Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Expressions and emotion detection from HMD sensors

Master's Thesis

Loïc Houmard

lhoumard@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Ard Kastrati, Dushan Vasilevski
Prof. Dr. Roger Wattenhofer

October 13, 2023

Acknowledgements

I would like to thank Ard and Dushan for their insigthful guidance and help during the whole project and the expertise they provided on the topic. Furthermore, I would like to thank Magic Leap for giving me the opportunity to have access to a great infrastructure and to meet very competent and friendly coworkers with a lot of knowledge in computer vision and machine learning. Finally, I would like to thank Prof. Roger Wattenhofer and the Distributed Computing Group at ETH Zürich for giving me the opportunity to work on this interesting topic and submit a paper to an international conference.

Abstract

Understanding emotions and expressions is vital for humans to communicate and understand each others mental state. With the growth of virtual and augmented reality and the aspiration to create realistic avatars representing humans in a virtual world, being able to correctly classify their feelings and facial movements from partial and multimodal data has become crucial. In this work, we investigate the role of facial and speech features in classifying emotional states and the benefits of combining both modalities in a single model. Moreover, we study the utility of partial facial features (eyes, mouth and head pose) and of different types of features, some being learnt for the specific task but computationally intensive and some being handcrafted and rather lightweight and efficient. Interestingly, our findings indicate that facial features and audio better work for different emotions, highlighting their complementarity and the advantages of a multimodal approach. Our results show that combining efficient features of both modalities give a nearly 10% accuracy improvement over the unimodal counterparts and our best model achieves an accuracy of 83.57% on the RAVDESS dataset, surpassing humans at emotion classification.

Contents

Chapter 1

Introduction

1.1 Motivation

Humans communicate through spoken language, but more importantly, through body language and facial expressions and perceiving the emotional state of an individual talking is crucial in order to understand him and create a true connection. Many different channels can be used to infer a person's current emotion, such as facial features and para-linguistic aspects of the voice like the pitch, and by combining all of these different clues, humans manage to accurately discern each others.

With the rise of virtual and augmented reality and the vast amount of online meetings carried out nowadays, some growing interests for predicting emotions and creating realistic avatars representing humans in a virtual world has emerged. Some companies, like Magic Leap or Meta, have developed precise head mounted displays (HMD) containing a lot of sensors gathering precious data about the people wearing them, which can be used to infer their emotional state, possibly through artificial intelligence, and either provide a more personalized live experience depending on their mood or animate an avatar representing them. A concrete example of such a model is the latest Magic Leap 2 device, released in September 2022, which contains, among others, eye tracking cameras, microphones, inertial measurement unit (IMU) sensors which can be used to infer the head pose and hand tracking cameras (see Figure 1.1). All of these different modalities carry some information about the person wearing them, but it is still not perfectly clear what the impact of each of them is for classifying the emotion of the subject and how to best combine these different data sources in order to increase the accuracy of the prediction.

1.2 Preview

In this work, we investigate the role of different modalities, namely speech, facial video and some of its sub-parts, specifically eyes, mouth and head pose for the

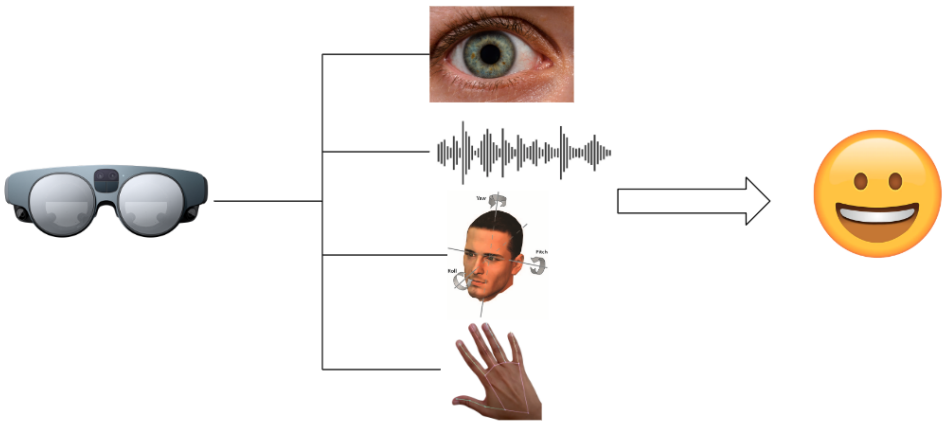

Figure 1.1: Magic Leap 2 device with some of its sensors used for emotion prediction.

task of emotion classification via deep learning. We also look at different type of features representing these modalities, some being very computationally intensive but famous for giving good results on other tasks and some rather lightweight which could be used in real time on HMDs. We first study the impact of each of the modalities for the prediction of particular emotions. Interestingly, our findings show that some emotions can better be determined by some different modalities, strengthening the idea that multimodal models can benefit from different sources. We then conduct some experiments on multimodal approaches, highlighting their great potential for emotion detection.

The rest of the document is structured as follows. In the second chapter, we provide a more detailed introduction to the elements that are important for understanding the problem in question. In the third chapter, we detail some of the existing datasets and give a clear description of the one we chose for this project and the reasons why it was chosen. In the fourth chapter, we summarize some of the works already carried out by the scientific community. In the fifth chapter, we include a detailed description of our approach, highlighting the different types of features and the models we tried out. In the sixth chapter, we present and discuss the experiments performed on the model and their corresponding results. In the seventh and last chapter, we draw some final conclusions and outline possible directions for future work.

Chapter 2

Preliminary Notions

In this chapter we include a brief introduction on notions related to our work. Firstly, we introduce the different theories on the representation of emotions. Then, we give a short introduction to the Facial Coding System (FACS) and the Action Units (AU), two important concept widely used in avatar animation and which are directly linked to emotion detection.

2.1 Emotional Models

Lately, two different theory on emotions, backed up by psychology, have been confronted: a categorical and a dimensional one. In the categorical framework, emotions are divided into different more or less fine-grained categories. Usually and as initially supported by Ekman [1], six basic and universal emotions, namely anger, disgust, fear, happiness, sadness and surprise are recognised as the main ones. Nevertheless, these emotions can be further divided into more specific subcategories as illustrated in Figure 2.1 on the left. On the other hand, Russell [2] argued that this categorical model no longer explained adequately the vast number of empirical observations from studies in affective neuroscience and therefore proposed a dimensional model based on two fundamental dimensions: valence which represents the pleasantness of the stimulus and arousal which refers to the level of energy as illustrated in Figure 2.1 on the right. This dimensional model has also been extended by other authors to more than 2 dimensions. Dominance is sometimes used as a third axis representing how much control one has over the emotion.

