Multimodal audio-visual emotion detection

Chaudhary, Nitesh Kumar

2021

Chaudhary, N. K. (2021). Multimodal audio-visual emotion detection. Master's thesis, Nanyang Technological University, Singapore. <https://hdl.handle.net/10356/153490>

<https://hdl.handle.net/10356/153490>

<https://doi.org/10.32657/10356/153490>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 11:03:48 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

Multimodal Audio-Visual Emotion Detection

Nitesh Kumar Chaudhary

School of Computer Science Engineering

2021

Multimodal Audio-Visual Emotion Detection

Nitesh Kumar Chaudhary

G1802997E

School of Computer Science Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirements for the degree of
Master of Engineering

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

23rd July 2021

.....

Date

NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU

.....

Nitesh Kumar Chaudhary

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

15/11/2021

.....

Date

NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU

Rajapakse

Professor Jagath C Rajapakse

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

23rd July 2021

.....

Date

NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
NTU NTU NTU NTU NTU NTU NTU NTU
.....

Nitesh Kumar Chaudhary

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Jagath C Rajapakse for his guidance, encouragement and supervision for my master research program. I am really fortunate to work under his supervision. He has constantly given me guidance, constructive suggestions and thoughtful discussion in the path of pursuing my research goals. He taught me how to explore research issue deeply to achieve novelty in work. I sincerely thank him for the research opportunity and appreciate research experience with his supervision.

Finally, I would like to express my special gratitude to my family for their love, support and encouragement throughout the course of this journey.

Table of Contents

LIST OF ABBREVIATIONS	VI
LIST OF FIGURES	VII
LIST OF TABLES	VIII
ABSTRACT	IX
CHAPTER 1 – INTRODUCTION	1
1.1 BACKGROUND AND MOTIVATION	2
1.2 LITERATURE REVIEW	4
1.2.1 <i>Audio emotion recognition (AER)</i>	4
1.2.2 <i>Video emotion recognition (VER)</i>	6
1.2.3 <i>Audio-Video emotion recognition (AVER)</i>	7
1.4 CONTRIBUTIONS	9
CHAPTER 2 – AUDIO EMOTION RECOGNITION (AER)	11
2.1 AUDIO SPECIFIC LOW-LEVEL FEATURES EXTRACTION AND CNN MODEL	12
2.1.1 <i>Spectral features</i>	12
2.1.2 <i>Prosodic features</i>	13
2.1.3 <i>Extracted AER features from raw audio</i>	14
2.1.4 <i>CNN embedding model using extracted low-level AER features</i>	16
2.2 SPATIO-TEMPORAL EMBEDDING FEATURES	17
2.2.1 <i>Audio features embeddings</i>	17
CHAPTER 3 – VISUAL EMOTION RECOGNITION (VER)	20
3.1 VIDEO SPECIFIC LOW-LEVEL FEATURES AND EMBEDDINGS FEATURES TO CNN MODEL	22
3.2 EMBEDDING FEATURES WITH CONTEXTUAL MODELLING	24
3.2.1 <i>Key frame Extraction</i>	24
3.2.2 <i>Deep features embedding of temporal key-frame stream</i>	25
4.1 AUDIO-VISUAL MODELLING USING LOW LEVEL FEATURES EXTRACTION	28
4.2 AUDIO-VISUAL CONTEXTUAL TEMPORAL MODELLING USING FEATURES EMBEDDINGS	29
4.2.1 <i>Gated Recurrent Units</i>	29
4.2.2 <i>GRU to audio-visual embeddings</i>	32
4.2.3 <i>Fusion of Audio and Video embeddings</i>	32
CHAPTER 5 – EVALUATION AND DISCUSSION	34
5.1 DATASETS	34
5.1.1 <i>RAVDESS</i>	34
5.1.2 <i>IEMOCAP</i>	35
5.2 DESCRIPTION OF DESIGN	36
5.3 EXPERIMENTAL RESULTS	38
5.3.1 <i>Results and comparison of AER</i>	39
5.3.2 <i>Results and comparison of VER</i>	41
5.3.3 <i>Results and comparison of AVER</i>	42
5.4 DISCUSSION	44
CHAPTER 6 – FUTURE WORK	46
6.1 <i>Attention fusion</i>	46
6.2 <i>Attention Mechanism</i>	47
6.3 <i>Visual features exploration</i>	48
6.3.1 <i>I3D: Spatio-temporal feature extractor</i>	48

REFERENCE.....	50
----------------	----

List of Abbreviations

AER	Audio emotion recognition
VER	Video emotion recognition
AVER	Audio-visual emotion recognition
SSD	Single shot detection
I3D	Inflated 3D ConvNet
GAN	Generative Adversarial Network
HCI	Human computer interaction
MFCC	Mel-frequency cepstral coefficient
LSTM	Long short-term memory
ELM	Extreme learning machine
VGGish	Visual geometric group (Audioset pretrained CNN from Google)
LPCC	Linear predictive cepstral coefficient
FLD	Face landmark detection
HOG	Histogram of gradient
ReLU	Rectified linear units
ELU	Exponential linear units
LUV	L stands for luminance, whereas U and V represent chromaticity values of colour images
GRU	Gated Recurrent Units
RAVDESS	The Ryerson Audio-Visual Database of Emotional Speech and Song
IEMOCAP	The Interactive Emotional Dyadic Motion Capture
DCT	Discrete cosine transform
TEO	Teager energy operator
MOCAP	Motion capture (in IEMOCAP dataset)
DNN	Deep Neural Network
ANN	Artificial Neural Network
SVM	Support vector machine

List of Figures

FIGURE 2.1 ILLUSTRATION OF AUDIO FEATURES EXTRACTION.	15
FIGURE 2.2 CONFIGURATION OF AUDIO CLASSIFIER NETWORK.....	16
FIGURE 2.3 ILLUSTRATION OF OBTAINING AUDIO REPRESENTATION IN THE AUDIO STREAM.....	18
FIGURE 2.4 THE VGGISH MODEL USED FOR GENERATING AUDIO EMBEDDINGS IN THE AUDIO STREAM.....	19
FIGURE 2.5 RNN TOPOLOGY FOR AER USING AUDIO VGGISH EMBEDDING	19
FIGURE 3.1.1 ILLUSTRATION OF VIDEO FEATURES EXTRACTION.	23
FIGURE 3.1.2 ILLUSTRATION OF VIDEO FEATURES EXTRACTION.	23
FIGURE 3.1.3 CONFIGURATION OF VIDEO CLASSIFIER NETWORK.....	24
FIGURE 3.2.1 VIDEO TEMPORAL KEY FRAMES EXTRACTION	25
FIGURE 3.2.2 ILLUSTRATION OF LEARNING VISUAL REPRESENTATIONS FROM THE VIDEO STREAM	26
FIGURE 4.1 OVERVIEW OF PROPOSED ARCHITECTURE FOR AUDIO-VISUAL MODELLING USING LOW LEVEL FEATURES.....	28
FIGURE 4.2 OVERVIEW OF PROPOSED CONTEXTUAL TEMPORAL MODELLING FOR AUDIO-VISUAL EMOTION DETECTION PIPELINE.....	29
FIGURE 4.3 GATED RECURRENT UNIT CELLS AND GATES	30
FIGURE 5.1 RAVDESS: AUDIO-VISUAL DYNAMIC DATABASE, MULTIMODAL SET	35
FIGURE 5.2 QUALITATIVE RESULTS ON IEMOCAP.....	36

List of Tables

TABLE 2.1 AUDIO SPECIFIC SPECTRAL AND PROSODIC FEATURES.....	15
TABLE 5.1 ACCURACIES OF EMOTION RECOGNITION USING AUDIO FEATURES ON RAVDESS DATASET	40
TABLE 5.2 ACCURACIES OF EMOTION RECOGNITION USING AUDIO FEATURES ON IEMOCAP DATASET.....	40
TABLE 5.3 EMOTION DETECTION USING VISUAL STREAM ON IEMOCAP DATASET	42
TABLE 5.4 EMOTION DETECTION USING VISUAL STREAM ON RAVDESS DATASET.....	42
TABLE 5.5 EMOTION DETECTION USING FUSION OF AUDIO AND VISUAL STREAMS ON RAVDESS DATASET	43
TABLE 5.6 EMOTION DETECTION USING FUSION OF AUDIO AND VISUAL STREAMS ON IEMOCAP DATASET	44

Abstract

Emotions are expressed when the brain uses our memories, ideas, and beliefs to understand what is going on around us. When anything happens around us, this sets off a chain of events that affects how we feel and act. These processes impact all our judgments and behaviors in some manner. Emotions may have an impact on not just the result of a decision, but also how swiftly you respond to it. Impatience and hasty decision-making can result from anger. When we are enthusiastic, we may make snap judgments without thinking about the consequences as we ride the wave of enthusiasm and hope for a better future. So, emotions contribute immensely in human decision handling, interaction and cognitive processes. Extraction and recognition of emotion has a high importance in human to human and human to machine interactions for a better interpersonal relationship and communication.

Looking at how technologies are progressing at a rapid pace with increased use of robotics and artificial intelligence, adding greater values and conveniences to human life. It became essential to have healthy communication with assistive technologies, but such inter linkage with machines lacks behavioral and emotional attributes during interactions. There is a growing need for smart interactive systems which can combine roles of emotions and logic in helping us to make pleasing and positive decisions. The most common way of detecting emotion cues are through verbal and nonverbal stats which humans send and receive constantly during conversation through facial expressions and vocal information (voice, tonality) associated with the context. In general, human emotions are a composition of audio, speech, and facial expressions. In terms of arousal and valence, speech carries vocal ways of communicating emotion. Whereas facial expression changes with facial landmarks exhibits emotions that carry temporal/contextual characteristics in nature.

Though audio and visual utterances in video are temporally and semantically dependent to each other so modeling of temporal and contextual characteristics take a major role in understanding of conflicting or supporting emotional cues in audio-visual emotion recognition (AVER). Yet, existing research works in audio-visual emotion recognition have mostly concentrated on emotion detection solely from audio or solely from video. However, because humans have a bimodal emotion production and perception system, visual and audio information are both necessary and complimentary for emotion detection. But realistic and robust emotion detection from audio and video is a challenging task despite its widespread applications in healthcare, robotics, human computer interactions, security, and surveillance.

