KLE Society's
KLE Technological University, Hubballi.

A Minor Project Report

On

Audio-Visual Emotion Recognition

submitted in partial fulfillment of the requirement for the degree of

Bachelor of Engineering

In

School of Computer Science and Engineering

Submitted By

Raghvendra Bhatt	01fe19bcs181
Aditya Vikram	01fe19bcs220
Rishab Jain	01fe19bcs228
Harshita Hiremath	01fe19bcs235

Under the guidance of
Dr. Shankar G

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

HUBBALLI – 580 031

Academic year 2022-23

KLE Society's
KLE Technological University, Hubballi.

2020 - 2021

SCHOOL OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Minor Project entitled **"Audio-Visual Emotion Recognition"** is a bonafied work carried out by the student team: Raghvendra Bhatt:01fe19bcs181, Aditya Vikram:01fe19bcs220, Rishab Jain:01fe19bcs228, Harshita Hiremath:01fe19bcs235, in partial fulfillment of completion of Sixth semester B. E. in Computer Science and Engineering during the year 2022-2023. The project report has been approved as it satisfies the academic requirement with respect to the project work prescribed for the above said program.

Guide	Head, SoCSE
Dr. Shankar G	Dr. Meena S. M

External Viva -Voce:

Name of the Examiners	Signature with date
1.	
2.	

Acknowledgement

We would like to thank our faculty and management for their professional guidance towards the completion of the project work. We take this opportunity to thank Dr. Ashok Shettar, Vice-Chancellor, Dr. N.H Ayachit, Registrar, and Dr. P.G Tewari, Dean Academics, KLE Technological University, Hubballi, for their vision and support.

We also take this opportunity to thank Dr. Meena S. M, Professor and Head, SoCSE for having provided us direction and facilitated for enhancement of skills and academic growth.

We thank our guide Dr.Shankar G, Professor, SoCSE for the constant guidance during interaction and reviews.

We extend our acknowledgement to the reviewers for critical suggestions and inputs. We also thank Project Co-ordinator Mr. Uday N. Kulkarni, Mr. Guruprasad Konnurmath and all the reviewers for their support during the reviews and course of completion.

We express gratitude to our beloved parents for constant encouragement and support.

Raghavendra Bhat - 01FE19BCS181
Aditya Vikram - 01FE19BCS220
Rishab Jain - 01FE19BCS228
Harshita Hiremath - 01FE19BCS235

ABSTRACT

Facial Emotion Recognition (FER) is perhaps the most recent challenge in human/computer communication. The vast majority of the past work on emotion recognition is on extracting features from visual or audio modalities independently or as a combination of multiple modalities using various learning techniques. In this work, we study FER in a cross-domain few-shot learning setting, where only a few frames of novel classes from the target domain are required as a reference. In particular, we aim to identify unseen emotions, in a 4-way one shot learning fashion. Our work follows few-shot learning principles that enables learning of an embedding network, which later used to recognize unseen emotions. We make use of multi-modal encoder architecture that is capable of processing audio and video inputs. Our embedding network is trained on four emotions, namely, happy, sad, angry and surprised, and tested on four unseen emotions, namely contempt, neutral, disgust and fear. During training, the goal is to construct a rich feature space of emotions, which enables the embedding network to better differentiate one emotion from the other. At test time, it uses that knowledge to differentiate between unseen emotions and recognize them.

Keywords : *Audio-visual, Emotion Recognition, Few-shot learning.*

Contents

Chapter 1

INTRODUCTION

Emotion, typically imparts our passionate state or demeanor to other people. They are communicated through verbal and non-verbal correspondence.Emotion recognition is a way of identifying human feelings, using deep learning based techniques.Individuals fluctuate broadly in their precision at perceiving the feelings of others. Utilization of innovation to assist individuals with emotion recognition is an active area of research. For the most part, the innovation works best assuming it involves various modalities in setting. Looks are for the most part broken down by utilizing cameras to recognize faces and catch ongoing human reactions to true situations. Every look that an individual showcases makes the facial muscles move in an unexpected way, and this makes the method involved with deciding a feeling more straightforward for the profound learning AI-based calculations. Until this point in time, the most work has been directed on recognizing the emotion of looks from video, expressed articulations from sound, composed articulations from text, and physiology as estimated by wearables.

Facial expression recognition (FER) has received a lot of interest in recent decades due to its wide range of applications in human-robot interaction, online education, driver monitoring, and so on (Corneanu et al. 2016).Facial expressions are often grouped into seven main expressions, according to Ekman and Friesen's study (Ekman and Friesen 1971), which include happiness, sadness, disgust, rage, fear, surprise, and neutral. The classification of these predefined basic expressions has been the subject of previous work on FER. As a result, many basic expression datasets have been gathered (Zhao et al. 2011), and significant progress has been achieved in addressing huge facial appearance changes caused by identity, position, occlusion, illumination, and other factors.

Regrettably, these fundamental expressions are unable to capture the whole range of human emotions found in nature. Beyond the above fundamental expressions, according to Du et al. (Du, Tao, and Martinez 2014), human emotions include compound expressions. By combining fundamental expressions, they increase the number of expressions available to 22. Later, with large-scale compound expression data, the EmotioNet dataset (Fabian Benitez-Quiroz, Srinivasan, and Martinez 2016) is built. Conventional deep learning-based approaches (Slimani et al. 2019; Guo et al. 2017) rely on a substantial amount of labelled compound expression training data to classify the aforementioned compound expressions. Collecting such data, on

the other hand, is time-consuming and frequently necessitates psychological training. We can readily recognise an unseen expression (a query) based on prior knowledge of numerous seen expressions given only a few reference images (a support set).

Recent research on few-shot learning (FSL) shows that it is possible to generalise to novel classes quickly with only a few labelled data from these classes, bridging the gap between humans and AI (Lu et al. 2020). The crossdomain FSL (CD-FSL) paradigm, which considerably reduces the cost of gathering large-scale labelled compound expression data, is investigated in this study. Specifically, rather than manually dividing a compound expression dataset into a base class set and a new class set, we investigate a more difficult but realistic setup that seeks to classify compound expressions from the beginning.