2.2 Facial Coding System

The Facial Coding System (FACS), originally developed by Hjortsjö [3], a Swedish anatomist, refers to a set of facial muscle activation used to display emotions. It originally contained 23 Action Units (AUs), which are the fundamental actions of

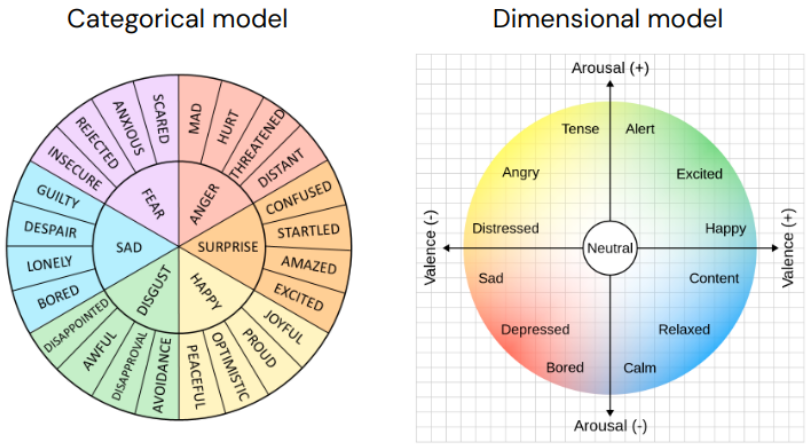

Figure 2.1: Emotional models.

individual muscles or groups of muscles, but was then adopted by Paul Ekman and Wallace V. Friesen [4] and further developed in 2002 [5]. Their temporal combinations is known to produce facial expressions and is widely used in psychology and in animation. An AU can either be represented by a continuous value representing how much the given muscle is activated or by a binary value describing whether or not the muscle is triggered and is usually annotated by professional human coders. Nowadays, some facial datasets are annotated with the FACS and some AI models have been trained to regress action units. Usually, only 12 to 18 of them are annotated as it requires a substantial amount of work and already provides a good approximation of the facial movements which can be used to reconstruct a real-looking face. Some examples of famous action units are the Inner Brow Raiser (AU1), the Brow Lowerer (AU4) and the Lip Corner Depressor (AU15) which, when they are all activated simultaneously, represent sadness (see Figure 2.2).

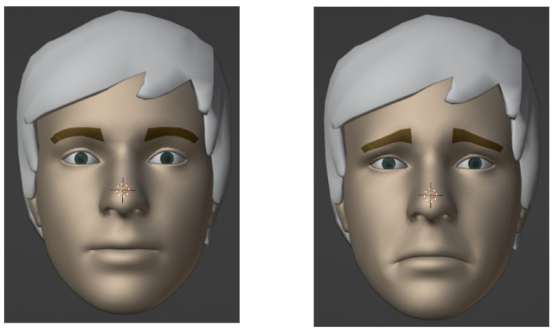

Figure 2.2: Example of the animation of an avatar with AUs. On the left, a neutral face without any AU activated. On the right, the same face with AU1, AU4 and AU15 fully activated, representing sadness.

Chapter 3

Datasets

In this chapter, we will first give an overview of some of the most common datasets for speech and facial emotion recognition without giving too many details. We will then dive deeper into the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [6], which was chosen for this project. The goal is to give an overview to anybody interested in the topic and a complete introduction to the data used in our experiments.

Before introducing the datasets, it might be interesting to explain what kind of data we are looking for in this project. First of all, since the main focus of the project is on emotion and expression recognition, we would need to have videos annotated with either categorical or dimensional emotions, ideally with a good balance between the samples of every classes. AU annotation would be beneficial as well, but is not strictly needed, especially because it can be extracted by other tools like OpenFace. Furthermore, since we are interested by HMD devices, having the eyes at a good resolution as one of the modality is needed, since it is one of the main sensor common to almost all of the models on the market. Some other modalities like hand and head tracking would be advantageous as well. We didn't find any datasets containing at the same time eyes and audio, hence we would need to extract the eyes ourselves from the full video and therefore the resolution should be good and the angle taken from the front. The size is important as well, as most deep AI models need quite a lot of samples to be trained. Finally, there should exist other studies on the dataset in order to make our results comparable. RAVDESS was one of the only datasets that fulfilled all of these requirements simultaneously and was freely available, therefore it was selected.

3.1 Overview

Audio datasets

– *EmoDB:* The Berlin Database of Emotional Speech (EmoDB) [7] contains 535 utterances in German of 7 emotions (6 basic emotions and neutral) vocalized by 10 professional speakers at 16kHz.

– *MSP-Podcast:* The MSP-Podcast dataset [8] contains 100 hours of audio podcast in English at 16kHz, split into shorter segments of 2.75 to 11 seconds. It was annotated with crowd-source for valence, arousal and dominance and for 9 emotions (7 basic emotions, neutral and other)

Facial datasets

– *JAFFE:* the Japanese Female Facial Expression (JAFFE) Dataset [9, 10] contains 213 grayscale images of size 256x256 of 7 facial emotions (6 basic facial emotions and neutral) played by 10 Japanese female expressers.
– *FER-2013:* The FER-2013 dataset [11] contains 35'887 natural grayscale images of only 48x48 pixels divided into a training and a testing set of 7 facial emotions (6 basic facial emotions and neutral) for many different subjects. It was used during the ICML 2013 Workshop on Challenges in Representation Learning.
– *CK+:* The CK+ dataset [12] contains 593 video sequences of 123 subjects from 18 to 50 years. It has a resolution of 30 frames per second and 640x480 pixels and contains the annotation of 7 emotions (no neutral) and 30 action units for the peak frame by professional FACs annotators.
– *RaFD:* The Radboud Faces Database [13] contains 5880 images of size 1024x681 of 8 emotions (7 basic facial emotions and neutral) played by 49 Caucasian Dutch models with 3 different gaze angles and 5 camera angles.
– *AffectNet:* The AffectNet dataset [14] contains 450'000 images of different sizes of 8 emotions (7 basic facial emotions and neutral) from many different subjects of different ages annotated with valence and arousal.

multimodal datasets

– *BAUM-2:* The BAUM-2 dataset[15] contains 1047 video sequences of 286 subjects with a wide range of ages extracted from movies, with various head poses, illumination conditions, accessories and temporary occlusions. It is labelled with 8 emotions (7 basic emotions and neutral) and might have multiple emotions at once. The audio is in English or in Turkish.
– *Aff-Wild2:* The Aff-Wild2 dataset [16] contains 564 video sequences (2.8M frames) from people of various ages and ethnicity taken from Youtube and cropped around the head. It was then manually labelled with 7 emotions (6 basic facial emotions and neutral), 12 AU, valence and arousal.
– *IEMOCAP:* The Interactive emotional dyadic motion capture (IEMOCAP) database [17] contains 12 hours of audiovisual data with scenes played (or improvised) by 10 actors. It contains 10 emotions (8 basic emotions, neutral and other), potentially more than one simultaneously and is also annotated with valence, arousal and dominance. Text, hand and head tracking are also present and might be treated as other modalities. The videos are quite low resolutions and taken slightly from the side.