To address the above-mentioned challenge in emotion recognition, in this research, we have focused on end-to-end AER where we have explored state-of-the-art methods by replacing the use of handcrafted features with the advent of deep learning-based features extraction embedding. In the similar fashion, we have explored VER independently by replacing handcrafted frame wise features extraction with temporal contextual windowed features and learning of embedding vectors. We have explored the dimensions of deep learning based on neural network architectures to capture the most accurate emotions from the audio and video feed. We continue to further explore emotions by combining audio and visual information for AVER.

Our research has emphasized on building an emotion recognition system applicable for real world applications with no constraint or specific requirements such as compulsion of wearing a device with sensors, recording in controlled environment, applying marker points, reading of specific annotation, number of user and linguistics constraints. We capture the emotion from data collected in a normal conversation without having any above-mentioned compulsions or limitations. Though some data considered for the experiment provides sensor-based marker points, we have used raw data as source of input for features extraction which avoids any sort

of compulsive requirements. We come up with a contextual temporal modular neural network for multi-modal end-to-end emotion recognition. The advantages of modular architecture allow us to do extensive research for each component of emotion cues. Contextual-temporal ways of extracting emotions capture long term dependencies to exactly convey deeper emotions associated with the context such as sarcastic conversation, relational conversation, ambiance sensing, etc. We have done extensive research on audio emotion pipeline, video emotion pipeline, and bi-modal fusion techniques to incorporate contextual and temporal characteristics. Our experimental research shows that the proposed contextual temporal modeling achieves a state-of-art performance recognition rate for AVER. This report presents the work performed on AER, VER, and AVER, and future work to be led by this thesis.

Keywords: Emotion recognition, Audio-visual, Multimodal, CNN, RNN, Features Embeddings

Chapter 1 – Introduction

An emotion entails a mental and physiological state of an individual, which manifests persons behaviors, actions, thoughts, and feelings. Emotions are true experiences that are linked to the activation of specific brain areas. Our brain is a very complex organ, which plays a crucial role in how you control and process your emotions. It turns out that some parts of the brain, particularly the limbic system, are linked to emotions. Happiness, for example, activates multiple portions of the brain, including the right frontal cortex, the precuneus, the left amygdala, and the left insula, and is related with different regions of limbic system activity than other emotions. Contrarily fear activates the amygdala on both sides, as well as the hypothalamus and parts of the left frontal cortex. These changes in brain signals have direct implications on changes in human behavior and the most common way of detecting emotion cues are through verbal and nonverbal stats which humans send and receive constantly during conversation through facial expressions and vocal information (voice, tonality) associated with the context.

With the experiences and time elapsed in our mind, we get trained to differentiate one type of sentiments/emotions from others. We get more trained and competent at interpreting emotions as we grow older. We can identify what we are feeling and put it into words and expressions rather than reacting like young kids. We improve our ability to understand what we are experiencing and why we are feeling it with time and practice. Emotional awareness aids us in determining what we require and desire (or do not desire!). It contributes to the development of stronger relationships. It is because being aware of feelings may help us communicate more clearly about our sentiments, avoid, or settle disputes more effectively, and get through uncomfortable feelings more quickly.

It has been observed that ambiance and surroundings have an influence on the emotional state of human that is known as Induced affect. Actually, Induced effect is the emotional influence of the surrounding environment on individuals. Arousal and valence (sentiment) are often used to classify induced emotional effect. The affinity for an object is measured in valence, which can be positive, neutral, or negative. The magnitude of the reaction to that stimulus is termed as arousal: calm, neutral, or active.

1.1 Background and Motivation

Today with the advent of machinery, we can accomplish things more rapidly and effortlessly due to digital platforms and technology. We are revolutionizing with technological innovations and progressing at a rapid pace with increased use of robotics and artificial intelligence, adding greater values and conveniences to human life. All of these robotics products are automated with minimal human intervention, despite the fact that emotion detection is critical in automated applications: for example, health-care rehabilitation robots that look after the comfort of the elderly and patients; machine teaching assistants, lie or anomaly detectors in security and surveillance, and digital marketing chatbot interactions with customers. One particular aspect that is not taken into account in these systems is emotion. Therefore, these kinds of human-computer interaction (HCI) lack naturalness, temperaments, situation awareness, which arise in consequent healthy communication. We may become anxious or impatient but access to contextual verbal/nonverbal emotion information would allow these systems to adapt to situation in order to sort out the underlying issue.

The emotional characteristics of users' experiences are associated to positive feelings, according to recent research by Tampere University in Finland (Partala & Kallinen) [1], and those frustrating and unsatisfying experiences leads to negative emotions. The study

shows how we feel when we interact with an automated system, the first time a user tries such a system, they may be perplexed and have minor difficulties, giving them some unpleasant impression. However, as they become more familiar with most of the features and how to access them, users will become emotionally attached to it, making each use a pleasant experience. As a result, emotional awareness is essential since it can assist people in better coping with emotionally challenging events or situations.

Several technical problems can impact the performance of a facial expressions-based algorithm [2]. For example, different groups of people may express the same feelings in various different ways. Furthermore, differing orientation of reference point result in uneven representations of emotion. Further, occlusions and variations in illumination may cause the identification system to be mistaken. If the emotion needs to be recognized by speech, ambient noise and variations in the voices of various individuals are important elements that might affect the final recognition performance. In order to correctly understand emotions, people utilize mainly auditory and visual information. Visual and audio information are both required and supportive for emotion detection due to the bimodal emotion generation and perception mechanisms of human beings. Example a scenario where only speech signals can be conveniently collected, such as conversation via telephone, in location with no cameras, or poor illumination. In that basal state, emotion can only be evident through audio context. Similarly, sometimes information is conveyed through nonverbal communication or expressions. In such cases it is logical to utilize facial behavioral cues to understand emotions.

1.2 Literature review

Emotion research has progressed tremendously in the recent decade. Emotion recognition from speech, visual, or audio-visual information from video can be done using a variety of ways. In recent years, analysis and recognition approaches for emotion detection from a unimodal input have gotten a lot of attention. However, emotion identification based solely on facial or vocal modality still has limitations. Exploring the data fusion technique for efficiently merging facial and voice cues could be a promising research topic for improving emotion recognition. Different groups of researchers coming from neuroscience, psychology, audiology, and computer science have tried to understand and predict emotions through facial and vocal cues.

1.2.1 Audio emotion recognition (AER)

Prosody characteristics such as pitch associated features, energy related features, zero crossing formant, teager energy operator (TEO), fundamental frequency (F0), speech rate, and spectral features for speech emotion identification were investigated by Z. Zeng et al. [3]. He focused his efforts on surveying the development of algorithms that can handle naturally occurring human affective behavior, hence GMM, HMM, Decision tree, MLP, and SVM were studied as classifiers to detect emotions. There are additional studies that indicate how pitch-related and energy-related audio characteristics are the most significant for emotion recognition. For emotion recognition, a study incorporating spectrum-related characteristics such as linear predictive coefficients [4], linear predictive cepstrum coefficients [5], Mel-frequency cepstrum coefficients, and their first derivative were employed. To detect emotions, Han et al. [6] employed both segment-level features including Mel-frequency cepstral coefficient, pitch period, and harmonic to noise ratio, as well as speech features. He utilized DNNs to extract high level utterance level features from

segment level features. These utterance level features are then fed into an extreme learning machine (ELM), a special simple and efficient single-hidden-layer neural network, to identify utterance-level emotions. To identify different emotions in a human, Tomba et al. [31] proposed to utilize speech, features like pitch (also referred as fundamental frequency), articulation rate, energy or MFCCs. They chose SVMs and ANNs as their algorithms, concluding that both algorithms produced outstanding results, with ANNs scoring slightly higher than SVMs. To learn prominent and discriminative features from spectrograms of speech signals, Mustaqeem et al. [32] suggested an artificial intelligence-assisted deep stride convolutional neural network (DSCNN) architecture utilizing the plain nets technique. On the IEMOCAP and RAVDESS datasets, the suggested technique improves accuracy by 7.85 percent and 4.5 percent, respectively, with a reduction in model size. Fayek et al. investigated different deep learning architectures for the Speech Emotion Recognition (SER) task in [34]. Experiments show how feed-forward and recurrent neural network architectures, as well as their modifications, may be used to recognize paralinguistic speech, notably emotion recognition. Convolutional Neural Networks (ConvNets) outperformed other architectures in terms of discriminative performance. The proposed SER system achieved 60.89 percent accuracy on the IEMOCAP dataset as a result of their experiment. Affective computing and sentiment analysis are critical for comprehending emotional intelligence. In [47] Cambria et al. proposed deep learning architectures to build a new version of SenticNet, a commonsense knowledge base for sentiment analysis. They have achieved state of art performance on 6 benchmark datasets for sentiment analysis. On IEMOCAP data, Yenigalla et al. discovered that a phoneme and spectrogram combination CNN model was the most accurate in distinguishing emotions [37]. When compared to existing state-of-the-art approaches, they improved overall accuracy and average class accuracy by more than 4%.