Essentially, Few Shot Learning is an illustration of meta-learning, where a student is prepared on a few related assignments, during the meta-training stage, so it can sum up well to concealed (yet related) undertakings, during the meta-testing stage. A compelling way to deal with the Few-Shot Learning issue is to become familiar with a common representations, i.e. an embedding network. We can readily recognise an unseen expression (a query) based on prior knowledge of numerous seen expressions when provided only a few reference images (a support set). Recent research on few-shot learning (FSL) shows that with only a few labelled data of these classes, it is possible to generalise to novel classes quickly, bridging the gap between humans and artificial intelligence (Lu et al. 2020). In this study, we look into compound FER in the context of the crossdomain FSL (CD-FSL) paradigm, which considerably reduces the time and effort required to collect huge amounts of labelled compound expression data. Specifically, rather than manually separating a compound expression dataset into a base class set and a new class set, we investigate a more difficult but feasible setup that seeks to classify compound expressions from the start.

Complex human way of behaving can be perceived by concentrating on actual elements from different modalities; essentially facial, vocal and actual signals. As of late, unconstrained multi-modular feeling acknowledgment has been broadly read up for human conduct examination. In this report, we propose another profound learning-based approach, for emotion recognition. Our methodology use ongoing advances in few-shot learning. An intermittent embedding network is trained to capture predominant elements. Our proposed approach is trained on one of the finest datasets, MEAD: Multiview Audio-Visual Dataset.

1.1 Motivation

Tech giants, as well as more modest new companies, have been putting resources into Emotion Recognition for over 10 years, utilizing either vision or voice investigation to perceive human feelings. A significant number of these organizations began with an emphasis on statistical surveying, investigating and catching human feelings in response to a product or TV ad. Business organizations are gradually arising in virtual personal assistant (VPAs), vehicles, call centers, robotics and smart gadgets.

Tech-giants predicts that by 2030, more than ten percent of individual gadgets will have emotion recognition capacities, either on-gadget or by means of cloud administrations.

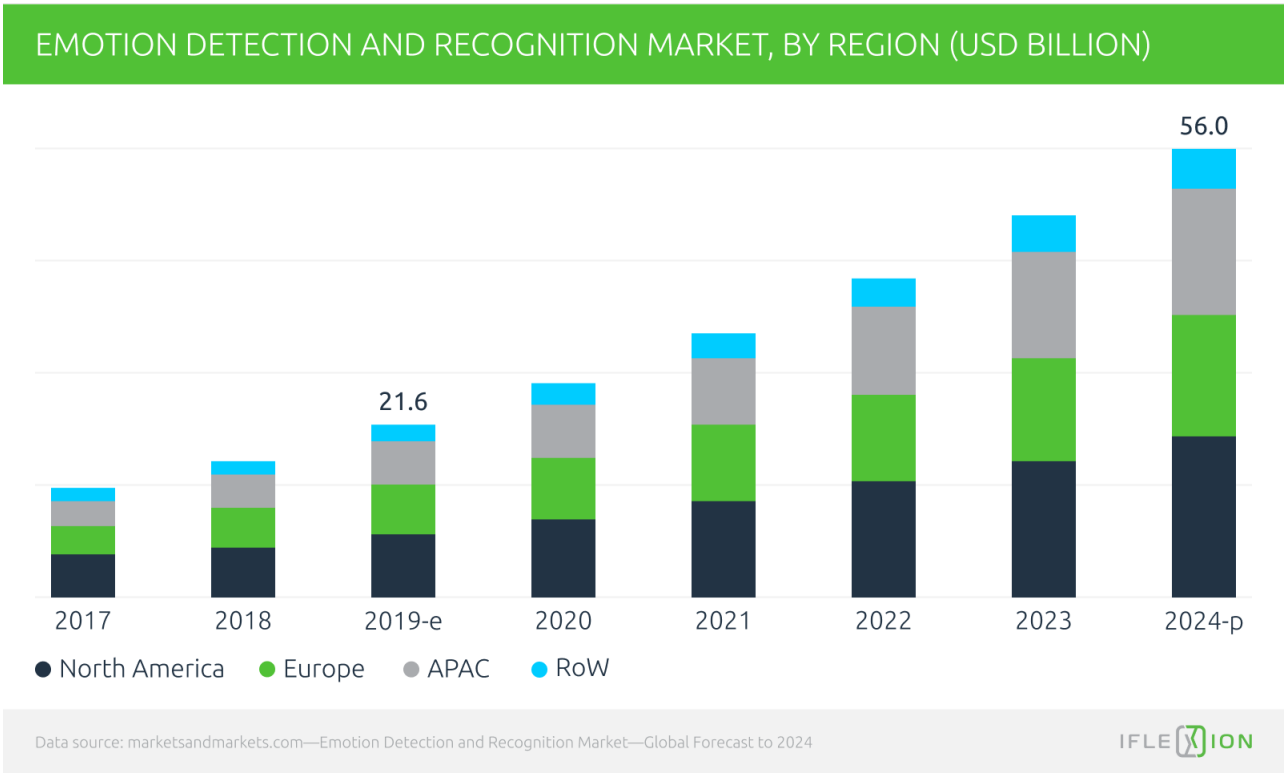

Figure 1.1: Growth of AVER in the upcoming years

1.2 Literature Survey

Authors of [1] propose a profound learning-based approach for AVER. The proposed methodology leverages the ongoing advances in deep learning like knowledge distillation and high-performing deep architecture models. The profound features of the audio-visual modalities are fused in model-level fusion. It mainly consists of three components, Facial expression embedding network, audio embedding network for emotion recognition and audio-visual fusion model.

Before knowledge distillation, the facial network detects and crops faces using MTCNN. The subsequent 140x140 RGB frames are then passed on to an Inception Resnet V1 until the Inception 4e block. This is followed by a 1x1 convolution layer (1x1 Conv), batch normalization (BN) and a ReLU activation. Then, five DenseNet blocks are applied. Finally, 1x1 Conv, BN and ReLU is applied. The result is then averaged over the spatial aspects, resulting in a vector of size Dface (in the figure, Dface = 128). Two separate linear layers then give us the last model results - a vector for the Google FEC triplet task, and class logits for AffectNet. The model is prepared to limit both the AffectNet and Google FEC loss at the same time.

