intensity levels to be predicted. The bag level representation of the instance level outputs is obtained using adaptive MIL pooling, which is obtained by averaging the outputs of selected frames (predicted as maximum intensity level) within the sequence, which is given by

$$P(\mathbf{X_i}) = \frac{1}{N_{t_i}} \sum_{j \in \max(1,..n_i)} G_{wl}(G_f(x_i^j)) \tag{3.5}$$

where $P(\mathbf{X_i})$ denotes the mean of the soft-max output responses of relevant frames predicted as maximum intensity levels and N_{t_i} represents the number of relevant frames of the corresponding sequence predicted as maximum intensity level.

3.3.3 Training Mechanism:

The deep network architecture consists of three major building blocks: feature mapping, label predictor, and domain classifier. In the proposed architecture, the feature mapping layers share the same weights between the source and target domains to ensure common feature space between source and target domains. It has been shown that the label prediction accuracy on

the target domain will be the same as that of the source domain by ensuring the similarity of distributions between source and target domains (Shimodaira, 2000). Next, an adversarial mechanism is deployed between the domain discriminator G_d, which is learned to discriminate the source and target domain samples, and feature extractor G_f, which is trained simultaneously to minimize the domain discrepancy between source and target domains. At the time of training, label prediction loss is minimized on the source domain by optimizing the parameters of G_f and G_l to learn the feature mapping given the labels, while simultaneously ensuring the features are domain-invariant. This is achieved by maximizing the loss of the domain classifier to minimize the discrepancy between the source and target domains while the parameters of G_d are learned by minimizing the loss of the domain classifier to discriminate between source and target domains. The label prediction loss (L_S) for the source domain is defined by

$$L_S = \frac{1}{N_s} \sum_{\substack{i=1 \\ d_i=0}}^{N_s} \sum_{j=1}^{n_i} ((G_l(G_f(x_i^j)) - y_i^j))^2 \tag{3.6}$$

where $d_i = 0$ represents the source domain, N_s denotes the number of video sequences in the source domain and n_i denotes the number of frames in the corresponding video sequence. In addition to source labels, the weak labels of the target domain are also used in the feature learning mechanism where the parameters of G_{wl} are optimized by minimizing the prediction loss pertinent to weak labels of the target data. The weak sequence level labels (ordinal intensity levels) of the target domain are encoded to soft Gaussian representations as mentioned in 3.3.1 instead of one-hot vectors to efficiently capture the ordinal relationship as well as to counterfeit the problem of limited data.

Contrary to MIL-based approaches for pain intensity estimation, which relies on the single frame with maximum intensity level (Gnana Praveen *et al.*, 2020), (Sikka, 2014), we have explored multiple frames with maximum predicted intensity levels to associate with the weak sequence level label, thereby improving the training mechanism due to the deployment of multiple relevant frames as described in 3.3.2. The prediction loss associated with the weak supervision of the

target domain is estimated as cross-entropy (CE) loss between the soft gaussian labels of target domain $\mathbf{Y_i}$ and predicted response $P(\mathbf{X_i})$, which is given by

$$L_T = -\frac{1}{N_T} \sum_{\substack{i=1 \\ d_i=1}}^{N_T} (\mathbf{Y_i}.\log(P(\mathbf{X_i}))) \tag{3.7}$$

where $\mathbf{Y_i}$ denotes the Gaussian representation vector of the ordinal level of the video sequence in the target domain, $P(\mathbf{X_i})$ denotes the vector of the predicted intensity level of $\mathbf{X_i}$ in the target domain, (.) denotes the dot product function and N_T represents the number of video sequence in the target domain.

Since domain classification is a typical binary classification problem, we have used logistic regression to diminish the domain differences between source and target domains, where the logistic loss function is given by

$$L_d = \frac{1}{N_s + N_T} \sum_{\substack{i=1 \\ d_i=0,1}}^{N_s+N_T} \sum_{j=1}^{n_i} [-d_i^j \log\left(G_d(G_f(x_i^j))\right) - (1 - d_i^j) \log\left(1 - G_d(G_f(x_i^j))\right)] \tag{3.8}$$

where d_i^j denotes the domain label of the j^{th} frame of the i^{th} video sequence.

The overall loss of the deep network architecture is given by

$$L = L_S + L_T - \lambda L_d \tag{3.9}$$

where λ is the trade-off parameter between the objectives of label prediction loss and domain prediction loss and the parameters of θ_l, θ_{wl}, θ_f and θ_d are jointly optimized using Stochastic Gradient Descent (SGD).

At the end of the training, the parameters of θ_l, θ_{wl}, θ_f and θ_d are expected to give a saddle point for the overall loss function as given by :

$$\hat{\theta}_f, \hat{\theta}_l, \hat{\theta}_{wl} = \arg\min_{\theta_f, \theta_l, \theta_{wl}} L(\theta_f, \theta_l, \theta_{wl}, \hat{\theta}_d) \tag{3.10}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} L(\hat{\theta}_f, \hat{\theta}_l, \hat{\theta}_{wl}, \theta_d) \tag{3.11}$$

At the saddle point, the feature mapping parameters θ_f minimize the label prediction loss to ensure discriminative features and maximizes the domain classification loss to constrain the features to be domain-invariant. To backpropagate through the negative term in our loss function, a special gradient reversal layer (GRL) is deployed in our SGD optimization framework, which is elaborated in detail in (Ganin & Lempitsky, 2015). The value of lambda is modified over successive epochs, such that the supervised prediction loss dominates at the early epochs of training. Further details on the training mechanism can be found in (Ganin & Lempitsky, 2015).

3.4 Results and Discussion

3.4.1 Experimental Setup:

The proposed approach has been evaluated on the UNBC-McMaster dataset (Lucey *et al.*, 2011), which is widely used for pain intensity level estimation in the context of MIL. Due to the availability of state-of-the-art results of the UNBC pain dataset in the context of MIL, we have primarily validated the proposed approach on the UNBC pain dataset. The dataset consists of 200 videos of pain expressions captured from 25 individuals, out of which 13 are female and 12 are male, resulting in 47,398 frames of size 320x240. Each video sequence is annotated using a PSPI score at frame level on a range of 16 discrete pain intensity levels (0-15). Due to the sparse nature of pain expressions and high-level imbalance among various intensity levels, we followed the widely adapted quantization strategy i.e., the pain levels are quantized to 5 ordinal levels as 0(0), 1(1), 2(2), 3(3), 4-5(4), 6-15(5). In our experiments, we followed the same experimental protocol as that of (Gnana Praveen *et al.*, 2020) to have a fair comparison with state-of-the-art

results, where Leave-One-Subject-Out (LOSO) cross-validation strategy is deployed i.e., 15 subjects have been used for training, 9 subjects for validation and 1 for testing in each cycle.

Due to the availability of labels for every frame, RECOLA (Ringeval *et al.*, 2013) dataset is used as the source domain, where each video sequence is obtained for a duration of 5 minutes and annotated with an intensity value between -1 to +1 for every 40 msec (same as the frame rate of 25fps) i.e., all the frames are annotated. The video sequences of UNBC (target) and RECOLA (source) datasets are converted to sub-sequences (bags) of 64 frames (instances) with a stride of 8 to generate more samples for the learning framework, resulting in 10496 sub-sequences for RECOLA and 2890 sub-sequences for UNBC dataset. To incorporate the setting of weakly supervised learning in the target domain (weakly labeled videos), only coarse labels of the sub-sequences of the UNBC dataset are considered i.e., the maximum intensity level within a sub-sequence is assigned as a coarse annotation to formulate the problem of MIL for ordinal regression (Hsu *et al.*, 2014). To use the Gaussian representation of the weak ordinal labels, the variance is considered to be 0.3.

The faces are detected, normalized, and cropped using MTCNN (Zhang *et al.*, 2016) and resized to 224 x 224. In our experiments, I3D architecture is used, where inception v-1 architecture is used as the base model, which is inflated from 2D pre-trained model on ImageNet to 3D CNN for videos of pain expressions. We have used Stochastic Gradient Descent (SGD) as our optimization technique for training the model with a momentum of 0.9, and a weight decay of 1e5. The initial learning rate is set to 0.001 and annealed according to a schedule pre-determined on the cross-validation set for every 5 epochs after 20 epochs. Due to the difference between the number of samples (sub-sequences) between the source and target domain, a batch size of 4 is used for the source domain and 2 for the target domain. Due to the huge imbalance among various intensity levels of the samples (sub-sequences) of the target domain, weighted random sampling is deployed for loading the data to counterfeit the problem of level imbalance. An early stopping strategy is used for model selection to avoid over-fitting.

3.4.2 Evaluation Measures:

Given the ordinal nature of pain intensity levels, the performance of the proposed approach is measured in terms of Pearson Correlation Coefficient (PCC), Intra class correlation (ICC(3,1)), and Mean-Absolute-Error (MAE). In most of the existing literature on MIL, the results are often reported for bag-level predictions. However, we have focused on instance-level prediction i.e., frame-level prediction of ordinal pain intensity levels for accurate localization of pain intensity levels in videos. PCC is invariant to linear transformations and efficiently captures the correlation between predictions and ground truth, which may differ in scale and location. The PCC measure between predictions (h_i) and ground truth values (y_i) of a sequence i is given by

$$PCC(y_i, h_i) = \frac{n_i \sum(y_i * h_i) - \sum y_i \sum h_i}{\sqrt{[n_i \sum y_i^2 - (\sum y_i)^2][n_i \sum h_i^2 - (\sum h_i)^2]}} \tag{3.12}$$

where n_i represents the number of frames in the sequence. Though the PCC measure captures the correlation between the two variables, it fails to capture the exact similarity measure i.e., an absolute agreement between ground truth and the predicted intensity levels. Therefore, we have used ICC(3,1) (Shrout & Fleiss, 1979), which is widely used to accurately measure the degree of correlation as it takes into account differences in scale and location. The ICC measure between the predictions (h_i) and ground truth (y_i) is computed using Between Mean Squares (BMS) and Error Mean Squares (EMS), as given by

$$ICC(y_i, h_i) = \frac{BMS_i - EMS_i}{BMS_i + EMS_i} \tag{3.13}$$

where BMS_i of sequence i is given by

$$BMS(y_i, h_i) = \frac{n_i \sum(y_i + h_i)^2 - (\sum y_i + \sum h_i)^2}{2n_i(n_i - 1)} \tag{3.14}$$

and EMS_i of sequence i is given by

$$EMS(y_i, h_i) = \frac{2\sum y_i^2 + 2\sum h_i^2 - \sum (y_i + h_i)^2}{2n_i} \tag{3.15}$$

We have also further provided the performance measure of Mean-Absolute-Error (MAE), which is widely used for continuous regression applications and accurately captures the error between the two measurements of predictions (h_i) and ground truth (y_i), which is given by

$$MAE(y_i, h_i) = \frac{\sum |y_i - h_i|}{n_i} \tag{3.16}$$

Table 3.1 PCC, ICC and MAE performance of proposed approach under various baseline scenarios

Training Scenario	**Frame-level**		
	PCC ↑	**ICC ↑**	**MAE ↓**
Supervised (source data only)	0.323	0.272	0.976
Supervised (target data only)	0.441	0.377	0.660
Supervised (source ∪ target)	0.570	0.448	0.539
Unsupervised DA	0.468	0.198	0.782
Transfer learning with weak labels	0.614	0.384	0.618
Supervised DA	0.750	0.724	0.440

3.4.3 Results with Baseline Training Models:

To analyze the impact of DA and availability of annotations of source and target domains, the performance of the proposed approach has been evaluated by conducting a series of experiments with various baseline models, where I3D training models are generated by varying the data ranging from using only source data with full labels to the entire dataset of source and target domains with full labels as shown in Table 3.1. In all these experiments, the performance of the training model has been validated on the test data of the target domain. Initially, we considered only the source domain with full labels without the target domain and generated the training model. Due to the domain differences between train data (source) and test data

(target), the generated training model exhibits poor performance. Next, we consider only the target data with full labels without source data and generate the training model, which shows improvement in performance as both train data and test data come from the same domain (target). Subsequently, we use both source data and target data with full labels without DA and found that the performance was further improved as training data spans a wide range of variations in source and target domains. Now we conduct another series of experiments with DA, where the

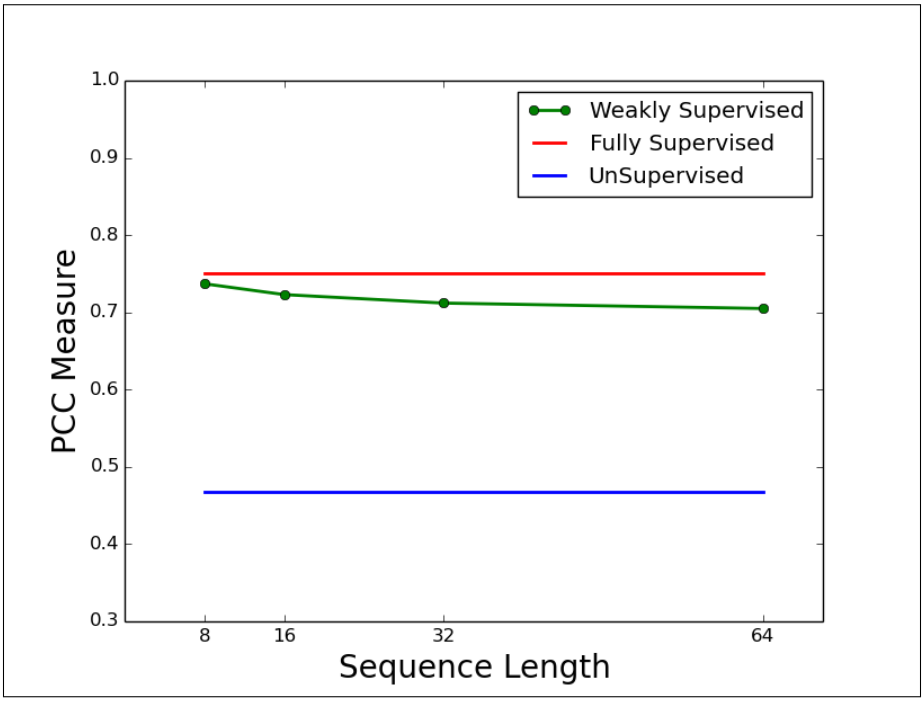

Figure 3.4 PCC accuracy of I3D model trained with deep WSDA-OR levels with decreasing level of weak supervision on target videos
Taken from Rajasekhar *et al.* (2021b)

training data is obtained from source data with full labels and target data with varying levels of supervision. By considering the full labels of the source domain, the level of supervision of the target domain is gradually reduced by decreasing the frequency of annotations i.e., labels are provided by increasing the duration of sequence lengths. Specifically, we have conducted experiments for sequence lengths of 8,16,32, and 64. In addition to varying sequence lengths, we have also conducted experiments of DA with no supervision of the target domain, which acts as the lower bound, and full supervision of the target domain, which acts as an upper bound.

As we gradually reduce the amount of labels of the target domain, we can observe that the performance of our approach gradually drops as shown in Fig 3.4. However, our approach still performs at par with full supervision as there is only minimal decline, which is attributed to the DA as we are leveraging source data to adapt to the target domain using adversarial DA. We have further evaluated the proposed approach with transfer learning, where the training model is first obtained only with the source domain, and then fine-tuned with the weak labels of the target domain. Since transfer learning does not try to diminish the domain differences and relies on the size of pretrained dataset, it shows lower performance compared to the proposed approach.

3.4.4 Ablation Study

Table 3.2 Performance of proposed approach with ablation study of individual modules in terms of PCC, ICC, and MAE

Training Scenario for WSDA-OR	**Frame-level**		
	PCC ↑	**ICC ↑**	**MAE ↓**
Baseline	0.511	0.498	0.632
Baseline + AMILP	0.627	0.597	0.740
Baseline + GM	0.598	0.599	0.617
Baseline + GM + AMILP	0.705	0.696	0.530

We have further analyzed the contribution of individual modules of the proposed approach: Gaussian modeling of ordinal levels (GM) and Adaptive MIL pooling (AMILP) as shown in Table 3.2. First, we have generated the training model with a baseline version without GM and AMILP i,e., we have used max operator for MIL pooling and conventional label smoothing (Szegedy, Vanhoucke, Ioffe, Shlens & Wojna, 2016). The performance of the baseline training model is low as the conventional max-pooling operation fails to leverage the significant information in the nearby relevant frames and traditional label smoothing is not able to efficiently capture the ordinal relationships. Next, we deployed AMILP of relevant frames without using GM of ordinal levels. This shows that AMILP efficiently leverages the information in the nearby relevant frames, thereby showing a significant improvement over conventional MIL pooling. To validate the contribution of GM of ordinal labels, we have further generated the training model

only with the GM of ordinal labels over the baseline version. Since GM effectively captures the ordinal relationship among the pain intensity levels, the performance of the approach has significantly improved over conventional label smoothing. Finally, we have enforced GM of ordinal levels along with AMILP of relevant frames. By combining both modules, there was a drastic improvement in the performance of the approach as it effectively leverages all the relevant frames for MIL pooling and captures the ordinal relationship among the pain intensity levels.

3.4.5 Comparison with State-of-the-Art Methods:

In most of the existing state-of-the-art approaches for weakly supervised learning, classical ML approaches have been explored due to the problem of limited data with limited annotations. However, we have used a deep model (I3D) along with source data to compensate for the problem of limited data with limited annotations using DA. Our work is closely related to that of MI-DORF (Ruiz *et al.*, 2018), which used graph-based models to capture the temporal dynamics of the frames and the ordinal relationship of labels. We have further compared the proposed approach with our previous work (Gnana Praveen *et al.*, 2020) and shown that exploring the ordinal relationships and relevant frames in adaptive MIL pooling significantly improves the performance of the system over a conventional regression framework. The estimation of frame level predictions of the proposed approach shows significant improvement, while sequence level estimation performs at par with that of our previous work (Gnana Praveen *et al.*, 2020). This is due to the fact that the sequence level performance in the proposed approach is obtained from the model trained for frame-level predictions whereas the sequence level performance of our previous work (Gnana Praveen *et al.*, 2020) is reported using the model trained for sequence-level predictions instead of frame level predictions. We have also provided visualization of some of our results for pain localization i.e., frame-level prediction for two subjects. Due to the efficient modeling of the ordinal labels and adaptive MIL pooling, the proposed approach accurately localizes pain better than (Gnana Praveen *et al.*, 2020) even though it was not captured in ground truth as shown in Figure 3.5.

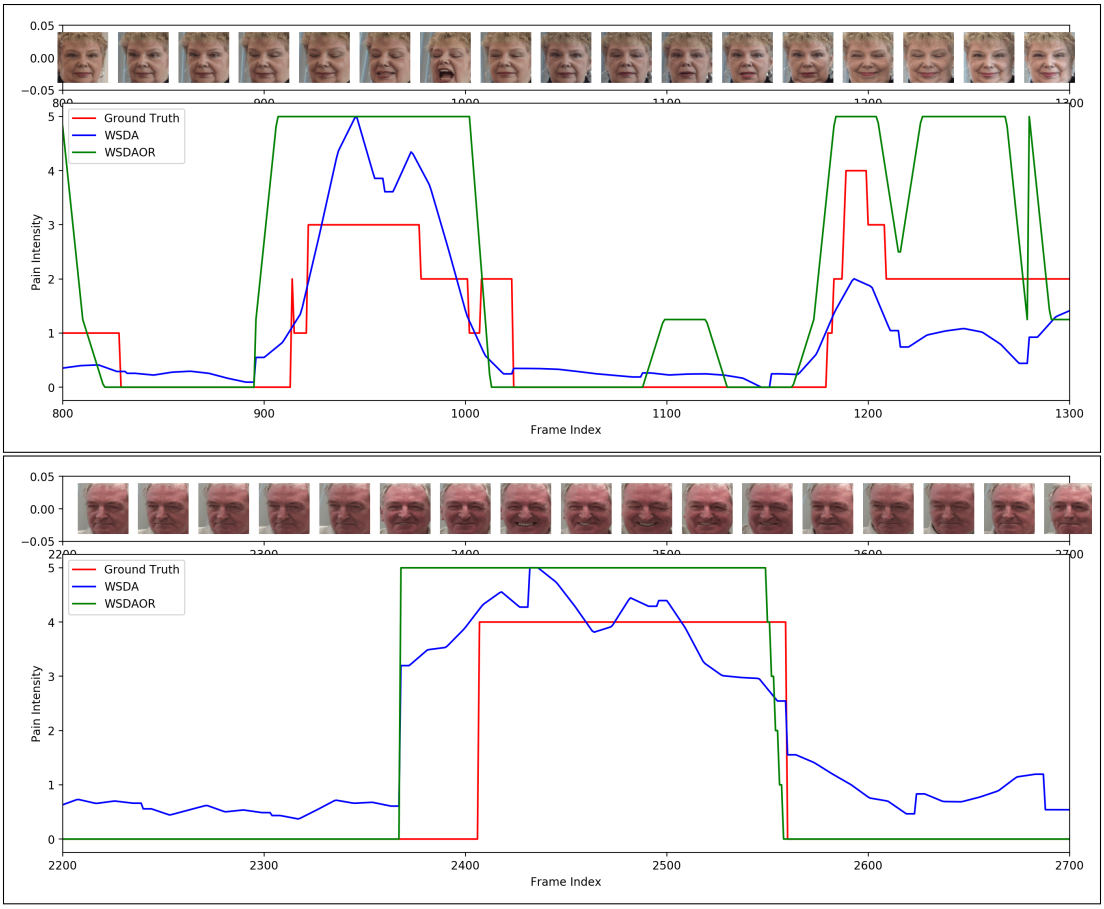

Figure 3.5 Visualization of pain localization on two different subjects in UNBC dataset. From top to bottom: Scenario with multiple peaks of pain expressions, Scenario where ground truth (GT) shows no pain, but our deep WSDA-OR approach correctly localizes pain better than WSDA
Taken from Rajasekhar *et al.* (2021b)

Since the problem of pain intensity estimation in the framework of ordinal regression with MIL is still at a rudimentary level, we have also compared our approach with that of (Sikka *et al.*, 2014), which was proposed for the classification of pain events at the sequence level. Unlike the existing classification approaches based on MIL, which estimates sequence level labels, we have estimated the ordinal intensity level of the individual frames using weak supervision of sequence level labels in the target domain. Compared to the classification-based approaches (Wu *et al.*, 2015b), (Sikka *et al.*, 2014), regression-based approaches (Ruiz *et al.*, 2018) are found to show superior performance due to the fact that intensity level estimation is closely related to

Table 3.3 Performance of the proposed WSDA-OR approach with state-of-the-art in terms of PCC, ICC, and MAE

Method	**Type of Supervision**	**Frame-level**			**Sequence-level**		
		PCC ↑	**ICC ↑**	**MAE ↓**	**PCC ↑**	**ICC ↑**	**MAE ↓**
MIR (Hsu *et al.*, 2014)	Weak	0.350	0.240	0.840	0.63	0.630	0.940
MILBOOST (Sikka *et al.*, 2014)	Weak	0.280	0.110	1.770	0.380	0.380	1.700
MI-DORF (Ruiz *et al.*, 2018)	Weak	0.400	0.460	0.190	0.670	0.660	0.800
WSDA (Gnana Praveen *et al.*, 2020)	Weak	0.630	0.567	0.714	0.828	0.762	0.647
WSDA-OR (ours)	Weak	**0.705**	**0.696**	**0.530**	**0.745**	**0.750**	**0.443**
BORMIR (Zhang *et al.*, 2018b)	Semi	0.605	0.531	0.821	-	-	-
LSTM (Rodriguez *et al.*, 2018)	Full	0.780	-	0.500	-	-	-
SCN (Tavakolian & Hadid, 2018)	Full	0.920	0.750	0.320 (MSE)	-	-	-

the problem of regression. However, failing to exploit the discrete nature of labels shows poor performance in accurately tracking the pain intensity levels as that of the ground truth as shown in Figure 3.5. Therefore, we have exploited the ordinal nature of the labels in our formulation, which significantly improves the performance of the approach and effectively tracks the pain intensity levels as reflected in the performance evaluation metrics as shown in Table 3.3.