1.2.2 Video emotion recognition (VER)

The CNN and recurrent neural network (RNN) have been applied to several domains of computer vision in recent decades, due to the considerable development of deep-learning techniques. In particular, the CNN has performed admirably in various studies, including face recognition, object identification, and video emotion recognition. For example, to extract facial features from static facial images as visual features [7]. Ding et al. used pre-trained AlexNet Deep CNN model using the FER2013 dataset. He proposed that when the size of available facial emotion dataset is small, the CNN fine tuning strategy is an effective way to find good features not only for image emotion recognition but also video emotion recognition. The overall accuracy of the proposed approach was compared to the challenge “Video based emotion recognition challenge is one of the sub- challenges of Emotion Recognition in the Wild Challenge”. It obtained 53.9 % accuracy on the EmotiW2016 dataset, which was higher than the challenge baseline of 40.47 %. In [38] two deep-learning networks were utilized to discern emotions from a facial image sequence. The first network, which combines CNN and LSTM models, captures temporal changes in appearance. The second network, which detects emotions, uses facial landmark points as input to the deep network. Finally, using the SoftMax function, the integrated model calculates the probabilities for emotions via weighted summation. In [8], dynamic sequences, 34 motion characteristics of the head, eyebrows, eyes, and lips were captured as visual features from each facial images frame. The system combines face and speech information at the feature and decision levels to recognize emotion accurately and reliably. They demonstrated that their strategy is equivalent to the optimally expected classifiers using statistical significance analysis. [9] employed image frame footage to build 18 face animation characteristics using a facial feature tracking technique. Triple HMM was designed and trained to recognize the primary emotion. A Triple HMM can be seen as a collection of HMMs, one for each

data stream. They were able to obtain 91 % accuracy using this method. [39] presented a method for extracting valuable features from raw data automatically for VER. They demonstrated two deep network models that interact with one another. The first network was built using many frames appearances, whereas the second network used raw face landmark points to extract useful temporal geometric properties. Their integration method gives more accurate results than traditional methods, such as a weighted summation and a feature concatenation method

1.2.3 Audio-Video emotion recognition (AVER)

Recently, there has been a lot of interest in combining neural networks that analyze audio and visual data for emotion recognition. Mehmet et al. [11] have used feature level fusion of audio and video information on RAVDESS. Further for training, the method utilized generative adversarial network (GAN) and variational auto-encoder to generate more samples to avoid the problem of overfitting and achieved validation accuracy around 87%. In [12], Tzirakis et al. using audio-visual data to compare emotion recognition on two types of model fusion technique, feature levels and decision level and concluded that feature level improves the recognition rate to 90% from separate classifier accuracy of 70.9% and 85%, respectively on audio and facial expressions. The accuracy of decision-level bimodal fusion ranges from 84% to 89%. Hoch et al. [13] obtained a high 90.7% recognition rate. However, this system was subject-dependent and could only classify three emotions. Kim et al. [14] proposed multi-modal emotion detection using Deep Belief Network (DBN) models and demonstrate these models show improvement in emotion classification over baselines, Support Vector Machines (SVM) with radial basis function, that do not employ deep learning. They underlined the relevance of complicated non-linear feature interactions in emotion classification tasks when utilising deep learning techniques. In the process, they

considered utterances with labels from the following set (IEMOCAP): Angry, Happy, Neutral, Sad. Zhang [15] presented audio and visual networks initialized with the pre-trained AlexNet and C3D-Sports-1M. DBN model used to fuse audio-visual cues. They demonstrated that deep learning audio-visual feature representation outperforms handcrafted features. In [16] demonstrated an audio-visual emotion recognition system based on MDR features for audio and ridgelet coefficients for the face. Unlike the MFCC feature's final step of discrete cosine transform (DCT) computation, MDR applies four-dimensional three-point linear regressions to the spectro-temporal pattern. They have used extreme machine learning (ELM) for both audio and video pipelines. Noorizi et al. [17] introduced a fusion network for audio-visual emotion recognition. For the audio data, they computed MFCC, Filter bank energies and statistics-acoustic features. For the video, first represented the data using key-frame, and then facial landmarks, geometric relations i.e., distance and angles were computed. In [40] Jannat et al. described a method for identifying emotions based on video and audio data from a mobile phone. For the picture data, they employed Haar features to detect the face in each image and crop and scale it to 256x256 once the face has been spotted. They overlay the raw audio signal onto the 2D picture plane to pre-process the audio data. To match the face data, the final waveform image is similarly scaled to 256x256. Finally, they used Inception V3 CNN to train the network. Though they achieved high accuracy, the performance of the fusion (audio and video) method degrades significantly when compared to the image-only method. Jaratrotkamjorn et al. proposed a method for bimodal emotion recognition using the RAVDESS database's deep belief network (DBN) in the classification of eight fundamental emotions with face and speech information in [41], which produced an overall accuracy rate of 97.92 %, which is better than unimodal emotion recognition (facial or speech expression). Majumder et al. [43] proposed a novel feature fusion technique that works in

a hierarchical order, fusing the modalities two in two first, and then all three modalities afterwards. Their hierarchical fusion improves accuracy by up to 2.4 %.

1.4 Contributions

Though audio and visual utterances in video are temporally and semantically dependent to each other, modeling of temporal and contextual characteristics plays a vital role in understanding of conflicting or supporting emotional cues in audio-visual emotion recognition (AVER). Existing research in audio-visual emotion recognition, on the other hand, has largely focused on emotion detection exclusively from audio or video. However, because humans have a bimodal emotion production and perception system, visual and aural information are both required and complimentary for emotion detection. Despite its widespread use in healthcare, robotics, security, and surveillance, accurate and consistent audio and video emotion identification remains a challenging task.

To address the above-mentioned challenge in emotion recognition, in this thesis, we have researched on an end-to-end Audio Emotion Recognition (AER) pipeline where we have explored the state-of-the-art methods by extending the directional classical machine learning approach with handcrafted features to the use of deep learning for learning features embedding. In the similar fashion, we have explored the Video Emotion Recognition (VER) independently in the direction of handcrafted frame wise features extraction, temporal contextual windowed features, and learning embedding vectors. We have explored in achieving modularity of neural network architectures to capture the most accurate emotions from the audio and video feed.

Our research emphasizes on building an emotion recognition system applicable for real world applications with no constraint or specific requirements such as compulsion of wearing a device with sensors, recording in controlled environment, applying marker points, reading of specific annotation, number of user and linguistics constraints. We capture the emotion from data collected in a normal conversation without having any above-mentioned compulsion. We come up with a contextual temporal modular neural network for multi model end-to-end emotion recognition. The advantages of modular architecture allow us to do extensive research for each component of emotion cues. Contextual-temporal ways of extracting emotions capture long-term dependencies to exactly convey deeper emotions associated with the context such as sarcastic conversation, relational conversation, ambiance sensing, etc. We have done extensive research on audio emotion pipeline, video emotion pipeline, and bi-modal fusion techniques to incorporate contextual and temporal characteristics. Our experimental research shows that the proposed contextual temporal modeling achieves a state-of-art performance recognition rate for AVER.

Chapter 2 – AUDIO EMOTION RECOGNITION (AER)

Audio Emotion Recognition (AER) is a technique for attempting to understand human emotion and effective states using audio parameters extracted from speech. This takes advantage of the fact that the tone and pitch of a person's voice often convey underlying emotion. Recently, AER, which tries to discern emotion states from speech signals, has gotten a lot of attention. With the growing use of Virtual Personal Assistants such as Siri, Alexa, and Google Assistant in our daily interactions, they serve the purpose of quickly and accurately answering our questions and performing out our requests. These assistants are sophisticated enough to understand our directions, but they are unable to recognise and respond to our emotions. As a result, developing an efficient emotion recognition system that can both increase the skills of such assistants and transform the firm is vital. Speech is a rich, dense medium of communication capable of delivering information effectively. Its basic qualities can be exploited to extract organised information, such as emotions.

In the last few years, AER has gotten much attention due to abundance of data sources and development of technologies in the areas of machine centric systems such as human machine interaction in hospitals, public safety, market consumer section, mobile communication etc. So, there have been plenty of studies on standalone audio emotion recognition. A feature extraction stage is usually followed by a classification task in AER. Two main methods for detecting emotion from audio have been described in this thesis. The first method concentrates on extracting various low-level information directly from speech or audio. The retrieved audio features were sent to CNN, which uses audio features to predict emotion classifications. In the second method, we obtained embedding representations by feeding patches of spectrogram features to the convolutional neural network VGGish [18]. The learnt embedding features were fed into Recurrent Neural

Networks (RNN), which were used to describe temporal relationships between audio feature patches and accurately predict emotions.

2.1 Audio Specific low-level features extraction and CNN model

The creation of audio speaking sounds requires the presence of a certain quantity of energy at a specific frequency (vocal fold vibration) over a certain period of time. In research people consider these parameters as

- Loudness, or prominence (varying between soft and loud)
- Pitch (The higher the pitch, the faster the vocal folds vibrate, and narrower vocal tract vibrate faster.) In general, children and women have a smaller larynx and shorter vocal folds than men, and their voices are higher pitched.
- Duration (length), longer duration contains more contextual information

Previously, the most of AER systems focused on spectral or prosodic properties. Log-mel spectrograms, Mel frequency cepstral coefficient (MFCC) [6], and linear predictive cepstral coefficient are some of the most used spectral features (LPCC) [5]. Formants, glottal characteristics, speaking pace, fundamental frequency, loudness, harmonic consonance, and the Teager energy operator (TEO) are all prosodic properties.

2.1.1 Spectral features

Spectral features obtained directly from spectra; spectral characteristics represent phonetic information. Speech systems based on these characteristics have achieved remarkably high degree of accuracy when capturing speech in a clean environment [19]. The absolute values of the fast Fourier transform (FFT) are calculated from the input samples during cepstral analysis, which determines spectral features of voice (following the Hamming window

segmentation and weighting). The power spectrum is computed next, and then the natural logarithm is applied.

Linear predictive cepstral coefficient (LPCC): Cepstral coefficients generated from linear prediction (LP) analysis capture emotion-specific information expressed via vocal tract characteristics [5]. The essential premise of linear predictive analysis is that the n th speech sample, $s(n)$, may be predicted using a linear combination of its preceding samples, as indicated in the equation below.

$$s(n) \approx a_1s(n-1) + a_2s(n-2) + a_3s(n-3) + \dots + a_ps(n-p)$$

Over a speech analysis frame $a_1, a_2, a_3, \dots$ are considered to be constants. Predictor coefficients, also known as linear predictive coefficients, are a type of predictor coefficient.

Mel frequency cepstral coefficients (MFCCs): MFCCs are a normally employed feature in automated speech and speaker recognition techniques. They were first presented in the 1980s by Davis and Mermelstein and have remained the state-of-the-art ever since. Windowing the signal, applying the DFT, obtaining the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT are all part of MFCC's feature extraction approach.