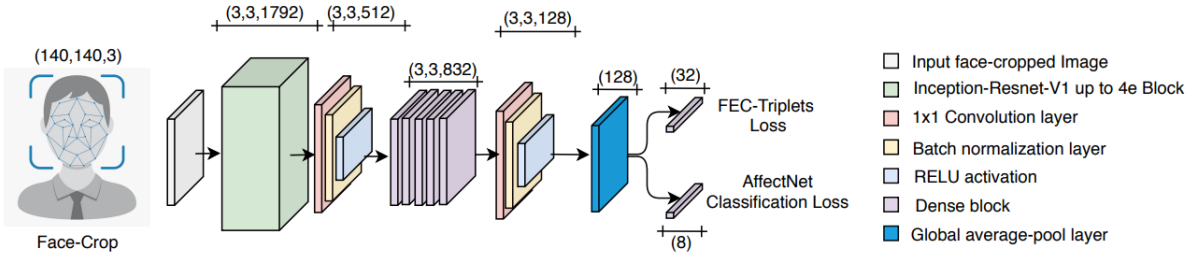

Figure 1.2: Facial Expression Recognition Network.The size of the feature maps (f) of each convolutional and fully-connected layer are shown above each block of operations. The kernel size (k) and stride (s) are specified below the convolution blocks.

Modified VGGish backbone is used for feature extractor in Speech Emotion Recognition. The Mel-Spectogram computed from raw audio signal is fed into modified VGGish backbone network comprising of 6 convolutional layers followed by 3 completely FC layers of size (4096, 4096 and 128) to yield an embedding vector of size 128.

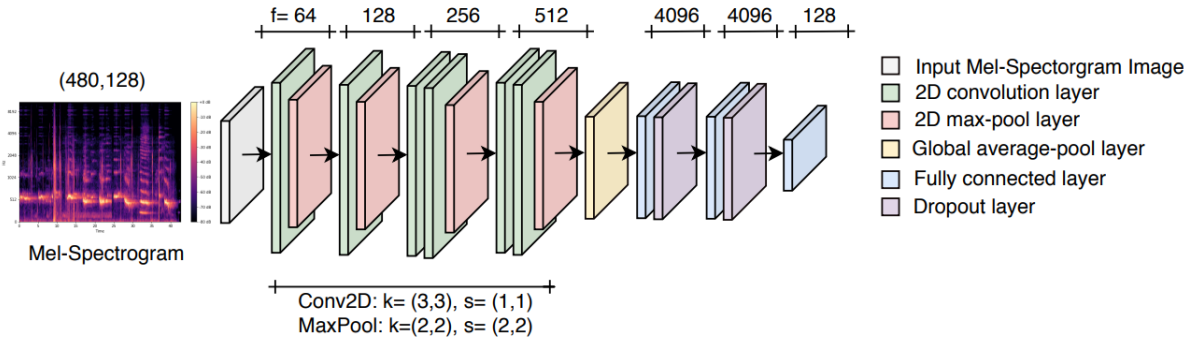

Figure 1.3: Modified Vggish backbone feature extraction for Speech Emotion Recognition.The size of the feature maps (f) of each convolutional and fully-connected layer are shown above each block of operations. The kernel size (k) and stride (s) are specified below the convolution blocks.

The audio and the visual embedding vectors are each fed to a small, independent convolutional network using model level fusion. This results in one tensor of size (9, 64) for each modality. We concatenate these two to give a tensor of size (9, 128), which is fed to a two-layer LSTM network with dimensionality of 256. Taking the final time step's output of this LSTM gives a single vector of size 256, which is pass through a single fully-connected layer with two outputs, which after a tanh activation gives our predictions between -1 and 1 for arousal and valance.

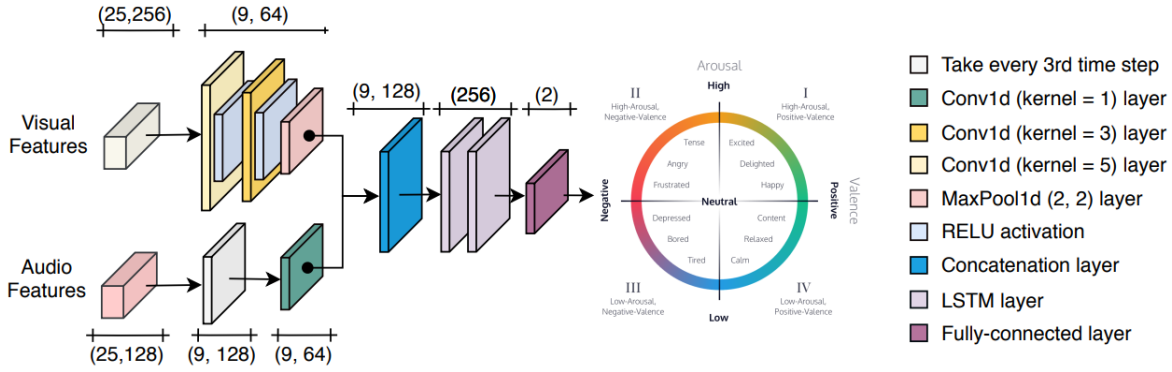

Figure 1.4: Audio-visual fusion network architecture.

Knowledge distillation for facial emotion recognition and VGGish backbone feature extractor presents a promising new course for anticipating feeling from audio and visual modalities. Besides, our deep neural network approach to multi-modal fusion has been shown to be effective in AVER, outperforming the state-of-art methods in predicting valence on the RECOLA dataset. For future work will be carried out to investigate the best strategy for continuous emotion encoding: classification with coarse categories, regression, label distribution learning or even ranking.