3.2 RAVDESS

The RAVDESS database is constituted of audio-visual videos of speech and song acted by 24 professional actors, equally distributed between the two genders. The data is split into two different sets, one containing speech and the other songs, and we decided to use only the speech set as it made more sense in our setting. The videos are rather short, varying between 2.9 and 5.2 seconds, and each one is played with one of the 8 emotions present in the dataset, namely the 6 basic emotions introduced by Ekman (anger, disgust, fear, happiness, sadness and surprise) as well as neutral and calm. Two neutral lexical statements are vocalized in a North American accent. Each expression is repeated twice and is produced at two levels of emotional intensity (normal and strong) except neutral, making the dataset balanced between all of the emotions, except for neutral which has half as much samples. In total, the speech set contains 1440 videos, 60 for each actor or 192 for each emotion (96 for neutral). The videos were each rated 10 times on emotional validity, intensity, and genuineness (read [6] for more information) ensuring an excellent overall quality. Humans raters achieved an accuracy of 80% for audio-video, 72% for video-only and 62% for audio-only. The videos are filmed from the front at a good resolution of 720x1080 pixels, at 30 frames per seconds and 48kHz sampling rate for the audio. The filming conditions (lighting, angle and distance) are always the same, without occlusions at any time making the database rather simple and good for overall studies about emotions. Nevertheless there is no guarantee that a model trained on it would work well in a real-life setting, where many more parameters can vary and must be taken into account.

Figure 3.1: Example of 2 frames for the happy emotion played by the first actor in the RAVDESS dataset.

Chapter 4

Previous Work

In this chapter we give an overview of the work of other scientists on the topic of emotion recognition. We first look at Speech Emotion Recognition (SER), which aims to predict emotions from linguistic and/or para-linguistic aspects of an audio signal. We then focus on Facial Emotion Recognition (FER) and finally on multimodal models proposed to enhance the accuracy of the unimodal ones.

4.1 Speech Emotion Recognition

Traditionally, feature engineering has been the foundation of speech emotion recognition as shown in [18] which compares and summarizes previous works. Some of the most common hand-crafted features included in the literature are the Mel Frequency Cepstral Coefficients (MFCCs) and the (Log) Mel Spectrogram which represent the overall envelope of a signal over time. Some other useful aspects of the speech are the pitch, the zero-crossing rate and the energy. In [19], the authors compared different audio features, mainly hand-crafted, and fed them to different well known models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Network (RNN) and they achieved an accuracy of 66% on the 7 emotions times 2 genders (14 classes) of the RAVDESS dataset, using 20 actors for training, 2 for validating and 2 for testing. They observed that the choice of the features had a bigger impact on the accuracy than the actual model, Log Mel Spectrogram yielding constantly better results than MFCCs and that the models they tried directly on the raw signal weren't very accurate, probably due to the lack of data for training. They also argue that separating individuals by gender might be beneficial due to the pitch and energy differences in the average male and female voice, which makes patterns in male emotions different from female emotions. In [20], the authors proposed a temporal model, called TIM-Net. It works by extracting temporal features at different scales from MFCCs using dilated convolutions and fusing them before passing them to a final Multi Layer Percepton (MLP). It reached 91.93% accuracy on the RAVDESS dataset (and comparable results on other datasets). However, the results were computed using speaker-dependent data (same users are part of both the training and testing set),

which makes the task much simpler and hence the results hardly comparable to the previous works cited above. Another study [21] on the same dataset showed the superiority of fine-tuning larger pretrained models. They achieved 76.58% accuracy using a 5 folds cross-validation speaker-independent strategy by feeding Mel Spectrogram to a 2D-CNN model pretrained on audio Spectrogram for sound classification.

With the advent of deep learning, the field has seen a shift towards models capable of operating on raw audio data or pre-trained features directly. Some works [22, 23, 24] tried to first reduce the dimension of the audio signal by learning a discriminative encoding of the signal with different encoder-decoder architectures and trained a separate classifier for emotion prediction. The main advantage being the use of unlabelled data, which is much easier to find online and makes the model more robust to different shifts (conditions of recording, languages, accents, noise) in the data. Lately, following the recent breakthrough in Natural Language Processing (NLP) using attention [25], Wav2vec [26] and HuBERT [27], two huge transformer models learning representations directly from raw signals were proposed. They were pretrained on a lot of audio data and were proven to achieve very good results on many different related tasks. [28] showed their superiority to predict valence and their overall better generalization in regards to CNNs. Interestingly, they also showed that these models were capable of dealing with the lexical parts of the speech. In [29], the authors fine-tuned the transformers module of Wav2vec on the RAVDESS dataset and by using the same 5 folds cross-validation speaker-independent strategy as cited before, achieved an accuracy of 81.82%.

4.2 Facial Emotion Recognition

Facial emotion recognition followed almost the same evolution as speech emotion recognition, namely using traditionally hand-crafted features but then shifting to an end-to-end learning with deeper models. Some datasets contain images only whereas others rather contain videos, which adds a temporal dimension to the task that must be addressed. However, the extraction of some kind of features, may it be learnt or engineered, is almost always present, either for a single image or for each frame of a video.

The traditional features can be divided into two main categories: geometric features which corresponds to geometric relations between landmarks such as the euclidean distance or the angles between them and appearance features which consists of statistic about the image and its texture such as the Histogram of Oriented Gradients (HOG). In [30], the authors computed the pairwise distance between 18 landmarks on the face (generated with a Kinect), which they normalized and used to compute a Structured Streaming Skeleton (SSS) to remove some intra-class variations. In order to make the tracking of points of interests on the face simpler,

some open toolkit like OpenFace [31, 32] have been developed. It works in real time and can accurately detect 2D and 3D landmarks from images, as well as estimating the AU activation, the head pose and the eye gaze. It was used in many papers, such as [29] where the authors used the action units as input to a bidirectional Long Short-Term Memory (bi-LSTM) network with attention and achieved an accuracy of 62.13% on the videos of the RAVDESS dataset. In [33], the authors rather compared the efficiency of different appearance features and came to the conclusion that different features work better for different emotions and different conditions of images, such as race, lighting or pose.

On the deep learning front, different models, some using transfer learning, some self-supervised learning or other techniques have been proposed. In [21], they fine-tuned a Spatial Transformer Networks (STN) on each individual frame of the RAVDESS dataset, where the label was inherited from the video. They then used a max-pooling strategy or a RNN to aggregate the frame classification at the video level. This approach has the drawback that the label given for each frame might be highly inaccurate, since the emotion might not be present at a precise frame of the video and hence the fine-tuning might be sub-optimal. They reached an accuracy of only 57.08% with their best model. Other researchers [34] also used transfer learning and compared 4 different well-known CNN-like architectures for emotion classification on images. They achieved high accuracy, especially with MobileNet [35] and ResNet [36]. Some models take advantage of unlabelled data by using self-supervised learning. FabNet [37] and the network from [38] are two examples of such models, trained to learn a compressed embedding representing the face and which can be used for downstream tasks such as emotion prediction.