2.1.2 Prosodic features

There are certain features that make speech sound more natural. Speech has some discernible melodic characteristics due to pitch fluctuation. The term "intonation" refers to the controlled variation of pitch. Speech has rhythmic characteristics because the sound units are shortened or stretched in accordance with the some underlying pattern. Linguistic stress can occur whenever some syllables or words are made more prominent than others. Melody, timing, and stress in speech provide information that improves the intelligibility

of spoken messages, allowing the listener to quickly separate continuous speech into phrases and words [20]. It can also express lexical and non-lexical information such as lexical tone, prominence, accent, and emotion, along with other things. Prosody is a term that refers to the features that cause us to experience these effects. Stress, tone, rhythm, and intonation are characteristics of prosodic cues. Each cue is a complex perceptual entity that is primarily expressed using three speech parameters: *pitch*, *energy*, and *duration*. Prosody can represent a variety of characteristics about the speaker or the utterance, including the speaker's emotional state, the form of the utterance (statement, inquiry, or command), the presence of irony or sarcasm, as well as emphasis, contrast, and focus. Formants, glottal characteristics, speaking pace, fundamental frequency, loudness, harmonic consonance, and the Teager energy operator (TEO) are all common prosodic qualities.

2.1.3 Extracted AER features from raw audio

We used audio to extract two types of information: spectral features and prosodic features. In speech emotion recognition, mel frequency cepstral coefficients (MFCCs) [6] and other spectral features are widely used [5] [6][16] [17]. We used 39-dimensional acoustic feature vector MFCCs as spectral features, which included 16 MFCCs, the normalized power, and first and second order derivatives. Formants, spectral Rolloff, speaking rate, fundamental frequency, spectral energy, short term energy, spectral centroid, and zero crossing rate were among the prosodic properties. Figure 2.1 depicts the spectral and prosodic properties obtained after pre-processing (emphasis, Frame blocking & windowing) while Table 2.1 lists the other audio-specific features.

FIGURE 2.1 ILLUSTRATION OF AUDIO FEATURES EXTRACTION.

Index	Name of features	Description
1.	Zero crossing rate	The frequency with which a signal's sign changes over the course of a frame.
2.	Spectral Centroid	obtained by evaluating the “center of gravity” using the Fourier transform’s frequency and magnitude information
3.	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale
4.	Fundamental frequency	The fundamental frequency of a speech signal, frequently abbreviated as F0, is the approximate frequency of spoken speech signals' periodic structure.
5.	Formants	Formants are frequency peaks in the spectrum which have a high degree of energy
6.	Speaking rate	The speed at which you talk, Speaking rate is often expressed in words per minute (wpm)
7.	Short term Energy	Sum of squares of the signal values, normalized by the respective frame length
8.	Spectral Energy	Short term energy calculated after DFT
9.	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.

TABLE 2.1 AUDIO SPECIFIC SPECTRAL AND PROSODIC FEATURES

2.1.4 CNN embedding model using extracted low-level AER features

The extracted audio features were fed to CNN, which uses audio features to predict emotional classifications. A CNN with a transfer encoding layer, 2-D convolution layers, max-pooling (after each convolution layer), dense layer 1 of size 512, and dense layer 2 (size equal to number of emotions classes) was fed into SoftMax to detect emotions from audio features shown in Figure 2.2. Rather of training the classifier from scratch using only audio input, we employed pre-trained encoded weights (shown on the left of Figure 2.2; it employs VGGish [18] embedding, which is also specified in detail in section 2.2.1). (shown on the right of Figure 2.2).

FIGURE 2.2 CONFIGURATION OF AUDIO CLASSIFIER NETWORK

2.2 Spatio-temporal embedding features

We use and analyse a variety of components in spoken communication without giving them much attention. Although the details are generally hard to communicate on paper, there are several conventional techniques of representing them in writing. But one of the major problems with the conventional approach is that much less effort has been made to make sense of how the extracted features change over time. Numerous studies in psychology support the importance of temporal information of the features, where it is demonstrated that the pattern of stress and intonation is highly associated with emotion status [21]. The majority of research on speech emotion recognition relies on low-level features extracted from the entire speech stream, which are then processed using a CNN model. CNNs perform remarkably well with huge and unanticipated image databases, but they demand a large volume of data and operating power. So, further in the research work we have embodied temporal correlation that can be revealed by using time-frequency analysis for non-stationary emotion-based speech signals.

2.2.1 Audio features embeddings

First, we extracted time-frequency spectrum information from the audio stream by splitting it into short time frames. We considered that audio frequencies remain constant over short periods of time while changes occur over longer periods of time. By performing Fourier transform over short-time frames, we obtained a good approximation of the frequency contours of the signal concatenated by adjacent frames. We computed 5 non overlapping mel-spectrogram patches for each audio file. The number of Mel filters was 64 and the frequency range of the Mel band is 125-7500 Hz and for each audio file, we generated a 5x64x96 tensor. Figure 2.3 illustrates the steps used to obtain temporal sequence of spectral features.

FIGURE 2.3 ILLUSTRATION OF OBTAINING AUDIO REPRESENTATION IN THE AUDIO STREAM.

After obtaining temporal sequence of spectral features, we extracted features embedding by using VGGish (Google audioset) trained model [18]. VGGish is a VGG model variant that generates high abstraction level feature embedding from audio recordings (AudioSet). In earlier research [18], a classification model has achieved better performances with VGGish generated embedding compared to those obtained with other encoders. In audio stream, to construct audio feature embedding from audio spectral characteristics, we utilized VGGish. The log mel spectrogram tensors are used by VGGish as an input. It generates a single embedding vector for every second of audio. To present 5x64x96 spectrogram tensor, we feed 5 frames one by one to generate temporal VGGish embedding features shown in Figure 2.4. We used the VGGish model from Google, which was trained on YouTube-8M.

FIGURE 2.4 THE VGGISH MODEL USED FOR GENERATING AUDIO EMBEDDINGS IN THE AUDIO STREAM

We employed Recurrent Neural Networks (RNN) to model contextual information from Vggish Embeddings. RNNs aggregate temporal information from previous audio embedding sequence to recognize them. Figure 2.5 shows a recurrent network topology composed of GRU layers. We proposed the integration of audio and visual streams in [section 4.2], where we included more detailed information.

FIGURE 2.5 RNN TOPOLOGY FOR AER USING AUDIO VGGISH EMBEDDING

Chapter 3 – VISUAL EMOTION RECOGNITION (VER)

Facial expressions are regarded as a fundamental nonverbal communication technique as well as the most intuitive way of expressing human emotions. Individuals make their sentiments known to others by using facial expressions. People can deduce an expresser's emotion through interpreting facial cues. Furthermore, emotions have been related to motivational inclinations, suggesting that perceivers may be able to predict the expresser's individual needs. Anticipating a perceiver's behavioural intentions can lead to a more suitable strategy to dealing with avoidance/withdrawal responses. As a result, identifying another's behavioural purpose through facial expressions is critical in social interaction.

Automatic emotion/sentiment identification varies over time, hence categorizing facial expressions into different kinds of emotion is vital. A system that can automatically understand human emotions is a desirable aim for a range of developing applications, including assistive technology, digital signage, audience measurement, and human-robot interaction [2][41][44]. The emotion recognition matrix is useful for determining how well assistive technology is perceived by users [3]. Due to a variety of challenges, designing such a system with a high recognition rate is a difficult challenge. [24][46] For example, scene illumination, reflection, refraction, camera viewpoint and quality of image, in some manner these can be addressed by applying pre-processing or acquisition setup. The diversity in appearance of various individuals, and even the same person exhibiting the same emotion, when combined with a bias factor, might alter the decision block of an emotion recognition system. Furthermore, an emotion recognition system based on facial information should account for intrinsic variances in how persons of varied ethnic, cultural, and geographical backgrounds execute the same expression [30]. To make the facial emotion recognition system to have human like interpretability and observance, it should also incorporate temporal aspect of emotion variation because emotions/change in

behavior are associated with contextual information varying over time. Keeping in mind the different aspect of challenges, there has been a lot of research addressing the above-mentioned problem to make the system robust and sophisticated [2, 3, 7, 8, 9, 21, 30, 38, 39, 41, 44].

For visual emotion recognition, deep neural networks such as convolutional neural networks (CNN) have been popular, for example, to extract facial features from static facial images as visual features [7]. Ding et al. used a long short-term memory (LSTM) to capture time dependency of facial characteristics over a period. In [8], dynamic sequences, 34 motion characteristics of the head, brows, eyes, and lips were captured as visual features from each facial image frame. In [22], facial feature tracking approach was used on image frame videos to obtain 18 facial animation parameters.

In our approach, first we propose to use 68 landmark features along with single shot detection (SSD) for face detection. We demonstrate superior performance of SSD compared to Faster R-CNN and YOLOv1 [10]. Further we have proposed to use the InceptionV3 as features embedding layer as it can achieve the lowest error rates compared to its contemporaries (VGGNet, Inception Networks (GoogLeNet/Inception v1)). Inception v3 is the third version of Google's Inception convolutional neural network, which was first unveiled at the ImageNet Recognition Challenge and has enhanced image analysis and object detection performance. Inception v3 has been trained on ImageNet database. We pushed temporal frames into inceptionv3 trained network to get penultimate visual representations. After getting the embedding features from Inceptionv3 we have used recurrent layer only to model contextual information.

3.1 Video specific low-level features and embeddings features to CNN model

The visual stream consists of three modules: (i) facial landmark detection, (ii) feature extraction module and (iii) a convolutional neural network (CNN) learning emotional embeddings. A frame containing the facial image is fed into the visual stream. Facial expressions are regarded as a fundamental nonverbal communication technique as well as the most natural way of expressing human emotions. [46] Face landmark detection (FLD), the technique of locating areas of interest in a picture of a human face, is the first step in detecting emotion from a facial image. Because of its appealing applications such as identifying emotions through facial gestures, assessing changing face appearance, and 3D visualization, FLD has lately experienced a significant development in the computer vision world [48]. Given an input image, FLD attempts to localize key points of interest inside the face. Meaning FLD algorithms try to automatically determine the positions of the facial important landmark points. The dominant points representing the distinctive location of a face component (e.g., eye corner, eyebrows, upper-lower lips, nose, jawline, mouth, etc.) or an interpolated point linking those dominant points around the facial components and facial contour are those important points. FLD can be seen a subproblem of facial shape prediction problem and starts with the detection of the face and its bounding box from the image. For detection of the box bounding the face, we have used a pre-trained single shot detection (SSD) network [10]. The use of SSD was motivated by its single stage object detection ability, i.e., it does not need object proposals in the generation step. This makes it, usually faster and more accurate than two stage approaches such as Faster R-CNN and YOLOv1. Facial landmarks are aimed at capturing rigid and non-rigid deformation of faces in a very compact description such as ethnicity,

expressions, gender, etc. Figure 3.1.1 below shows detection of bounding box and FLD points on RAVDESS sample dataset.