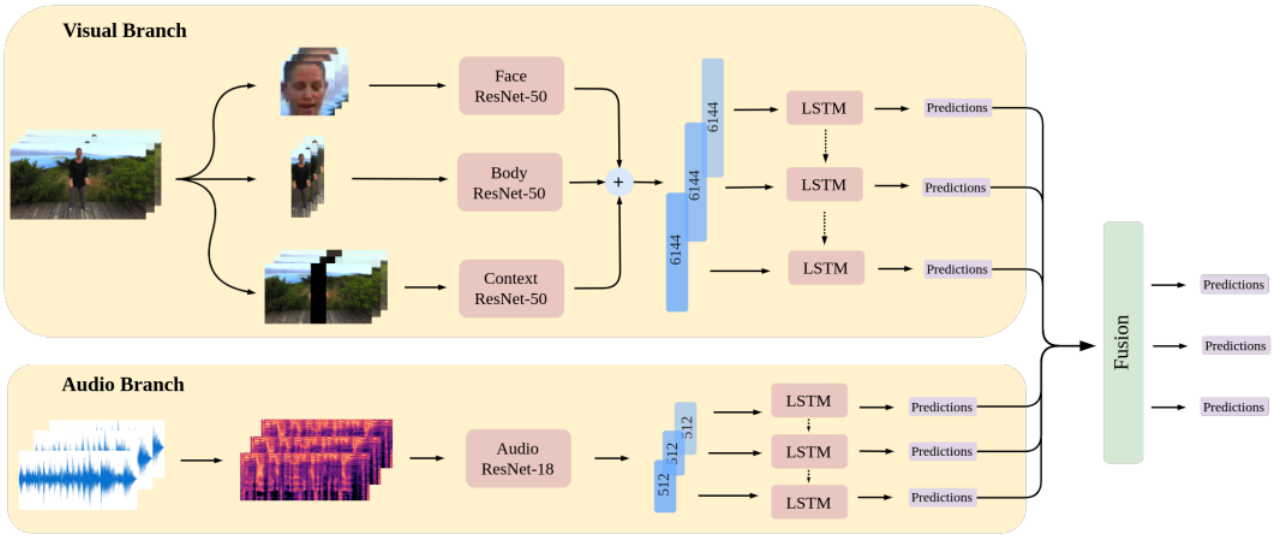

Figure 1.5: Overview of Self-Supervised Training Pipeline

Self supervised learning has as of late drawn in a great deal of exploration interest for both the sound and visual modalities. In any case, most works commonly center around a specific methodology or element alone and there has been extremely restricted work that concentrates on the association between the two modalities for learning self administered portrayals. Authors of [4] propose a system for learning audio features directed by the visual methodology by utilizing a generative audio to-video training scheme plan in which we invigorate a still picture comparing to a given audio segment and optimize the generated video to be just about as close as conceivable to the genuine video of the speech segment.

In [5], the network is composed of an audio stream, and a video stream. The architecture of the audio stream is based on the VGG-M network, with modified filter sizes (13x20) to account for inputs of different dimensions. To the audio stream, input is provided in the form of MFCC values extracted from a 0.2 seconds long audio clip, at a sampling rate of 100 Hz. The architecture of the video stream is based on the Early Fusion model, with a modified filter to take in 5-channel input, which is a sequence of five gray-scale images of the mouth region taken from a 0.2 seconds long video clip.

During the training phase, the objective is to maximize the Euclidean distance between each of the network's outputs for a genuine pair (audio and video clips in sync) and minimize the same for a false pair (audio and video clips not in sync). Hence, the contrastive loss function is used. The networks' weights are updated simultaneously, using stochastic gradient descent with momentum, using contrastive loss as the loss function. Each of the audio and video

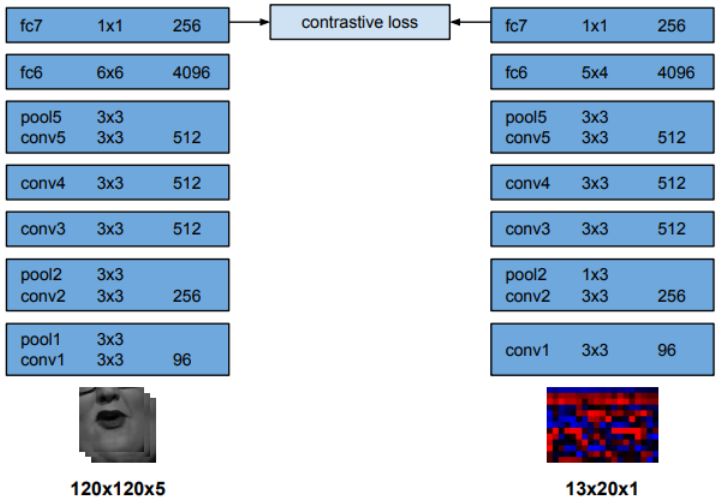

Figure 1.6: SyncNet Architecture

networks gives embedding of size 1x256, which is then concatenated and is fed to two fully connected layers, which is finally fed to a softmax layer. The probability obtained from the softmax is used to decide whether the given frames are in sync with the audio or not.

[6] uses the knowledge distillation in which one model teaches the other model.The idea behind this method is to train a small model which is also called as student model to match large pretrained teacher model.Knowledge is transferred from teacher to student model by minimizing a loss function which is aimed at matching softened teacher logits as well as ground-truth labels. The authors of [6] make use of this idea to solve the problem in cross modal approach. The teacher model is trained using the frames for visual emotion classification. The student model uses the speech modality for speech based emotion classification,the knowledge transfer is done from visual model to speech model to match the logits and classify emotions.