The proposed approach shows superior performance in terms of PCC compared to ICC and MAE, thereby effectively tracking the pain intensity levels of the individual frames as shown in Figure 3.5. Since ICC is more reliable than PCC for sequence-level estimation as it efficiently captures the intra-class correlation, the proposed approach exhibits the higher performance of ICC for sequence-level estimation compared to frame-level estimation. We have further compared the performance of the approach to that of state-of-the-art fully supervised (Tavakolian & Hadid, 2018) as well as partially annotated scenarios (Zhang *et al.*, 2018b). The proposed approach performs better than (Zhang *et al.*, 2018b) even without the requirement of any prior information pertinent to peak and valley frames.

3.4.6 Results with Additional Datasets:

The WSDA-OR model was also validated on Biovid and Fatigue (private) datasets. In our experiments, we have used Biovid Part A, which has 8700 videos of 87 subjects, which are labeled with pain levels of 0 to 4 at sub-sequence level. The Fatigue dataset is obtained from

Table 3.4 PCC, ICC and MAE performance of proposed WSDA-OR approach under different scenarios

Training Scenario	**Sequence-Level**		
	PCC ↑	**ICC ↑**	**MAE ↓**
Biovid Dataset			
Supervised (source data only)	0.026	0.003	1.424
Transfer learning	0.246	0.240	1.242
DA (proposed approach)	0.341	0.317	1.162
Fatigue (Private) Dataset			
Supervised (source data only)	0.028	0.007	1.645
DA (proposed approach)	0.436	0.367	0.363

18 participants in the Rehabilitation center, who are suffering from fatigue-related issues. A total of 27 video sessions are captured from 18 participants with a duration of 40 - 45 minutes and the videos are labeled at sequence level on a scale of 0 to 10 for every 10 to 15 minutes. However, due to the lack of frame-level labels and the availability of state-of-the-art results for these datasets in the context of weak supervision, we have validated with other baseline models with transfer learning for sequence level estimation. With Biovid, 20 subjects are randomly selected for testing out of 87 subjects, whereas with Fatigue, on each trial we alternate testing on one subject out of 18 subjects.

We have conducted experiments on three scenarios and the results are shown in Table 3.4. In the first scenario, the model was trained using the Recola dataset as source domain data. Since the model is trained only on the source domain, the model shows poor performance on the test (target) data since there is a significant shift between the source and target domains. In the second case, the model trained using Recola as source domain data is further fine-tuned on Biovid as target domain data. This model shows improvement when compared to the first case, where the model is trained only on source data. Finally, we train a model using our WSDA-OR proposed approach, which shows significant improvement compared to the previous scenarios since it minimizes the domain differences and leverages the variability of both domains to improve the generalization capability.

3.5 Conclusion

In this work, we have proposed a generic DL framework of weakly-supervised DA with ordinal regression for pain level assessment in videos. To address the problem of variations across different operational conditions, we have explored deep DA to leverage the performance of deep models by overcoming the problem of limited representative training data. We have formulated the framework of deep DA in the context of limited annotations, where the source domain is assumed to have fully supervised labels and the target domain is assumed to have weak sequence level labels. Contrary to the existing approaches for pain intensity estimation, which explored DL models for regression, we have shown that exploiting ordinal relationships among the intensity levels significantly improves the performance of the system to accurately track and localize the pain intensity levels in videos. The ordinal intensity levels are modeled using a Gaussian distribution, which efficiently captures the ordinal relationships among the intensity levels.

In the conventional MIR approach (Hsu *et al.*, 2014), the weak sequence-level label is associated only with the frame pertinent to the frame with the maximum intensity level. However, we have improved the performance of the system by deploying multiple frames relevant to the sequence level label instead of a single frame. We have conducted an extensive set of experiments with various baseline models using various combinations of source and target datasets and further analyzed the performance of the proposed approach under varying levels of supervision for the target data. Finally, we have compared the performance of the proposed approach with the state-of-the-art approaches and shown that the proposed approach significantly outperforms over the state-of-the-art approaches.

CHAPTER 4

AUDIO-VISUAL FUSION FOR EMOTION RECOGNITION IN THE VALENCE-AROUSAL SPACE USING JOINT CROSS-ATTENTION

Gnana Praveen Rajasekhar[a] , Patrick Cardinal[b] , Eric Granger[a]

[a]Department of Systems Engineering, École de technologie supérieure,
[b]Department of Software and IT Engineering, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in IEEE Transactions on Biometrics, Behavior, and Identity Science, January 2023

Abstract

Automatic emotion recognition (ER) has recently gained lot of interest due to its potential in many real-world applications. In this context, multimodal approaches have been shown to improve performance (over unimodal approaches) by combining diverse and complementary sources of information, providing some robustness to noisy and missing modalities. In this paper, we focus on dimensional ER based on the fusion of facial and vocal modalities extracted from videos, where complementary audio-visual (A-V) relationships are explored to predict an individual's emotional states in valence-arousal space. Most state-of-the-art fusion techniques rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of A-V modalities. To address this problem, we introduce a joint cross-attentional model for A-V fusion that extracts the salient features across A-V modalities, that allows to effectively leverage the inter-modal relationships, while retaining the intra-modal relationships. In particular, it computes the cross-attention weights based on correlation between the joint feature representation and that of the individual modalities. By deploying the joint A-V feature representation into the cross-attention module, it helps to simultaneously leverage both the intra and inter modal relationships, thereby significantly improving the performance of the system over the vanilla cross-attention module. The effectiveness of our proposed approach is validated experimentally on challenging videos from the RECOLA and AffWild2 datasets. Results indicate that our joint cross-attentional A-V fusion model provides a cost-effective

solution that can outperform state-of-the-art approaches, even when the modalities are noisy or absent.

4.1 Introduction

Automatic recognition and analysis of human emotions have drawn much attention over the past few decades. It has been extensively researched in various fields such as neuroscience, psychology, cognitive science, and computer science, leading to the advancement of a wide range of applications in various fields, such as health care (e.g., assessment of anger, fatigue, depression, and pain), robotics (human-machine interaction), driver assistance (assessment of driver's state), etc (Kołakowska, Landowska, Szwoch, Szwoch & Wróbel, 2014). Emotion recognition (ER) is a challenging problem since the expressions linked to human emotions are extremely diverse in nature across individuals and cultures. Ekman conducted a cross-cultural study on human emotions and categorized the basic emotions into six categories – anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992). Subsequently, contempt has been added to these six basic emotions (Matsumoto, 1992). The categorical model of ER has been explored extensively in the field of affective computing due to its simplicity and universality (Anagnostopoulos, Iliou & Giannoukos, 2015).

Recently, real-world applications have driven a shift of affective computing research from laboratory-controlled environments to more realistic natural settings. This shift has further led to the analysis of a wide range of subtle, continuous emotional and health states that are elicited in real-world settings. Conventionally, the estimation of continuous ER states is formulated as the dimensional ER problem, where complex human emotions can be represented in a dimensional valence-arousal space. Figure 4.1 illustrates the use of a two-dimensional space to represent emotional states, where valence and arousal are employed as dimensional axes (Schlosberg, 1954). Valence reflects the wide range of emotions in the dimension of pleasantness from being negative (sad) to positive (happy), whereas arousal spans the range of intensities from passive (sleepiness) to active (high excitement) (Nicolaou, Gunes & Pantic, 2011). Recognizing such fine-grained emotional states is beneficial in various applications, such as assessing driver

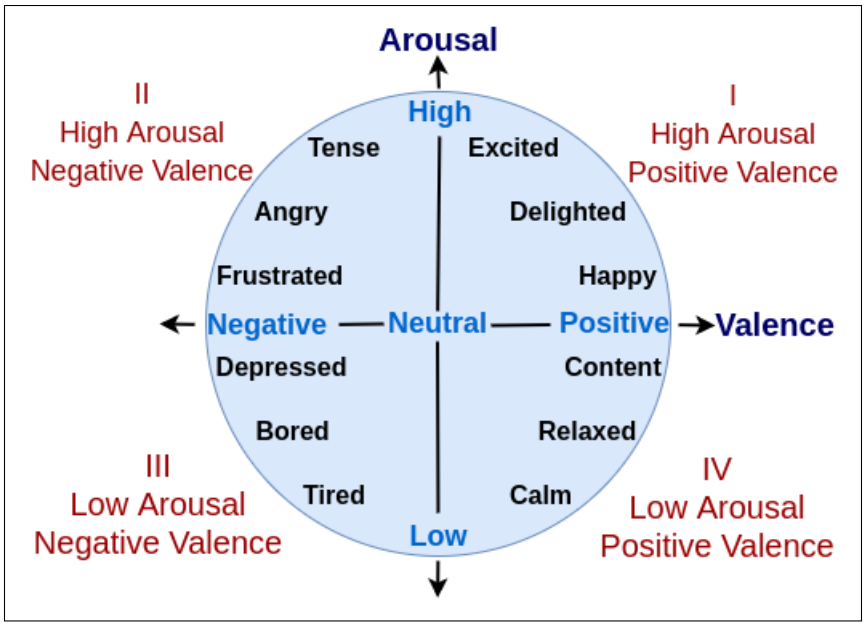

Figure 4.1 The valence-arousal space. Valence denotes the range of emotions from being very sad (negative) to very happy (positive) and arousal reflects the energy or intensity of emotions from very passive to very active
Taken from Praveen *et al.* (2023a)

fatigue, estimating the level of depression or pain in health care, assessing customer engagement in marketing, etc. Given the growing need for continuous ER in real-world applications, this paper focused on dimensional ER in the valence-arousal space.

Human emotions can be conveyed through various modalities such as face, voice, text, and physiology (electroencephalogram, electrocardiogram, etc.), which typically carry complementary information among them. Although human emotions can be expressed through various modalities, vocal and facial modalities are the predominant contact-free channels, which carry complementary information (Shivappa *et al.*, 2010). Audio-visual (A-V) fusion has also been widely explored for various applications including identity verification (Ben-Yacoub, Luttin, Jonsson, Matas & Kittler, 1999), event localization (Duan *et al.*, 2021), action recognition (Lee *et al.*, 2021), etc. Efficiently leveraging the complementary nature of A-V relationships captured in videos can play a crucial role in improving the performance of multimodal systems

over unimodal systems (Shon, Oh & Glass, 2019). Techniques for multimodal fusion can be broadly categorized as model-agnostic or model-based (Baltrušaitis, Ahuja & Morency, 2019). In model-based approaches, fusion is performed using specialized models to cope with the diverse information in multimodal data. Depending on the type of model used for fusion, these techniques are typically classified further as kernel methods, graphical models, or neural networks (Baltrušaitis *et al.*, 2019). Unlike model-based fusion, model-agnostic fusion can be achieved using almost any uni-modal classifier or regressor. They do not rely on any specialized model for fusion. Most of the existing fusion models belong to this category, where fusion is often performed by concatenating the features or individual modal predictions. Model agnostic approaches can be further classified as three major strategies: decision-, feature-, and hybrid-level fusion (Wu *et al.*, 2014). In decision-level fusion (late fusion), multiple modalities are trained end-to-end independently, and then the predictions obtained from the individual modalities are fused to obtain the final predictions. Although decision-level fusion is easy to implement and requires less training, it neglects the interactions across the individual modalities, thereby resulting in limited improvement over uni-modal approaches. Conventionally, feature-level fusion (early fusion) is achieved by concatenating the features of A-V modalities immediately after they are extracted, which is further used for predicting the final outputs. Hybrid fusion takes advantage of both decision-level and feature-level fusion by combining outputs from both feature-level fusion and decision-level fusion. Though feature level fusion is conventionally done by aggregating or concatenating the features immediately after they are extracted, it can also be performed by learning the interactions between the modalities for better feature representations before concatenating the features (Zhang *et al.*, 2021b; Rajasekhar *et al.*, 2021a). In this work, we explore feature-level fusion based on joint cross-attention, where the A and V features extracted from videos are further modeled using a joint cross-attention model before concatenation.

Deep learning (DL) models provide state-of-the-art performance in many V recognition applications, such as image classification, object detection, action recognition, etc. Inspired by their performance, several ER approaches have been proposed for video-based dimensional ER using CNNs to obtain the deep features, and a recurrent neural network to capture the

temporal dynamics (Schoneveld *et al.*, 2021; Tzirakis *et al.*, 2017). Deep models have also been widely explored for vocal emotion recognition, typically using spectrograms with 2D-CNNs (Schoneveld *et al.*, 2021; Wang *et al.*, 2021), or raw waveforms with 1D-CNNs (Tzirakis *et al.*, 2017). In most of the existing approaches (Tzirakis *et al.*, 2017, 2021) for dimensional ER, A-V fusion is performed by concatenating the deep features extracted from individual facial and vocal modalities and fed to LSTM for predicting valence and arousal. Although LSTM-based fusion models can improve the system performance by leveraging the intra-modal relationships, it does not effectively capture the inter-modal relationships across the individual modalities. We, therefore, investigate the prospect of extracting more comprehensive salient features that can effectively exploit the complementary relationships across the A and V modalities.

Attention mechanisms have recently gained much interest in the areas of computer vision and machine learning as they allow extracting task-relevant features, thereby improving system performance. This has been extensively explored for various applications, such as event/action recognition (Shi, Dai, Mu & Wang, 2020), ER (Lee *et al.*, 2018), etc. Most of the existing attention-based approaches for dimensional ER explore the intra-modal relationships (Lee *et al.*, 2018). Although a few approaches (Parthasarathy & Sundaram, 2021; Tzirakis *et al.*, 2021) attempt to capture the cross-modal relationships using cross-attention based on transformers, they fail to effectively leverage the complementary relationship of A-V modalities. Indeed, their computation of attention weights does not consider the correlation across the A and V features.

A preliminary version of the cross-attentional (CA) A-V fusion model for dimensional ER was presented in our previous work (Rajasekhar *et al.*, 2021a). In this work, we further extend our previous work, where a joint A-V feature representation is deployed in the CA model to jointly capture both intra and inter-modal relationships. In this previous paper (Rajasekhar *et al.*, 2021a), the attention weights are computed based on the correlation across A and V modalities, which depends only on intermodal relationships. Instead of using individual feature representations across the modalities to generate the attention weights, we introduce joint A-V feature representations to capture the relationships within the same modality as well as other modalities, thereby leveraging both inter- and intra-modal relationships to obtain the attention

weights. Using the joint feature representation drastically reduces the heterogeneity across the A and V features, which further helps to provide robust A-V feature representations. Specifically, we obtain the cross-correlation matrix across the deep joint feature representation and features of individual modalities to obtain the attention weights for the A and V modalities. Therefore, the attention weights of each modality are obtained not only using the features of itself but also from the other modality, resulting in more informative features. Besides providing improved performance over individual modalities, a benefit of our joint A-V representation is its ability to perform well even when a modality is noisy or absent. Finally, we have also explored the impact on JCA performance of the feature-level fusion, where multiple diverse backbones are combined for the A and V modalities.

The main contributions of the paper are: (1) A joint cross-attentional (JCA) model for A-V fusion is introduced to effectively exploit the complementary relationship across modalities for dimensional ER in valence-arousal space. Contrary to the prior approaches, the proposed model simultaneously leverages both intra and inter-modal relationships to effectively capture the complementary relationships. (2) Deploying the joint feature representation also helps to reduce the heterogeneity across A and V features, thereby resulting in robust AV feature representations. (3) An extensive set of experiments on the challenging RECOLA and Affwild2 datasets indicate that our proposed JCA fusion model can outperform related state-of-the-art fusion models for dimensional ER. Our visual interpretation of the fusion process shows that JCA can efficiently leverage the complementary intermodal relationships while retaining the intramodal relationships.

The rest of this paper is organized as follows. Section II provides a critical analysis of the relevant literature on dimensional ER and attention models for A-V fusion. Section III describes the proposed JCA A-V fusion model in detail. Section IV presents the experimental methodology for the backbones of the individual modalities and the experimental settings used in our fusion model. Finally, the results obtained with the proposed approach with RECOLA and Affwild2 datasets are presented and discussed in Section V.

4.2 Related Work

4.2.1 Audio-Visual Fusion for Dimensional Emotion Recognition:

One of the early approaches using DL models for A-V fusion-based dimensional ER was proposed by (Tzirakis *et al.*, 2017), where A and V features are obtained using ResNet50 and 1D-CNN respectively. The obtained features are then concatenated and fed to a Long short-term memory model (LSTM) for the prediction of valence and arousal. (Ortega *et al.*, 2019) investigated an empirical study of fine-tuning pretrained CNN models by freezing various convolutional layers. (Schoneveld *et al.*, 2021) explored knowledge distillation using the teacher-student model for V modality and the CNN model for A modality using spectrograms. The deep feature representations are combined using a model-based fusion strategy, where RNNs are used to capture the temporal dynamics. Inspired by the deep auto-encoders, (Nguyen *et al.*, 2021) investigated the prospect of how to simultaneously learn compact representative features from A and V modalities using deep auto-encoders. They have proposed a deep model of two-stream auto-encoders and LSTM for efficiently integrating V and A streams for dimensional ER.

(Deng, Wu & Shi, 2021) proposed an iterative self-distillation method for modeling the uncertainties in the labels in a multi-task framework. They have trained a model with multiple task labels, which is further used to distill iteratively to several student models. They have shown that iterative distillation significantly improves the performance of the system. (Kuhnke *et al.*, 2020) proposed two stream A-V network, where V features are extracted from R(2plus1)D model (Tran *et al.*, 2018) pretrained from action recognition dataset and A features are obtained from Resnet18 model (He *et al.*, 2016). The obtained features are further concatenated for the final prediction of valence and arousal. (Wang *et al.*, 2021) further improved their approach (Kuhnke *et al.*, 2020) by introducing teacher-student model in a semi-supervised learning framework. The teacher model is trained on the available labels, which are further used to obtain pseudo labels for unlabeled data. The pseudo labels are finally used to train the student model, which is used for the final prediction. Though the above-mentioned approaches have shown significant improvement for dimensional ER, they fail to effectively capture the inter-modal relationships

and relevant salient features specific to the task. Therefore, we have focused on capturing the comprehensive features in a complementary fashion using attention mechanisms.

4.2.2 Attention Models for Audio-Visual Fusion:

Attention mechanisms are widely used in the context of multimodal fusion with various modalities such as A and text (Lee, Yoon & Jung, 2020; Krishna & Ankita, 2020), V and text (Ma *et al.*, 2018a; Wei, Zhang, Li, Zhang & Wu, 2020), etc. (Zhao *et al.*, 2020) proposed an end-to-end architecture for emotion classification by integrating spatial, channel-wise and temporal attention into V network and temporal attention into A network. (Ghaleb *et al.*, 2020) explored attention to weigh the time windows of a video sequence to efficiently exploit the temporal interactions between the A-V modalities. They used transformer (Vaswani *et al.*, 2017) based encoders to obtain the attention weights through self-attention for emotion classification. (Lee *et al.*, 2018) proposed spatiotemporal attention for the V modality to focus on emotional salient parts using Convolutional LSTM (ConvLSTM) modules and a temporal attention network using deep networks for A modality. Then the attended features are concatenated and fed to the regression network for the prediction of valence and arousal. However, these approaches focused on modeling the intra-modal relationships and failed to effectively exploit the inter-modal relationship of the A-V modalities.

(Wang *et al.*, 2020) investigated the prospect of exploiting the implicit contextual information along with the A and V modalities. They have proposed an end-to-end architecture using cross-attention based on transformers for A-V group ER. (Parthasarathy & Sundaram, 2021) also explored transformers with cross-modal attention for dimensional ER, where cross-attention is integrated along with self-attention. (Tzirakis *et al.*, 2021) investigated various fusion strategies along with attention mechanisms for A-V fusion-based dimensional ER. They have further explored self-attention as well as cross-attention fusion based on transformers to enable the extracted features of different modalities to attend to each other. Although these approaches have explored cross-modal attention with transformers, they fail to leverage semantic relevance among the A-V features based on cross-correlation.

(Zhang *et al.*, 2021b) investigated the prospect of improving the fusion performance over individual modalities and proposed leader-follower attentive fusion for dimensional ER. The obtained features are encoded and attention weights are obtained by combining the encoded A and V features. The attention weights are further attended on the V features and concatenated to the original V features for final prediction. (Zhang *et al.*, 2020) proposed an attentive fusion mechanism, where V features are obtained from 3D-CNNs and A features from spectrograms fed to 2D-CNN. The obtained A and V features are further re-weighted using weights, obtained from scoring functions based on the relevant information in the individual modalities. (Wang *et al.*, 2020a) addressed the problem of multi-modal feature fusion along with frame alignment issues between A and V modalities using cross-attention for speech recognition. (Luo *et al.*, 2018) investigated the potential of joint representation learning using Convolutional Recurrent Neural Networks (CRNN) for vocal ER. They have also shown that the impact of time intervals significantly impacts the performance of the system. (Hu *et al.*, 2019) proposed dense multi-modal fusion by densely integrating the representation at multiple shared layers to capture hierarchical correlations across the modalities. (Vukotić *et al.*, 2016) proposed a cross-modal deep network architecture, where the weights of two deep networks are enforced to be symmetry, yielding joint representation in a common feature space. In this work, we have used a simple joint representation of feature concatenation of A and V modalities in our JCA framework.

Unlike prior approaches, we advocate for a simple yet efficient JCA model based on joint modeling of intra and inter-modal relationships between A and V modalities. Cross-attention has been successfully applied in several applications, such as weakly-supervised action localization (Lee *et al.*, 2021), and few-shot classification (Hou, Chang, MA, Shan & Chen, 2019). A similar idea of exploiting the complementary relationships for better audiovisual fusion has also been explored for person verification (Shon *et al.*, 2019), where an attention layer is used for the fusion of A and V modalities. In most of these cases, cross-attention has been applied across the individual modalities. However, we have explored joint attention between individual and combined AV features. By deploying the joint AV feature representation, we can effectively capture the intra and inter-modal relationships simultaneously by allowing interactions across

the modalities as well as within oneself. Recently, joint co-attention has been explored by (Duan *et al.*, 2021) recursively for A-V event localization. They have shown that recursive training of joint co-attention yields more discriminant and robust feature representations for multimodal fusion. In this paper, joint (combined) A-V features are extracted through cross-attention (instead of co-attention) for dimensional ER. Specifically, the features of each modality attend to themselves, as well as those of the other modality, through cross-correlation of the concatenated A-V features, and features of individual modalities. By effectively leveraging the joint modeling of intra- and inter-modal relationships, the proposed approach can significantly improve system performance.

4.3 Proposed Approach

4.3.1 Visual Network:

Facial expressions from videos involve both appearance and temporal dynamics of video sequences. Efficient modeling of these spatial and temporal dynamics plays a crucial role in extracting discriminant and robust features, which in turn improves the overall system performance. State-of-the-art performance is typically achieved using 2D-CNN in combination with Recurrent Neural Networks (RNN) to capture the effective latent appearance representation, along with temporal dynamics (Kim *et al.*, 2019). Several approaches have been explored for dimensional facial ER based on 2D-CNNs and LSTMs (Nicolaou *et al.*, 2011; Wöllmer, Kaiser, Eyben, Schuller & Rigoll, 2013). However, 3D-CNNs are found to be efficient in capturing the spatiotemporal dynamics in videos (Rajasekhar *et al.*, 2021b), and have also been explored for dimensional facial ER. For instance, in (Kuhnke *et al.*, 2020), they have shown that R3D (Tran *et al.*, 2018) pretrained on the Kinetics-400 action recognition dataset (Kay *et al.*, 2017) has outperformed conventional 2D-CNNs for dimensional ER on Affwild2 dataset. Inspired by the performance of 3D-CNNs, we consider Inflated 3D-CNN (Carreira & Zisserman, 2017), to extract spatiotemporal features of the facial clips from a video sequence. Initially, proposed by (Carreira & Zisserman, 2017) for action recognition, the Inflated 3D (I3D) CNN model

can efficiently capture the spatiotemporal dynamics of the V modality while optimizing fewer parameters than that of conventional 3D-CNNs. The I3D model is obtained by inflating the filters and pooling kernels of 2D ConvNet, expanding to 3D CNN. Therefore, it allows leveraging existing common pretrained 2D-CNNs, which are trained on large-scale image datasets for facial expressions, thereby improving the spatial discrimination for videos. Though I3D model has been primarily explored for action recognition, it has also been used for other applications in the field of affective computing, like in video-based pain localization (Gnana Praveen *et al.*, 2020), etc. In the proposed approach, we train the I3D model to extract spatiotemporal features for the facial modality (see implementation details in Section 4.4.2).