FIGURE 3.1.1 ILLUSTRATION OF VIDEO FEATURES EXTRACTION.

Further we employed HOG features to count the incidence of gradient orientations in localized regions of an image, given the face area, to reconstruct the form of facial muscles using edge analysis. The use of HOG to define facial expression characteristics outperform the majority of frequently used facial emotion detection algorithms, according to [23]. The HOG descriptor is based on pooling gradient directions across a pixel in a small spatial region known as a “cell”. We used facial landmark point extraction as features using dlib library [24]. Video feature extraction pipeline shown below in Figure 3.1.2.

FIGURE 3.1.2 ILLUSTRATION OF VIDEO FEATURES EXTRACTION.

Facial features are ordered and concatenated into a unique HOG histogram feature vector, which is used as input to CNN shown in Figure 3.1.3. The CNN architecture composed of 2-D convolutions followed by max pooling and a dense layer has been proposed to capture semantic meaning of emotions from feature vectors collected from image frame of video data. The activation functions of neurons were rectified linear units (ReLU) and exponential linear units (ELU) in alternate layers.

FIGURE 3.1.3 CONFIGURATION OF VIDEO CLASSIFIER NETWORK

3.2 Embedding features with Contextual modelling

3.2.1 Key frame Extraction

In the video stream, we first identify key frames conveying good visual representation. Algorithms for extracting key-frames from videos ensure that important aspects of the video, such as the chronological order, are maintained. We used a cluster-based frame extraction technique as cluster-based methods are mostly applied in video content analysis research [25]. Figure 3.2.1 illustrates the five step process to get sequences of key frames:

(a) RGB data is converted into LUV color space where absolute difference in LUV color space filter is applied; (b) brightness score filter applied (brightness tone measure the distinctiveness of region from its subsequent frame); c) entropy/contrast value filtering of retrieved frames; d) K-means clustering of frames utilizing image histogram; e) optimal frame selection from clusters based on and variance of Laplacian (image blur detection). Identified key frames give us visual temporal sequences for AVER. In video pipeline, for key frames selection we did experiments with different choice of keyframes for a same length of video segment. Varying from $2^1, 2^2, 2^3, \dots, 2^6$. We found that $K=16$ clusters preserved all the necessary inherited information, when we trained the video pipeline (i.e. increasing the number of keyframes doesn't further improve the video emotion pipeline accuracy).

FIGURE 3.2.1 VIDEO TEMPORAL KEY FRAMES EXTRACTION

3.2.2 Deep features embedding of temporal key-frame stream

We extracted feature embeddings using the Inceptionv3 architecture after extracting the temporal visual streams [26]. To get penultimate visual representations, we pushed temporal frames into an inceptionv3 trained network shown in Figure 3.2.2. Actually Inceptionv3 is the third version of Google's Inception convolutional neural network, which was initially unveiled during the ImageNet recognition Challenge and improves image

analysis and object recognition. Inception3 was trained on over a million images from the ImageNet database for 1000 different classes.

FIGURE 3.2.2 ILLUSTRATION OF LEARNING VISUAL REPRESENTATIONS FROM THE VIDEO STREAM.

Chapter 4 – AUDIO VISUAL EMOTION RECOGNITION (AVER)

The vast number of recent emotion detection algorithms have focused on recognizing emotions from either audio or visual information. Audio visual emotion recognition, on the other hand, necessitates both visual and auditory information due to the bimodal nature of human emotion generation and perception systems. Although several researchers have focused on merging both aural and visual signals (referenced in literature review section 1.2.3), the temporal aspects and contextual information sharing across two pipelines is frequently inadequate due to early fusion, missing temporal information in video-audio sequence.

A detailed description of two strategies for recognizing emotion from audio-visual data has been provided. The first way, as shown in Figure 4.1, focuses on directly collecting various low-level information from audio and video. CNN received the audio and video features and used them to generate emotional classifications using audio-visual features.

The second approach, which is our main method, contains the contextual information of the temporal frames depicted in Figure 4.2. We first construct key-frames of images/spectrograms from audio and video streams for data temporalization, and then extract feature embeddings as indicated in the previous section Figure 2.4 and Figure 3.2.2 for audio and video streams, respectively. These two embeddings were then fed into a recurrent neural network to obtain contextual data of temporal embeddings. Further we applied dimensionality equalization, concatenation and multi-class classification using SoftMax layer to predict different emotions labels associated with inputs.

4.1 Audio-visual modelling using low level features extraction

FIGURE 4.1 OVERVIEW OF PROPOSED ARCHITECTURE FOR AUDIO-VISUAL MODELLING USING LOW LEVEL FEATURES

Fusion module combines CNN outputs of the audio stream and video stream. Let a and b be the audio and video feature vectors, respectively, and y be the output emotional label. The output of the CNN is the emotion class conditional probability given the inputs and model parameters. We used probabilistic fusion to fuse the output of the two stream. The output emotion class label c^* is given by

$$c^* = \underset{c}{\operatorname{argmax}} \{ (p(y = c|a, A))^{w_1} (p(y = c|v, V))^{w_2} \}$$

where A and V are parameters of CNN in audio and video streams, respectively, and fusion parameters (w_1, w_2) are to be estimated using cross-validation. The maximum in the above equation is taken over all the emotion labels.

4.2 Audio-visual Contextual temporal modelling using features embeddings

FIGURE 4.2 OVERVIEW OF PROPOSED CONTEXTUAL TEMPORAL MODELLING FOR AUDIO-VISUAL EMOTION DETECTION PIPELINE

4.2.1 Gated Recurrent Units

Accurate capturing of contextual information temporally has improved speech recognition [27]. We have used Recurrent Neural Networks (RNN) to model contextual information from spatio-temporal audio and video streams. RNNs captures the temporal information from preceding video/audio frames and aggregate temporal information for recognition. Particularly, we used Gated Recurrent Units (GRUs) [28] that are more robust and better handle long term

dependencies of signals. GRU (Gated Recurrent Unit) is a recurrent neural network that tries to address the vanishing gradient problem which is inherent to standard RNN.

A GRU contains two gates: a reset gate that determines how the current input and prior memory should be combined, and an update gate that determines how much of the previous information should be kept. In essence, they are two vectors that determine what data should be sent to the output. They are remarkable for being able to be trained to remember knowledge from the past without wiping it out over time or removing information that is irrelevant to the predictions.

FIGURE 4.3 GATED RECURRENT UNIT CELLS AND GATES

It begins by applying the formula to calculate the update gate z_t for time step t :

$$z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1})$$

When x_t is fed into the network unit as an input, its own weight $W(z)$ is multiplied. The same may be said for h_{t-1} , which stores information for earlier $t - 1$ units and is multiplied by its own weight $U(z)$. Both outputs are combined, and the result is squeezed between 0 and 1 using a sigmoid activation function. The update gate aids the model in determining how much past feature information (from previous time steps) should be passed on to future time steps. This

is extremely useful since the model may choose to replicate all of the data from the past, eliminating the risk of vanishing gradients problem. In essence, the reset gate is utilized by the model to determine how much information from the past should be erased. We use the following formula to compute it:

$$r_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1})$$

This is the same formula as for the update gate. The difference is in the weights and how the gate is used, which we'll discuss later. We use the sigmoid function to multiply h_{t-1} and x_t with their respective weights, add the results, then apply the *sigmoid* function.

In addition, a novel memory content was created that would employ the reset gate to gather important data from the past. The following is how it is calculated:

$$h'_t = \tanh(Wx_t + r_t \odot Uh_{t-1})$$

Multiply the inputs x_t and h_{t-1} by a weight W and a weight U . Calculate the Hadamard (elementwise) product of the reset gate r_t and the input gate Uh_{t-1} . This will decide what data from earlier time steps should be discarded shown above in Figure 4.3. Considering our case of emotion recognition for determining the persons intensions from audio and video signals. Sometime a chunk of audio and frame of video doesn't exhibit correct prediction because of lack of sequential features in small intervals. Contrarily, many times long sequence of information carries unrelated or repeated information. Let's look at it from the perspective of the sentiment analysis issue, which is used to determine one's opinion about a book based on a review. The statement begins, "This is a motivating book that demonstrates..." and concludes, "I didn't quite enjoy the book since it catches too many unrealistic information." The last section of the paragraph is all that is required to establish the overall level of satisfaction from the book. As the neural network reaches the conclusion of the first phrase, it will learn to assign r_t vectors near to 0, clearing away the previous sentences and focusing solely on the last ones.

In the next step it sums up the results Wx_t and $r_t \odot U h_{t-1}$ then apply the nonlinear activation function $\tanh$. It can be seen Figure 4.3 that element-wise multiplication of h_{t-1} and r_t is applied and then sum the result with the input x_t . Finally, $\tanh$ is used to produce h'_t .

The network will calculate h_t — a vector that stores information for the current unit and transmits it down to the network as the final memory step at the current time step. The update gate is required in order to do so. It decides what to gather from the present memory content h'_t and what to collect from earlier steps h_{t-1} . The following is how it is done:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t$$

4.2.2 GRU to audio-visual embeddings

Input audio and video temporal frame embedding $v_{in} \in R^{N \times d_V}$ and $a_{in} \in R^{N \times d_A}$ are passed through two GRU to obtain context aware features v_1 and a_1 , respectively, as

$$a_1 = GRU(a_{in})$$

$$v_1 = GRU(v_{in})$$

Where d_A and d_V denotes the dimensions of signals in audio and video streams, respectively and N is the maximum number of chunk/utterances in the video data.