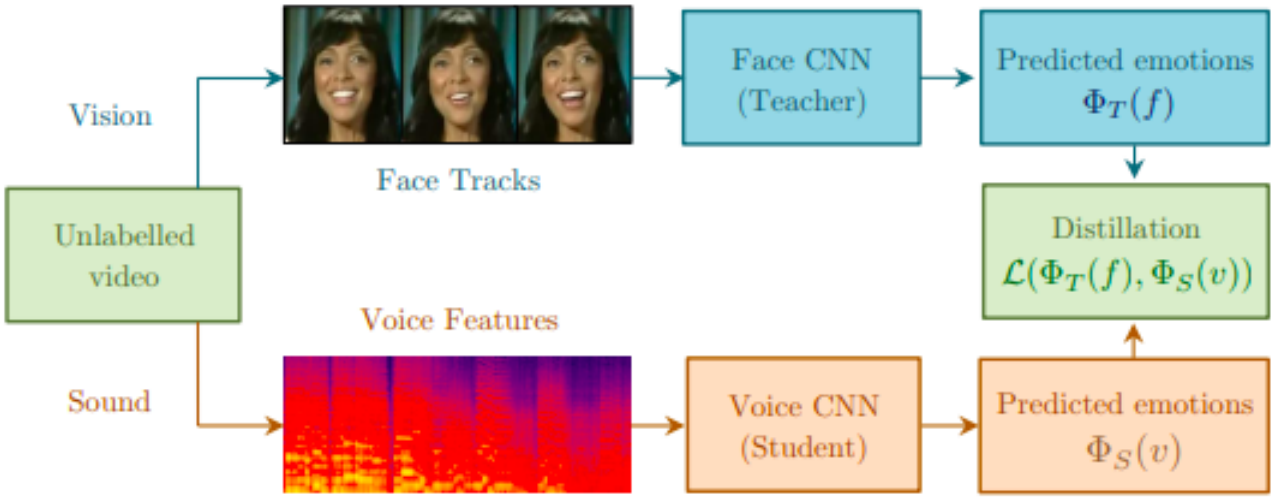

Figure 1.7: Cross-Modal Transfer

Figure 1.7: The teacher model i.e Face CNN uses Resnet-50 kind of architecture and is pre-trained on large scale VGG-Face-2 dataset.The student model which is voice CNN is trained for emotion recoginition using voice which uses architecture based on VGG-M. Mel Spectrograms extracted from speech are used as voice features to train the network.Voice CNN is trained using VoxCeleb dataset. The dataset consits of talking faces and is preprocessed to extract speech .There are no labels for emotions thus student model must learn to classify emotions by knowledge transfer from teacher model.

Cross modal knowledge transfer in this case is possible becacuse emotion content in the speech is correlated with the facial expression.Thus the student model learns to classify emotions using speech entirely by learning from teacher model.Since the visual-emotion data is readily available compared to speech-domain data which lacks the emotion annotations and by exploiting the correlation that exists between visual and speech domain the problem of emotion recognition on speech is solved.

Fig 1.8: There are two branches to the EGS-Net: an emotion branch and a similarity branch. (a)During the training phase, EGS-Net is gradually taught using a two-stage learning architecture.We employ a multi-task technique to accomplish cooperative learning of the two branches in stage one.In stage 2, we alternate between the two branches and learn. (b) During the testing phase, the performance is evaluated using the learned similarity branch on the compound expression dataset.

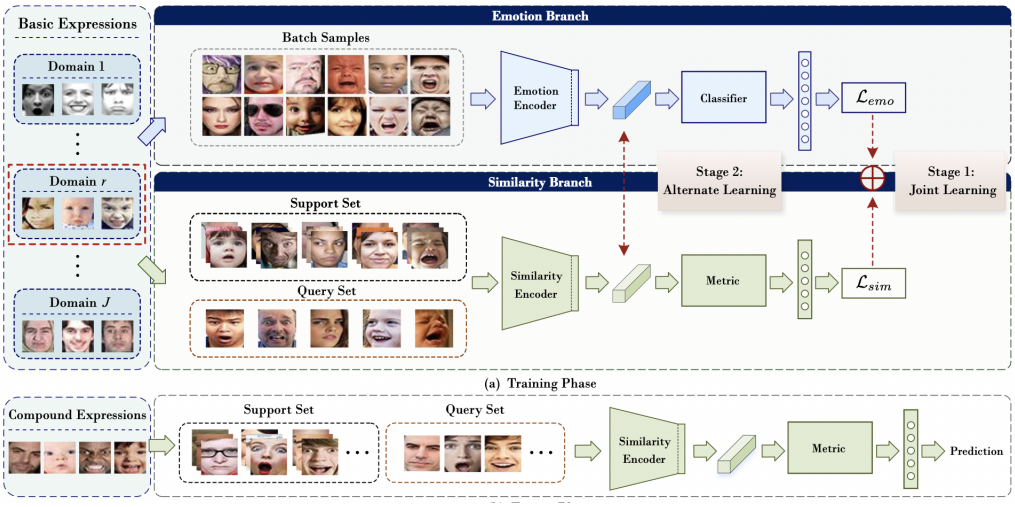

Figure 1.8: Architecture of EGS-Net.

The authors employed compound FER in the CD-FSL context, where the training set (the base class set) and the test set (the novel class set) include disjoint classes and come from different domains. To boost the variety of base classes and bridge the domain gap between the training and test sets, we employ different source domains (i.e., multiple basic expression datasets) for training.

The EGS-Net is made up of an emotion and a similarity branch.To categorise all basic expressions, the emotion branch learns global feature representations, while the similarity branch learns a transferable similarity metric between two expressions. During the training phase, mini-batch training is utilised to learn the emotion branch.Meanwhile, the L2M setting is used to train the similarity branch in an episodic manner. A meta-task is done in each episode by picking a support set and a query set from a source domain at random, and then changing the model parameters based on the classification errors on the sampled query set. For the testing phase, we build comparable meta-tasks using the compound expression dataset.In each meta-task, a query picture is categorised into the closest category in the support set based on the learned similarity branch.

1.3 Problem Statement

Given the audio and the video of a person as input, we address the problem of emotion recognition.