4.3.2 Audio Network:

The para-lingual information of vocal signals was found to convey significant information on the emotional state of a person. Even though vocal ER has been widely explored using conventional handcrafted features, such as Mel-frequency cepstral coefficients (MFCCs) (Sethu *et al.*, 2015), there has been a significant improvement over the recent years with the introduction of DL models. Though deep vocal ER models can be explored using spectrograms with 2D-CNNs (Schoneveld *et al.*, 2021; Wang *et al.*, 2021), as well as raw A signal with 1D-CNNs (Tzirakis *et al.*, 2017), spectrograms are found to carry significant para-lingual information about the affective state of a person (Ma *et al.*, 2018b; Satt *et al.*, 2017). Spectrograms have been explored with various 2D-CNNs in the literature for ER (Slimi, Hamroun, Zrigui & Nicolas, 2020; Albanie, Nagrani, Vedaldi & Zisserman, 2018). Therefore, we consider spectrograms in the proposed framework along with 2D-CNN models to extract A features. In particular, Resnet18 (He *et al.*, 2016) was used for Affwild2 dataset, and the A model is shown in Table 4.2 for RECOLA dataset. Given the differences in the size of the datasets, we have used different 2D-CNN models for RECOLA and Affwild2 to avoid over-fitting. (see implementation details in Section 4.4.2).

4.3.3 Feature-Level Fusion of Multiple Backbones:

We have also explored the fusion of features extracted from multiple backbones for both A and V modalities. Deploying multiple backbones for each modality can allow for capturing diverse information for the same modality. Specifically, we have extracted V features from I3D, R3D, and 2D CNN in conjunction with Long Short-Term Memory (LSTM). I3D and R3D are 3D CNN models, used to simultaneously capture the spatiotemporal relationships, which is efficient at capturing the short-term temporal relations. 2D CNN with LSTM extracts spatial features, and performs temporal modeling, which captures the long-term temporal relationships. Similarly, for A modality, combined features from a 2D CNN trained on spectrograms, and conventional handcrafted MFCC features, are widely used in speech processing for many applications.

Then we have considered two different feature-level fusion strategies to obtain a feature representation for each modality. First, we concatenate the features from all the backbones, followed by a fully connected layer to produce a compact joint representation based on multiple diverse backbones. Feature concatenation followed by a fully connected layer has been widely used in the literature for many applications. The second strategy is a more specialized feature stacking approach, where the features extracted from multiple diverse backbones and a sequence are assembled into a block of features, and then processed by the A-V fusion model. This approach eliminates the need for training an additional fully connected layer to combine features, as all features are trained within the fusion model.

4.3.4 Joint Cross-Attentional Audio-Visual Fusion:

Though A-V fusion can be achieved through unified multimodal training, it was found that multimodal performance often declines over that of individual modalities (Wang *et al.*, 2020b). This has been attributed to several factors, such as differences in learning dynamics for A and V modalities (Wang *et al.*, 2020b), different noise topologies, with some modality streams containing more or less information for the task at hand, as well as specialized input representations (Nagrani *et al.*, 2021). Therefore, we have trained DL models for the individual

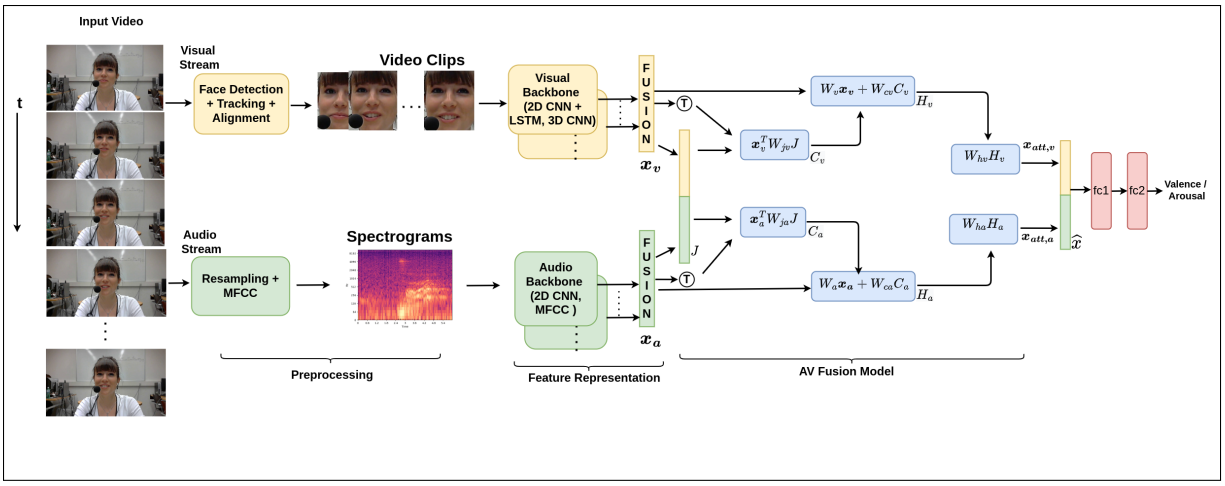

Figure 4.2 Joint cross-attention model proposed for A-V fusion (in testing mode) Taken from Praveen *et al.* (2023a)

A and V modalities independently to extract A and V features, which is further fed to the JCA fusion model for A-V fusion that outputs final valence and arousal prediction.

For a given video sequence, the V modality carries relevant information in some video clips, whereas the A modality might be more relevant for others. Since multiple modalities convey more diverse and complementary information for valence and arousal than a single modality, their complementarity can be effectively explored through A and V fusion. To reliably fuse these modalities, we rely on a cross attention based fusion mechanism to efficiently encode the inter-modal information while preserving the intra-modal characteristics. Though cross-attention has been conventionally applied across the features of individual modalities, we have explored cross-attention in a joint framework. Specifically, our joint A-V feature representation is obtained by concatenating the A and V features to attend to the individual A and V features. By using the joint representation, features of each modality attend to oneself, as well as the other modality, helping to capture the semantic inter-modal relationships across A and V. The heterogeneity among the A and V modalities can also be drastically reduced by using the combined feature representation in the cross-attentional module, which further improves system performance. A block diagram of the proposed model is shown in Figure 4.2.

A) Training mode: Let $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$ represent two sets of deep feature vectors extracted for the A and V modalities, in response to a given input video sub-sequence $\boldsymbol{S}$ of fixed size, where $\boldsymbol{X_a} = \{\boldsymbol{x}_\mathbf{a}^1, \boldsymbol{x}_\mathbf{a}^2, ..., \boldsymbol{x}_\mathbf{a}^L\} \in \mathbb{R}^{d_a \times L}$ and $\boldsymbol{X_v} = \{\boldsymbol{x}_\mathbf{v}^1, \boldsymbol{x}_\mathbf{v}^2, ..., \boldsymbol{x}_\mathbf{v}^L\} \in \mathbb{R}^{d_v \times L}$. L denotes the number of non-overlapping fixed-size clips sampled uniformly from $\boldsymbol{S}$, d_a and d_v represents the dimension of the A and V feature representations, respectively, and $\boldsymbol{x}_\mathbf{a}^l$ and $\boldsymbol{x}_\mathbf{v}^l$ denotes the A and V feature vectors, respectively, for $l = 1, 2, ..., L$ clips. Instead of applying cross-attention across the features of individual A and V modalities, we use cross-attention in a joint learning framework. The joint representation of A-V features, $\boldsymbol{J}$, is obtained by concatenating the A and V feature vectors:

$$\boldsymbol{J} = [\boldsymbol{X_a}; \boldsymbol{X_v}] \in \mathbb{R}^{d \times L} \tag{4.1}$$

where $d = d_a + d_v$ denotes the feature dimension of concatenated features.

The concatenated A-V feature representations ($\boldsymbol{J}$) of the given video sub-sequence ($\boldsymbol{S}$) are now used to attend to unimodal feature representations $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$. The joint correlation matrix $\boldsymbol{C_a}$ across the A features $\boldsymbol{X_a}$, and the combined A-V features $\boldsymbol{J}$ are given by:

$$\boldsymbol{C_a} = \tanh\left(\frac{\boldsymbol{X_a^T W_{ja} J}}{\sqrt{d}}\right) \tag{4.2}$$

where $\boldsymbol{W_{ja}} \in \mathbb{R}^{L \times L}$ represents learnable weight matrix across the A and combined A-V features, and T denotes transpose operation. Similarly, the joint correlation matrix for V features is given by:

$$\boldsymbol{C_v} = \tanh\left(\frac{\boldsymbol{X_v^T W_{jv} J}}{\sqrt{d}}\right) \tag{4.3}$$

The joint correlation matrices $\boldsymbol{C_a}$ and $\boldsymbol{C_v}$ for A and V modalities provide a semantic measure of relevance not only across the modalities but also within the same modality. A higher correlation coefficient of the joint correlation matrices $\boldsymbol{C_a}$ and $\boldsymbol{C_v}$ shows that the corresponding samples are strongly correlated within the same modality as well as other modality. Therefore, the proposed approach can efficiently leverage the complementary nature of A and V modalities (i.e., inter-modal relationship) as well as intra-modal relationships, thereby improving the performance

of the system. After computing the joint correlation matrices, the attention weights of the A and V modalities are estimated.

Since the dimensions of joint correlation matrices ($\mathbb{R}^{d_a \times d}$) and the features of corresponding modality ($\mathbb{R}^{L \times d_a}$) differ, we rely on different learnable weight matrices corresponding to features of the individual modalities, and the corresponding joint correlation matrices, to compute attention weights of the modalities. For the A modality, the joint correlation matrix $\boldsymbol{C_a}$ and the corresponding A features $\boldsymbol{X_a}$ are combined using the learnable weight matrices $\boldsymbol{W_{ca}}$ and $\boldsymbol{W_a}$ respectively to compute the attention weights of A modality, which is given by

$$\boldsymbol{H_a} = ReLu(\boldsymbol{W_a X_a} \; + \; \boldsymbol{W_{ca} C_a^T}) \tag{4.4}$$

where $\boldsymbol{W_{ca}} \in \mathbb{R}^{k \times d}$, $\boldsymbol{W_a} \in \mathbb{R}^{k \times L}$ and $\boldsymbol{H_a}$ represents the attention maps of the A modality. Similarly, the attention maps ($\boldsymbol{H_v}$) of V modality are obtained as

$$\boldsymbol{H_v} = ReLu(\boldsymbol{W_v X_v} \; + \; \boldsymbol{W_{cv} C_v^T}) \tag{4.5}$$

where $\boldsymbol{W_{cv}} \in \mathbb{R}^{k \times d}$, $\boldsymbol{W_v} \in \mathbb{R}^{k \times L}$. In our experiments, we have considered k to be 32.

Finally, the attention maps are used to compute the attended features of the A and V modalities. These features are obtained as:

$$\boldsymbol{X_{att,a}} = \boldsymbol{W_{ha} H_a} + \boldsymbol{X_a} \tag{4.6}$$

$$\boldsymbol{X_{att,v}} = \boldsymbol{W_{hv} H_v} + \boldsymbol{X_v} \tag{4.7}$$

where $\boldsymbol{W_{ha}} \in \mathbb{R}^{k \times L}$ and $\boldsymbol{W_{hv}} \in \mathbb{R}^{k \times L}$ denote the learnable weight matrices, respectively. The attended A and V features, $\boldsymbol{X_{att,a}}$ and $\boldsymbol{X_{att,v}}$ are further concatenated to obtain the A-V feature representation, which is given by:

$$\widehat{\mathbf{X}} = [\boldsymbol{X_{att,v}}; \boldsymbol{X_{att,a}}] \tag{4.8}$$

Finally, the A-V features are fed to the fully connected layers for the predictions of valence and arousal.

The Concordance Correlation Coefficient (ρ_c) has been widely used in the literature to measure the level of agreement between the predictions (x) and ground truth (y) annotations for dimensional ER (Tzirakis *et al.*, 2017). Let μ_x and μ_y represent the mean of predictions and ground truth, respectively. Similarly, if σ_x^2 and σ_y^2 denote the variance of predictions and ground truth, respectively, then ρ_c between the predictions and ground truth is:

$$\rho_c = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4.9}$$

where σ_{xy}^2 denotes the predictions – ground truth covariance. Although MSE has been widely used as a loss function for regression models, we use $\mathcal{L} = 1 - \rho_c$ since it is a standard and conventional loss function in the literature of dimensional ER literature (Tzirakis *et al.*, 2017). The parameters of our A-V fusion model ($\boldsymbol{W}_{\mathbf{ca}}$, $\boldsymbol{W}_{\mathbf{a}}$, $\boldsymbol{W}_{\mathbf{cv}}$, $\boldsymbol{W}_{\mathbf{v}}$, $\boldsymbol{W}_{\mathbf{ha}}$, and $\boldsymbol{W}_{\mathbf{hv}}$) are optimized according to this loss. **B) Test mode:** As shown in Figure 4.2, we assume that a continuous video sequence is an input to our model during inference. Feature representations $\boldsymbol{x}_{\mathbf{a}}$ and $\boldsymbol{x}_{\mathbf{v}}$ are extracted by A and V backbones for successive input clips and spectrograms, and fed to the fusion model for the prediction of valence and arousal.

4.4 Experimental Methodology

4.4.1 Datasets:

The proposed architecture is validated on REmote COLlaborative and Affective (RECOLA) (Ringeval *et al.*, 2013) and AffWild2 (Kollias *et al.*, 2019).

RECOLA: In total, this dataset consists of 9.5 hours of multimodal recordings, which are recorded by 46 French-speaking participants, performing a collaborative task during a video conference. Among the participants, 17 are French, 3 are German and 3 are Italian. The video

sequences are divided into sequences of 5 minutes each, which are annotated with a regressed intensity value for every 40 msec by 6 French-speaking annotators (three male and three female). The dataset is split into three partitions: train (16 subjects), validation (15 subjects), and test (15 subjects) by balancing the age and gender of the speakers. Due to the uncontrolled spontaneous nature of expressions of the subjects, the dataset has been widely used by the research community in affective computing for various challenges such as AVEC 2015 (Ringeval *et al.*, 2015a), AVEC 2016 (Valstar *et al.*, 2016), etc. Most of the existing approaches in the literature, e.g., (Schoneveld *et al.*, 2021; Tzirakis *et al.*, 2017), have been validated on the dataset used for the AVEC 2016 (Valstar *et al.*, 2016) challenge, which consists of 9 subjects for training, and 9 subjects for validation. Therefore, we have also validated our proposed approach on the dataset used in AVEC 2016 challenge.

Affwild2: Affwild2 is the largest dataset in the field of affective computing, consisting of 564 videos collected from YouTube, all captured in the wild (Kollias *et al.*, 2019). Sixteen of these videos display two subjects, both of which have been annotated. The annotations are provided by four experts using a joystick and the final annotations are obtained as the average of the four raters. In total, there are $2,816,832$ frames with 455 subjects, out of which 277 are male and 178 female. The annotations for valence and arousal are provided continuously in the range of $[-1, 1]$. The dataset is split into training, validation, and test sets. The partitioning is done in a subject-independent manner so that every subject's data will present in only one subset. The partitioning produces 341, 71, and 152 videos for the training, validation, and test sets respectively.

4.4.2 Implementation Details:

RECOLA: For the **V modality**, the faces are extracted and pre-processed from the video sequences of the dataset using MTCNN model (Zhang *et al.*, 2016), a deep cascaded multi-task framework of face detection and alignment. Faces are resized to 224×224 to be fed to the I3D model (Carreira & Zisserman, 2017). To generate more samples, the videos of the dataset are converted to sequences of 128 frames with a subsequence length of 16, resulting in 21,284

training samples and 16,177 validation samples. I3D used the Inception_v1 architecture as the base model as shown in Table 4.1, which is pre-trained on Kinetics-400 dataset (Kay *et al.*, 2017), and then inflated to a 3D-CNN using RECOLA videos of facial expressions. Typically, the pooling operation is performed on the last convolutional layer($512 \times 7 \times 7$) to reduce the spatial dimension to size 1 ($7 \rightarrow 1$), however, it may leave out useful information. Therefore, the scaled dot product of audio and visual features are performed to smoothly reduce the dimension of raw visual features as in (Duan *et al.*, 2021). For regularizing the network, dropout is used with $p = 0.8$ on the linear layers. The initial learning rate of the network was set to be $1e-4$ and the momentum of 0.9 is used for Stochastic Gradient Descent (SGD). Also, weight decay of $5e-4$ is used. Due to hardware limitations and memory constraints, the batch size of the network is set to 8. Data augmentation is performed on the training data by random cropping, which produces a scale-invariant model. The number of epochs is set to 50 and early stopping is used to obtain the best network weights.

Table 4.1 Deep NN (I3D) for V Model.
"Conv : 64, $7 \times 7 \times 7$, $2 \times 2 \times 2$" : 3D conv layer of 64 filters,
each of kernel size $7 \times 7 \times 7$ and stride $2 \times 2 \times 2$.
"Pool : $3 \times 3 \times 3$, $1 \times 2 \times 2$" : kernel size $3 \times 3 \times 3$ and
stride $1 \times 2 \times 2$. "Linear: in = 1024, out = 256":
fully connected layer of input size 1024 and output size 256

Stage	**Layers**	**Output size**
Input	-	3 x 8 x 224 x 224
Block 1	Conv : 64, 7x7x7, 2x2x2 Max pool : 1x3x3, 1x2x2	64 x 7 x 112 x 112
Block 2	Conv : 192, 3x3x3, 1x2x2 Max pool : 3x3x3, 1x2x2	192 x 7 x 56 x 56
Block 3	2 x Inception Module Max pool : 3x3x3, 1x2x2	480 x 6 x 28 x 28
Block 4	5 x Inception Module Max pool : 3x3x3, 1x2x2	832 x 2 x 14 x 14
Block 5	2 x Inception Module Avg pool : 2x7x7, 1x2x2	1024 x 1 x1 x 1
Block 6	Linear : in = 1024, out = 256	256x1
Block 7	Linear : in = 256, out = 1	1x1

The **A network** is composed of 3 blocks of conv. layers: the first block has conv. layer followed by the max pooling layer, the second block has two conv. layers followed by max pooling layer, and the third block has two conv. layers followed by the average pooling layer, which then outputs the feature vectors. Finally, the feature vectors are fed to the linear layers to obtain the final predictions. All the conv. and linear layers are followed by ReLU activation functions. The vocal signal is extracted from video sub-sequences and re-sampled to 16KHz, which is further segmented into short segments. First, we split the extracted vocal signal to 5.12 sec, which corresponds to the sequence of 128 frames of the V network. Next, the spectrogram is obtained using Discrete Fourier Transform (DFT) of length 1024 for each short vocal segment of 5.12 sec, where the window and shift length are both 40 msec to match with the granularity of annotation frequency. Following aggregation of short-time spectra, we obtain the spectrogram of 128×129. The spectrogram is converted to log-power-spectrum, expressed in dB. Finally, mean and variance normalization is performed on the spectrogram. Apart from mean and variance normalization, no other voice-specific processing such as silence removal, noise filtering, etc is performed. These spectrograms are then fed to the deep NN described in Table 4.2. The A

Table 4.2 Deep NN for A Model. "Conv: 64, 5×5, 1×2" denotes a conv layer of 64 filters, each of kernel size 5×5 and stride of 1×2. "Pool : 4×4, 4×4" denotes kernel size of 4×4 and stride of 4×4. "Linear: in = 1024, out = 256" denotes linear fully connected layer of input size 1024 and output size 256.

Stage	**Layers**	**Output size**
Input	-	1 x 128 x 129
Block 1	Conv : 64, 5x5, 1x2 Max pool : 4x4, 4x4	64 x 31 x 15
Block 2	Conv : 128, 5x5, 1x2 Conv : 256, 3x3, 1x1 Max pool : 2x2, 2x2	256 x 15 x 4
Block 3	Conv : 512, 5x5, 1x1 Conv : 1024, 3x3, 1x1 Avg pool : 13x2, 1x1	1024 x 1 x 1
Block 4	Linear : in = 1024, out = 256	256x1
Block 5	Linear : in = 256, out = 1	1x1

network is trained from scratch, where the initial weights of the network are initialized with values from a normal distribution. The number of epochs is set to 100, and early stopping is used. The network is optimized using SGD with a momentum of 0.9. The initial learning rate is set to be 0.001 and the batch size is fixed to be 16. Due to the limited data, the network might be prone to over-fitting. Therefore, to prevent the network from over-fitting, dropout is used with p = 0.5 after the last linear layer. Also, weight decay of $5e-4$ is used for all the experiments.

For the **A-V fusion network**, the size of A-V features is set to be 1024. In the joint cross-attention module, the initial weights of the cross-attention matrix are initialized with Xavier method (Glorot & Bengio, 2010), and the weights are updated using Adam optimizer. The initial learning rate is set to be 0.001 and batch size is fixed to be 16. Also, a dropout of 0.5 is applied on the attended A-V features and weight decay of $5e-4$ is used for all the experiments. Due to the spontaneity of the expressions, the annotations are also found to be highly stochastic in nature. Therefore, post-processing steps are applied to predictions and labels. A rigorous analysis of some of the post-processing steps for annotations appears in (Huang *et al.*, 2015). (Tzirakis *et al.*, 2017) explored a series of post-processing steps for validating their architecture on the RECOLA. Inspired by their approach, we have followed similar post-processing steps to validate our architecture: (i) median filtering with the window size ranging from 0.4sec to 20sec; (ii) centering the predicted values by computing the bias between annotated (ground truth) values and predicted values; (iii) matching the scaling of predicted values and annotations using the ratio of the standard deviation of annotated values and predicted values. (iv) time shifting the annotations forward in time with values ranging from 0.04 to 10sec to compensate for the delay in human annotations (delay in correspondence between the annotated values and the video frames). The details regarding the complexity of the code is provided in Appendix II.

Affwild2: For the **V modality**, we have used the cropped and aligned images provided by the challenge organizers (Kollias & Zafeiriou, 2021a). For the missing frames in the V modality, we have considered black frames (i.e., zero pixels). Faces are resized to 224x224 to be fed to the I3D network. The subsequence length and the sequence length of the videos are considered to be 8 and 64 respectively, obtained by down-sampling a sequence of 256 frames by 4. Therefore,

we have 8 sub-sequences in each sequence, resulting in 1,96,265 training samples and 41,740 validation samples, and 92,941 test samples. Similar to the RECOLA dataset, the I3D model was pre-trained on the Kinetics-400 dataset (Kay *et al.*, 2017), and inflated to a 3D-CNN using Affwild2 videos of facial expressions. Instead of a conventional pooling layer after the last convolutional layer, we have used scaled dot product of audio and visual features similar to that of (Duan *et al.*, 2021). To regularize the network, dropout is used with $p = 0.8$ on the linear layers. The initial learning rate was set to be $1e - 3$, and the momentum of 0.8 is used for SGD. Weight decay of $5e - 4$ is used. Here again, the batch size of the network is set to be 8. Data augmentation is performed on the training data by random cropping, which produces a scale-invariant model. The number of epochs is set to 50 and early stopping is used to obtain the best weights of the network.

For the **A modality**, the vocal signal is extracted from the corresponding video and re-sampled to 44100Hz, which is further segmented to short vocal segments corresponding to a sub-sequence of 256 frames of the V network. The spectrogram is obtained using Discrete Fourier Transform (DFT) of length 1024 for each short segment, where the window length is considered to be 20 msec and the hop length to be 10 msec. Following aggregation of short-time spectra, we obtain the spectrogram of 64 x 107 corresponding to each sub-sequence of the V modality. Now a normalization step is performed on the obtained spectrograms. The spectrogram is converted to log-power-spectrum, expressed in dB. Finally, mean and variance normalization is performed on the spectrogram. Now the obtained spectrograms are fed to the Resnet18 (He *et al.*, 2016) to obtain the A features. Due to the availability of a large number of samples in the Affwild2 dataset, we trained the Resnet18 model from scratch. To adapt to the number of channels of the spectrogram, the first conv. layer in the Resnet18 model is replaced by a single channel. The network is trained with an initial learning rate of 0.001 and weights are optimized using the Adam optimizer. The batch size is considered to be 64 and early stopping is used to obtain the best prediction model. For the **A-V fusion network**, we have used a similar training strategy as with the RECOLA dataset.

4.5 Results and Discussion

4.5.1 Ablation Study:

Table 4.3 Performance of our approach with various components on the RECOLA dataset. The 2D-CNN in Table 4.2 is used to extract A features in all experiments.