Finally, very few studies have been carried out on the impact of sub-parts of the face such as the eyes or the mouth, which are meaningful when people wear HMDs. In [39], they removed the eyes on images by replacing them with black stripes and still managed to achieve a high accuracy of 95.9% on the RaFD dataset with their best CNN, demonstrating that the lower face carry enough information for precise emotion classification.

4.3 multimodal Emotion Recognition

multimodal fusion aims to improve the accuracy of unimodal models by combining different modalities in a common model. As explained in [40], which summarizes the current literature, the fusion of the modalities can usually be done at 3 different levels. The first one is at classification level (late fusion), where specialized models using only one modality are trained individually and then, a new module taking the predictions of the different models as input is trained to make a final decision, which is also known as an ensemble of models. This has the advantage to be fairly simple and that the best unimodal models can be used without any change, but it has the drawback that relationships between the different modalities can

not be learnt jointly. This technique was used in [29] and [21], where they made an ensemble of their best models cited in the 2 previous sections, reaching an accuracy of 86.70% (respectively 80.08%) on RAVDESS. The second type of fusion is at feature level, where features are first extracted for each modality and then concatenated and used as input to the multimodal model. In [41], they extracted MFCC for the audio and facial features using EfficientFace [42] at each frame and then used 1D convolutions and transformer blocks to fuse them. They also proposed the modality dropout, which consists of removing (setting to zero) one of the modality for some samples during the training, so that the model also works when only one modality is present, which is quite common in real settings and also adds regularization to the model. They achieved an accuracy of 81.58% on RAVDESS. The last fusion method consists to create directly an hybrid model which handles all the data sources jointly. The last type of fusion seems to be the most promising as it takes the full advantage of multiple sources, however the exact architecture to use is hard to find and still subject to researches, the training more expensive and more inclined to ovefitting if proper regularization is not used.

Chapter 5

Approach

In this chapter, we will explain our approach thoroughly, starting with a detailed explanation on the extraction of new modalities and of different types of features for each of them, which is one of the main contribution of this work in comparison to what had previously been proposed. We will then explain the different models we tried, some coming from the literature and one being newly engineered.

5.1 Modalities

As stated in chapter 3.2, RAVDESS initially only provides two modalities, audio and full face video. Nevertheless, eyes and head pose are almost always present on HMD devices, thus learning more about their influence is one of the goal of this project. For that reason, we decided to implement a small pipeline to extract them, as well as the mouth (which is a good complement to the overall understanding of facial emotions, even though it is less likely to be tracked in real applications). In order to do so, we used OpenFace [31, 32], which offers a precise tracking of key-points on the face. It provides both 2D and 3D locations of 68 landmarks distributed over the full face (Figure 5.1) and of 28 more precise landmarks on each eyes (Figure 5.2). OpenFace also computes the head pose at each frame, constituted of the 3D location of the head with respect to the camera in millimeter and the rotation in radians around the 3 axes (pitch, yaw and roll). The only preprocessing applied for this modality was to remove the mean vector of each video in order to replace the origin and have values closer to 0 which is often beneficial for AI models.

In order to extract the eyes and the mouth, the following procedure was applied. For the mouth, we computed the 20 2D key-points related to it (numbers 48 to 67 in Figure 5.1) and for each of the eyes the 28 precise corresponding landmarks. We then filtered the maximal and minimal values along both axes and computed the middle point between them, giving us the center of the mouth (respectively the eyes). Note that we might have only used the 4 landmarks at the extremities and it would probably have yielded the exact same results, however

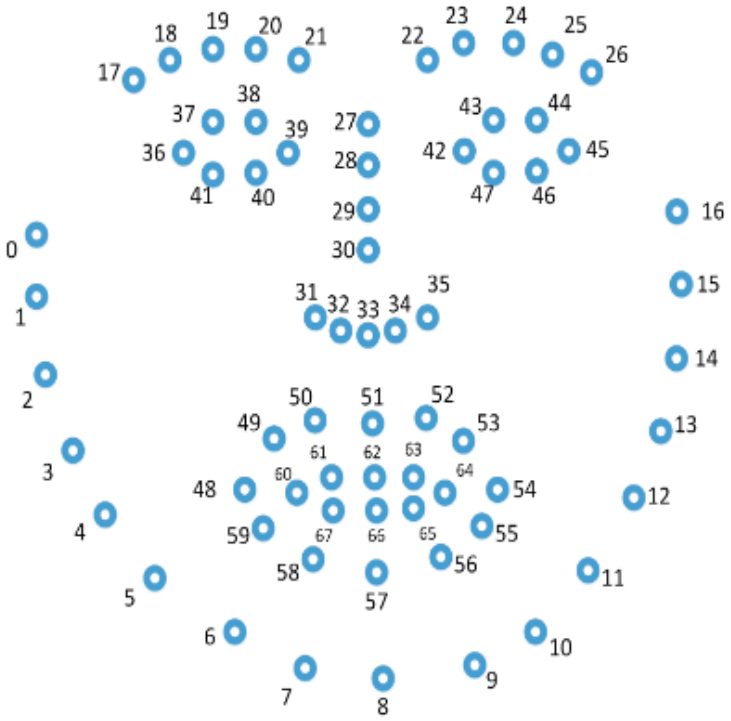

Figure 5.1: Face key-points tracked by OpenFace

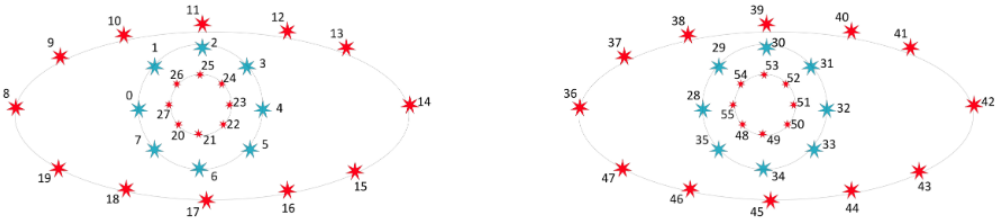

Figure 5.2: Precise eyes key-points tracked by OpenFace

we took the maximal and minimal values across all the corresponding landmarks to be more resilient to OpenFace's potential errors. We then extracted a patch around them. The patch was of size 128x128 pixels for each of the eyes (which is not as precise as the usual tracking cameras on HMD devices which is for example 400x400 for the Magic Leap 2 device, but enough from an human perspective) and of size 110x180 for the mouth. The sizes of the patches were chosen empirically, so that they are big enough to fully contain the mouth (respectively the eyes) at each frame of each video in the dataset, but not too much more information about the face. Note that one of the video was removed from the mouth dataset as the actor was moving his head a lot and his mouth disappeared from the camera for a few frames. Figure 5.3 shows the result of the extraction for the eyes and the mouth.

Figure 5.3: Example of the extraction of eyes and mouth for an happy video frame (top 2 rows) and for a fear video frame (bottom 2 rows) for two different actors.