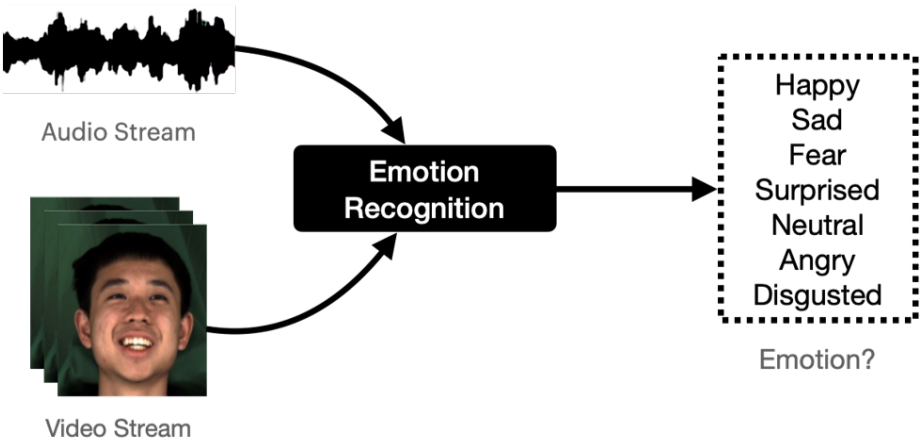

Figure 1.9: Block Diagram.

1.4 Applications

- Marketing companies use ER to visualize customer reactions to their campaigns
- Surveillance Solutions utilize ER softwares to monitor unusual behavior.
- Healthcare industries utilize ER softwares to monitor patient's emotional response to medicines.

1.5 Objectives and Scope of the Project

1.5.1 Objectives

- To make a study of the different approaches of emotion recognition.
- To build a few-shot learning based pipeline for recognition of emotions.
- To evaluate the performance and comparison with the existing SOTA methods.

1.5.2 Scope

- Our model is capable of recognizing emotions for frontal view only.
- Our model is capable of recognizing eight emotions, namely happy, sad, surprise, disgust, neutral, contempt, fear and anger.

Chapter 2

REQUIREMENT ANALYSIS

Requirement Analysis, otherwise called Requirement Engineering, is the method involved with characterizing client assumptions for another product being assembled or altered. In programming, it is now and then alluded to freely by names, for example, necessities get-together or prerequisites catching. Necessities investigation envelops those errands that go into deciding the requirements or conditions to meet for a new or adjusted item or undertaking, assessing the potentially clashing prerequisites of the different partners, dissecting, archiving, approving and overseeing programming or framework necessities.

2.1 Functional Requirements

- The system shall be able to detect faces.
- The system must be able to learn to categorize new, unseen emotions.
- The system shall be able to recognize seven different emotions in total.
- The system shall be able to express the confidence with which it recognized a particular emotion.

2.2 Non Functional Requirements

- The system shall accept videos not more than 7 seconds in length.
- The system shall be implemented using PyTorch library.
- The system shall not take more than 10 seconds to recognize the emotion in the video.

2.3 Hardware & Software Requirements

- The system shall be trained on a GPU and it will be trained on Nvidia P100 GPU.
- Python is required to run the model.

Chapter 3

SYSTEM DESIGN

3.1 Architecture Design

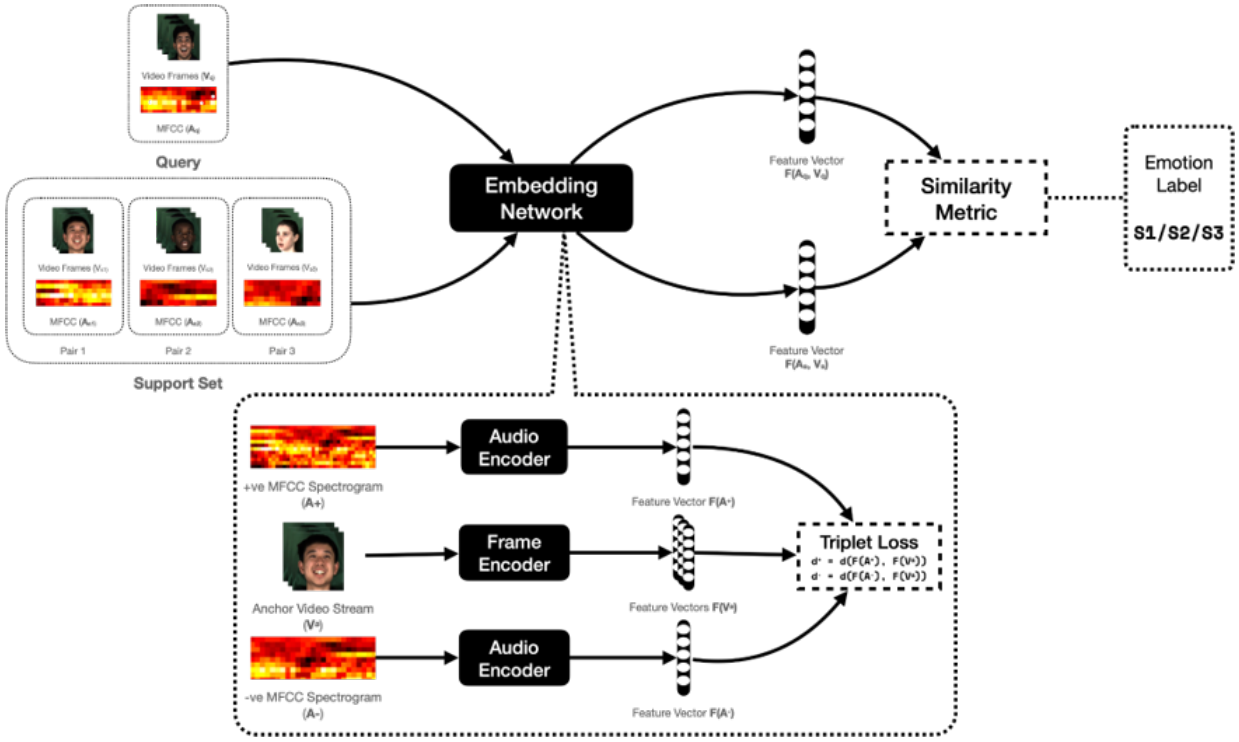

Figure 3.1: Block Diagram.