Method: V + Fusion	**Valence**	**Arousal**
2D-CNN + Feature Concatenation	0.538	0.680
2D-CNN + LSTM	0.552	0.697
I3D + Feature Concatenation	0.579	0.732
I3D + Cross-Attention (Rajasekhar *et al.*, 2021a)	0.687	0.831
I3D + Joint Cross-Attention (JCA)	0.728	0.842
I3D; R3D; 2DCNN + JCA	0.762	0.891

RECOLA: Table 4.3 presents the results of our ablation study on the RECOLA validation dataset. To analyze the performance of our joint cross-attention model for A-V fusion, we compare it with various fusion strategies widely used in the literature. One of those fusion strategies is LSTM-based fusion, where the A and V features are concatenated and fed to the LSTM followed by linear layers. We have extracted V features (frame-level) using VGG 2D-CNN architecture, pretrained on FER dataset similar to (Ortega *et al.*, 2019), and further fine-tuned on RECOLA. Initially, we compare the proposed approach without LSTM, where the A and V features are concatenated and directly fed to linear layers. LSTM model-based fusion is evaluated by feeding the concatenated features to LSTM layer followed by fully connected layers. Given the temporal modeling of the concatenated features, the fusion performance improves over the non-LSTM based fusion strategy. We also compare the performance to I3D using baseline concatenation, where the A-V features are concatenated without attention, and fed to linear layers for valence/arousal prediction (similar to that of the fusion in (Ortega *et al.*, 2019)). We have further compared the performance improvement of joint cross-attention fusion over that of conventional CA fusion (Rajasekhar *et al.*, 2021a). In the case of conventional CA fusion, attention weights are computed based on the cross-correlation across the A and V modalities. The attention weights encode the semantic relevance across the A and V features. However,

they do not allow the features to interact within the same modality, thereby failing to capture the temporal modeling within the same modality. Though temporal modeling across the modalities captures inter-modal relationships and can improve state-of-art accuracy, retaining the temporal modeling within the same modality also plays a pivotal role to capture intra-modal relationships. Therefore, we have integrated the modeling within A and V modalities, along with modeling of inter-modal relationships, and further improve system performance. Since we have introduced joint feature representation in the proposed JCA fusion model, it simultaneously captures both intra- and inter-modal relationships and thereby outperforms the conventional CA fusion in (Rajasekhar *et al.*, 2021a), along with most of the widely used fusion strategies.

Table 4.4 Performance of our approach with various components on the Affwild2 dataset. Resnet18 (He *et al.*, 2016) is used to extract A features in all experiments.

Method: V + Fusion	**Valence**	**Arousal**
TSAV (Ortega *et al.*, 2019) + Feature Concatenation	0.531	0.493
TSAV (Ortega *et al.*, 2019) + Joint Cross-Attention (Ours)	0.642	0.592
I3D + Feature Concatenation	0.498	0.452
I3D + Leader-Follower Fusion (Schoneveld *et al.*, 2021)	0.592	0.521
I3D + Cross-Attention (Rajasekhar *et al.*, 2021a)	0.541	0.517
I3D + Joint Cross-Attention (Ours)	0.657	0.580
I3D; R3D; 2DCNN + JCA	0.725	0.614

Affwild2: Table 4.4 presents the results of our ablation study on the Affwild2 validation dataset. The performance of our proposed JCA fusion was compared using various A and V backbones and A-V fusion strategies. Since we used I3D for the V modality, we have compared it against a V backbone based on 3D (2plus1d) CNNs (Ortega *et al.*, 2019). First, we implemented the backbone of TSAV (Ortega *et al.*, 2019) with simple feature concatenation, where the extracted A and V features are concatenated, and fed to fully connected layers for valence and arousal prediction. Our proposed model provides a significant performance improvement. We have also analyzed when our V backbone (I3D) is used with baseline feature concatenation and leader-follower fusion-based attention (Schoneveld *et al.*, 2021) with our backbones and found that there is a significant improvement in the performance over that of baseline feature concatenation. We have also implemented the conventional CA fusion (Rajasekhar *et al.*,

2021a) with the I3D backbone. Although its performance improves over that of baseline feature concatenation, it shows lower performance than leader-follower attention (Schoneveld *et al.*, 2021). Finally, we have compared the proposed JCA fusion with I3D and found that it outperforms other fusion strategies in the literature on Affwild2. By allowing the features of each modality to interact with itself and other modalities, we can effectively capture the semantic relevance of intra- and inter-modal relationships of A and V modalities for dimensional ER. We can also observe that the performance of our proposed A-V fusion model with TSAV (Ortega *et al.*, 2019) slightly outperforms that of JCA fusion with I3D. We have further validated the proposed fusion model with multiple backbones of V and A modalities and showed further improvement in the performance of the system.

We have also explored multiple backbones for A and V modalities along with the proposed fusion model and further improved the performance of the system. As discussed in Section 4.3.1 and 4.3.2 for V and A modalities respectively, we have used I3D, R3D, and 2D CNN in conjunction with LSTM to obtain spatiotemporal features for V modality. Similarly, we have used MFCCs and spectrograms with 2D CNNs for A modality. The features of multiple backbones are fused using feature concatenation followed by a fully connected layer and stacking of features to obtain comprehensive features for both A and V modalities. Features from multiple backbones help to obtain diverse information about each modality and thereby improve the performance of the system. The proposed AV JCA fusion model is validated with the fusion of features from multiple backbones for both A and V modalities and the results are shown in Table 4.5.

Table 4.5 Performance of the proposed AV fusion model using the fusion of features from multiple backbones for A and V modalities. FC denotes a fully connected layer.

Dataset	**Fusion Method**	**Valence**	**Arousal**
RECOLA	Concatenation + FC	0.762	0.891
	Stacking	0.754	0.865
Affwild2	Concatenation + FC	0.725	0.614
	Stacking	0.712	0.595

We have also evaluated our proposed approach for the case where a growing proportion of A is replaced by background noise in test mode. Specifically, we have randomly replaced some segments/spectrograms to reflect background noise in the video. We have tested our system on Affwild2 with a video named "16-30-1920x1080.mp4" with 5475 frames and varied the percentage of missing spectrograms by 10, 25, 50 and 100%. Even though spectrograms are noisy and absent, we can observe that there is a modest minimal decline in CCC performance (see Fig 4.3). In particular, since we can effectively encode the complementary relationship across modalities (by jointly modeling intra- and inter-modal relationships), our models can sustain a high level of performance for valence.

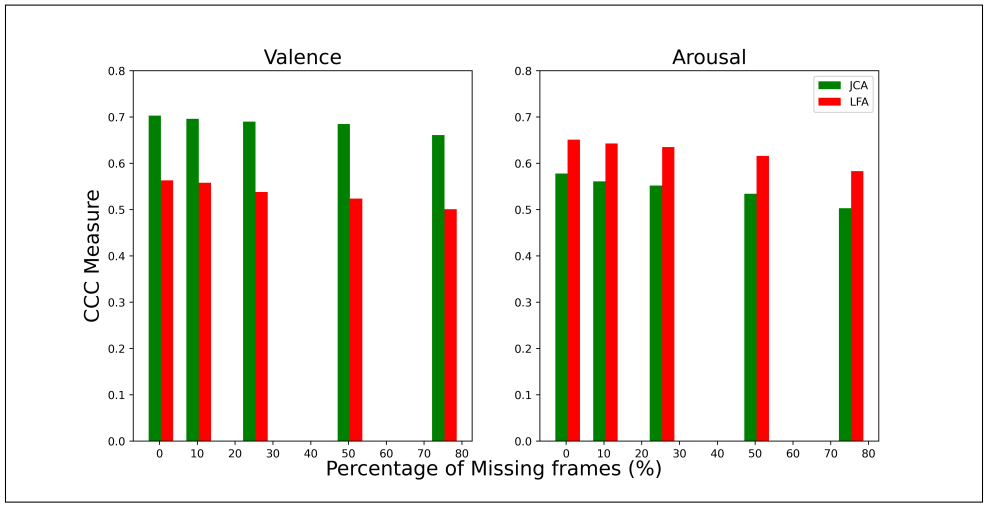

Figure 4.3 Performance of our proposed A-V fusion (JCA) and Leader-Follower Attention (LFA) (Zhang *et al.*, 2021b) models with a growing proportion of missing A modality
Taken from Praveen *et al.* (2023a)

4.5.2 Comparison to State-of-the-Art:

RECOLA: Table 4.6 presents our comparative results against state-of-the-art A-V fusion models on the RECOLA development set. (He *et al.*, 2015) explored handcrafted LPQ-TOP features for V, and low-level descriptors (LLDs) such as MFCC, energy, etc. for A, along with physiological modalities like electrocardiogram (ECG) and electro-dermal activity (EDA). Given the use of additional physiological modalities, and more LLD descriptors in A, as well as additional

geometric features of V, the fusion performance provides significant improvement. (Han, Zhang, Cummins, Ringeval & Schuller, 2017) explored LLD features for A, and facial landmark features (only geometric) for V, combined hierarchically to leverage the individual advantages of support vector regressor (SVR) and Bidirectional Long Short-Term Memory Networks (BLSTM), and improved the performance for valence. Inspired by the performance of DL models, (Tzirakis *et al.*, 2017) explored Resnet50 2D-CNN for V and 1D-CNN on raw data for A. However, the features are directly concatenated, and fed to LSTMs. This results in a decline in CCC performance over individual modalities. The performance has been further improved by (Ortega *et al.*, 2019), where they pre-train a CNN on FER for V and LLD for A. Recently, (Schoneveld *et al.*, 2021) used knowledge distillation for V, and a VGG network on spectrograms for A. Instead of direct concatenation, they rely on two independent CNNs before concatenating them, and showed that their fusion outperforms individual modalities. Though deep models have improved the performance over handcrafted features, they fail to effectively leverage the complementary nature of the A-V modalities. By effectively leveraging the intra and inter-modal relationships of A and V features, the proposed model outperforms state-of-the-art approaches using joint cross-attention.

Table 4.6 CCC performance of proposed and state-of-art methods for A-V fusion on the RECOLA development set. (SM represents strength modeling of SVR + BLSTM.)

Method – A/V backbone	**Valence**			**Arousal**		
	Audio	**Visual**	**Fusion**	**Audio**	**Visual**	**Fusion**
(He *et al.*, 2015) – A: LLDs; V: LLDs	0.400	0.441	0.609	0.800	0.587	0.747
(Han *et al.*, 2017) – A: LLDs + SM; V: geometric features + S.M.	0.480	0.592	0.554	0.760	0.350	0.685
(Tzirakis *et al.*, 2017) – A: 1D-CNN; V: Resnet50	0.428	0.637	0.502	0.786	0.371	0.731
(Ortega *et al.*, 2019) – A:LLDs; V: 2D-CNN	-	-	0.565	-	-	0.749
(Schoneveld *et al.*, 2021) – A: Finetuned VGGish; V: Distilled CNN	0.460	0.550	0.630	0.800	0.570	0.810
(Rajasekhar *et al.*, 2021a) – A: 2D-CNN; V: I3D	0.463	0.642	0.687	0.822	0.582	0.831
Joint Cross-Attention (Ours) – A: 2D-CNN; V: I3D	0.463	0.642	**0.695 ± 0.033**	0.822	0.582	**0.801 ± 0.041**

Affwild2: Table 4.7 shows our comparative results against relevant state-of-the-art A-V fusion models on the Affwild2 dataset. In recent years, most of the existing work on the Affwild2 dataset has been submitted to the Affective Behavior Analysis in-the-wild (ABAW) challenges (Kollias, Schulc, Hajiyev & Zafeiriou, 2020; Kollias & Zafeiriou, 2021a). Therefore, we compare

Table 4.7 CCC performance of the proposed and state-of-the-art methods for A-V fusion on the Affwild2 development set. (TCN denotes Temporal Convolutional Network.)

Method – A/V backbone	**Valence**			**Arousal**		
	Audio	**Visual**	**Fusion**	**Audio**	**Visual**	**Fusion**
(Kuhnke *et al.*, 2020) – A: Resnet18; V: R(2plus1)D	0.355	0.463	0.493	0.359	0.570	0.613
(Zhang *et al.*, 2021b) – A: VGGish; V: Resnet50 + TCN	-	0.425	0.469	-	**0.647**	**0.649**
(Rajasekhar *et al.*, 2021a) – A: Resnet18; V: I3D	0.355	0.412	0.541	0.359	0.534	0.517
Joint Cross-Attention (Ours) – A: Resnet18; V: I3D	0.355	0.412	**0.625 ± 0.032**	0.359	0.534	0.541 ± 0.039

our proposed approach with that of the top relevant approaches appearing in ABAW challenges for A-V fusion.

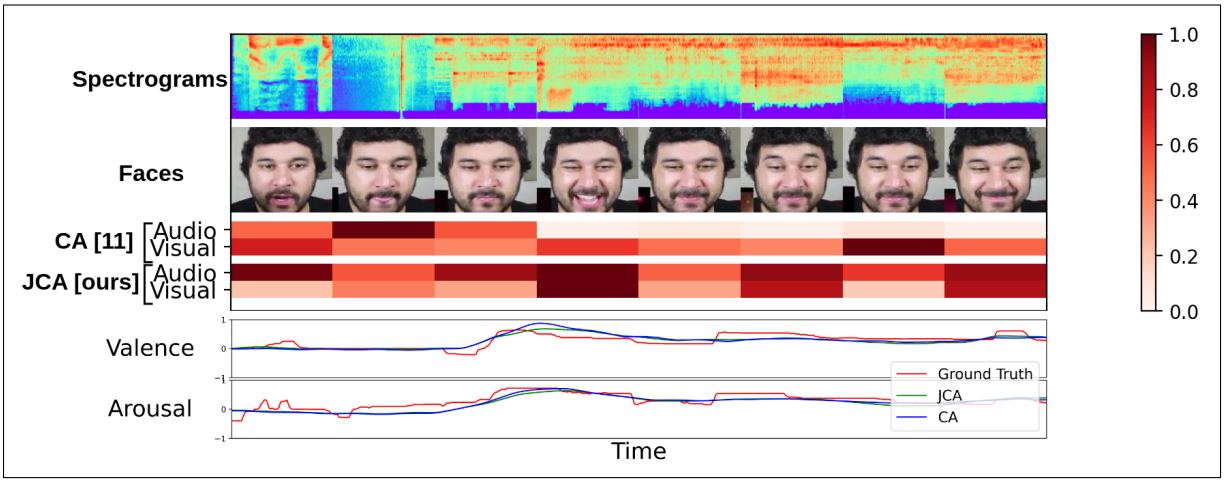

Figure 4.4 Visualization of attention scores of our proposed A-V fusion (JCA) and CA (Rajasekhar *et al.*, 2021a) models on a video named "317" of Affwild2 dataset Taken from Praveen *et al.* (2023a)

However, the experimental protocol and training data vary widely among these approaches. We, therefore, re-implemented these approaches according to our experimental protocol and analyzed the results on the Affwild2 validation set for a fair comparison. Similar to our A and V backbones, (Kuhnke *et al.*, 2020) also used 3D-CNNs, where R(2plus1)D model is used for visual modality and Resnet18 is used for audio modality. However, they perform simple feature concatenation without any specialized fusion model for the prediction of valence and arousal. So the fusion performance was not significantly improved over the uni-modal performance. (Zhang

et al., 2021b) explored the leader-follower attention model for fusion and showed minimal improvement in fusion performance over uni-modal performances. Though they have shown significant performance for arousal than valence, it is highly attributed to the visual backbone. In our proposed approach, we have shown significant improvement for fusion, especially for valence than arousal. Even with vanilla CA fusion (Rajasekhar *et al.*, 2021a), we have shown that fusion performance for valence has been improved better than (Zhang *et al.*, 2021b) and (Kuhnke *et al.*, 2020). By deploying joint representation into the cross attentional fusion model, the fusion performance of valence has been significantly improved further. In the case of arousal, though the fusion performance is lower than that of (Zhang *et al.*, 2021b) and (Kuhnke *et al.*, 2020), we can observe that it has been improved better than that of uni-modal visual performance. Therefore, the proposed approach is more effective in capturing the variations spanning over a wide range of emotions (valence) than that of the intensities of the emotions (arousal).

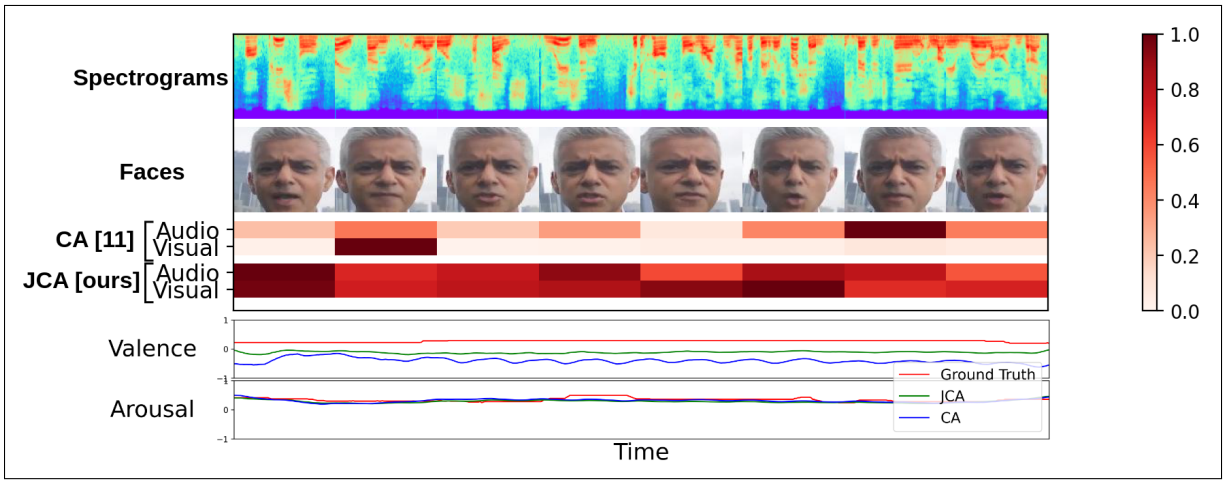

Figure 4.5 Visualization of attention scores of our proposed A-V fusion (JCA) and CA (Rajasekhar *et al.*, 2021a) models on a video named "video92" of Affwild2 validation data Taken from Praveen *et al.* (2023a)

Table 4.8 shows the results of our approach against relevant state-of-the-art A-V fusion models on the Affwild2 test set. In recent years, several challenges such as FG2020 (Kollias *et al.*, 2020), ICCV2021 (Kollias & Zafeiriou, 2021a) have been performed on the Affwild2 dataset as it has been the largest in-the-wild dataset in the field of affective computing. (Ortega *et al.*, 2019) proposed a two stream A-V network by using R(2plus1)D (Tran *et al.*, 2018) for V

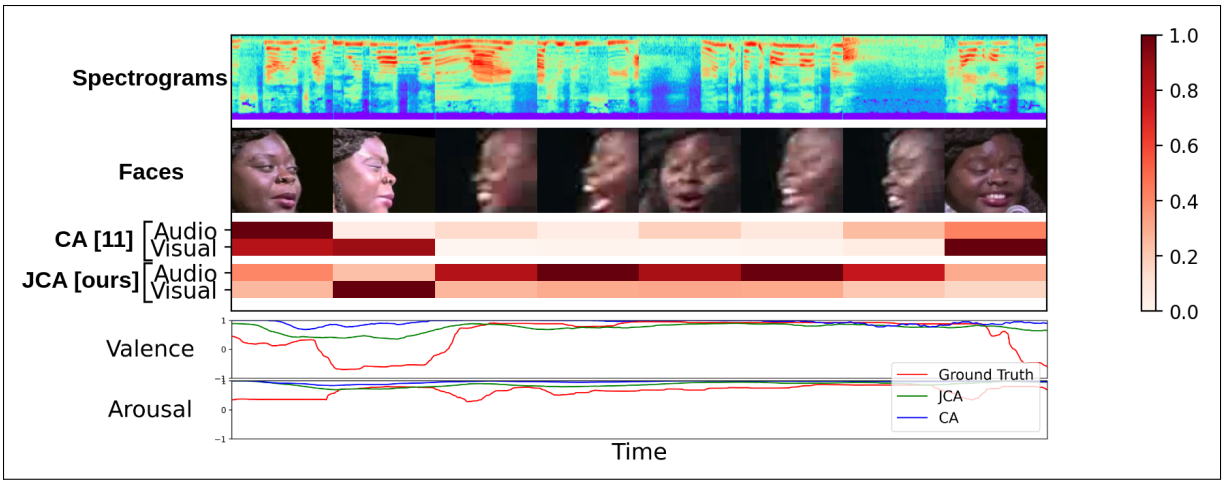

Figure 4.6 Visualization of attention scores of our proposed A-V fusion (JCA) and CA (Rajasekhar *et al.*, 2021a) models on a video named "21-24-1920x1080" of Affwild2 validation dataset. Negative example where the proposed approach fails to focus on semantic information
Taken from Praveen *et al.* (2023a)

stream, and Resnet18 (He *et al.*, 2016) for A stream. They have also used additional masks as external inputs to guide the spatial attention of the V modality and label filtering based on multi-task labels to deal with the noisy annotations of valence and arousal. (Wang *et al.*, 2021) further extended their approach to perform semi-supervised learning. However, they use the annotations of other ABAW challenge tasks (expression classification and action unit classification) to filter the noisy labels of valence and arousal, as well as to estimate pseudo labels for the unlabeled samples. (Deng *et al.*, 2021) proposed an iterative distillation method for modeling the uncertainty of annotations of valence and arousal and showed significant improvement in the performance. However, they have used iterative distillation of student models, which is computationally expensive as well as labels of other tasks to model the uncertainty of valence/arousal labels. (Zhang *et al.*, 2021b), (Meng *et al.*, 2022) and (Karas, Tellamekala, Mallol-Ragolta, Valstar & Schuller, 2022) are the only approaches, which does not use the labels of additional tasks. (Meng *et al.*, 2022) has shown significant improvement in the performance by using three external datasets along with multiple backbones of A and V modalities, whereas (Zhang *et al.*, 2021b) and (Karas *et al.*, 2022) use only Affwild2 dataset similar to ours. The

proposed approach performs at par with that of (Zhang *et al.*, 2021b) and is better than that of (Karas *et al.*, 2022) in terms of valence.

Table 4.8 CCC of the proposed approach compared to state-of-the-art methods for A-V fusion on Affwild2 test set.

Method	**Valence**	**Arousal**	**Mean**
(Meng *et al.*, 2022)	0.606	0.596	0.601
(Kuhnke *et al.*, 2020)	0.448	0.417	0.432
(Zhang *et al.*, 2021b)	0.463	0.492	0.477
(Wang *et al.*, 2021)	0.478	0.498	0.488
(Deng *et al.*, 2021)	0.533	0.454	0.493
(Karas *et al.*, 2022)	0.418	0.407	0.413
JCA (Ours)	**0.451**	**0.389**	**0.420**

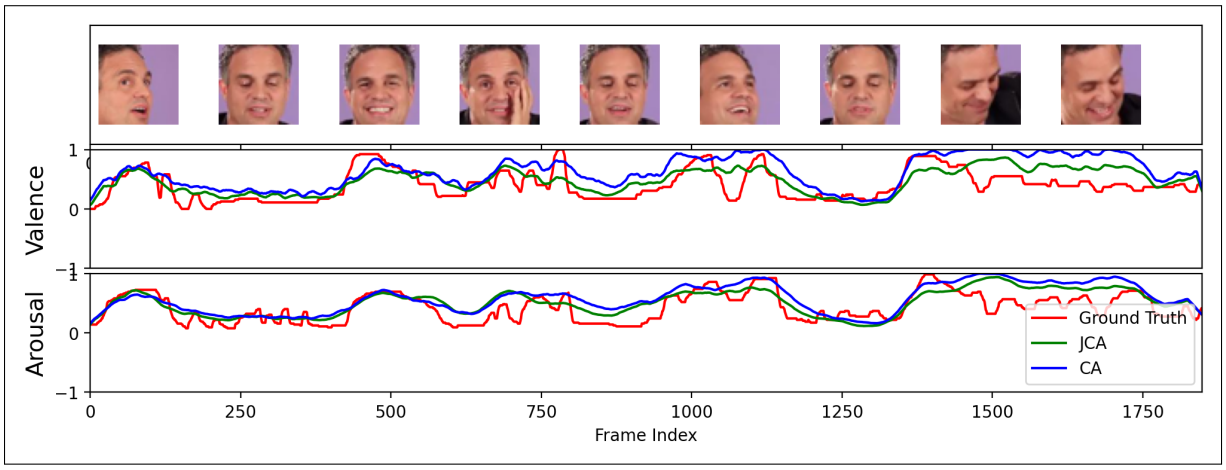

Figure 4.7 Visualization of valence and arousal predictions over time for our proposed A-V fusion (JCA) and Cross-Attention (CA) (Rajasekhar *et al.*, 2021a) on video named "video67" of Affwild2 validation dataset
Taken from Praveen *et al.* (2023a)

4.5.3 Visual Analysis

We have further validated the proposed approach using interpretability analysis by visualizing the attention scores of the A and V modalities. In the proposed approach, we have primarily exploited the temporal attention within the same modality, as well as across the modalities. So the clip-level attention scores help us to intuitively understand the semantic clips in the video, where the fusion attention model focused on the temporal sequence of A and V modalities. To highlight the improvement of the proposed approach w.r.t. that of the vanilla CA model (Rajasekhar *et al.*, 2021a), we have also plotted the attention scores of the proposed JCA model along with that of (Rajasekhar *et al.*, 2021a) including the predictions and ground truth of valence and arousal. It can be observed that the proposed JCA model can effectively capture the importance of modalities, as well as the temporal significance within the modalities. For instance, as shown in Fig 4.4, the proposed JCA model focused on V modality when the person smiles as the facial muscles around his nose and mouth significantly change over time. Similarly, the proposed model assigns a high attention score for A modality when the person exhibits significant modulation of vocal expressions. From Fig 4.5, we observe that the proposed model assigns a higher attention score for clips when the person elicits knitted brows and significant facial muscle movement near his mouth, whereas (Rajasekhar *et al.*, 2021a) fails to capture those important clips of the V modality. In both cases, we can observe that JCA assigns a higher attention score to the corresponding modality when there is significant temporal variation (i.e., facial expression or tone changes), whereas the vanilla CA model (Rajasekhar *et al.*, 2021a) fails to focus on the some of the important clips of V modality. Since the proposed JCA model leverages both the intramodal and intermodal relationships, it can effectively leverage the contextual information among A and V modalities. Therefore, the proposed model can efficiently exploit the importance of modalities as well as temporal importance within the modalities, resulting in better performance than that of (Rajasekhar *et al.*, 2021a), which has also been reflected in the predictions of valence and arousal.