5.2 Features

We will here present different types of features used in our experiments. Features are usually any kind of lower dimensional embeddings of the full modality representing and summarising it. We will first talk about hand-crafted features, i.e. features engineered by humans based on their knowledge on what matters most in a signal in order to solve a task. Then we will talk about learnt features, i.e. extracted directly by AI modules by minimising a given loss.

5.2.1 Hand-Crafted

MFCC

The Mel Frequency Cepstral Coefficients (MFCCs) of a signal are a small set of features which concisely describe the overall shape of a spectral envelope. They are often the default features used in many acoustic experiments as they were proven to work well for many applications such as speech recognition, emotion recognition, gear fault detection, Electrocardiogram (ECG) and Electroencephalogram (EEG) classification [43] or music information retrieval [44].

They can be computed using 5 consecutive steps, namely signal framing, computation of the power spectrum via the Fourier transform, mapping it to a mel scale, which is a perceptual scale of pitches based on the way human perceive sounds, i.e. in a non-linear fashion, eventually computing the logarithm (this step was skipped in our experiments as we used the default implementation from Pytorch which uses the DB-scaled Mel spectrogram) and finally applying a discrete cosine transform (hence MFCC can be seen as spectrum of a spectrum). For further mathematical explanations, refer to [43].

These features are 2 dimensional (Figure 5.4), the first axis representing time (the exact length depends on the window and hop length used to compute it) and the second one being a compact representation of spectral features. It can hence be seen as an image and can be used by models such as conventional 2D CNNs or as a time-series and rather be used by temporal models such as LSTMs.

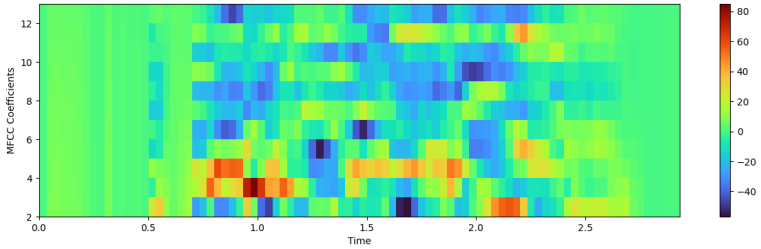

Figure 5.4: Example of the MFCCs of a signal with 13 coefficients. Two first coefficients have been removed for visualization purpose.

Log Mel Spectrogram

The Log Mel Spectrogram (LMS) is very related to the MFCCs as it is also derived from the mel scale. It is computed using the 4 same initial steps as the MFCCs (hence just skipping the final discrete cosine transform). Since it is a spectrogram, its interpretation is however more straightforward as it represents a signal in its popular time-frequency paradigm. It can also be treated as an image or as a time-series (Figure 5.5).

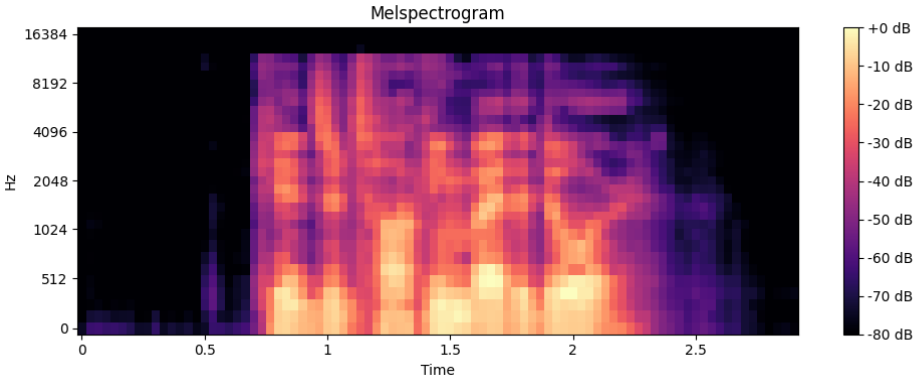

Figure 5.5: Example of a Mel Spectrogram

Action Units

Action Units have already been introduced in section 2.2. OpenFace allows the extraction of both binary and continuous values via two different models (hence their values might not always match). The binary values are either 0 or 1 for each frame of the video for 18 AUs. The continuous values originally range from 0 to 5 and were remapped to the 0-1 range and are present for 17 AUs at each frame. We investigated the use of both types of AUs and their combinations. A table of all the AUs used in our experiments is present in appendix A

Facial Key-Points Distances

The facial key-points distances have been used in some other studies such as [45, 30], as they give quite a lot of information about the changes on the face and can efficiently be used by temporal neural networks to extract emotions. Once again, we relied on OpenFace to extract the 68 facial 3D key-points shown in Figure 5.1. Since many landmarks were close to each other and in order to reduce the feature dimensionality, we decided to select 40 of them and computed the pairwise distance between them. It yielded a 780 dimensional vector (equation 5.1 with $N = 40$) for each frame of the video.

$$\text{Dim} = \frac{N \times (N-1)}{2} \tag{5.1}$$

In order to remove the inter-people variations such as the size of facial attribute (nose, mouth, eyes, ...), we applied some normalization to these distances. We tried 3 different types of normalization (plus no normalization at all). The first method consisted to choose the most neutral frame of the same video, and dividing each distance (for each frame) by the equivalent distance in that particular neutral frame. The second method was quite similar, except that we took the most neutral frame of the video labelled with the neutral emotion of the same actor vocalizing the same statement. It has the advantage that the same normalization is used between different videos of the same actors but the drawback that it requires the model to have access to a neutral video of the same actor, which was not a problem in our case since the dataset is annotated, but which would require an additional step in a real-life setting where the person wearing the HMD device would first be asked to pose with a neutral face. However, the results were significantly better using this approach making this additional step worth taking. We used equation 5.3 to compute the most neutral frame, which used both the binary and continuous AU values and choose the frame with the least activation by penalizing large continuous values for any of them. The last method we considered was inspired by [30], where we represented the facial landmarks as a connected graph and normalized each pairwise distance by the distance of the path between the two key-points on the graph. Note that we used only 18 key-points (153 dimensional vector using equation 5.1 with $N = 18$) as they did in the mentioned paper in order to use the same graph.

$$\begin{aligned} \text{neutral\_score}_f = &\sum_{i=1}^{18} \text{binary\_AU}_{i,f} + \\ &\sum_{i=1}^{17} \text{continuous\_AU}_{i,f} + \\ &\sum_{i=1}^{17} \mathbb{1}(\text{continuous\_AU}_{i,f} > 1) + \\ &\sum_{i=1}^{17} \mathbb{1}(\text{continuous\_AU}_{i,f} > 2) \times 3 \end{aligned} \tag{5.2}$$

$$\text{most\_neutral\_frame} = \arg\min_f \{\text{neutral\_score}_f\} \tag{5.3}$$

Eyes Key-Points Distances

The eyes key-points distances follow the same idea as the facial key-points and the same normalization schemes. The only difference resides in the landmarks chosen. For these features, 12 precise eyes key-points (Figure 5.2) from each eyes were selected and the pairwise distance for each individual eye was computed. Then 10 landmarks on the eyes and eyebrows (Figure 5.1) were taken in order to compute some inter-eyes distances. The eye-gaze (angle for both axes in radians in world coordinates averaged for both eyes) was added yielding a vector of dimension 179 ($179 = 66 \times 2 + 45 + 2$, where 66 and 45 come from equation 5.1 with $N = 12$, respectively 10).