Figure 3.1 illustrates the overall pipeline of our proposed method. We make use of audio and frame encoders for the task of feature extraction. Triplet loss is used as the loss function at training time, it computes the distances between the anchor video stream, the positive and the negative MFCC spectrogram. Cosine similarity function is used at test time to compute the similarity between each sample from the support set, and the sample in the query set.

Chapter 4

IMPLEMENTATION

Given the audio and the video of a person as input, our model should be able to recognize the emotion from it. We propose a new method wherein we use few-shot learning to solve the aforementioned problem.

4.1 Overview

Our embedding network consists of two encoders. We have used a frame encoder and an audio encoder. During training phase, the encoders are responsible for extracting feature vectors from the inputs fed to them. Then, Triplet loss is used to compute the distance between the feature vectors, updating the encoders' weights in the process. During testing phase, we set-up a few-shot learning pipeline by generating query and support sets that contain frames and spectrograms sampled from unseen emotion labels. The embedding network extracts the feature vectors for each input sample in the support set, and computes the similarity between it and the sample in the query set. The sample in the support set that yields the highest similarity score is the emotion that is recognized in the video.

4.2 Few Shot Learning

Using past knowledge, Few-Shot Learning (FSL) may quickly generalise to new tasks having only a few samples with supervised information. For each input xi, few-shot classification learns a classifier h that predicts the label yi. The N-way-K-shot classification is often used, in which $DTrain$ has $I = KN$ instances from N classes, each containing K examples. Few-shot regression estimates a regression function h using a small number of input-output example pairs, where output yi is the observed value of the dependent variable y and xi is the input that records the observed value of the independent variable x.

4.3 Audio Encoder

MFCC values are used as audio input. On a non-linear mel scale of frequency, this is a depiction of a sound's short-term power spectrum. At each time step, 13 mel frequency bands are employed. For a 0.2-second input signal, the features are calculated at a sampling rate

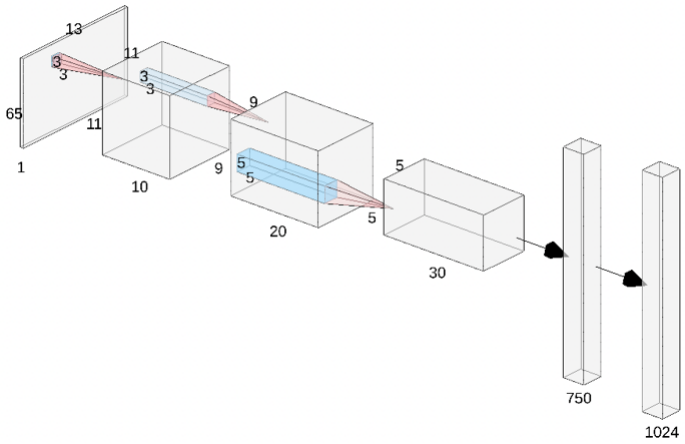

Figure 4.1: Audio Encoder

of 100Hz, yielding 20 time steps. The audio is shown as a heatmap graphic, with MFCC values for each time step and mel frequency range. We employ a convolutional neural network influenced by image recognition systems. Our layer design is based on VGG- M, but with changed filter sizes to absorb unusually sized inputs.The input size is 20 steps in the time-direction, and 13 steps in the other direction, making the input's shape be 13x80.

4.4 Frame Encoder

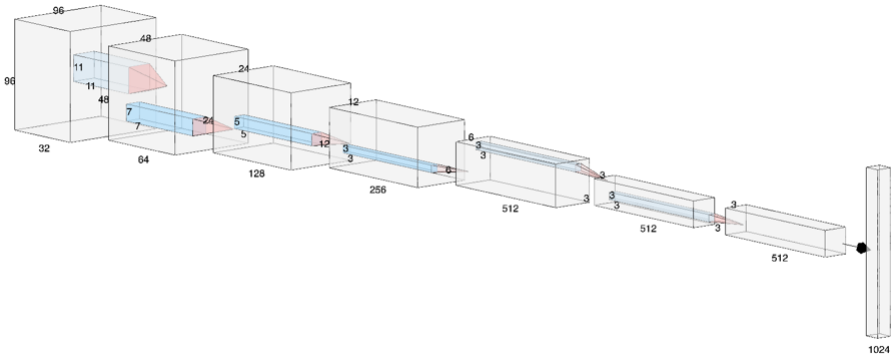

Figure 4.2: Frame Encoder

The input to the frame encoder is a sequence of face images. The input dimensions are 96x96x15 for 5 frames, which corresponds to 0.167 seconds of audio at a 30Hz frame rate. We base our architecture of the encoder on that of [9] which is designed for the task of visual speech recognition. The ConvNet layer configuration is shown in Fig 4.2. The conv1 filter has been modified to ingest the 5-channel input. The input is converted into a vector of size 512.

4.5 Loss Function

The training objective is to construct a feature space wherein samples belonging to the same class are close to each other and samples belonging to different classes are far apart. The loss function is the triplet loss, which is defined as the sum of the squared distances between the anchor, positive and negative samples. The triplet loss is used to train the encoders. The triplet loss is defined as:

$$\mathcal{L}\left(A, P, N\right) = \max\left(\| \mathrm{f}\left(A\right) - \mathrm{f}\left(P\right) \|^2 - \| \mathrm{f}\left(A\right) - \mathrm{f}\left(N\right) \|^2 + \alpha, 0\right) \tag{4.1}$$

Where A is an anchor input, P is a positive input of the same class A, and N is a negative input of a different class from A. α is the margin between positive and negative pairs, and f is an embedding.

We randomly sample five frames from a video and make it the anchor, P, and we take the audio segment that corresponds to the anchor frame as the positive sample, A. We then consider a sample from a different emotion, and extract the audio segment and that becomes our negative sample, N. We then compute the triplet loss between A, P and N. The loss value is back-propagated and used to update the weights of the encoders.