Though the proposed JCA fusion model can outperform (Rajasekhar *et al.*, 2021a), we observe lower performance for arousal than with valence, on both RECOLA and Affwild2 datasets. Since

the proposed model considers the intra and inter-variations in computing the attention scores, JCA fusion sometimes becomes misleading by assigning higher attention scores for neutral frames, and lower attention scores for more relevant clips when there is significant occlusion, blur or pose variations in the temporal sequence of the V modality. For instance, as shown in Fig 4.6, the proposed model assigns higher attention scores for neutral clips, but lower attention scores for clips with more relevant facial expressions due to blur and strong pose variations. In addition to the attention scores of A and V modalities, we also visualize the valence and arousal predictions over time for videos of the Affwild2 dataset. The proposed JCA model is able to capture the contextual relationships between A and V modalities better than that of (Rajasekhar *et al.*, 2021a), which helps to achieve better performance. As shown in Fig 4.7, we can observe that both the JCA and vanilla cross-attention models (Rajasekhar *et al.*, 2021a) can track the ground truth for valence and arousal. Yet, when a fully frontal face is not available (due to pose variations), the predictions of the proposed JCA model closely follow the ground truth more closely than that of (Rajasekhar *et al.*, 2021a), especially for valence.

4.6 Conclusion

In this paper, JCA A-V fusion model is explored for video-based dimensional ER. Contrary to the prior approaches, we leverage the intra- and inter-modal relationships across the A and V features in a unified framework. Specifically, the complementary relationship between A and V features is efficiently captured based on the correlation between the joint A-V feature representations and individual A and V features while retaining the intra-modal relationships. By jointly modeling the inter and inter-modal relationships, features of each modality attend to the other modality as well as itself, resulting in robust A and V feature representations. With the proposed model, A and V backbones are first trained individually for facial (V) and vocal (A) modalities. Then, an attention mechanism based on the correlation between joint and individual features is applied to obtain the attended A and V features. Finally, the attention-weighted features are concatenated, and fed to linear connected layers to predict valence and arousal values. The proposed A-V fusion model is validated experimentally on the challenging RECOLA and

Affwild2 video datasets, using different A and V backbones, and different proportions of missing A segments during the testing mode. Results show that the proposed model is a cost-effective approach that can outperform the state-of-the-art. It encodes inter-modal relationships, while sustaining a high level of performance, even when A segments are noisy and absent. Although the JCA AV fusion model has been proposed for dimensional emotion recognition, it can also be explored for other applications pertinent to audio-visual fusion such as identity verification, event localization, etc.

CONCLUSION AND RECOMMENDATIONS

5.1 Summary of Contributions

Emotions play an important part in human communication. Human emotions are often conveyed through multiple modalities such as audio, visual, text, etc. So it is obvious to exploit the diverse and complementary information available in multiple modalities to build a robust system for emotion recognition. Although emotion recognition has been widely explored for many decades in the field of affective computing, there are still many open challenges associated with developing a robust emotion recognition system in real-world scenarios. This thesis aims to design a robust emotion recognition system by addressing the challenges pertinent to weakly annotated videos for pain intensity estimation and A-V fusion for dimensional emotion recognition and validated on pain and fatigue datasets.

Chapter 2 presented a detailed review of weakly supervised learning models for facial behavior analysis. The existing approaches have been rigorously analyzed for weakly supervised learning models of facial behavior analysis and provided a taxonomy of these approaches along with their insights and limitations. In addition to that, the challenges associated with developing weakly supervised learning models have been discussed for facial behavior analysis along with potential research directions.

Chapter 3 introduced a novel framework of weakly supervised DA for pain intensity estimation using weakly labeled videos by exploiting the DL models. The proposed model enforces ordinal relationships among the intensity levels assigned to target sequences and associates multiple relevant frames to sequence-level labels (instead of a single frame). Specifically, the proposed model learns discriminant and domain-invariant feature representations by integrating MIL with deep adversarial DA, where soft Gaussian labels are used to efficiently represent the weak ordinal sequence-level labels from the target domain. Experimental results on pain and fatigue

datasets indicate that our proposed approach can significantly improve performance over the state-of-the-art models, allowing us to achieve a greater pain or fatigue localization accuracy.

Chapter 4 presented a novel attention model for effective A-V in dimensional emotion recognition. The proposed model investigated the prospect of leveraging the complementary relationship across the A and V modalities to predict the individual's emotional states in valence-arousal space. In particular, the proposed model leverages the inter-modal relationships while still retaining the intra-modal relationships by computing the cross-attention weights based on the correlation between the joint feature representation and that of the individual modalities. By deploying the joint A-V feature representation into the cross-attention module, it helps to simultaneously leverage both the intra and inter-modal relationships, thereby significantly improving the performance of the system. The proposed model has been evaluated for dimensional emotion recognition and detection of fatigue levels. Results indicate that our joint cross-attentional A-V fusion model provides a cost-effective solution that can outperform state-of-the-art approaches, even when the modalities are noisy or absent.

5.2 Recommendations

Based on the systematic study of ER pertinent to WSL and A-V fusion, the following research directions are found to be worth exploring in the future to further improve the performance of the system.

- **Modeling the relationship between the video clip and its corresponding high-level annotation.** Since the high-level annotation in MIL for regression is associated with one of the frames in the video clip, estimating the frame in the video clip that associates with the high-level annotation helps us to effectively capture the relationship between the video clip (bag) and the corresponding annotation.

- **Exploring the prospect of leveraging weakly supervised domain adaptation from weakly labeled source domain to unlabeled target domain.** Investigating the prospect of leveraging the proposed framework of Weakly supervised domain adaptation (WSDA) to an unlabeled target domain or weakly labeled source domain to an unlabeled target domain can minimize the burden of the complex process of obtaining the annotations.
- **Explore the framework of WSDA for other tasks such as classification or continuous regression or even for other applications**. The proposed framework of WSDA was found to promising in leveraging the source domain to deal with the challenges of weak annotations in the target domain. So, the proposed framework can be explored further to extend to other classification tasks or even for other applications.
- **Exploring gating mechanism to leverage the adaptive fusion of A and V modalities.** Fusion of A and V modalities may not always help in improving the performance of the system as the A and V modalities may contradict each other in some of the video clips. So exploring techniques like gating mechanisms can be worth exploring to adaptively fusion A and V modalities based on their semantic relationship.
- **To further investigate the challenges associated with missing modalities.** In this thesis, the performance of the proposed attention model is evaluated for A-V fusion for missing A modality. However, it will be worth exploring to further investigate the challenges of missing modalities to further improve the performance of the system.

APPENDIX I

RECURSIVE JOINT ATTENTION FOR AUDIO-VISUAL FUSION IN REGRESSION-BASED EMOTION RECOGNITION

Gnana Praveen Rajasekhar[a] , Eric Granger[a] , Patrick Cardinal[b]

[a] Department of Systems Enginering, École de technologie supérieure,
[b] Department of Software and IT Engineering, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Paper published in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023

Abstract

In video-based emotion recognition (ER), it is important to effectively leverage the complementary relationship among audio (A) and visual (V) modalities, while retaining the intra-modal characteristics of individual modalities. In this paper, we present a recursive joint attention model along with long short-term memory (LSTM) modules for fusion of vocal and facial expressions in regression-based ER. Specifically, we investigated the possibility of exploiting the complementary nature of A and V modalities using joint cross attention model in a recursive fashion and LSTMs to capture the intra-modal temporal dependencies within the same modalities as well as among the A-V feature representations. By integrating LSTMs with recursive joint cross attention, our model can efficiently leverage both intra- and inter-modal relationships for fusion of A and V modalities. The results of extensive experiments performed on the challenging Affwild2 and Fatigue (private) datasets indicate that the proposed A-V fusion model can significantly outperform state-of-the-art-methods.

1. Introduction

Automatic emotion recognition (ER) is a challenging problem due to the complex and extremely diverse nature of expressions across individuals and cultures. In most of the real-world applications, emotions are exhibited over a wide range of emotional states besides the six

basic categorical expressions - anger, disgust, fear, happiness, sad, and surprise (Ekman, 1992). For instance, emotional states can be expressed as intensities of fatigue, stress, and pain over discrete levels. Similarly, the wide range of continuous emotional states is often formulated as dimensional ER, where diverse and complex human emotions are represented along the dimensions of valence and arousal. Valence denotes the range of continuous emotional states pertinent to pleasantness, spanning from being very sad (negative) to very happy (positive). Similarly, arousal spans the range of emotional states related to intensity, from being very passive (sleepiness) to extremely active (high excitement). In this paper, we have focused on developing a robust model for regression-based ER in valence-arousal space, as well as for fatigue.

A and V modalities often carry complementary relationships among themselves, which is crucial to be exploited to build an efficient A-V fusion system for regression-based ER. In addition to the inter-modal relationships across A and V modalities, temporal dynamics in videos carry significant information pertinent to the evolution of facial and vocal expressions over time. Therefore, effectively leveraging both the inter-modal association across the A and V modalities and temporal dynamics (intra-modal) within A and V modalities plays a major role in building a robust A-V recognition system. In this paper, we have investigated the prospect of leveraging these inter- and intra-modal characteristics of A and V modalities in a unified framework. In most of the existing approaches for regression-based ER, LSTMs have been used to model the intra-modal temporal dynamics in videos (Schoneveld *et al.*, 2021; Kuhnke *et al.*, 2020) due to their efficiency in capturing the long-term temporal dynamics (Karas *et al.*, 2022). On the other hand, cross-attention models (Rajasekhar *et al.*, 2021a) have been explored to model the inter-modal characteristics of A and V modalities for dimensional ER.

In this work, we have proposed a unified framework for A-V fusion, which effectively leverages both the intra- and inter-modal information in videos using LSTMs and joint cross attention respectively. To further improve the A-V feature representations of the joint cross-attention model, we have also explored the recursive attention mechanism. Training the joint cross-attention model recursively allows refining the A and V feature representations, thereby improving the system performance. The main contributions of the paper are as follows. (1) A recursive joint

cross-attentional model for A-V fusion is introduced to effectively exploit the complementary relationship across modalities while deploying a recursive mechanism to further refine the A-V feature representations. (2) LSTMs are further integrated to effectively capture the temporal dynamics within the individual modalities, as well as within the A-V feature representations. (3) An extensive set of experiments are conducted on the challenging Affwild2 and Fatigue (private) datasets, showing that the proposed A-V fusion model outperforms the related state-of-the-art models for regression-based ER.

2. Related Work

An early DL approach for A-V fusion-based dimensional ER was proposed by (Tzirakis *et al.*, 2017), where the deep features (obtained with Resnet-50 for V and 1D CNN for A) are concatenated and fed to an LSTM. Recently, (Karas *et al.*, 2022) investigated the effectiveness of attention models and compared them with recurrent networks. They have shown that LSTMs are quite efficient in capturing the temporal dependencies when compared to attention models for dimensional ER. (Kuhnke *et al.*, 2020) proposed a two-stream A-V network, where deep models are used to extract A and V features, and further concatenated for dimensional ER. Most of these approaches fail to effectively capture the intermodal semantics across A and V modalities. In (Tzirakis *et al.*, 2021) and (Parthasarathy & Sundaram, 2021), authors focused on cross-modal attention using transformers to exploit the inter-modal relationships of A and V modalities for dimensional ER. (Rajasekhar *et al.*, 2021a) explored cross-attention models to leverage the inter-modal characteristics based on cross-correlation across the A and V features. They improved their approach by introducing joint feature representation into the cross-attention model to retain the intra-modal characteristics (Praveen *et al.*, 2023a). In most of these approaches, they cannot effectively leverage intra-modal relationships. (Chen & Jin, 2016) modeled A and V features using LSTMs, and the unimodal predictions are combined using attention weights from conditional attention based on LSTMs. (Darshana, F, Simon & Clinton, 2019) also explored LSTMs for V features and used DNN-based attention on the concatenated features of A and V modalities for the final output predictions. (Beard *et al.*, 2018) proposed a recursive recurrent attention model, where LSTMs are augmented using an additional shared memory

state to capture the multi-modal relationships recursively. In contrast with these approaches, we focus on modeling the A-V relationships by allowing the A and V features to interact and measure the semantic relevance across and within the modalities recursively before feature concatenation. LSTMs are employed for temporal modeling of both uni-modal and multimodal features to further enhance the proposed framework. Therefore, our proposed model effectively leverages the intra- and complementary inter-modal relationships, resulting in a higher level of performance.

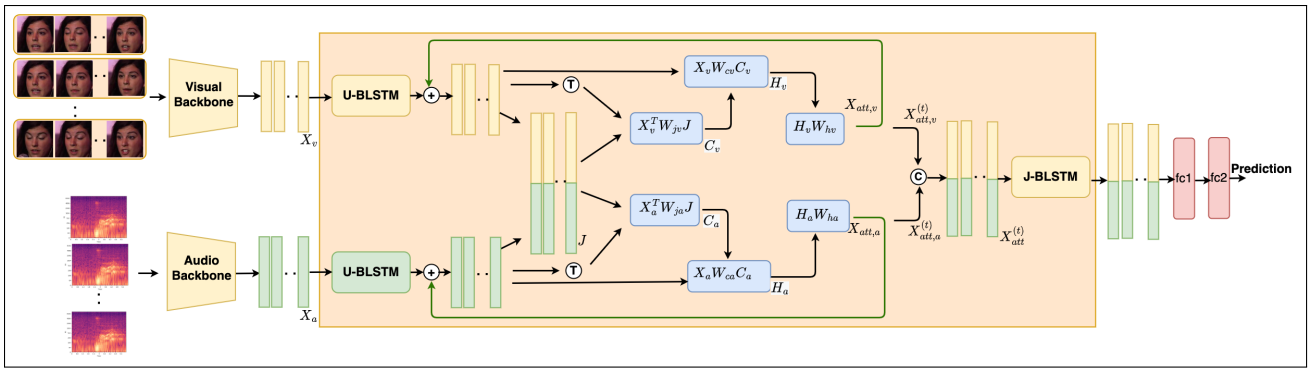

Figure-A I-1 Block diagram of the proposed recursive joint attention model with BLSTMs
Taken from Praveen *et al.* (2023b)

3. Proposed Approach

The block diagram of the proposed approach is shown in Figure I-1.

A) Problem Formulation: Given an input video sub-sequence S, L non-overlapping video clips are uniformly sampled and deep feature vectors $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$ are extracted for the individual A and V modalities respectively from pre-trained networks. Let $\boldsymbol{X_a} = \{\boldsymbol{x_a^1}, \boldsymbol{x_a^2}, ..., \boldsymbol{x_a^L}\} \in \mathbb{R}^{d_a \times L}$ and $\boldsymbol{X_v} = \{\boldsymbol{x_v^1}, \boldsymbol{x_v^2}, ..., \boldsymbol{x_v^L}\} \in \mathbb{R}^{d_v \times L}$ where d_a and d_v represent the dimensions of the A and V feature representations, respectively, and $\boldsymbol{x_a^l}$ and $\boldsymbol{x_v^l}$ denotes the A and V feature vectors of the video clips, respectively, for $l = 1, 2, ..., L$ clips. The objective of the problem is to estimate the regression model $F : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ from the training data $\boldsymbol{X}$, where $\boldsymbol{X}$ denotes the set of A and V feature vectors of the input video clips and $\boldsymbol{Y}$ represents the regression labels of the corresponding video clips.

B) Audio and Visual Networks: Spectrograms has been found to be promising with various 2D-CNNs (Resnet-18 (He *et al.*, 2016)) for ER (Slimi *et al.*, 2020; Albanie *et al.*, 2018). Therefore, we have explored spectrograms in the proposed framework. In order to effectively leverage the temporal dynamics within the A modality, we have also explored LSTMs across the temporal segments of the A sequences. Finally, the A feature vectors of L video clips are shown as $\boldsymbol{X_a} = (\boldsymbol{x}_{\mathbf{a}}^1, \boldsymbol{x}_{\mathbf{a}}^2, ..., \boldsymbol{x}_{\mathbf{a}}^L) \in \mathbb{R}^{d_a \times L}$.

Facial expressions exhibit significant information pertinent to both visual appearance and temporal dynamics in videos. LSTMs are found to be efficient in capturing the long-term temporal dynamics while 3DCNNs are effective in capturing the short-term temporal dynamics (Fan *et al.*, 2016). Therefore, we have used LSTMs with 3D CNNs (R3D (Tran *et al.*, 2018)) to obtain the V features for the fusion model. In most of the existing approaches, the output of the last convolution layer is 512 x 7 x 7, which is further passed through a pooling operation to reduce the spatial dimensions to 1 ($7 \rightarrow 1$). This reduction in spatial dimension was found to leave out useful information as the stride is big. Therefore, inspired by the idea of (Duan *et al.*, 2021), we use the A feature representation to smoothly reduce the spatial dimensions of raw V features for each video clip similar to that of (Duan *et al.*, 2021). Finally, we obtain a matrix of V feature vectors of the video clips as $\boldsymbol{X_v} = (\boldsymbol{x}_{\mathbf{v}}^1, \boldsymbol{x}_{\mathbf{v}}^2, ..., \boldsymbol{x}_{\mathbf{v}}^L) \in \mathbb{R}^{d_v \times L}$.

C) Recursive Joint Attention Model: Given the A and V feature representations $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$, the joint feature representation is obtained by concatenating the A and V feature vectors $\boldsymbol{J} = [\boldsymbol{X}_{\mathbf{a}}; \boldsymbol{X}_{\mathbf{v}}] \in \mathbb{R}^{d \times L}$, where $d = d_a + d_v$ denotes the feature dimension of concatenated features. The concatenated A-V feature representations ($\boldsymbol{J}$) of the given video sub-sequence ($\boldsymbol{S}$) are now used to attend to unimodal feature representations $\boldsymbol{X}_{\mathbf{a}}$ and $\boldsymbol{X}_{\mathbf{v}}$. The joint correlation matrix $\boldsymbol{C}_{\mathbf{a}}$ across the A features $\boldsymbol{X}_{\mathbf{a}}$, and the combined A-V features $\boldsymbol{J}$ are given by:

$$\boldsymbol{C}_{\mathbf{a}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{a}}^T \boldsymbol{W}_{\mathbf{ja}} \boldsymbol{J}}{\sqrt{d}}\right) \tag{A I-1}$$

where $\boldsymbol{W}_{\mathbf{ja}} \in \mathbb{R}^{L \times L}$ represents learnable weight matrix across the A and combined A-V features, and T denotes transpose operation. Similarly, the joint correlation matrix for V features is given

by:

$$\boldsymbol{C}_{\mathbf{v}} = \tanh\left(\frac{\boldsymbol{X}_{\mathbf{v}}^{T}\boldsymbol{W}_{\mathbf{jv}}\boldsymbol{J}}{\sqrt{d}}\right) \tag{A I-2}$$

The joint correlation matrices capture the semantic relevance across the A and V modalities as well as within the same modalities among consecutive video clips, which helps in effectively leveraging intra- and inter-modal relationships. After computing the joint correlation matrices, the attention weights of the A and V modalities are estimated. For the A modality, the joint correlation matrix $\boldsymbol{C}_{\mathbf{a}}$ and the corresponding A features $\boldsymbol{X}_{\mathbf{a}}$ are combined using the learnable weight matrices $\boldsymbol{W}_{\mathbf{ca}}$ to compute the attention weights of A modality, which is given by $\boldsymbol{H}_{\mathbf{a}} = ReLu(\boldsymbol{X}_{\mathbf{a}}\boldsymbol{W}_{\mathbf{ca}}\boldsymbol{C}_{\mathbf{a}})$ where $\boldsymbol{W}_{\mathbf{ca}} \in \mathbb{R}^{d_a \times d_a}$ and $\boldsymbol{H}_{\mathbf{a}}$ represents the attention maps of the A modality. Similarly, the attention maps ($\boldsymbol{H}_{\mathbf{v}}$) of V modality are obtained as $\boldsymbol{H}_{\mathbf{v}} = ReLU(\boldsymbol{X}_{\mathbf{v}}\boldsymbol{W}_{\mathbf{cv}}\boldsymbol{C}_{\mathbf{v}})$ where $\boldsymbol{W}_{\mathbf{cv}} \in \mathbb{R}^{d_v \times d_v}$. Then, the attention maps are used to compute the attended features of A and V modalities as:

$$\boldsymbol{X}_{\mathbf{att,a}} = \boldsymbol{H}_{\mathbf{a}}\boldsymbol{W}_{\mathbf{ha}} + \boldsymbol{X}_{\mathbf{a}} \tag{A I-3}$$

$$\boldsymbol{X}_{\mathbf{att,v}} = \boldsymbol{H}_{\mathbf{v}}\boldsymbol{W}_{\mathbf{hv}} + \boldsymbol{X}_{\mathbf{v}} \tag{A I-4}$$

where $\boldsymbol{W}_{\mathbf{ha}} \in \mathbb{R}^{d \times d_a}$ and $\boldsymbol{W}_{\mathbf{hv}} \in \mathbb{R}^{d \times d_v}$ denote the learnable weight matrices for A and V respectively. After obtaining the attended features they are fed again to the joint cross-attentional model to compute the new A and V feature representations as:

$$\boldsymbol{X}_{\mathbf{att,a}}^{(t)} = \boldsymbol{H}_{\mathbf{a}}^{(t)}\boldsymbol{W}_{\mathbf{ha}}^{(t)} + \boldsymbol{X}_{\mathbf{a}}^{(t-1)} \tag{A I-5}$$

$$\boldsymbol{X}_{\mathbf{att,v}}^{(t)} = \boldsymbol{H}_{\mathbf{v}}^{(t)}\boldsymbol{W}_{\mathbf{hv}}^{(t)} + \boldsymbol{X}_{\mathbf{v}}^{(t-1)} \tag{A I-6}$$

where $\boldsymbol{W}_{\mathbf{ha}}^{(t)} \in \mathbb{R}^{d \times d_a}$ and $\boldsymbol{W}_{\mathbf{hv}}^{(t)} \in \mathbb{R}^{d \times d_v}$ denote the learnable weight matrices of t^{th} iteration for A and V respectively. Finally, the attended A and V features after t iterations are further concatenated and fed to BLSTM to obtain the temporal dependencies within the refined A-V feature representations, which is fed to fully connected layers for final prediction.