These features are particularly meaningful in the context of HMD devices, because the eyes landmarks are already computed for other related tasks and hence they don't add much overhead (only normalization must be applied).

5.2.2 Learnt

Wav2Vec

Wav2Vec [26] is a model working directly on raw audio. It was trained in a self-supervised manner by learning representation from unlabelled audio signal. It is composed of a feature encoder which consists of several blocks of convolutions followed by layer normalization and non-linear activation functions. The output of this feature encoder is then fed to a Transformer module [25, 46, 47]. For pre-training, a contrastive loss was employed, which requires to identify the true quantized latent speech representation for a masked time step within a set of distractors. See Figure 5.6 for a full sketch of the model.

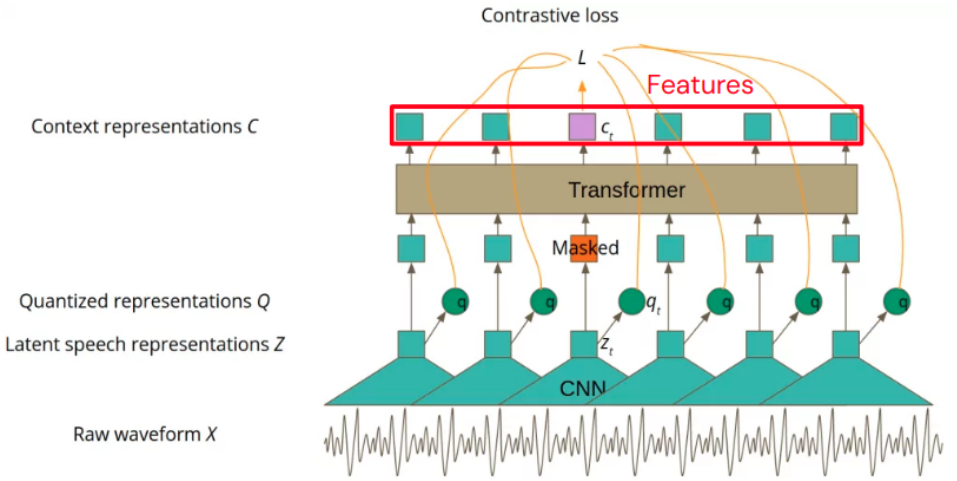

Figure 5.6: Wav2Vec architecture. Original image comes from [48].

Wav2Vec is present with two different versions: a base model having 95M parameters and a large one having 317M parameters. In our experiments, we used the base model for the binary experiments as it was enough to have a good overall picture and we didn't aim to achieve the very best accuracy, but the large one for the multi-class setting in order to make our results more comparable to the literature. We froze the feature encoder which had been trained on enough data and fine-tuned only the transformer modules by using the context representations (of shape 146×1024 for the large model, 146×768 for the base one) as features and fine-tuning the model by feeding them to our temporal model. Note that these features already contain temporal information learnt through the transformer modules and therefore using a simple linear layer on a time-average of the context representation (as it was done in [29]) instead of a full temporal model for the fine-tuning works just as fine (might even prevent some overfitting), but we decided to use the temporal model to have a common framework in regards to the others features. To really test the efficiency of the temporal model, we might rather have tried to use the latent speech representation as input features, but it would probably have lowered the accuracy.

ResNet

ResNet [36] is a deep convolutional network pretrained for image classification on the extensive ImageNet dataset[49]. It is a very deep network working efficiently by adding residual connections between blocks of convolutions in order to propagate the loss signal more easily into the network (see Figure 5.7). It was proven to work very well for a lot of different visual tasks and is a common choice for transfer learning applications.

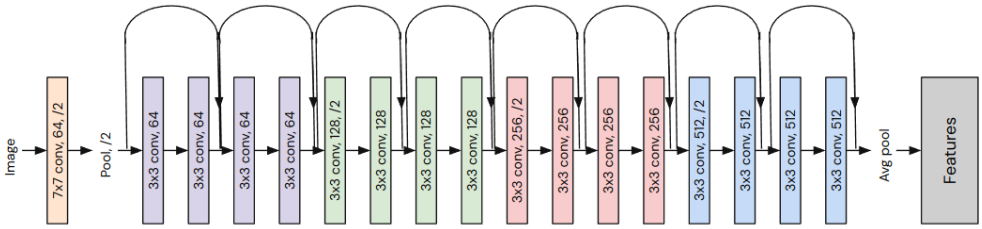

Figure 5.7: ResNet18 architecture.

In our experiments, we used the smallest ResNet architecture (ResNet18) which is less resource consuming than its bigger counterparts. To fine-tune it, we removed the classification head and treated the output of the last layer (a 512 dimensional vector) as a feature vector. By computing the aforementioned feature vector for different frames of the video, we end up with a time-series that can be fed to our temporal model.

Autoencoder

For the eyes, we also trained a fairly simple autoencoder (AE). The hope was to be able to learn features specific to the eyes with a small model, potentially by leveraging unlabelled data, since there isn't any famous pretrained model for eyes data specifically. We tried to train our model on the RAVDESS data only using a simple reconstruction loss and to then use the latent code (hence removing the decoder part) as features. We tried to both fine-tune the encoder for emotion classification and to freeze it. We also pretrained our model with more data using a private eye dataset from Magic Leap, which had however quite a big domain shift due to different orientation and quality of the data. Our preliminary results were however significantly worse than those obtained with ResNet and because of time constraints, we didn't explore this approach further. The main explanations for the quite disappointing results are the simplicity of the model, which was just constituted of two convolutional layers with batch normalization and max-pooling followed by two linear layers for the encoder part and the equivalent opposite layers for the decoder; and the small amount of data (with a strong domain shift) used to train it. We still believe that this approach might be interesting, but also that using a stronger model than a simple AE might be beneficial as it rarely gives state-of-the-art (SOTA) results in current times and that crawling more eyes data from the internet would also help.

FabNet

FabNet [37] is a model leveraging self-supervised learning in order to learn facial attributes from videos of a human performing common tasks. It is composed of an encoder mapping the image to a lower embedding space. For the pre-training, two frames (source and target frames) were fed so that their embeddings were computed and concatenated. Then a decoder learnt an offset which was used by a bilinear sampler on the source frame to reconstruct the target frame and a reconstruction loss was applied. It forced the encoder to learn useful facial attributes that can be used for other downstream tasks. The authors claimed to have state-of-the-art results at that time (2018) for self-supervised methods.