4.2.3 Fusion of Audio and Video embeddings

Let the outputs of GRUs be $v_1 \in R^{N \times d_1}$ and $a_1 \in R^{N \times d_2}$. In order for both audio and video streams to have same dimensions for fusion, we use fully connected layers. Let the outputs of fully connected layers be $v_2 \in R^{N \times d}$ and $a_2 \in R^{N \times d}$ with equal dimension d . Then,

$$a_2 = \tanh(W_A a_1 + b_A)$$

$$v_2 = \tanh(W_V v_1 + b_V)$$

where weights and biases for fully connected networks are $W_A \in R^{d_2 \times d}$, $b_A \in R^d$, $W_V \in R^{d_1 \times d}$, and $b_V \in R^d$.

The fusion module concatenates the audio and video outputs and combined in another $\tanh$ layer and passed through another GRU to obtain combined representations h .

$$h = GRU(\tanh(W_1[a_2; v_2] + b_2))$$

The fused representations h is passed through a softmax layer for multiclass emotion classification. The output activation y of the softmax layer is given by

$$y = \text{softmax}(W_2 h + b_2)$$

The output class label is assigned to the class receiving maximum activation.

Chapter 5 – Evaluation and Discussion

5.1 Datasets

Emotion recognition for model training and assessment relies heavily on audiovisual emotion databases. Although several emotion databases have been compiled for various applications, the large number of mechanisms of emotion databases focus on voice, facial expressions, texts, bio-signals, and body postures. However, in this thesis, we concentrated on emotion databases that contain both auditory and visual information, i.e., audiovisual information. We experimented our method for AVER on RAVDESS [29] and IEMOCAP [30] datasets which comprise of auditory and visual input information synced simultaneously.

5.1.1 RAVDESS

The RAVDESS database of emotional speech and music is a verified multimodal database. RAVDESS includes 7356 audio-only (16bit, 48kHz), audio-video (720p, 48kHz, 30fps), and video-only (720p, 48kHz, 30fps) files for the 24 actors (12 girls, 12 males) (no sound). The database collects two lexically matched sentences from 24 professional actors (12 females, 12 males) in a neutral North American accent. It comprises expressions including calm, happiness, sadness, anger, fear, surprise, and disgust, shown in Figure 5.1. There were 104 unique vocalizations made by all performers, with 60 spoken utterances.

On emotional validity, intensity, and authenticity, each file was scored ten times. 247 people that fit the profile of untrained adult study participants from North America submitted ratings. A total of 72 individuals contributed intra-participant test-retest data for the reliability task. Participants were asked to identify the expressed emotion in order to validate it. Neutral (calm + neutral), happy, sad, angry, fearful, surprise and disgust are the seven categories of main emotions that we attempt to anticipate. The baseline circumstances were calm and neutral,

while the remaining states were chosen to represent the set of six basic or fundamental emotions that are considered to be culturally universal.

FIGURE 5.1 RAVDESS: AUDIO-VISUAL DYNAMIC DATABASE, MULTIMODAL SET

5.1.2 IEMOCAP

In the study of expressive human communication, the IEMOCAP corpus can be quite useful. Gestures and voice from various subjects may be studied and used to model personal styles using the IEMOCAP database. Speaker-independent emotion identification systems can be built, for example, by learning inter-personal correlations.

IEMOCAP database contains audio-visual data from 10 speakers (five sessions of male-female pairs). In dyadic sessions, ten actors were captured (5 sessions with 2 subjects each). They were given three texts with strong emotional content to perform. In addition to the scripts, the participants were instructed to improvise dialogues in hypothetical settings that were meant to provoke certain emotions (happiness, anger, sadness, frustration and neutral state). It has about 12 hours of data, which includes video, audio, face motion capture, and text transcription. The audio-video data is broken down into short utterances ranging from 3 to 15 seconds in duration,

which are subsequently analyzed by three to four evaluators. After the collection of datasets, subjective test was conducted to evaluate the emotional content of dataset in terms of their emotional categories of the database. So, in total 6 persons have evaluated the dataset. Anger, sadness, happiness, frustration, excitement, surprise, and neutral states were all included in the dataset. Figure 5.2 below shows a representative sample of data.

FIGURE 5.2 QUALITATIVE RESULTS ON IEMOCAP

5.2 Description of design

In the dataset, we have varying sizes of video files from 3s to 15s and we segmented the file in segments of 5s as an analysis segment. Segment less than or equal to 5s was padded/clipped at 5 seconds. In visual pipeline, for key frames selection, we did experiments with different choice of keyframes for a same length of video segment. Varying from $2^1, 2^2, 2^3, \dots, 2^6$ for clusters,

We found that $K=16$ clusters preserved all the necessary inherited information, when we trained the video pipeline (i.e., only increasing the number of frames doesn't further improve the video emotion pipeline accuracy rather representative frames (keyframes) makes significant improvement). In audio pipeline we are using audioset VGGish embedding topology, VGGish takes a spectrogram "patch" of 96 frames x 64 mel channels as input. This corresponds to 96 successive columns from a spectrogram calculated with 10ms frames (based on 25ms windows), so it covers ~ 0.96 s of input per patch (actually 0.975s if you include the whole of the 25 ms window "at the end"). Considering the entire 5 seconds length segment, we have 5 (96x64) spectrogram patterns. If the size of file is smaller, we have padded it and similarly bigger segment of chunk has been clipped to get 5x96x64. Before the fusion, Dense connected layer has been used for features dimensionality equalization. Features inherited from audio and video domain doesn't learn temporal aspect in its individual modularity if they are fused in the early stage which leads to performance degradation for multimodal audio-visual emotion detection. Idea behind late fusion was to incorporate inter and intra domain contextual information propagation.

5.3 Experimental results

The proposed method was first trained for audio and video separately on RAVDESS and IEMOCAP dataset. After getting the best parameters for unimodal recognition, we trained audio-visual embeddings simultaneously using the proposed architecture. For the training, we used 10-fold cross-validation, which divides the original sample into 10 equal-sized subsamples at random. A single subsample from the 10 is kept as validation data for testing the model, while the remaining 9 are used as training data. The cross-validation process is then performed 10 times (the folds), with each of the 10 subsamples serving as validation data exactly once. The 10-fold results were then averaged to generate model accuracy. This method has the advantage of using all observations for both training and validation, and each observation is only utilised for validation once. We selected a batch size of 48 for training, as well as a stochastic gradient descent with 0.9 stochastic momentum and an adaptive learning rate. Validation loss of the network saturated to the optimal value after training epoch 86.

Tables 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6 provide a comparison of the results to the state of the art. Our findings on temporal hand-crafted features (frame of log mel spectrogram and images key-frame) with contextual modelling and temporal embedding feature with contextual modelling have been presented. The table demonstrate the importance of embeddings, temporal features, and contextual modelling. For both audio and visual datasets, our proposed technique yields state-of-the-art accuracies. MDR, MFCC, spectrogram, and DCT are examples of handcrafted audio features, while HOG, landmark features, sensor fused spatial features, geometric relations, and other visual features are examples of handcrafted visual features. We compared our methods to the state-of-the-art methods listed in the tables below, and the proposed method greatly outperformed them. Both handcrafted and embedded characteristics are used to test our

method. Throughout each experiment, the embedded features outperformed the non-embedded features.

5.3.1 Results and comparison of AER

On the RAVDESS and IEMOCAP datasets, we compared the performance of our AER methods to that of existing AER approaches. The accuracy of our methodologies to the state-of-the-art methodology is shown in Tables 5.1 and 5.2. Tomba et al. [31] proposed SVMs and ANN technique, which is shown in Table 5.1. They found that both gave excellent outcomes, with ANNs scoring slightly higher than SVMs. Mustaqeem et al. [32] proposed deep stride convolutional neural network (DSCNN) architecture improves accuracy by 7.85 percent and 4.5 percent on the IEMOCAP and RAVDESS datasets, respectively. In [34], Fayek et al. looked at various deep learning architectures for the Speech Emotion Recognition (SER) problem. CNN beat alternative architectures in terms of discriminative performance, according to experiments. In their trial, the proposed SER system achieved 60.89 percent accuracy on the IEMOCAP dataset. In [37] Yenigalla et al. combined spectrogram and phoneme-based CNN model that is able to achieve a 4% higher average class accuracy compared to the existing state of art method. We have presented our results in the last two columns of Table 5.1 and 5.2 which shows that both of our models work adequately, as evidenced by the results, but the second method beats all others. That is, when trained on RNN network with sequential CNN Embedding inputs in proposed method 2, the combination of spectrogram and VGGish embedding features (5 frames one by one to generate temporal VGGish embedding features shown in Figure 2.4) leads to an improvement in AER.

TABLE 5.1 ACCURACIES OF EMOTION RECOGNITION USING AUDIO FEATURES ON RAVDESS DATASET

RAVDESS Audio emotion recognition (AER)		
Approach	Input	Avg. Accuracy
Stress Detection Through Speech Analysis [31]	Spectrogram	78.75
Mustaqeem [32]	Spectrogram	70
Mustaqeem [32]	Clean Spectrogram	81
Recognizing Emotion from Singing and Speaking Using Shared Models [33]	Spectrogram	83.12
Recognizing Emotion from Singing and Speaking Using Shared Models [33]	openSMILE(LLD & MFCC)	81.61
Proposed contextual method 1*	Spectrogram	81.2
Proposed contextual method 2*	Spectrogram to VGGish Embedding	86.3

TABLE 5.2 ACCURACIES OF EMOTION RECOGNITION USING AUDIO FEATURES ON IEMOCAP DATASET

IEMOCAP Audio emotion recognition (AER)		
Approach	Input	Avg. Accuracy
Fayek et al. [34]	Spectrogram	60.89
Luo et al [35]	Spectrogram	63.98
Tripathi et al. [36]	Spectrogram	62.72
Yenigalla et al. [37]	Phoneme and spectrogram	73.9
Mustaqeem et al. [32]	Spectrogram	77
Mustaqeem et al. [32]	Clean Spectrogram	84
Proposed contextual method 1*	Spectrogram	72.2
Proposed contextual method 2*	Spectrogram to VGGish Embedding	84.12

5.3.2 Results and comparison of VER

We have taken into account the proposed approach to evaluate our VER, and Table 5.3 and Table 5.4 shows a comparison of VER performance validated on the IEMOCAP and RAVDESS datasets. For VER, our proposed strategy outperforms the state-of-the-art method.