4. Results and Discussion

A) Dataset: Affwild2 is among the largest dataset in affective computing, consisting of 564 videos collected from YouTube, all captured in-the-wild (Kollias *et al.*, 2019). The annotations are provided by four experts using a joystick and the final annotations are obtained as the average of the four raters. In total, there are $2,816,832$ frames with 455 subjects, out of which 277 are male and 178 female. The annotations for valence and arousal are provided continuously in the range of $[-1, 1]$. The dataset is split into training, validation, and test sets. The partitioning is done in a subject-independent manner so that every subject's data will present in only one subset. The partitioning produces 341, 71, and 152 videos for the training, validation, and test sets respectively.

B) Ablation Study: Table I-1 presents the results of the experiments conducted on the validation set for the ablation study. The performance of the approach is evaluated using Concordance Correlation Coefficient (CCC). In this section, we have analyzed the contribution of BLSTMs in the proposed model, where we have performed experiments with and without BLSTMs. Firstly, we have conducted experiments without using BLSTM for both the individual A and V representations as well as the A-V feature representations. Then, we included BLSTMs only for the individual A and V modalities before feeding to the joint attention fusion model i.e., only Unimodal-BLSTMs (U-BLSTMs). By including U-BLSTMs to capture the temporal dependencies within the individual modalities, we can observe the improvement in the performance of the system. Therefore, BLSTMs are found to be promising in capturing the intra-modal temporal dynamics better than that of correlation-based intra-modeling in the joint attention model. After that, we have also included joint BLSTM (J-BLSTM) in order to capture the temporal dynamics across the joint A-V feature representations, which further improved the performance of the system. It is worth mentioning that in all the above experiments, we have not performed recursive attention. The Fatigue dataset is obtained from 18 participants in a Rehabilitation center, suffering from degenerative diseases inducing fatigue. A total of 27 video sessions are captured with a duration of 40 - 45 minutes and labeled at sequence level on a scale of 0 to 10 for every 10 to 15 minutes. We have considered 80% of data as training data (50,845 samples) and 20% as validation data (21,792 samples).

Table-A I-1 Performance of our approach with components of BLSTM and recursive attention on Affwild2 dataset.

Method	**Valence**	**Arousal**
JA Fusion w/o recursion		
Fusion w/o U-BLSTM	0.670	0.590
Fusion w/o J-BLSTM	0.691	0.646
Fusion w/ U-BLSTM and J-BLSTM	0.715	0.688
JA Fusion w/ recursion		
JA Fusion w/o BLSTMs $t = 2$	0.703	0.623
JA Fusion with BLSTMs $t = 2$	**0.721**	**0.694**
JA Fusion with BLSTMs $t = 3$	0.706	0.652
JA Fusion with BLSTMs $t = 4$	0.685	0.601

In addition to the impact of U-BLSTM and J-BLSTMs, we have also conducted a few more experiments to investigate the impact of the recursive behavior of the joint attention model. First, we did recursion without LSTMs and found some improvement due to recursion. Then we included LSTMs and conducted several experiments by varying the number of recursions (iterations) in the fusion model. As we increase the number of recursive times, the model performance increases and starts to decrease after a certain recursion number. A similar trend of the model performance is also observed in the test set. Therefore, this can be attributed to the fact that recursion also works as a regularizer which improves the generalization ability of the model. In our experiments, we found that $t = 2$ gives the best performance i.e, we have achieved the best performance of our model with two recursive iterations.

C) Comparison to state-of-the-art:

Table I-2 shows our comparative results against relevant state-of-the-art A-V fusion models on the Affwild2 dataset. Recently, Affwild2 dataset has been widely used for Affective Behavior Analysis in-the-wild (ABAW) challenges (Kollias *et al.*, 2020; Kollias & Zafeiriou, 2021a). Therefore, we compare our approach with that of the relevant approaches in the ABAW challenges. (Kuhnke *et al.*, 2020) used a simple feature concatenation using Resnet-18 for A and R3D for V modalities and showed better performance for arousal than valence. (Zhang *et al.*, 2021b) proposed leader-follower attention model for fusion and improved the performance (arousal) of the model proposed by (Kuhnke *et al.*, 2020). (Rajasekhar *et al.*, 2021a) explored the

Table-A I-2 CCC performance of the proposed and state-of-the-art methods for A-V fusion on the Affwild2 dataset.

Method	**Type of Fusion**	**Valence**	**Arousal**
Validation Set			
(Kuhnke *et al.*, 2020)	Feature Concatenation	0.493	0.613
(Zhang *et al.*, 2021b)	Leader Follower Attention	0.469	0.649
(Rajasekhar *et al.*, 2021a)	Cross Attention	0.541	0.517
(Praveen *et al.*, 2023a)	Joint Cross Attention	0.657	0.580
Ours	LSTM + Transformers	0.628	0.654
Ours	Recursive JA + BLSTM	**0.721**	**0.694**
Test Set			
(Meng *et al.*, 2022)	LSTM + Transformers	0.606	0.596
(Karas *et al.*, 2022)	LSTM + Transformers	0.418	0.407
(Praveen *et al.*, 2023a)	Joint Cross Attention	0.451	0.389
Ours	Recursive JA + BLSTM	**0.467**	**0.405**

cross-attention model by leveraging only the inter-modal relationships of A and V modalities and showed improvement for valence but not so efficient for arousal. (Praveen *et al.*, 2023a) further improved the performance of the model by introducing joint feature representation to the cross-attention model. The proposed model performs even better than that of vanilla JCA (Praveen *et al.*, 2023a) by introducing LSTMs as well as a recursive attention mechanism. For test set results, the winner of the latest ABAW challenge (Meng *et al.*, 2022) has shown improvement using A-V fusion, however using three external datasets and multiple backbones. We have also compared the performance of our A and V backbones with the ensembling of LSTMs and transformers (Meng *et al.*, 2022) on the validation set. (Karas *et al.*, 2022) used LSTMs to capture intra-modal dependencies and explored transformers for cross-modal attention, however, they fail to effectively capture the inter-modal relationships across the consecutive video clips. (Praveen *et al.*, 2023a) further improved the performance (valence) using joint cross-attention. The proposed model outperforms both (Karas *et al.*, 2022) and (Praveen *et al.*, 2023a).

Table I-3 shows the performance of the proposed approach on the Fatigue dataset. We have shown the performance of individual modalities along with feature concatenation and cross-attention

Rajasekhar *et al.* (2021a). The proposed approach outperforms cross-attention Rajasekhar *et al.* (2021a) and baseline feature concatenation.

Table-A I-3 CCC performance on the Fatigue dataset

Method	**Fatigue Level**
Audio only (2D-CNN: Resnet-18)	0.312
Visual only (3DCNN: R3D)	0.415
Feature Concatenation	0.378
Cross Attention (Rajasekhar *et al.*, 2021a)	0.421
Recursive JA + BLSTM (Ours)	**0.447**

5. Conclusion

This paper introduces a recursive joint attention model along with BLSTMs that allows effective spatiotemporal A-V fusion for regression-based ER. In particular, the joint attention model is trained in a recursive fashion, allowing for the refinement of A-V features. We further investigated the impact of BLSTMs for capturing the intra-modal temporal dynamics of individual A and V modalities, as well as A-V features for regression-based ER. By effectively capturing the intra-modal relationships using BLSTMs, and inter-modal relationships using recursive joint attention, the proposed approach is able to outperform the related state-of-the-art approaches.

APPENDIX II

COMPLEXITY OF CODE

We have also provided the complexity of the code for the proposed joint cross-attentional (JCA) A-V fusion model for dimensional emotion recognition. Initially we compute the joint feature representation ($\boldsymbol{J}$) of the A-V features. Since the dimension of the joint representation is the summation of the dimensions of the individual A and V feature vectors, it consumes more number of parameters. After that, we compute the parameters required by the joint cross correlation matrices $\boldsymbol{C}_{\mathbf{a}}$ and $\boldsymbol{C}_{\mathbf{v}}$. The dimensions of the learnable matrices to compute the joint cross correlation matrices is LxL, where L denotes the sequence length. So it consumes less number of parameters for the learnable weights related to joint cross correlation matrices. The learnable weights required to estimate the joint cross attention weights depends of the dimension of the individual features representations. Finally, learnable weights are used to compute the final attended features to reduce the dimension of the attention weights to that of the dimension of the original feature vectors. The number of flops required for each module is proportional to the number of parameters required in each module. The total number of parameters in the proposed JCA fusion model is 2.6M parameters and the number of flops are 2701M. The number of parameters required in each module of the proposed JCA fusion modal along with the number of flops are shown in Table II-1.

Table-A II-1 Number of Parameters and flops of the proposed JCA A-V fusion model

Module Name	**Number of Parameters**	**Number of Flops**
Joint Representation ($\boldsymbol{J}$)	1049600	1073741824
Joint Correlation Matrix across A ($\boldsymbol{C}_{\mathbf{v}}$)	272	8388608
Joint Correlation Matrix across V ($\boldsymbol{C}_{\mathbf{a}}$)	272	8388608
Attention Weights of V ($\boldsymbol{X}_{\mathbf{att,v}}$)	262656	268435456
Attention Weights of A ($\boldsymbol{X}_{\mathbf{att,a}}$)	262656	268435456
Weights of Attended features ($\boldsymbol{X}_{\mathbf{att,v}}$)	524800	536870912
Weights of Attended features ($\boldsymbol{X}_{\mathbf{att,a}}$)	524800	536870912
	Total = 2625056	Total = 2701131776

APPENDIX III

PUBLICATIONS DURING PH.D. STUDY

Journal Articles

- **Gnana Praveen Rajasekhar**, Eric Granger, Patrick Cardinal. "Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos", Image and Vision Computing, Volume 110, 2021.
- **Gnana Praveen Rajasekhar**, Patrick Cardinal, Eric Granger. "Audio-Visual Fusion for Emotion Recognition in the Valence-Arousal Space Using Joint Cross-Attention", IEEE Transactions on Biometrics, Behavior, and Identity Science, 2023.
- **Gnana Praveen Rajasekhar**, Eric Granger, Patrick Cardinal, . "Weakly Supervised Learning for Facial Behavior Analysis: A Review", (Submitted to IEEE Transactions on Affective Computing, 2022)

Conference Articles

- **R. Gnana Praveen**, E. Granger and P. Cardinal, "Deep Weakly Supervised Domain Adaptation for Pain Localization in Videos," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 473-480.
- **R. G. Praveen**, E. Granger and P. Cardinal, "Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition," 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 2021, pp. 1-8.
- **R Gnana Praveen** and de Melo, Wheidima Carneiro and Ullah, Nasib and Aslam, Haseeb and Zeeshan, Osama, and Denorme, Théo and Pedersoli, Marco, and Koerich, Alessandro L. and Bacon, Simon and Cardinal, Patrick and Granger, Eric, "A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2486-2495, 2022.
- Madhu Kiran, **R Gnana Praveen**, Le Thanh Nguyen-Meidine, Soufiane Belharbi, Louis-Antoine Blais-Morin, Eric Granger, "Holistic Guidance for Occluded Person Re-Identification" British Machine and Vision Conference (BMVC), 2021.

- **R. Gnana Praveen**, E. Granger and P. Cardinal, "Recursive Joint Attention for audio-visual fusion in regression based emotion recognition," 48th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

BIBLIOGRAPHY

Abdelwahab, M. & Busso, C. (2017). Ensemble feature selection for domain adaptation in speech emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5000-5004.

Abdulrahman, M. & Eleyan, A. (2015). Facial expression recognition using Support Vector Machines. *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pp. 276-279.

Adão Martins, N. R., Annaheim, S., Spengler, C. M. & Rossi, R. M. (2021). Fatigue Monitoring Through Wearables: A State-of-the-Art Review. *Frontiers in Physiology*, 12.

Albanie, S., Nagrani, A., Vedaldi, A. & Zisserman, A. (2018). Emotion Recognition in Speech Using Cross-Modal Transfer in the Wild. *26th ACM Multimedia*, pp. 292–301.

Almaev, T. R. & Valstar, M. F. (2013). Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition. *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356-361.

Anagnostopoulos, C., Iliou, T. & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artif Intell Rev*, 43(6), 155-177.

Andrews, S., Tsochantaridis, I. & Hofmann, T. (2002). Support Vector Machines for Multiple-Instance Learning. *Advances in Neural Information Processing Systems (NIPS)*, 15.

Aung, M. S. H., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., Valstar, M., Meng, H., Kemp, A., Shafizadeh, M., Elkins, A. C., Kanakam, N., de Rothschild, A., Tyler, N., Watson, P. J., d. C. Williams, A. C., Pantic, M. & Bianchi-Berthouze, N. (2016). The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset. *IEEE Tran. on Affective Computing*, 7(4), 435-451.

Ayral, T., Pedersoli, M., Bacon, S. & Granger, E. (2021). Temporal Stochastic Softmax for 3D CNNs: An Application in Facial Expression Recognition. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3028-3037.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C. & Baskurt, A. (2012). Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification. *British Machine Vision Conference (BMVC)*, pp. 124.1-124.12.

Badshah, A. M., Ahmad, J., Rahim, N. & Baik, S. W. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. *International Conference on Platform Technology and Service (PlatCon)*, pp. 1-5.

Baltrušaitis, T., Ahuja, C. & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. PAMI*, 41(2), 423-443.

Bargal, S. A., Barsoum, E., Ferrer, C. C. & Zhang, C. (2016). Emotion Recognition in the Wild from Videos Using Images. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 433–436.

Barsoum, E., Zhang, C., Ferrer, C. C. & Zhang, Z. (2016). Training Deep Networks for Facial Expression Recognition with Crowd-sourced Label Distribution. *Proc. of 18th ACM ICMI*, pp. 279–283.

Bayerl, S. P., Wagner, D., Baumann, I., Bocklet, T. & Riedhammer, K. (2023). Detecting Vocal Fatigue with Neural Embeddings. *Journal of Voice*.

Beard, R., Das, R., Ng, R. W. M., Gopalakrishnan, P. G. K., Eerens, L., Swietojanski, P. & Miksik, O. (2018). Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition. *CoNLL*.

Bellantonio, M., Haque, M. A., Rodriguez, P., Nasrollahi, K., Telve, T., Escalera, S., Gonzalez, J., Moeslund, T. B., Rasti, P. & Anbarjafari, G. (2017). Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images. *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*.

Ben-Yacoub, S., Luttin, J., Jonsson, K., Matas, J. & Kittler, J. (1999). Audio-visual person verification. *CVPR*, pp. 580-585.

Benitez-Quiroz, C. F., Srinivasan, R. & Martinez, A. M. (2016). EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. *CVPR*, pp. 5562-5570.

Benitez-Quiroz, C. F., Wang, Y. & Martinez, A. M. (2017). Recognition of Action Units in the Wild with Deep Nets and a New Global-Local Loss. *ICCV*, pp. 3990-3999.

Benitez-Quiroz, C. F., Srinivasan, R., Feng, Q., Wang, Y. & Martínez, A. M. (2017). EmotioNet Challenge: Recognition of facial expressions of emotion in the wild. *arXiv*.

Bouzakraoui, M. S., Sadiq, A. & Alaoui, A. Y. (2020). Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques. *Advances in Science, Technology and Engineering Systems*, 5(4), 594–599.

Bozorgtabar, B., Mahapatra, D. & Thiran, J.-P. (2020). ExprADA: Adversarial domain adaptation for facial expression analysis. *Pattern Recognition*, 100, 107-111.

Brabham, D. C. (2008). Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence*, 14(1), 75-90.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32.

Breuer, R. & Kimmel, R. (2017). A Deep Learning Perspective on the Origin of Facial Expressions. *arXiv*, abs/1705.01842.

Brousmiche, M., Rouat, J. & Dupont, S. (2019). Audio-Visual Fusion And Conditioning With Neural Networks For Event Recognition. *IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S. & Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42, 335.

Calvo, R., D'Mello, S., Gratch, J. & Kappas, A. (2015). *The Oxford Handbook of Affective Computing*. Oxford University Press.

Carbonneau, M. A., Cheplygina, V., Granger, E. & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77, 329 - 353.

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A. & Karpouzis, K. (2006). Modeling Naturalistic Affective States via Facial and Vocal Expressions Recognition. *Proceedings of the 8th International Conference on Multimodal Interfaces*, pp. 146–154.

Carneiro de Melo, W., Granger, E. & Lopez, M. B. (2020). Encoding Temporal Information For Automatic Depression Recognition From Facial Analysis. *ICASSP*.

Carreira, J. & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CVPR*, pp. 4724-4733.

Chamoli, A., Semwal, A. & Saikia, N. (2017). Detection of emotion in analysis of speech using linear predictive coding techniques (L.P.C). *International Conference on Inventive Systems and Control (ICISC)*, pp. 1-4.

Chapelle, O., Schlkopf, B. & Zien, A. (2010). *Semi-Supervised Learning* (ed. 1st). The MIT Press.

Chaudhari, A., Bhatt, C., Krishna, A. & Mazzeo, P. L. (2022). ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation*, 5(4).

Chen, C. & Jack, R. E. (2017). Discovering cultural differences (and similarities) in facial expressions of emotion. *Current Opinion in Psychology*, 17, 61-66.

Chen, J., Guo, C., Xu, R., Zhang, K., Yang, Z. & Liu, H. (2022a). Toward Children's Empathy Ability Analysis: Joint Facial Expression Recognition and Intensity Estimation Using Label Distribution Learning. *IEEE Transactions on Industrial Informatics*, 18(1), 16-25.

Chen, J., Chen, Z., Chi, Z. & Fu, H. (2014). Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning. *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 508–513.

Chen, M., He, X., Yang, J. & Zhang, H. (2018). 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Processing Letters*, 25(10), 1440-1444.

Chen, P., Gao, Y. & Ma, A. J. (2022b). Multi-level Attentive Adversarial Learning with Temporal Dilation for Unsupervised Video Domain Adaptation. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 776-785.

Chen, S. & Jin, Q. (2016). Multi-Modal Conditional Attention Fusion for Dimensional Emotion Prediction. *Proc. of ACM'M*.

Chen, S., Sun, Y., Zhang, H., Liu, Q., Lv, X. & Mei, X. (2022c). Speech Fatigue Detection Based on Deep Learning. 2224(1), 012023.

Chen, Y. & Wang, J. Z. (2004). Image Categorization by Learning and Reasoning with Regions. *J. Mach. Learn. Res.*, 5, 913–939.

Chen, Z., Ansari, R. & Wilkie, D. J. (2022d). Learning Pain from Action Unit Combinations: A Weakly Supervised Approach via Multiple Instance Learning. *IEEE Tran. on Affective Computing*, 13(1), 135-146.

Cheng, Y., Jiang, B. & Jia, K. (2014). A Deep Structure for Facial Expression Recognition under Partial Occlusion. *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 211-214.

Chu, W. & Ghahramani, Z. (2005). Gaussian Processes for Ordinal Regression. *Journal of Machine Learning Research*, 6(35), 1019-1041.

Cohen, I., Sebe, N., Cozman, F. G. & Huang, T. S. (2003). Semi-supervised Learning for Facial Expression Recognition. *Proc. of 5th ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, pp. 17–22.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297.

Darshana, P., F, T., Simon, S. & Clinton, F. (2019). Learning Salient Features for Multimodal Emotion Recognition with Recurrent Neural Networks and Attention Based Fusion. *Proc. of AVSP*.

de Melo, W. C., Granger, E. & Hadid, A. (2019). Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions. *FG*.

de Santana Correia, A. & Colombini, E. L. (2022). Attention, please! A survey of neural attention models in deep learning. *Artificial Intelligence Review*.

Deb, S. & Dandapat, S. (2019a). Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions. *IEEE Transactions on Affective Computing*, 10(3), 360-373.

Deb, S. & Dandapat, S. (2019b). Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification. *IEEE Transactions on Cybernetics*, 49(3), 802-815.

Deng, D., Wu, L. & Shi, B. E. (2021). Iterative Distillation for Better Uncertainty Estimates in Multitask Emotion Recognition. *ICCVW*, pp. 3550-3559.

Dhall, A., Goecke, R., Lucey, S. & Gedeon, T. (2012). Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia*, 19(3), 34-41.

Dhall, A., Kaur, A., Goecke, R. & Gedeon, T. (2018). EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. *Proc. of 20th ACM ICMI*, pp. 653–656.

Ding, H., Zhou, S. K. & Chellappa, R. (2017). FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition. *IEEE FG*, pp. 118-126.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Du, S., Tao, Y. & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15), E1454-E1462.

Duan, B., Tang, H., Wang, W., Zong, Z., Yang, G. & Yan, Y. (2021). Audio-Visual Event Localization via Recursive Fusion by Joint Co-Attention. *WACV*, pp. 4012-4021.

Dwivedi, K., Biswaranjan, K. & Sethi, A. (2014). Drowsy driver detection using representation learning. *2014 IEEE International Advance Computing Conference (IACC)*, pp. 995-999.

Díaz, R. & Marathe, A. (2019). Soft Labels for Ordinal Regression. *CVPR*.

Egede, J., Valstar, M. & Martinez, B. (2017). Fusing Deep Learned and Hand-Crafted Features of Appearance, Shape, and Dynamics for Automatic Pain Estimation. *FG*, pp. 689-696.

Ekman, P. (2002). Facial Action Coding System (FACS). *A Human Face*.

Ekman, P., Friesen, W. & Hager, J. (2002). Facial action coding system: Research Nexus. *Network Research Information*, 1.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), 169-200.

Ekman, P. & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

Ekman, P. & Friesen, W. V. (1976). *Pictures of Facial Affect*. Consulting Psychologists Press.

Fan, Y., Lu, X., Li, D. & Liu, Y. (2016). Video-Based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. *ACM ICMI*, pp. 445–450.

Fang, Y. & Chang, L. (2015). Multi-instance Feature Learning Based on Sparse Representation for Facial Expression Recognition. *MultiMedia Modeling*, pp. 224–233.

Fayek, H. M., Lech, M. & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, 92, 60-68. Advances in Cognitive Engineering Using Neural Networks.

Feng, Z., Shu, K., Charless, C. F., Tao, C. & Baiying, L. (2020). Fine-grained facial expression analysis using dimensional emotion model. *Neurocomputing*, 392, 38 - 49.

Florea, C., Badea, M., Florea, L., Racoviteanu, A. & Vertan, C. (2020). Margin-Mix: Semi-Supervised Learning for Face Expression Recognition. *ECCV*.

Foulds, J. & Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1), 1–25.

Freund, Y. & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, pp. 148–156.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189 – 1232.

Ganin, Y. & Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. *Proc. of the 32nd ICML*, 37, 1180 - 1189.

Gavade, P. A., Bhat, V. & Pujari, J. (2021). Facial Expression Recognition in Videos by learning Spatio-Temporal Features with Deep Neural Networks. *2021 Sixth International Conference on Image Information Processing (ICIIP)*, 6, 359-363. doi: 10.1109/ICIIP53038.2021.9702545.

Gehrig, T. & Ekenel, H. K. (2013). Why is Facial Expression Analysis in the Wild Challenging? *Proc. of EmotiW Challenge and Workshop*, pp. 9–16.

Ghaleb, E., Niehues, J. & Asteriadis, S. (2020). Multimodal Attention-Mechanism For Temporal Emotion Recognition. *ICIP*, pp. 251-255.

Ghazal, M., Abu Haeyeh, Y., Abed, A. & Ghazal, S. (2018). Embedded Fatigue Detection Using Convolutional Neural Networks with Mobile Integration. *6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 129-133.

Ghosh, S., Laksana, E., Morency, L.-P. & Scherer, S. (2016). Representation Learning for Speech Emotion Recognition. *Interspeech*.

Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H., Palm, G. & Schwenker, F. (2011). Multiple Classifier Systems for the Classification of Audio-Visual Emotional States. *Affective Computing and Intelligent Interaction*, pp. 359–368.

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Int. Conf. on Artificial Intelligence and Statistics*, 9, 249–256.

Gnana Praveen, R., Granger, E. & Cardinal, P. (2020). Deep Weakly Supervised Domain Adaptation for Pain Localization in Videos. *FG*.

Gnana Praveen, R., Eric, G. & Patrick, C. (2021). Weakly Supervised Learning for Facial Behavior Analysis: A Review. *arXiv*.

Greeley, H. P., Friets, E., Wilson, J. P., Raghavan, S., Picone, J. & Berg, J. (2006). Detecting Fatigue From Voice Using Speech Recognition. *IEEE International Symposium on Signal Processing and Information Technology*, pp. 567-571.

Green, C. & Guo, K. (2018). Factors contributing to individual differences in facial expression categorisation. *Cognition and Emotion*, 32(1), 37-48.