In our experiments, we tried to fine-tune the encoder on RAVDESS for emotion prediction. The preliminary results obtained were also lower than the results obtained by ResNet and therefore we didn't extend our researches using it. It is still interesting to mention that the results were not much worse, especially if we take into account the fact that the model is more than 2 times smaller than ResNet18.

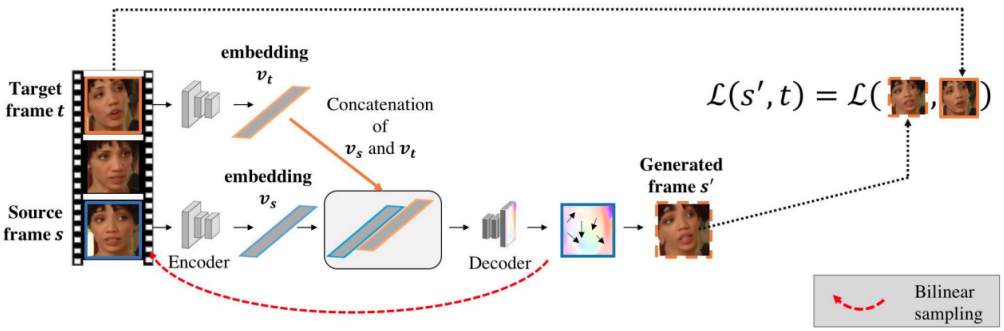

Figure 5.8: FabNet architecture from the original paper.

5.3 Models

In this section, we will introduce the models we tried for the actual emotion classification. All of them take one or more features mentioned in the previous section as input and output the probability of each of the 8 emotions. The 2 first models (5.3.1, 5.3.2) were tried mainly to reproduce results from the literature in a first stage and the last two models (5.3.3, 5.3.4) were the ones actually used as main components for our experiments.

5.3.1 CNN-14

2D-CNN models have been proven to work well for many application and audio is no exception. Nevertheless, a significant amount of data is required to train them. For that reason and in order to partially reproduce (we didn't use the exact same preprocessing steps and hyperparameters) the best results obtained in [21], we decided to fine-tune the CNN-14 model from PANNs [50]. This model was pre-trained on the large-scale AudioSet dataset, which is composed of million of audio events designed to classify sounds. It works by first computing a spectrogram from the audio and treat it as an image by its convolutional layers.

5.3.2 Frame Classification and Pooling

Following the work from [21], we also tried to classify the emotion at the frame level and use a max pooling strategy to aggregate the results for the full video. To do so, each frame inherited the label from its parent video and we fine-tuned ResNet, FabNet or trained a very simple self-made CNN (constituded of 4 layers of convolutions, ReLU, max pooling and batch normalization followed by an average pooling and a final linear layer) to classify each frame individually. A max-pooling strategy was then applied to all the frames to predict the emotion which was most present in the video as the final prediction (Figure 5.9). The main drawback

from this method is however the labelling of the frames which is quite inaccurate, because the emotion present in the full video might not be present at a given frame (actually, most of the frames are rather neutral) and by inheriting the label of the video, it gives a wrong learning signal to the model. In order to reduce this labelling problem, we also tried to classify groups of 30 frames (1 second of video) together using our simple CNN by concatenating the different frames along the channel axis and then applying the max-pooling operator. This method happened to give better results than the classification at frame level (even thought the CNN model was much simpler than a full ResNet) and also better than classification of the full video with frames concatenated as channels with the same small CNN, showing the potential advantage of averaging over different intervals of the video and classify more than one frame in order to reduce the labelling problem.

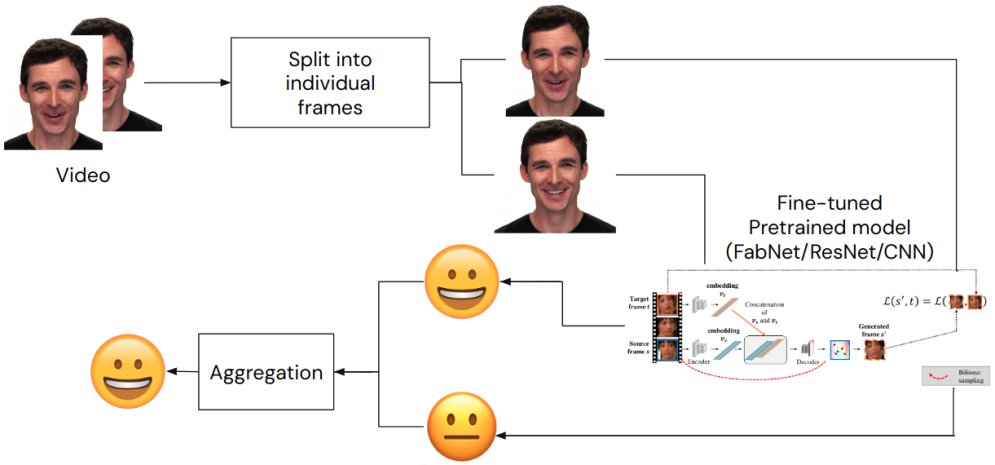

Figure 5.9: Frame classification and pooling framework.

5.3.3 TIM-Net

TIM-Net [20] is a recent temporal model which was proven to achieve a high accuracy on different emotional datasets, including RAVDESS. The original implementation used the MFCCs as input features, but any other time-series could work as its main purpose is to learn temporal dependencies at different scales. In our experiments, it was used as the main unimodal model with different features from different modalities. Figure 5.10 shows the full architecture in detail.

The model works by taking the time-series as input and feeding it to two similar branches, one reversed in time, in order to learn bi-directional relationships (from the past and from the future). The temporal features are then fed to n of so called Temporal-Aware Blocks (TABs), whose purpose is to capture temporal dependencies at different scales. They are constituted of different layers (originally 2) of causal 1D convolution, batch normalization, ReLU activation and spatial

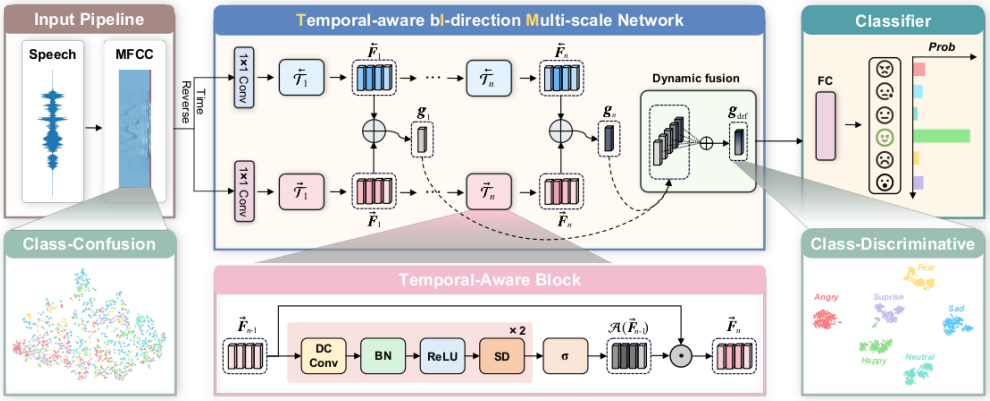

Figure 5.10: TIM-Net model from the original paper.

dropout followed by a final sigmoid activation. The receptive field of each TAB is specified by the dilation value of its convolution, which is an increasing power of 2 (from to 2^0 to 2^{n-1}). The features with the same receptive fields in the two directions are then combined and averaged over time, providing n features of different temporal scales. They are then fused in the multi-scale fusion component where a weight is learnt for each of the scale and the features are added proportionally to these weights. It outputs a final feature vector which compiles the full modality and which is discriminative for emotion classification. This feature vector is only processed by a final linear layer and a softmax for the final prediction.