For IEMOCAP dataset, all the results so far have been reported utilizing IEMOCAP speech data or MOCAP in the existing publications (motion capture head or face key points data). Motion capture (MOCAP) is a sort of technology that uses sensor equipment to record the changing pattern of a participant's movement in real time so that the participant's movement track can be differentiated. However, none of the studies used raw video from the IEMOCAP data. As a result, Table 5.3 only depicts a VER ablation research utilizing IEMOCAP data.

We compared our VER method using the RAVDESS dataset to the previous approaches in Table 5.4. The results showed that the proposed strategy outperforms the prior study overall. Human emotion recognition based on the weighted 53 integration approach combining image sequences and auditory characteristics obtained 84.6 % accuracy, according to Byun and Lee [38]. Jung et al. developed two deep network models that interact with each other in their paper [39]. The first network was created by combining numerous frames appearances, whereas the second network was created by extracting useful temporal geometric characteristics from raw facial landmark points. Their approach of integration produces more precise findings than standard methods. The overall accuracy of our proposed approach, which combines the idea of CNN features embeddings with RNN to model contextual information to recognise emotions, is the optimum. We trained our technique with the same architecture but with I3D embedding as an exploration of features embedding methods. We found that while I3D gives additional information of low-level motion features, it cannot outperform Inception V3 embedding combined with RNN network design.

TABLE 5.3 EMOTION DETECTION USING VISUAL STREAM ON IEMOCAP DATASET

IEMOCAP Video emotion recognition (VER)		
Approach	Input	Avg. Accuracy
Our proposed method 1*	Face detection (SSD) + Key-points + HOG	51.2
Our proposed method 2*	Spatio-temporal embedding tensor (Inception V3)	50.6

TABLE 5.4 EMOTION DETECTION USING VISUAL STREAM ON RAVDESS DATASET

RAVDESS Video emotion recognition (VER)		
Approach	Input	Avg. Accuracy
Weighted joint fine-tuning [38]	CNN-LSTM features + FLD	84.6
Joint fine tuning [39]	Deep network + FLD	82.1
Our proposed method 1*	Face detection (SSD) + Key-points + HOG	71.2
Our proposed method 2*	Spatio-temporal embedding tensor (Inception V3)	86.6
Our proposed method 2* (with I3D)	Spatio-temporal embedding tensor I3D	85.2

5.3.3 Results and comparison of AVER

The main purpose of this research thesis is to demonstrate that adopting the proposed multi model method increases the performance of an emotion identification system in the discrimination of 8 emotions. Bimodal classifiers trained on both facial expressions and speech data outperform single-modal classifiers, as expected. Table 5.5 and Table 5.6 shows the performance of systems that were trained using a mix of the two modalities (audio and video) on RAVDESS and IEMOCAP dataset. Compared against previous state of art method where

[13] used a generative adversarial network (GAN) and a variational auto-encoder to produce more samples to minimise overfitting and achieved validation accuracy of roughly 87% on the RAVDESS dataset, as shown in Table 5.5. The proposed technique by Jannat et al. [40] obtained good accuracy, however the fusion (audio and video) method's performance degrade significantly when compared to the image-only method. Jaratrotkamjorn et al. method for bimodal emotion recognition in the classification of 8 basic emotions using the RAVDESS database's deep belief network (DBN) produced an overall accuracy rate of 97.92 %, which is better than unimodal emotion recognition (facial or speech expression). Majumder et al. [43] developed feature fusion technique that works in a hierarchical order improves by up to 2.4 %. Samarth et al. [36] attempted to use neural networks' effectiveness to perform multimodal emotion recognition on the IEMOCAP dataset, and they achieved an overall accuracy of 71.04 %. AVER outperforms the VER and AER scenarios due to better utilisation of spatiotemporal information residing in the audio-visual data, as expected with our proposed approach that combines the idea of CNN features embeddings with RNN to model contextual information to recognise emotions. Also on both the IEMOCAP and RAVDESS datasets, our multi-model AVER beat the state-of-the-art approach..

TABLE 5.5 EMOTION DETECTION USING FUSION OF AUDIO AND VISUAL STREAMS ON RAVDESS DATASET

RAVDESS Audio-visual emotion recognition (AVER)		
Approach	Input	Avg. Accuracy
Mehmet et al. [11]	DNN	87.33
Ubiquitous Emotion Recognition Using Audio and Video Data [40]	Handcrafted features	96.09
Bimodal Emotion Recognition using Deep Belief Network [41]	DBN and handcrafted features	97.92

Our proposed multi-model method 1*	Handcrafted features	93.2
Our proposed multi-model method 2*	Embedded features	98.14

TABLE 5.6 EMOTION DETECTION USING FUSION OF AUDIO AND VISUAL STREAMS ON IEMOCAP DATASET

IEMOCAP Audio-visual emotion recognition (AVER)		
Approach	Input	Avg. Accuracy
3DCLS model to establish spatio-temporal emotion recognition [42]	Open smile and 3D-CNN	69.49
N. Majumder et al.[43]	Handcrafted features	76.8
Samarth Tripathi et al. [36]	Handcrafted features	71.04
Our proposed multi-model method 1*	Handcrafted features	71.2
Our proposed multi-model method 2*	Embedded features	86.6

5.4 Discussion

In our proposed contextual method 1, we focused mostly on handcrafted features pertaining to audio and video emotion detection throughout our preliminary study for audio and video emotion recognition. For the classification objective, we employed CNN network topology to train our emotion recognition classifier. The output of the CNN from audio and video features streams were fused together using probabilistic fusion.

On the RAVDESS dataset, our proposed contextual method 1 achieves an accuracy of around 81.2 %, which is comparable to prior state-of-the-art method accuracy, whereas on the IEMOCAP dataset, we were unable to achieve similar accuracy to prior art, achieving an accuracy of around 72.2 %, which is 12 % less than the prior state-of-the-art method. For the RAVDESS dataset, VER recommended contextual technique 1 has an accuracy of roughly 71.2 percent, which is significantly lower than state-of-the-art accuracy, implying that VER's

hand-crafted features are not as good as deep learning-based Embedding. We were only able to get roughly 51 % with the EMOCAP dataset. We found that the model performs poorly on “surprise” with only 36% accuracy, and often misidentifies these samples as “anger” and “disgust”. The accuracy rate for the remaining emotions is above 52%. One thing to note is that none of the studies used the IEMOCAP data's raw video. As a result, [Table 5.3] only shows a VER ablation study based on IEMOCAP data. Overall, AER outperforms VER in the first method, although AVER as a whole outperforms each stream individually. We got 93.2 percent accuracy, which is little less than the previous AVER technique.

Later in the research, we concentrated on end-to-end AVER, where we researched cutting-edge methodologies by replacing handmade features with deep learning-based feature extraction embedding. In general, CNN is good at picking up spatial information, but RNN is good at understanding temporal dynamics in data. As a result, we've opted to use an RNN network with sequential CNN Embedding inputs to improve accuracy. We used pre-trained CNN networks for the features embedding layer, such as VGGish for audio and InceptionV3 for video. To demonstrate the benefits of our proposed method 2, we directly compare our results to those of earlier studies on these datasets that used hand-crafted features and a CNN technique. Our model outperformed the state-of-the-art technique on both the IEMOCAP and RAVDESS datasets. We've seen a similar pattern with IEMOCAP datasets, where the Surprise emotion class gets mixed up with the Angry and Disgust emotion class. With a proposed approach that combines the idea of CNN features embeddings with Recurrent Neural Networks (RNN) to model contextual information to recognise emotions, we demonstrated that we achieved state-of-the-art results.

Chapter 6 – Future work

In order to take this thesis, work forward, our future work would emphasis in accomplishing the following objectives:

- Extending our fusion approach to utilize an attention mechanism to explicitly model interactions between the two streams.
- To improve the proposed technique even more, we'll investigate a reinforcement learning-based approach that can help video transcription algorithms using emotion/sentiment cues.
- we would explore some heuristics to extract more on the state-of-the-art visual feature extraction technique such as I3D (spatial-temporal feature extractor) and compare with our inception embedding features extraction pipeline.

6.1 Attention fusion

To fuse data from different sources, previous research approaches employed tensor concatenation. These methods find joint features as the concatenation of inputs; and several methods such as [11], [12] even eliminate the modalities' temporal connection. Because such approaches were incorporated from the start, interactions inside the modal are repressed, causing the modalities to lose their overall correlation or even temporal interdependence.

An ideal fusion method would synergistically combine the two modalities and ensure that the resultant product reflects the salient features of input modalities by using the complementary information of multimodal data. Attention module would also have the ability to capture the global and local connections. So, as part of our future research, we're focusing on the attention process for audio and video fusion. Attention fusion has been a subject of growing interest of video caption creation (using autoencoder for audio and video by sharing weight and memory). The first appearance of an attention mechanism was in the work of image caption processing,

where it was used to dynamically pick more essential characteristics and disregard redundant information in order to obtain the most accurate description [44].

6.2 Attention Mechanism:

By integrating different modalities, multimodal fusion aims to take use of the complimentary information included in multimodal data. Parallelizable computation, lightweight structure, and the capacity to capture both long-range and local correlations are all characteristics of neural networks with attention. Assessing the connection between audio and visual modules would be at the root of the attention mechanism approach. The score k is generated by the compatibility function $g(v_2, a_2)$, which might represent the dependency between v_2 and a_2 . Function softmax transforms the score to a probability, which is then utilised as a weight.

$$k = [g(v_2, a_2)]_{n=1}^N$$

$$P(y|v_2, a_2) = \text{softmax}(k)$$

$$\mathcal{S} = \sum_{n=1}^N P(y = n|v_2, a_2) \cdot v_2$$

Where $v_2 \in R^{N \times d}$, $a_2 \in R^{N \times d}$ and k is represented as a vector of n correlation scores. We obtain a probability distribution for attention $P(y|v_2, a_2)$ by applying k to the function softmax. And for query a_2 , $\mathcal{S}$ is the output vector.