Gudi, A., Tasli, H. E., den Uyl, T. M. & Maroulis, A. (2015). Deep learning based FACS Action Unit occurrence and intensity estimation. *FG*, 06, 1-5.

Gunes, H. & Schuller, B. (2013). Categorical and Dimensional Affect Analysis in Continuous Input: Current Trends and Future Directions. *Image Vision Comput.*, 31(2), 120–136.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R. R., Cheng, M.-M. & Hu, S.-M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3), 331-368.

Han, J., Zhang, Z., Cummins, N., Ringeval, F. & Schuller, B. (2017). Strength Modelling for Real-World automatic Continuous Affect Recognition from Audiovisual Signals. *Image Vision Comput.*, 65(C), 76–86.

Han, K., Yu, D. & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. *INTERSPEECH*.

Happy, S. L., Dantcheva, A. & Bremond, F. (2019). A Weakly Supervised learning technique for classifying facial expressions. *Pattern Recognition Letters*, 128, 162 - 168.

Hasani, B. & Mahoor, M. H. (2017). Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. *CVPRW*, pp. 2278-2288.

Hassan, T., Seuß, D., Wollenberg, J., Weitz, K., Kunz, M., Lautenbacher, S., Garbas, J. & Schmid, U. (2019). Automatic Detection of Pain from Facial Expressions: A Survey. *IEEE Trans. on PAMI*, 1-1.

Hayat, M., Khan, S. H., Werghi, N. & Goecke, R. (2017). Joint Registration and Representation Learning for Unconstrained Face Identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1551-1560.

He, H. & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Tran. on Knowledge and Data Engineering*, 21(9), 1263-1284.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778.

He, L., Jiang, D., Yang, L., Pei, E., Wu, P. & Sahli, H. (2015). Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. *AVEC*.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *NIPS Deep Learning and Representation Learning Workshop*.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.

Hongyi Zhang,Moustapha Cisse, Y. N. D. & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. *ICLR*.

Hou, R., Chang, H., MA, B., Shan, S. & Chen, X. (2019). Cross Attention Network for Few-shot Classification. *NIPS*.

Hsu, K., Lin, Y. & Chuang, Y. (2014). Augmented Multiple Instance Regression for Inferring Object Contours in Bounding Boxes. *IEEE Trans. on Image Processing*, 23(4), 1722-1736.

Hu, D., Wang, C., Nie, F. & Li, X. (2019). Dense Multimodal Fusion for Hierarchically Joint Representation. *ICASSP*, pp. 3941-3945.

Hu, R., Zhou, S., Tang, Z. R., Chang, S., Huang, Q., Liu, Y., Han, W. & Wu, E. Q. (2021). DMMAN: A two-stage audio–visual fusion framework for sound separation and event localization. *Neural Networks*, 133, 229-239.

Huang, D.-Y., Zhang, Z. & Ge, S. S. (2014). Speaker State Classification Based on Fusion of Asymmetric Simple Partial Least Squares (SIMPLS) and Support Vector Machines. *Comput. Speech Lang.*, 28(2), 392–419.

Huang, M. X., Ngai, G., Hua, K. A., Chan, S. C. F. & Leong, H. V. (2016). Identifying User-Specific Facial Affects from Spontaneous Expressions with Minimal Annotation. *IEEE Tran. on Affective Computing*, 7(4), 360-373.

Huang, X., Deng, W., Shen, H., Zhang, X. & Ye, J. (2020). PropagationNet: Propagate Points to Curve to Learn Structure Information. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263-7272.

Huang, Z., Dang, T., Cummins, N., Stasak, B., Le, P., Sethu, V. & Epps, J. (2015). An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction. *Proc. of the 5th Int. Workshop on Audio/Visual Emotion Challenge*, pp. 41-48.

Hui Chen, Jiangdong Li, Fengjun Zhang, Yang Li & Hongan Wang. (2015). 3D model-based continuous emotion recognition. *CVPR*, pp. 1836-1845.

Ilse, M., Tomczak, J. & Welling, M. (2018). Attention-based Deep Multiple Instance Learning. *ICML*.

Issa, D., Fatih Demirci, M. & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894.

Jaderberg, M., Simonyan, K., Zisserman, A. & kavukcuoglu, k. (2015). Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 28.

Jaiswal, S., Egede, J. & Valstar, M. (2018). Deep Learned Cumulative Attribute Regression. *IEEE FG*, pp. 715-722.

Jamal, A., Namboodiri, V. P., Deodhare, D. & Venkatesh, K. S. (2018). Deep Domain Adaptation in Action Space. *BMVC*.

Jan, A., Ding, H., Meng, H., Chen, L. & Li, H. (2018). Accurate Facial Parts Localization and Deep Learning for 3D Facial Expression Recognition. *Proc. of 13th IEEE FG*, pp. 466-472.

Ji, Y., Hu, Y., Yang, Y. & Shen, H. (2023). Region Attention Enhanced Unsupervised Cross-Domain Facial Emotion Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 35(04), 4190-4201.

Jiang, H. & Learned-Miller, E. (2017). Face Detection with the Faster R-CNN. *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 650-657.

Kalischek, N., Thiam, P., Bellmann, P. & Schwenker, F. (2019). Deep Domain Adaptation for Facial Expression Analysis. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 317-323.

Karas, V., Tellamekala, M., Mallol-Ragolta, A., Valstar, M. & Schuller, B. (2022). Time-Continuous Audiovisual Fusion with Recurrence vs Attention for In-The-Wild Affect Recognition. *CVPRW*, pp. 2381-2390.

Karmakar, P., Teng, S. W. & Lu, G. (2021). Thank you for Attention: A survey on Attention-based Artificial Neural Networks for Automatic Speech Recognition. *arXiv*.

Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L. & Kollias, S. (2007). Modeling Naturalistic Affective States Via Facial, Vocal, and Bodily Expressions Recognition. *Artifical Intelligence for Human Computing*, pp. 91–112.

Kaur, A., Mustafa, A., Mehta, L. & Dhall, A. (2018). Prediction and Localization of Student Engagement in the Wild. *Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-8.

Kawashima, T., Nomiya, H. & Hochin, T. (2021). Facial Expression Intensity Estimation Using Deep Convolutional Neural Network. *Proceedings of the the 8th International Virtual Conference on Applied Computing and Information Technology*, pp. 7–12.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M. & Zisserman, A. (2017). The Kinetics Human Action Video Dataset. *arXiv*.

Kaya, H., Grpnar, F. & Salah, A. A. (2017). Video-based Emotion Recognition in the Wild Using Deep Transfer Learning and Score Fusion. *Image Vision Comput.*, 65, 66–75.

Kazakos, E., Nagrani, A., Zisserman, A. & Damen, D. (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5491-5500.

Kim, D. H., Lee, M. K., Choi, D. Y. & Song, B. C. (2017). Multi-Modal Emotion Recognition Using Semi-Supervised Learning and Multiple Neural Networks in the Wild. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 529–535.

Kim, D. H., Baddar, W. J., Jang, J. & Ro, Y. M. (2019). Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition. *IEEE Trans. on Affective Computing*, 10(2), 223-236.

Kim, D., Tsai, Y.-H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K. & Chandraker, M. (2021, October). Learning Cross-Modal Contrastive Features for Video Domain Adaptation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13618-13627.

Kim, H.-C., Pang, S., Je, H.-M., Kim, D. & Bang, S.-Y. (2002). Support Vector Machine Ensemble with Bagging. *Pattern Recognition with Support Vector Machines*, pp. 397–408.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W. & Wróbel, M. R. (2014). Emotion Recognition and Its Applications. *Human-Computer Systems Interaction: Backgrounds and Applications 3*.

Kollias, D., Schulc, A., Hajiyev, E. & Zafeiriou, S. (2020). Analysing Affective Behavior in the First ABAW Competition. *FG 2020*.

Kollias, D. (2022). ABAW:Valence-Arousal Estimation, Expression Recognition, Action Unit Detection and Multi-Task Learning Challenges. *CVPRW*, pp. 2327-2335.

Kollias, D. & Zafeiriou, S. (2018). A Multi-component CNN-RNN Approach for Dimensional Emotion Recognition in-the-wild. *arXiv*.

Kollias, D. & Zafeiriou, S. (2021a). Analysing Affective Behavior in the Second ABAW2 Competition. *ICCVW*, pp. 3652-3660.

Kollias, D. & Zafeiriou, S. (2021b). Analysing affective behavior in the second abaw2 competition. *ICCVW*, pp. 3652–3660.

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I. & Zafeiriou, S. (2019). Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *IJCV*, 127, 907–929.

Kollias, D., Psaroudakis, A., Arsenos, A. & Theofilou, P. (2023). FaceRNET: a Facial Expression Intensity Estimation Network.

Kong, Y. S., Suresh, V., Soh, J. & Ong, D. C. (2021). A Systematic Evaluation of Domain Adaptation in Facial Expression Recognition. *CoRR*, abs/2106.15453.

Korkmaz, O. E. & Atasoy, A. (2015). Emotion recognition from speech signal using mel-frequency cepstral coefficients. *9th International Conference on Electrical and Electronics Engineering (ELECO)*, pp. 1254-1257.

Krajewski, J., Wieland, R. & Batliner, A. (2008). An Acoustic Framework for Detecting Fatigue in Speech Based Human-Computer-Interaction. *Computers Helping People with Special Needs*, pp. 54–61.

Krajewski, J., Sommer, D., Schnupp, T., Laufenberg, T., Heinze, C. & Golz, M. (2010). Applying Nonlinear Dynamics Features for Speech-Based Fatigue Detection. *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research*.

Krishna, D. N. & Ankita, P. (2020). Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. *INTERSPEECH*, pp. 4243–4247.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.

Kuhnke, F., Rumberg, L. & Ostermann, J. (2020). Two-Stream Aural-Visual Affect Analysis in the Wild. *FG Workshop*, pp. 600-605.

Lee, J. & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Proc. Interspeech*, pp. 1537–1540.

Lee, J., Kim, S., Kim, S. & Sohn, K. (2018). Audio-Visual Attention Networks for Emotion Recognition. *Workshop on Audio-Visual Scene Understanding for Immersive Multimedia*, pp. 27–32.

Lee, J.-T., Jain, M., Park, H. & Yun, S. (2021). Cross-Attentional Audio-Visual Fusion for Weakly-Supervised Action Localization. *ICLR*.

Lee, J.-T., Yun, S. & Jain, M. (2022). Leaky Gated Cross-Attention for Weakly Supervised Multi-Modal Temporal Action Localization. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 817-826.

Lee, K. K. & Xu, Y. (2003). Real-time estimation of facial expression intensity. *2003 IEEE International Conference on Robotics and Automation (ICRA)*, 2, 2567-2572.

Lee, Y., Yoon, S. & Jung, K. (2020). Multimodal Speech Emotion Recognition Using Cross Attention with Aligned Audio and Text. *INTERSPEECH*, pp. 2717–2721.

Lewis, M., Haviland-Jones, J. & Barrett, L. (2010). Handbook of Emotions. Guilford press.

Li, C., Wen, C. & Qiu, Y. (2023). A Video Sequence Face Expression Recognition Method Based on Squeeze-and-Excitation and 3DPCA Network. *Sensors*, 23(2).

Li, H., Sun, J., Xu, Z. & Chen, L. (2017). Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network. *IEEE Tran. on Multimedia*, 19(12), 2816-2831.

Li, H., Lin, Z., Shen, X., Brandt, J. & Hua, G. (2015). A convolutional neural network cascade for face detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5325-5334.

Li, S. & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. *IEEE Tran. on Affective Computing*, 1-1.

Li, S., Deng, W. & Du, J. (2017). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. *IEEE CVPR*, pp. 2584-2593.

Li, W., Abtahi, F. & Zhu, Z. (2017). Action Unit Detection with Region Adaptation, Multi-labeling Learning and Optimal Temporal Fusing. *IEEE CVPR*, pp. 6766-6775.

Li, W., Abtahi, F., Zhu, Z. & Yin, L. (2018). EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 40(11), 2583-2596.

Li, X. & Ji, Q. (2005). Active affective State detection and user assistance with dynamic bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 35(1), 93-105.

Li, X., Xia, J., Cao, L., Zhang, G. & Feng, X. (2021). Driver fatigue detection based on convolutional neural network and face alignment for edge computing device. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 235(10-11), 2699-2711.

Li, Y., Wu, B., Ghanem, B., Zhao, Y., Yao, H. & Ji, Q. (2016). Facial action unit recognition under incomplete data based on multi-label learning with missing labels. *Pattern Recognition*, 60, 890 - 900.

Li, Y., Wu, B., Zhao, Y., Yao, H. & Ji, Q. (2019). Handling missing labels and class imbalance challenges simultaneously for facial action unit recognition. *Multimedia Tools and Applications*, 78, 20309-20332.

Liu, D., Zhou, Y., Sun, X., Zha, Z. & Zeng, W. (2017). Adaptive Pooling in Multi-instance Learning for Web Video Annotation. *ICCVW*.

Liu, H., Xu, M., Wang, J., Rao, T. & Burnett, I. (2016). Improving Visual Saliency Computing With Emotion Intensity. *IEEE Tran. on Neural Networks and Learning Systems*, 27(6), 1201-1213.

Liu, J.-T., Wu, F.-Y., Lu, W.-J. & Zhang, B.-L. (2019). Domain Adaption for Facial Expression Recognition. *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 1-6.

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z. & Chen, X. (2014). Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild. *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 494–501.

Liu, S., Quan, W., Liu, Y. & Yan, D.-M. (2022). Bi-Directional Modality Fusion Network For Audio-Visual Event Localization. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4868-4872.

Liu, Y., Zeng, J., Shan, S. & Zheng, Z. (2018). Multi-Channel Pose-Aware Convolution Neural Networks for Multi-View Facial Expression Recognition. *IEEE FG*, pp. 458-465.

Liu, Y., Wang, F. & Kong, W. A. (2019). Probabilistic Deep Ordinal Regression Based on Gaussian Processes. *ICCV*.

Liu, Y., Liu, Y., Zhong, S. & Chan, K. C. (2011). Semi-Supervised Manifold Ordinal Regression for Image Ranking. *ACMM*.

Long, C., Guojiang, X., Yuling, L. & Junwei, H. (2021). Driver Fatigue Detection Based on Facial Key Points and LSTM. *Security and Communication Networks*, 2021.

Lu, G. & Zhang, W. (2019). Happiness Intensity Estimation for a Group of People in Images using Convolutional Neural Networks. *3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, pp. 1707-1710.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *IEEE CVPRW*, pp. 94-101.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E. & Matthews, I. (2011). Painful data: The UNBC-McMaster shoulder pain expression archive database. *IEEE FG*, pp. 57-64.

Luengo, I., Navas, E. & Hernáez, I. (2010). Feature Analysis and Evaluation for Automatic Emotion Identification in Speech. *IEEE Transactions on Multimedia*, 12(6), 490-501.

Luo, D., Zou, Y. & Huang, D. (2018). Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. *Interspeech*, pp. 152–156.

Lv, J., Shao, X., Xing, J., Cheng, C. & Zhou, X. (2017). A Deep Regression Architecture with Two-Stage Re-initialization for High Performance Facial Landmark Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3691-3700.

Lynch, M. (2001). Pain as the fifth vital sign. *Journal Intraven Nursing Off Publ, Intraven Nurses Soc*, 24(2), 85-94.

Ma, C., Shen, C., Dick, A., Wu, Q., Wang, P., Hengel, A. v. d. & Reid, I. (2018a). Visual Question Answering with Memory-Augmented Networks. *CVPR*, pp. 6975-6984.

Ma, F., Sun, B. & Li, S. (2021). Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing*, 1-1.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H. & Cai, L. (2018b). Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. *INTERSPEECH*, pp. 3683–3687.

Mao, Q., Dong, M., Huang, Z. & Zhan, Y. (2014). Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Transactions on Multimedia*, 16(8), 2203-2213.

Mao Xu, Wei Cheng, Qian Zhao, Li Ma & Fang Xu. (2015). Facial expression recognition based on transfer learning from deep convolutional networks. *Int. Conf. on Natural Computation (ICNC)*, pp. 702-708.

Marina Martinez, C., Heucke, M., Wang, F., Gao, B. & Cao, D. (2018). Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 19(3), 666-676.

Martinez, B. & Valstar, M. F. (2016). Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition. *Advances in Face Detection and Facial Image Analysis*, 63-100.

Martinez, D. L., Rudovic, O. & Picard, R. W. (2017). Personalized Automatic Estimation of Self-reported Pain Intensity from Facial Expressions. *arXiv*.

Matsumoto, D. (1992). More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(363-368), 363-368.

Matsumoto, D. & Hwang, H. S. (2011). Reading facial expressions of emotion. *Psychological Science Agenda*.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P. & Cohn, J. F. (2013). DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Tran. on Affective Computing*, 4(2), 151-160.

McKeown, G., Valstar, M., Cowie, R., Pantic, M. & Schroder, M. (2012). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Tran. on Affective Computing*, 3(1), 5-17.

Mehrabian, A. (2017a, 09). Nonverbal Communication.

Mehrabian, A. (2017b, 09). Communication Without Words.

Meng, L., Liu, Y., Liu, X., Huang, Z., Jiang, W., Zhang, T., Liu, C. & Jin, Q. (2022). Valence and Arousal Estimation based on Multimodal Temporal-Aware Features for Videos in the Wild. *CVPRW*, pp. 2344-2351.

Miao, S., Tony, X. H., Ming-Chang, L. & Khodayari-Rostamabad, A. (2016). Multiple Instance Learning Convolutional Neural Networks for Object Recognition. *arXiv*, abs/1610.03155.

Miyato, T., Maeda, S., Ishii, S. & Koyama, M. (2018). Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *IEEE Trans. on PAMI*, 1-1.

Mollahosseini, A., Chan, D. & Mahoor, M. H. (2016a). Going deeper in facial expression recognition using deep neural networks. *IEEE WACV*, pp. 1-10.

Mollahosseini, A., Hassani, B., Salvador, M. J., Abdollahi, H., Chan, D. & Mahoor, M. H. (2016b). Facial Expression Recognition from World Wild Web. *IEEE CVPRW*, pp. 1509-1516.

Mollahosseini, A., Hasani, B. & Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Tran. on Affective Computing*, 10(1), 18-31.

Moller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4), 525 - 533.

Monwar, M. M. & Rezaei, S. (2006). Pain Recognition Using Artificial Neural Network. *IEEE International Symposium on Signal Processing and Information Technology*, pp. 28-33.

Muhlenbach, F., Lallich, S. & Zighed, D. A. (2004). Identifying and Handling Mislabelled Instances. *J. Intell. Inf. Syst.*, 22(1), 89–109.

Munro, J. & Damen, D. (2019). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3723-3726.

Nagrani, A., Yang, S., Arnab, A., Schmid, C. & Sun, C. (2021). Attention Bottlenecks for Multimodal Fusion. *NIPS*.

Neumann, M. & Vu, T. (2017). Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech.

Newell, A., Yang, K. & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. *Computer Vision – ECCV 2016*, pp. 483–499.

Nguyen, D., Nguyen, D. T., Zeng, R., Nguyen, T. T., Tran, S., Nguyen, T. K., Sridharan, S. & Fookes, C. (2021). Deep Auto-Encoders with Sequential Learning for Multimodal Dimensional Emotion Recognition. *IEEE Trans. on Multimedia*, 1-1.

Nguyen, K., Ng, T. & Nguyen, L. (2012). Adaptive boosting features for automatic speech recognition. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4733-4736.

Nicholas, C., Vidhyasaharan, S., Julien, E., Sebastian, S. & Jarek, K. (2015). Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, 75, 27 - 49.

Nicholas, C., Alice, B. & Björn, W. S. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151, 41 - 54.

Nicolaou, M. A., Gunes, H. & Pantic, M. (2011). Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. *IEEE Trans. on Affective Computing*, 2, 92-105.

Niu, X., Han, H., Shan, S. & Chen, X. (2019). Multi-label Co-regularization for Semi-supervised Facial Action Unit Recognition. In *NIPS* (vol. 32, pp. 909-919).

Niu, Y., Zou, D., Niu, Y., He, Z. & Tan, H. (2017). A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks. *arXiv*.

Niu, Z., Zhou, M., Wang, L., Gao, X. & Hua, G. (2016). Ordinal Regression with Multiple Output CNN for Age Estimation. *CVPR*.

Noor, J., Daud, M., Rashid, R., Mir, H., Nazir, S. & Velastin, S. A. (2020). Facial Expression Recognition using Hand-Crafted Features and Supervised Feature Encoding. *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1-5. doi: 10.1109/ICECCE49384.2020.9179473.

Ortega, J. D. S., Cardinal, P. & Koerich, A. L. (2019). Emotion Recognition Using Fusion of Audio and Video Features. *SMC*, pp. 3847-3852.

Ouyang, X., Kawaai, S., Goh, E. G. H., Shen, S., Ding, W., Ming, H. & Huang, D.-Y. (2017). Audio-Visual Emotion Recognition Using Deep Transfer Learning and Multiple Temporal Models. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 577–582.

Pan, B. & Wang, S. (2018). Facial Expression Recognition Enhanced by Thermal Images Through Adversarial Learning. *Proc. of 26th ACMM*, pp. 1346–1353.

Pantic, M., Valstar, M., Rademaker, R. & Maat, L. (2005). Web-based database for facial expression analysis. *Proc. of IEEE Int. Conf. on Multimedia and Expo*, pp. 5.

Parkhi, O. M., Vedaldi, A. & Zisserman, A. (2015). Deep Face Recognition. *British Machine Vision Conference*.

Parthasarathy, S. & Sundaram, S. (2021). Detecting Expressions with Multimodal Transformers. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 636-643.

Pei, E., Jiang, D., Alioscha-Perez, M. & Sahli, H. (2019). Continuous affect recognition with weakly supervised learning. *Multimedia Tools and Applications*, 78, 19387 - 19412.

Pei, W., Dibeklioglu, H., Baltrusaitis, T. & Tax, D. M. J. (2017). Attended End-to-end Architecture for Age Estimation from Facial Expression Videos. *arXiv*.

Peng, G. & Wang, S. (2018, June). Weakly Supervised Facial Action Unit Recognition Through Adversarial Training. *IEEE CVPR*, pp. 2188-2196.

Peng, G. & Wang, S. (2019). Dual Semi-Supervised Learning for Facial Action Unit Recognition. *Proc. of the AAAI*, 33, 8827–8834.

Praveen, R. G., de Melo, W. C., Ullah, N., Aslam, H., Zeeshan, O., Denorme, T., Pedersoli, M., Koerich, A. L., Bacon, S., Cardinal, P. & Granger, E. (2022). A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition. *CVPRW*, pp. 2485-2494.

Praveen, R. G., Cardinal, P. & Granger, E. (2023a). Audio-Visual Fusion for Emotion Recognition in the Valence-Arousal Space Using Joint Cross-Attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Praveen, R. G., Granger, E. & Cardinal, P. (2023b). Recursive Joint Attention for Audio-Visual Fusion in Regression based Emotion Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Pérez Rosas, V., Mihalcea, R. & Morency, L.-P. (2013). Multimodal Sentiment Analysis of Spanish Online Videos. *IEEE Intelligent Systems*, 28(3), 38-45.

Qiang Ji, Zhiwei Zhu & Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Transactions on Vehicular Technology*, 53(4), 1052-1068.

Qiang Ji, Lan, P. & Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 36(5), 862-875.

Rajasekhar, G. P., Granger, E. & Cardinal, P. (2021a). Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition. *FG*, pp. 1-8.

Rajasekhar, G. P., Granger, E. & Cardinal, P. (2021b). Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos. *Image and Vision Computing*, 110, 104167.

Rasipuram, S., Bhat, J. H. & Maitra, A. (2020). Multi-modal Sequence-to-sequence Model for Continuous Affect Prediction in the Wild Using Deep 3D Features. *IEEE FG*, pp. 611-614.

Ray, S. & Page, D. (2001). Multiple Instance Regression. *Proc. of the 18th ICML.*

Reddy, B., Kim, Y., Yun, S., Seo, C. & Jang, J. (2017). Real-Time Driver Drowsiness Detection for Embedded System Using Model Compression of Deep Neural Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 438-445.

Ren, S., He, K., Girshick, R. & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, 28.

Ren, Y., Hu, J. & Deng, W. (2017). Facial Expression Intensity Estimation Based on CNN Features and RankBoost. *4th Asian Conference on Pattern Recognition (ACPR)*, pp. 488-493.