Chapter 5

RESULTS AND DISCUSSION

5.1 Dataset Description

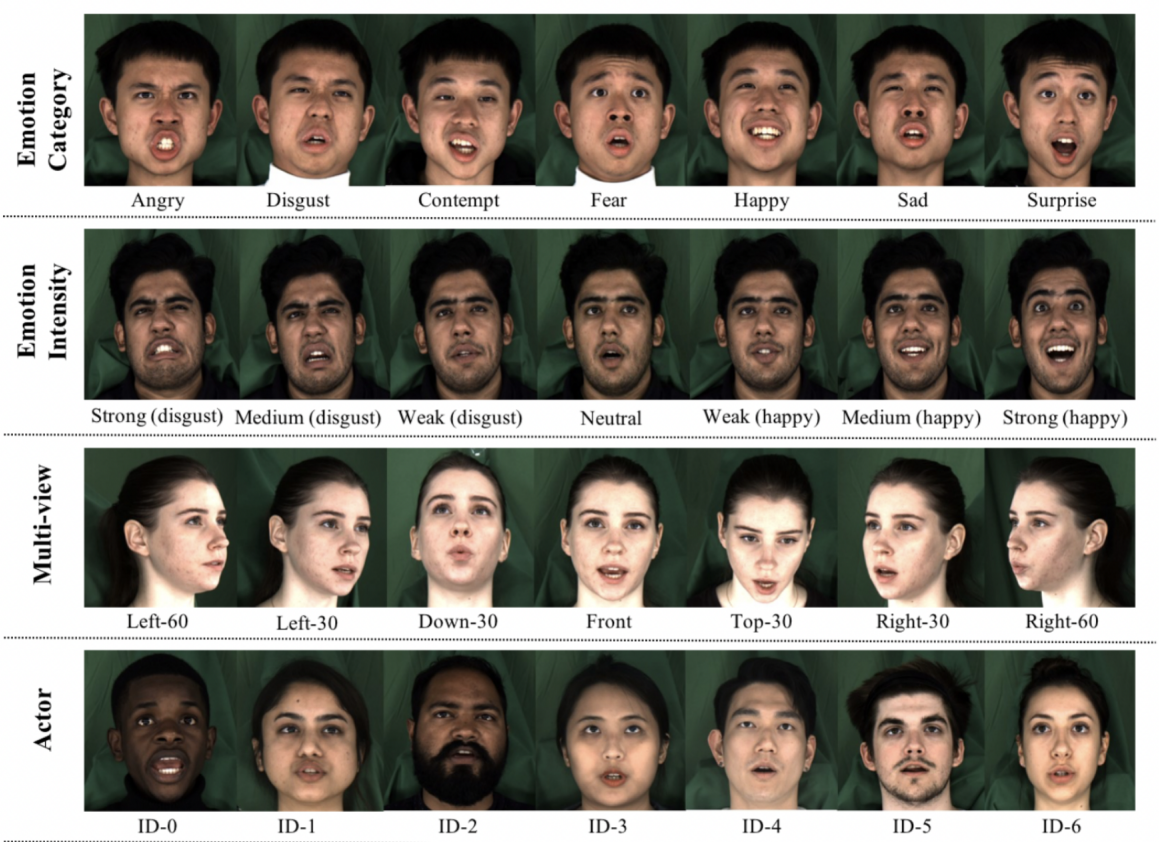

Figure 5.1: MEAD Dataset

We have used MEAD dataset for training and testing purposes. MEAD is a talking-face video corpus that includes 60 actors and actresses expressing eight different emotions at three different intensities. In a tightly regulated setting, high-quality audio-visual clips are taken at seven various view points. Weak is the initial level, which describes subtle but perceptible face motions. The usual state of the emotion, which represents the typical manifestation of the emotion, is the second level medium. Strong is the third degree, which reflects the most excessive representations of this emotion and necessitates significant facial movements.

5.2 Comparison of Results

We trained the model on a subset of the dataset consisting of just happy, sad emotions, and on performing testing on the same emotions, we achieved an accuracy of 65.7%. However, on training the model on happy, sad, angry, surprised emotions, and then performing testing on neutral, disgusted, fear, contempt, we achieved an accuracy of just 34.3%.

Chapter 6

CONCLUSION AND FUTURE SCOPE

We proposed a new approach to the problem of audio-visual emotion recognition, by incorporating a pipeline based on few-shot learning. We use two encoders, one for extracting features from frames, and the other for doing the same from audio. We randomly sample an anchor window consisting of five frames for a particular emotion, and the audio segment that corresponds to that window becomes the positive sample. Further, we select an audio segment of a different emotion label, and that constitutes our negative sample. The anchor window, positive sample, and the negative sample is then used to calculate the triplet loss, which is back-propagated through the two encoders. We end up with a feature space wherein samples sharing the same emotion label lie closer together than samples that belong to different emotion labels. At testing time, we generate support and query sets consisting of samples with emotions labels that are disjoint to the emotion labels used for training the model. For each sample in the support set, we calculate the cosine similarity with respect to the query set, and the one that yields the highest similarity, that is the emotion detected.

Future works may implement the proposed method and subject it to experimentation, such as changing the architecture of the encoders, or using a hybrid loss function in an attempt to build an even more effective system.

REFERENCES

[1] Liam Schoneveld, Alice Othmani, *Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition*, 2021.

[2] Abhinav Shukla, Konstantinos Vougioukas, *Visually Guided Self-Supervised Learning of Speech Representations*, 2021.

[3] Chung, J. S. and Zisserman, A., *Out of time: automated lip sync in the wild*, 2018.

[4] Zou, Xinyi and Yan, Yan and Xue, Jing-Hao and Chen, Si and Wang, Hanzi, *When Facial Expression Recognition Meets Few-Shot Learning: A Joint and Alternate Learning Framework*, 2022.

[5] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, Andrew Zisserman, *Emotion Recognition in Speech using Cross-Modal Transfer in the Wild*, 2018.

[6] A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild, *Out of time: automated lip sync in the wild*, 2020.

[7] Prajwal K R and Rudrabha Mukhopadhyay and Jerin Philip and Abhishek Jha and Vinay Namboodiri and C V Jawahar, *Towards Automatic Face-to-Face Translation*, 2019.