Choosing distinct compatibility functions $g(v_2, a_2)$ in the attention mechanism will yield varied experimental outcomes. Different types of compatibility functions also lead to different types of attention processes. The attention mechanism of our technique employs the dot product attention compatibility function as a future work development in this thesis work:

$$g(v_2, a_2) = \langle w^{d_1} v_2, w^{d_2} a_2 \rangle$$

$\langle \circ, \circ \rangle$ indicates the inner product, where w^{d_1} and w^{d_2} are learnable parameters.

6.3 Visual features exploration

Further in the research we would apply some heuristic approach to explore more on the state-of-the-art visual feature extraction technique such as I3D (Inflated 3D ConvNet: spatio-temporal feature extractor), reinforcement learning and compare with our inception embedding features extraction pipeline.

6.3.1 I3D: Spatio-temporal feature extractor

The temporal element is one of the key distinctions between information in a single image and information in a video. And 3D ConvNets, which are similar to regular convolutional networks but with spatiotemporal filters, are seen as a natural approach to video modelling. However, there is an underlying issue with 3D ConvNets: the architecture prevents them from becoming more complex. I3D [45] simply converts successful image (2D) classification models into 3D ConvNets rather than redoing the procedure for spatiotemporal models. The method begins with a two-dimensional design and inflates all of the filters and pooling kernels. They provide an additional dimension to be considered as time progresses by inflating them. Filters in 2D models are square $N \times N$, but they become cubic $N \times N \times N$ when they are inflated..

In video processing, I3D is one of the most often used feature extraction algorithms. It's used to capture low-level motion characteristics and extract spatial features from neighbouring frames. To acquire high-level temporal characteristics, I3D's output feature vectors are fed into a GRU network. For extracting spatial information and collecting low-level motion data among adjacent frames, I3D has been used. The underlying I3D model's output feature will then be fed into the GRU network, which is primarily responsible for modelling high-level spatial features. As a consequence, input video characteristics may be learnt and represented

effectively, and our proposed approach is capable of learning both low-level and high-level features. Experimental results of I3D used in VER shown in the Table 5.4, what we have observed that I3D provides an additional information of low-level motion features which can be combined with our proposed VER pipeline to capture better insight.

REFERENCE

- [1] T. Partala and A. Kallinen, “Understanding the most satisfying and unsatisfying user experiences: Emotions, psychological needs, and context,” *Interact. Comput.*, vol. 24, no. 1, pp. 25–34, Jan. 2012, doi: 10.1016/j.intcom.2011.10.001.
- [2] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, “Audio-Visual Emotion Recognition in Video Clips,” *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, Jan. 2019, doi: 10.1109/TAFFC.2017.2713783.
- [3] Zhihong Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009, doi: 10.1109/TPAMI.2008.52.
- [4] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974, doi: 10.1121/1.1914702.
- [5] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition,” *IEEE Trans Acoust Speech, Signal Process*, vol. ASSP-28, no. 4, pp. 357–366, 1980.
- [6] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 223–227, 2014.
- [7] W. Ding *et al.*, “Audio and face video emotion recognition in the wild using deep neural networks and small datasets,” *ICMI 2016 - Proc. 18th ACM Int. Conf. Multimodal Interact.*, no. October, pp. 506–513, 2016, doi: 10.1145/2993148.2997637.

- [8] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech," *Multimed. Tools Appl.*, vol. 49, no. 2, pp. 277–297, 2010, doi: 10.1007/s11042-009-0344-2.
- [9] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, no. 10–12, pp. 1913–1920, Jun. 2008, doi: 10.1016/j.neucom.2007.07.041.
- [10] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2.
- [11] M. Dedeoglu, J. Zhang, and R. Liang, "Emotion classification based on audiovisual information fusion using deep learning," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 2019-Novem, pp. 131–134, 2019, doi: 10.1109/ICDMW.2019.00029.
- [12] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1301–1309, 2017, doi: 10.1109/JSTSP.2017.2764438.
- [13] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. II, p. ii/1085-ii/1088 Vol. 2, 2005, doi: 10.1109/ICASSP.2005.1415597.
- [14] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 3687–3691, 2013, doi: 10.1109/ICASSP.2013.6638346.
- [15] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning Affective Features with

- a Hybrid Deep Model for Audio-Visual Emotion Recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, 2018, doi: 10.1109/TCSVT.2017.2719043.
- [16] M. S. Hossain and G. Muhammad, “Emotion recognition using deep learning approach from audio–visual emotional big data,” *Inf. Fusion*, vol. 49, no. November 2017, pp. 69–78, 2019, doi: 10.1016/j.inffus.2018.09.008.
- [17] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, “Audio-Visual Emotion Recognition in Video Clips,” *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, 2019, doi: 10.1109/TAFFC.2017.2713783.
- [18] S. Hershey *et al.*, “CNN architectures for large-scale audio classification,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 131–135, 2017, doi: 10.1109/ICASSP.2017.7952132.
- [19] J. Přibíl, A. Přibílová, and D. Ďuračková, “Evaluation of spectral and prosodic features of speech affected by orthodontic appliances using the GMM classifier,” *J. Electr. Eng.*, vol. 65, no. 1, pp. 30–36, 2014, doi: 10.2478/jee-2014-0004.
- [20] L. Mary and B. Yegnanarayana, “Extraction and representation of prosodic features for language and speaker recognition,” *Speech Commun.*, vol. 50, no. 10, pp. 782–796, 2008, doi: 10.1016/j.specom.2008.04.010.
- [21] I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” *J. Acoust. Soc. Am.*, vol. 93, no. 2, pp. 1097–1108, Feb. 1993, doi: 10.1121/1.405558.
- [22] M. Song, M. You, N. Li, and C. Chen, “A robust multimodal approach for emotion recognition,” *Neurocomputing*, vol. 71, no. 10–12, pp. 1913–1920, 2008, doi:

- 10.1016/j.neucom.2007.07.041.
- [23] P. Carcagnì, M. Del Coco, M. Leo, and C. Distantè, “Facial expression recognition and histograms of oriented gradients: a comprehensive study,” *Springerplus*, vol. 4, no. 1, 2015, doi: 10.1186/s40064-015-1427-3.
 - [24] Y. Wu and Q. Ji, “Facial Landmark Detection: A Literature Survey,” *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, 2019, doi: 10.1007/s11263-018-1097-z.
 - [25] Wikipedia, “Key_Frame.”
 - [26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
 - [27] M. W. Kinch, M. J. C. Wim, and S. Keates, “The Benefits of Contextual Information for Speech Recognition Systems,” in *2018 10th Computer Science and Electronic Engineering Conference, CEEC 2018 - Proceedings*, 2019, doi: 10.1109/CEEC.2018.8674204.
 - [28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” pp. 1–9, 2014.
 - [29] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.
 - [30] C. Busso *et al.*, “IEMOCAP: interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

- [31] K. Tomba, J. Dumoulin, E. Mugellini, O. Abou Khaled, and S. Hawila, “Stress Detection Through Speech Analysis,” in *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, 2018, pp. 394–398, doi: 10.5220/0006855803940398.
- [32] Mustaqeem and S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors (Switzerland)*, vol. 20, no. 1, 2020, doi: 10.3390/s20010183.
- [33] B. Zhang, G. Essl, and E. M. Provost, “Recognizing emotion from singing and speaking using shared models,” in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, 2015, pp. 139–145, doi: 10.1109/ACII.2015.7344563.
- [34] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, vol. 92, pp. 60–68, Aug. 2017, doi: 10.1016/j.neunet.2017.02.013.
- [35] D. Luo, Y. Zou, and D. Huang, “Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition,” in *Interspeech 2018*, 2018, pp. 152–156, doi: 10.21437/Interspeech.2018-1832.
- [36] S. Tripathi, S. Tripathi, and H. Beigi, “Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning,” Apr. 2018.
- [37] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, “Speech Emotion Recognition Using Spectrogram and Phoneme Embedding,” in *Interspeech 2018*, 2018, pp. 3688–3692, doi: 10.21437/Interspeech.2018-1811.
- [38] S.-W. Byun and S.-P. Lee, “Human emotion recognition based on the weighted

- integration method using image sequences and acoustic features,” *Multimed. Tools Appl.*, Sep. 2020, doi: 10.1007/s11042-020-09842-1.
- [39] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2983–2991, doi: 10.1109/ICCV.2015.341.
 - [40] R. Jannat, I. Tynes, L. La Lime, J. Adorno, and S. Canavan, “Ubiquitous Emotion Recognition Using Audio and Video Data,” in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*, 2018, pp. 956–959, doi: 10.1145/3267305.3267689.
 - [41] A. Jaratrotkamjorn and A. Choksuriwong, “Bimodal Emotion Recognition using Deep Belief Network,” in *2019 23rd International Computer Science and Engineering Conference (ICSEC)*, 2019, pp. 103–109, doi: 10.1109/ICSEC47112.2019.8974707.
 - [42] G. Xu, W. Li, and J. Liu, “A social emotion classification approach using multi-model fusion,” *Futur. Gener. Comput. Syst.*, vol. 102, pp. 347–356, Jan. 2020, doi: 10.1016/j.future.2019.07.007.
 - [43] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, “Multimodal sentiment analysis using hierarchical fusion with context modeling,” *Knowledge-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018, doi: 10.1016/j.knosys.2018.07.041.
 - [44] W. Hao, Z. Zhang, and H. Guan, “Integrating both visual and audio cues for enhanced video caption,” *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 6894–6901, 2018.
 - [45] J. Carreira and A. Zisserman, “Quo vadis, action recognition,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6299–6308, 2017.

- [46] Wu, Y., & Ji, Q. (2019). Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision*, 127(2), 115–142. <https://doi.org/10.1007/s11263-018-1097-z>
- [47] Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. *International Conference on Information and Knowledge Management, Proceedings*, 105–114. <https://doi.org/10.1145/3340531.3412003>
- [48] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 Faces In-The-Wild Challenge: database and results. *Image and Vision Computing*, 47, 3–18. <https://doi.org/10.1016/j.imavis.2016.01.002>