Rhevanth, M., Ahmed, R., Shah, V. & Mohan, B. R. (2022). Deep Learning Framework Based on Audio–Visual Features for Video Summarization. *Advanced Machine Intelligence and Signal Processing*, pp. 229–243.

Ringeval, F., Sonderegger, A., Sauer, J. & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *FG*.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R. & Pantic, M. (2015a). AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. *AVEC*.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R. & Pantic, M. (2015b). AVEC 2015: The 5th International Audio/Visual Emotion Challenge and Workshop. *Proc. of 23rd ACMM*, pp. 1335-1336.

Ringeval, F., Pantic, M., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N. & Schmitt, M. (2017, 10). AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. *Proc. of 7th Annual Workshop Audio/Visual Emotion Challenge*, pp. 3-9.

Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B. & Roca, F. X. (2018). Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. *IEEE Transactions on Cybernetics*, 1-11.

Rouast, P. V., Adam, M. & Chiong, R. (2019). Deep Learning for Human Affect Recognition: Insights and New Developments. *IEEE Tran. on Affective Computing*, 1-1.

Ruiz, A., d. Weijer, J. V. & Binefa, X. (2015). From Emotions to Action Units with Hidden and Semi-Hidden-Task Learning. *Proc. of IEEE ICCV*, pp. 3703-3711.

Ruiz, A., Rudovic, O., Binefa, X. & Pantic, M. (2018). Multi-Instance Dynamic Ordinal Random Fields for Weakly Supervised Facial Behavior Analysis. *IEEE Tran. on Image Processing*, 27(8), 3969-3982.

Ruiz, A., Van de Weijer, J. & Binefa, X. (2014). Regularized Multi-Concept MIL for weakly-supervised facial behavior categorization. *BMVC*.

Ruiz, A., Rudovic, O., Binefa, X. & Pantic, M. (2016). Multi-Instance Dynamic Ordinal Random Fields for Weakly-Supervised Pain Intensity Estimation. *Proc. of ACCV*, pp. 171–186.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei. (2015). ImageNet Large Scale Visual Recognition Challenge. 115, 211 - 252.

Sabri, M. & Kurita, T. (2018). Facial expression intensity estimation using Siamese and triplet networks. *Neurocomputing*, 313, 143 - 154.

Samal, A. & Iyengar, P. A. (1992). Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recognition*, 25(1), 65 - 77.

Sandbach, G., Zafeiriou, S., Pantic, M. & Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10), 683 - 697.

Sangineto, E., Zen, G., Ricci, E. & Sebe, N. (2014). We Are Not All Equal: Personalizing Models for Facial Expression Analysis with Transductive Parameter Transfer. *ACM Multimedia*, pp. 357–366.

Sariyanidi, E., Gunes, H. & Cavallaro, A. (2015). Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Tran. on PAMI*, 37(6), 1113-1133.

Satt, A., Rozenberg, S. & Hoory, R. (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Proc. of INTERSPEECH*.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, 61(363-368), 81-88.

Schoneveld, L., Othmani, A. & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-Visual emotion recognition. *Pattern Recognition Letters*, 146, 1-7.

Schuller, B. W. & Rigoll, G. (2009). Recognising interest in conversational speech - comparing bag of frames and supra-segmental features. *INTERSPEECH*.

Sethu, V., Epps, J. & Ambikairajah, E. (2015). Speech Based Emotion Recognition. *Speech and Audio Processing for Coding, Enhancement and Recognition*.

Settles, B., Craven, M. & Ray, S. (2007). Multiple-Instance Active Learning. *NIPS*, pp. 1289–1296.

Shah Fahad, M., Ranjan, A., Yadav, J. & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, 110, 102951.

Shangfei, W., Quan, G. & Qiang, J. (2017). Expression-assisted facial action unit recognition under incomplete AU annotation. *Pattern Recognition*, 61, 78-91.

Shao, Z., Liu, Z., Cai, J., Wu, Y. & Ma, L. (2019). Facial Action Unit Detection Using Attention and Relation Learning. *IEEE Tran. on Affective Computing*, 1-1.

Shao, Z., Liu, Z., Cai, J. & Ma, L. (2018). Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment. *ECCV*.

Shao, Z., Cai, J., Cham, T., Lu, X. & Ma, L. (2019). Weakly-Supervised Unconstrained Action Unit Detection via Feature Disentanglement. *arXiv*.

Shen, Z. & Wei, Y. (2021). A high-precision feature extraction network of fatigue speech from air traffic controller radiotelephony based on improved deep learning. *ICT Express*, 7(4), 403-413.

Shi, B., Dai, Q., Mu, Y. & Wang, J. (2020). Weakly-Supervised Action Localization by Generative Attention Modeling. *CVPR*, pp. 1006-1016.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2), 227 - 244.

Shiomi, T., Nomiya, H. & Hochin, T. (2022). Facial Expression Intensity Estimation Considering Change Characteristic of Facial Feature Values for Each Facial Expression. *23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Summer)*, pp. 15-21.

Shivappa, S. T., Trivedi, M. M. & Rao, B. D. (2010). Audiovisual Information Fusion in Human–Computer Interfaces and Intelligent Environments: A Survey. *Proc. of the IEEE*, 98(10), 1692-1715.

Shon, S., Oh, T.-H. & Glass, J. (2019). Noise-tolerant Audio-visual Online Person Verification Using an Attention-based Neural Network Fusion. *ICASSP*, pp. 3995-3999.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*.

Sikander, G. & Anwar, S. (2019). Driver Fatigue Detection Systems: A Review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6), 2339-2352.

Sikka, K., Sharma, G. & Bartlett, M. (2016). LOMo: Latent Ordinal Model for Facial Analysis in Videos. *IEEE CVPR*, pp. 5580-5589.

Sikka, K. (2014). Facial Expression Analysis for Estimating Pain in Clinical Settings. *ICMI*, pp. 349–353.

Sikka, K., Dhall, A. & Bartlett, M. S. (2014). Classification and weakly supervised pain localization using multiple segment representation. *Image and Vision Computing*, 32(10), 659 - 670.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. pp. 1-14.

Slimi, A., Hamroun, M., Zrigui, M. & Nicolas, H. (2020). Emotion Recognition from Speech Using Spectrograms and Shallow Neural Networks. *Int. Conf. on Advances in Mobile Computing & Multimedia*, pp. 35–39.

Song, X., Zhao, S., Yang, J., Yue, H., Xu, P., Hu, R. & Chai, H. (2021, June). Spatio-temporal Contrastive Domain Adaptation for Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9787-9795.

Song, Y., McDuff, D., Vasisht, D. & Kapoor, A. (2015). Exploiting sparsity and co-occurrence structure for action unit recognition. *IEEE FG*, 1, 1-8.

Sun, L., Chen, J., Xie, K. & Gu, T. (2018a). Deep and shallow features fusion based on deep convolutional neural network for speech emotion recognition. *International Journal of Speech Technology*, 21, 931-940.

Sun, N., Li, Q., Huan, R., Liu, J. & Han, G. (2019a). Deep spatial-temporal feature fusion for facial expression recognition in static images. *Pattern Recognition Letters*, 119, 49-61.

Sun, X., Wu, P. & Hoi, S. C. (2018b). Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*, 299, 42-50.

Sun, Y., Wen, G. & Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control*, 18, 80-90.

Sun, Y., Wang, X. & Tang, X. (2013). Deep Convolutional Network Cascade for Facial Point Detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3476-3483.

Sun, Y., Nomiya, H. & Hochin, T. (2019b). Automatic Evaluation of Motion Picture Contents by Estimation of Facial Expression Intensity. *20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 227-232.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *CVPR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9.

Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI*, pp. 4278–4284.

Tamulevičius, G., Karbauskaitė, R. & Dzemyda, G. (2017). Selection of fractal dimension features for speech emotion classification. *2017 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pp. 1-4.

Tan, Z., Zhou, S., Wan, J., Lei, Z. & Li, S. Z. (2017). Age Estimation Based on a Single Network with Soft Softmax of Aging Modeling. *ACCV*.

Tang, Y. (2013). Deep Learning using Linear Support Vector Machines. *arXiv: Learning*.

Tao, H., Liang, R., Zha, C., Zhang, X. & Zhao, L. (2016). Spectral Features Based on Local Hu Moments of Gabor Spectrograms for Speech Emotion Recognition. *IEICE Transactions on Information and Systems*, E99.D(8), 2186-2189.

Tavakolian, M. & Hadid, A. (2018). Deep Spatiotemporal Representation of the Face for Automatic Pain Intensity Estimation. *Proc. of 24th ICPR*, pp. 350-354.

Tavakolian, M., Bordallo Lopez, M. & Liu, L. (2020). Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recognition Letters*, 140, 26-33.

Thuseethan, S., Rajasegarar, S. & Yearwood, J. (2019). Emotion Intensity Estimation from Video Frames using Deep Hybrid Convolutional Neural Networks. *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-10.

Tian, Y., Shi, J., Li, B., Duan, Z. & Xu, C. (2018). Audio-Visual Event Localization in Unconstrained Videos. *ECCV*.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang & Xiaogang Wang. (2015). Learning from massive noisy labeled data for image classification. *IEEE CVPR*, pp. 2691-2699.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *Proc. of the IEEE ICCV*, pp. 4489–4497.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. & Paluri, M. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. *CVPR*, pp. 6450-6459.

Trigeorgis, G., Snape, P., Nicolaou, M. A., Antonakos, E. & Zafeiriou, S. (2016). Mnemonic Descent Method: A Recurrent Process Applied for End-to-End Face Alignment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4177-4187.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W. & Zafeiriou, S. (2017). End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. of Selected Topics in Signal Processing*, 11(8), 1301-1309.

Tzirakis, P., Chen, J., Zafeiriou, S. & Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68, 46-53.

Valstar, M. F. & Pantic, M. (2010). Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database. *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pp. 65–70.

Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M. & Cohn, J. F. (2015). FERA 2015 - second Facial Expression Recognition and Analysis challenge. *IEEE FG*, 06.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R. & Pantic, M. (2013). AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R. & Pantic, M. (2016). AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. *AVEC*, pp. 3–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. & Polosukhin, I. (2017). Attention is All you Need. *NIPS*, 30.

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1, I-I.

Viola, P. A., Platt, J. C. & Zhang, C. (2006). Multiple Instance Boosting for Object Detection. In *Proc. of NIPS* (vol. 18, pp. 1417–1424).

Vukotić, V., Raymond, C. & Gravier, G. (2016). Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications. *ICMR*, pp. 343–346.

Walecki, R., Rudovic, O., Pavlovic, V., Schuller, B. & Pantic, M. (2017). Deep Structured Learning for Facial Action Unit Intensity Estimation. *CVPR*, pp. 5709-5718.

Wang, C., Ding, J., Yan, H. & Shen, S. (2022, December). A Prototype-Oriented Contrastive Adaption Network For Cross-domain Facial Expression Recognition. *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 4194-4210.

Wang, F., Xiang, X., Liu, C., Tran, T. D., Reiter, A., Hager, G. D., Quon, H., Cheng, J. & Yuille, A. L. (2017). Regularizing face verification nets for pain intensity regression. *Proc. of IEEE ICIP*.

Wang, H., Gao, F., Zhao, Y. & Wu, L. (2020a). WaveNet With Cross-Attention for Audiovisual Speech Recognition. *IEEE Access*, 8, 169160-169168.

Wang, J., Ma, Y., Zhang, L., Gao, R. X. & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48, 144-156.

Wang, L., Wang, S., Qi, J. & Suzuki, K. (2021). A Multi-task Mean Teacher for Semi-supervised Facial Affective Behavior Analysis. *ICCVW*, pp. 3596-3601.

Wang, M. & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 135 - 153.

Wang, S. & Peng, G. (2019). Weakly Supervised Dual Learning for Facial Action Unit Recognition. *IEEE Tran. on Multimedia*, 1-1.

Wang, S., Pan, B., Chen, H. & Ji, Q. (2018a). Thermal Augmented Expression Recognition. *IEEE Tran. on Cybernetics*, 48(7), 2203-2214.

Wang, S., Peng, G., Chen, S. & Ji, Q. (2018b). Weakly Supervised Facial Action Unit Recognition With Domain Knowledge. *IEEE Tran. on Cybernetics*, 48(11), 3265-3276.

Wang, S., Peng, G. & Ji, Q. (2018c). Exploring Domain Knowledge for Facial Expression-Assisted Action Unit Activation Recognition. *IEEE Tran. on Affective Computing*, 1-1.

Wang, S., Pan, B., Wu, S. & Ji, Q. (2019). Deep Facial Action Unit Recognition and Intensity Estimation from Partially Labelled Data. *IEEE Tran. on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2019.2914654.

Wang, W., Tran, D. & Feiszli, M. (2020b). What Makes Training Multi-Modal Classification Networks Hard? *CVPR*, pp. 12692-12702.

Wang, X., Bo, L. & Fuxin, L. (2019). Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6970-6980.

Wang, X., Wang, X. & Ni, Y. (2018). Unsupervised Domain Adaptation for Facial Expression Recognition Using Generative Adversarial Networks,. *Computational Intelligence and Neuroscience*.

Wang, Y., Li, J. & Metze, F. (2019). A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. *ICASSP*.

Wang, Y., Ma, J., Hao, B., Hu, P., Wang, X., Mei, J. & Li, S. (2020). Automatic Depression Detection via Facial Expressions Using Multiple Instance Learning. *Proc. of IEEE 17th Int. Symp. on Biomedical Imaging (ISBI)*, pp. 1933-1936.

Wang, Y., Wu, J., Heracleous, P., Wada, S., Kimura, R. & Kurihara, S. (2020). Implicit Knowledge Injectable Cross Attention Audiovisual Model for Group Emotion Recognition. *ACM ICMI*, pp. 827–834.

Wang, Z., Wang, S. & Ji, Q. (2013). Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition. *IEEE CVPR*, pp. 3422-3429.

Wei, Q., Bozkurt, E., Morency, L.-P. & Sun, B. (2019). Spontaneous smile intensity estimation by fusing saliency maps and convolutional neural networks. *Journal of Electronic Imaging*, 28(2), 023031.

Wei, X., Zhang, T., Li, Y., Zhang, Y. & Wu, F. (2020). Multi-Modality Cross Attention Network for Image and Sentence Matching. *CVPR*.

Wen, Y., Zhang, K., Li, Z. & Qiao, Y. (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. *Computer Vision – ECCV 2016*, pp. 499–515.

Wu, B., Lyu, S., Hu, B.-G. & Ji, Q. (2015a). Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 48(7), 2279 - 2289.

Wu, C., Wang, S. & Ji, Q. (2015b). Multi-instance Hidden Markov Model for facial expression recognition. *Proc. of 11th IEEE FG*, 1, 1-6.

Wu, C.-L., Liu, S.-F., Yu, T.-L., Shih, S.-J., Chang, C.-H., Yang Mao, S.-F., Li, Y.-S., Chen, H.-J., Chen, C.-C. & Chao, W.-C. (2022). Deep Learning-Based Pain Classifier Based on the Facial Expression in Critically Ill Patients. *Frontiers in Medicine*, 9.

Wu, C.-H., Lin, J.-C. & Wei, W.-L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA Trans. on Signal and Information Processing*, 3, e12.

Wu, N. & Sun, J. (2022). Fatigue Detection of Air Traffic Controllers Based on Radiotelephony Communications and Self-Adaption Quantum Genetic Algorithm Optimization Ensemble Learning. *Applied Sciences*, 12(20).

Wu, S., Wang, S., Pan, B. & Ji, Q. (2017). Deep Facial Action Unit Recognition from Partially Labeled Data. *IEEE ICCV*, pp. 3971-3979.

Wu, S., Falk, T. H. & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5), 768-785.

Wu, W., Yin, Y., Wang, X. & Xu, D. (2019). Face Detection With Different Scales Based on Faster R-CNN. *IEEE Transactions on Cybernetics*, 49(11), 4017-4028.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B. & Rigoll, G. (2013). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2), 153-163.

Xiang, X., Ye, T., Gregory, D. H. & Trac, D. T. (2018). Assessing Pain Levels from Videos Using Temporal Convolutional Networks. *IEEE WACV workshops*.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S. & Kassim, A. (2016). Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks. *Computer Vision – ECCV 2016*, pp. 57–72.

Xie, L., Tao, D. & Wei, H. (2019). Early Expression Detection via Online Multi-Instance Learning With Nonlinear Extension. *IEEE Tran. on Neural Networks and Learning Systems*, 30(5), 1486-1496.

Xie, S. & Hu, H. (2019). Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks. *IEEE Tran. on Multimedia*, 21(1), 211-220.

Xinggang, W., Yongluan, Y., Peng, T., Xiang, B. & Wenyu, L. (2018). Revisiting multiple instance neural networks. *Pattern Recognition*, 74, 15 - 24.

Xu, L. & Mordohai, P. (2010). Automatic Facial Expression Recognition using Bags of Motion Words. *Proc. of BMVC*, pp. 13.1–13.13.

Xue, C., Zhong, X., Cai, M., Chen, H. & Wang, W. (2023). Audio-Visual Event Localization by Learning Spatial and Semantic Co-Attention. *IEEE Transactions on Multimedia*, 25, 418-429.

Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., Zong, Y. & Sun, N. (2016). Multi-Clue Fusion for Emotion Recognition in the Wild. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 458–463.

Yang, G., Lin, Y. & Bhattacharya, P. (2010). A driver fatigue recognition model based on information fusion and dynamic Bayesian network. *Information Sciences*, 180(10), 1942-1954.

Yang, J., Wang, K., Peng, X. & Qiao, Y. (2018). Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. *ACM ICMI*, pp. 594–598.

Yang, J., Liu, Q. & Zhang, K. (2017). Stacked Hourglass Network for Robust Facial Landmark Localisation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2025-2033.

Yang, L., Huang, Y., Sugano, Y. & Sato, Y. (2022). Interact before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14702-14712.

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S. & Vepa, J. (2018). Speech Emotion Recognition Using Spectrogram and Phoneme Embedding. *Proc. Interspeech 2018*, pp. 3688–3692.

Yin, L., Chen, X., Sun, Y., Worm, T. & Reale, M. (2008). A high-resolution 3D dynamic facial expression database. *IEEE FG*, pp. 1-6.

Zeng, J., Chu, W., De la Torre, F., Cohn, J. F. & Xiong, Z. (2016). Confidence Preserving Machine for Facial Action Unit Detection. *IEEE Tran. on Image Processing*, 25(10), 4753-4767.

Zeng, J., Shan, S. & Chen, X. (2018). Facial Expression Recognition with Inconsistently Annotated Datasets. *ECCV*, pp. 227–243.

Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth, D. & Levinson, S. (2007). Audio-Visual Affect Recognition. *IEEE Transactions on Multimedia*, 9(2), 424-428.

Zhang, F., Su, J., Geng, L. & Xiao, Z. (2017). Driver Fatigue Detection Based on Eye State Recognition. *2017 International Conference on Machine Vision and Information Technology (CMVIT)*, pp. 105-110.

Zhang, F., Xu, M. & Xu, C. (2021a). Weakly-Supervised Facial Expression Recognition in the Wild with Noisy Data. *IEEE Transactions on Multimedia*, 1-1.

Zhang, J., Shan, S., Kan, M. & Chen, X. (2014a). Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. *Computer Vision – ECCV 2014*, pp. 1–16.

Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10), 1499-1503.

Zhang, L., Verma, B., Tjondronegoro, D. & Chandran, V. (2018). Facial Expression Analysis Under Partial Occlusion: A Survey. *ACM Comput. Surv.*, 51(2).

Zhang, Q., Goldman, S. A., Yu, W. & Fritts, J. (2002). Content-Based Image Retrieval Using Multiple-Instance Learning. *ICML*.

Zhang, S., Ding, Y., Wei, Z. & Guan, C. (2021b). Continuous Emotion Recognition with Audio-visual Leader-follower Attentive Fusion. *ICCVW*, pp. 3560-3567.

Zhang, W., Murphey, Y. L., Wang, T. & Xu, Q. (2015). Driver yawning detection based on deep convolutional neural learning and robust nose tracking. *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P. & Girard, J. M. (2014b). BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10), 692 - 706.

Zhang, Y., Dong, W., Hu, B. & Ji, Q. (2018a). Classifier Learning with Prior Probabilities for Facial Action Unit Recognition. *IEEE CVPR*, pp. 5108-5116.

Zhang, Y., Zhao, R., Dong, W., Hu, B. & Ji, Q. (2018b). Bilateral Ordinal Relevance Multi-instance Regression for Facial Action Unit Intensity Estimation. *IEEE CVPR*, pp. 7034-7043.

Zhang, Y., Fan, Y., Dong, W., Hu, B. & Ji, Q. (2019a). Semi-Supervised Deep Neural Network for Joint Intensity Estimation of Multiple Facial Action Units. *IEEE Access*, 7, 150743-150756.

Zhang, Y., Jiang, H., Wu, B., Fan, Y. & Ji, Q. (2019b). Context-Aware Feature and Label Fusion for Facial Action Unit Intensity Estimation With Partially Labeled Data. *Proc. of IEEE ICCV*, pp. 733-742.

Zhang, Y., Wu, B., Dong, W., Li, Z., Liu, W., Hu, B. & Ji, Q. (2019c). Joint Representation and Estimator Learning for Facial Action Unit Intensity Estimation. *IEEE CVPR*, pp. 3452-3461.

Zhang, Y., Doughty, H., Shao, L. & Snoek, C. M. (2022). Audio-Adaptive Activity Recognition Across Video Domains. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13781-13790.

Zhang, Y., Dong, W., Hu, B.-G. & Ji, Q. (2018, June). Weakly-Supervised Deep Convolutional Neural Network Learning for Facial Action Unit Intensity Estimation. *IEEE CVPR*.

Zhang, Y.-H., Huang, R., Zeng, J. & Shan, S. (2020). Multi-Modal Continuous Valence-Arousal Estimation in the Wild. *IEEE FG*, pp. 632-636.

Zhao, B., Gong, M. & Li, X. (2021). AudioVisual Video Summarization. *IEEE Transactions on Neural Networks and Learning Systems*, 1-8.

Zhao, G. & Pietikainen, M. (2007). Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Tran. on PAMI*, 29(6), 915-928.

Zhao, J., Mao, X. & Chen, L. (2019). Speech emotion recognition using deep 1D and 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47, 312-323.

Zhao, K., Chu, W.-S. & Martinez, A. M. (2018, June). Learning Facial Action Units From Web Images With Scalable Weakly Supervised Clustering. *IEEE CVPR*.

Zhao, R., Gan, Q., Wang, S. & Ji, Q. (2016). Facial Expression Intensity Estimation Using Ordinal Information. *IEEE CVPR*, pp. 3466-3474.

Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H. & Keutzer, K. (2020). An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos. *AAAI*, 34(01), 303-311.

Zhao, Z., Zheng, P., Xu, S. & Wu, X. (2019). Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212-3232.

Zheng, K., Yang, D., Liu, J. & Cui, J. (2020). Recognition of Teachers' Facial Expression Intensity Based on Convolutional Neural Network and Attention Mechanism. *IEEE Access*, 8, 226437-226444.

Zheng, W. Q., Yu, J. S. & Zou, Y. X. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 827-831.

Zhong, Y., Chen, J. & Huang, B. (2017). Toward End-to-End Face Recognition Through Alignment Learning. *IEEE Signal Processing Letters*, 24(8), 1213-1217.

Zhou, J., Hong, X., Su, F. & Zhao, G. (2016a). Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video. *arXiv*.

Zhou, X., Guo, J. & Bie, R. (2016b). Deep Learning Based Affective Model for Speech Emotion Recognition. *Int. Conf. on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*, pp. 841-846.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44-53.

Zhu, H., Ji, Y., Wang, B. & Kang, Y. (2022). Exercise fatigue diagnosis method based on short-time Fourier transform and convolutional neural network. *Frontiers in Physiology*, 13.

Zhu, R., Sang, G. & Zhao, Q. (2016). Discriminative Feature Adaptation for cross-domain facial expression recognition. *Int. Conf. on Biometrics*, pp. 1-7.

Zhu, X., Lei, Z., Yan, J., Yi, D. & Li, S. Z. (2015). High-fidelity Pose and Expression Normalization for face recognition in the wild. *IEEE CVPR*, pp. 787-796.

Zuopeng, Z., Nana, Z., Lan, Z., Hualin, Y., Yi, X. & Zhongxin, Z. (2020). Driver Fatigue Detection Based on Convolutional Neural Networks Using EM-CNN. *Computational Intelligence and Neuroscience*, 2020.