5.3.4 Multi-Net

We implemented our own multimodal model called Multi-Net (see Figure 5.11). Its architecture is highly inspired by TIM-Net which is one of its main component. It works by first extracting features for each of the modalities independently, then feeding them to a TIM-Net module to extract discriminative features from all of the different sources. These features are then concatenated and used as input to a Multi-Layer Perceptron (MLP) made of linear layers, ReLU activations, batch normalizations and dropouts. Finally a softmax is applied in order to output the probability of each emotion. This model has the advantage to learn jointly from different modalities in comparison to an ensemble of unimodal models where each model is trained independently and relationships between them are harder to make. On the other hand, this model is more likely to overfit to the training set since it is quite big and receives a lot of input signals. Therefore it would benefit from more training data.

It is interesting to mention that we also tried alternate architectures for our model. Two main modifications were considered. The first one was to fuse the

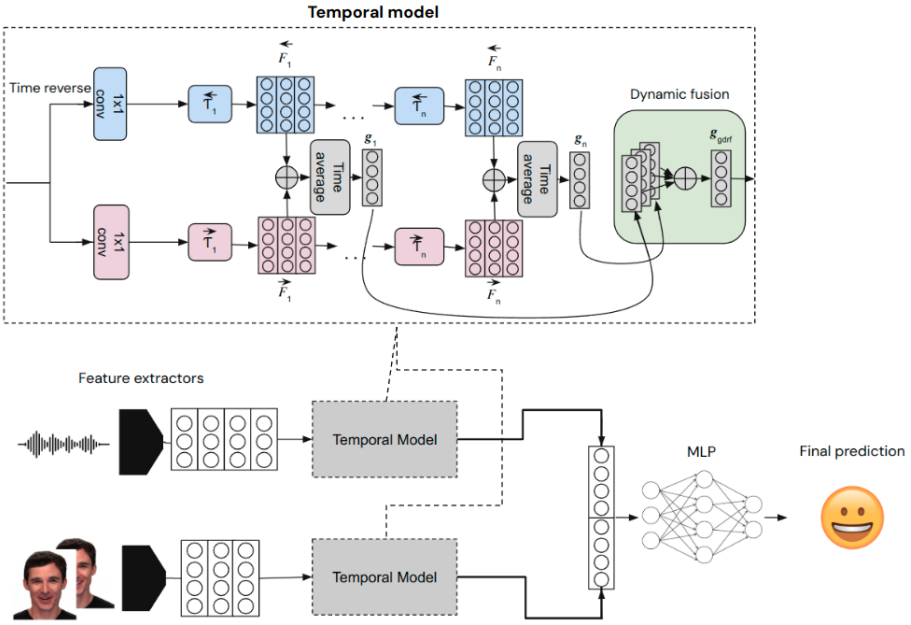

Figure 5.11: Multi-Net architecture

features at different scales (g_i on the diagram) of different modalities together before passing them to a single dynamic fusion module. For that, we tried to either concatenate the features of the same scale of different modalities and learn a common weight (see Figure 5.12) or to just apply the dynamic fusion on all of the features at different scales of all of the modalities without concatenation.

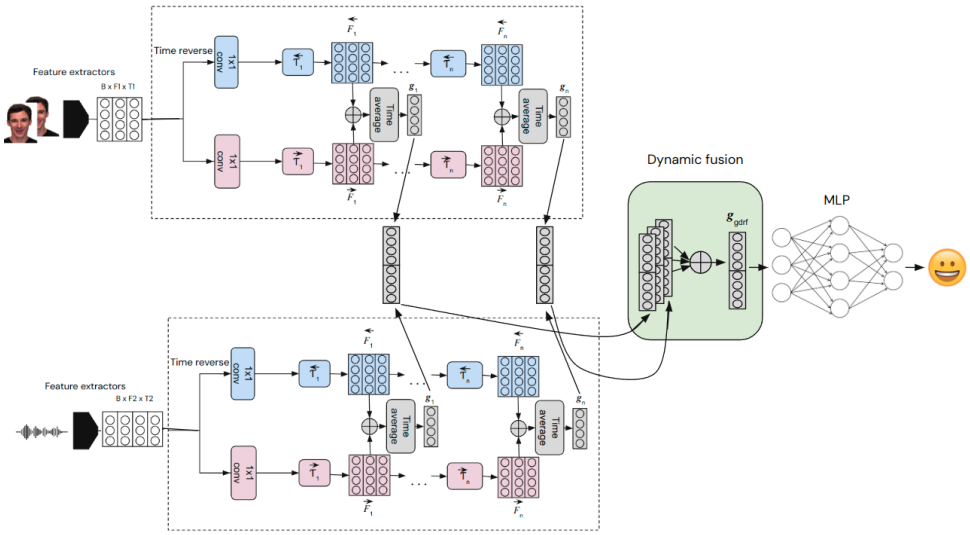

Figure 5.12: Alternative architecture tried for Multi-Net

The second idea was to add a multi-head attention module before the dynamic fusion, in order to let the model adapt itself and rely more or less on one of the modality depending on the input sample by scaling the features with attention.

None of these modifications improved the results significantly (the results were overall very close or slightly worse, but we didn't make any throughout statistical analysis), so we stuck to the model from Figure 5.11 and used it as the main multimodal model in the experiments.

Chapter 6

Experiments and Results

In this section, we will discuss about the different experiments we tried out and their results.

6.1 Settings

In this section, we will shortly explain some of the settings that were used in our experiments.

First of all, it is important to mention the way the dataset was split into different sets. In the first part of the project, we mainly wanted to be able to train many models and just have an overview of what seemed to work well and what not, without any stronger guarantees. Nevertheless, we made the choice to train our models by splitting the dataset by actors, meaning that all the videos of one actor either end up in one set or in the other, ensuring a certain level of robustness without adding any computational cost. This also implies that the results would be lower since the models will never have seen the actors on which they will be tested during training, but it is closer to what is usually experienced in a real-life setting where the models must work with new users without having to be retrained. For our preliminary experiments, we hence used the 20 first actors as a training set, actors 21 and 22 as a validation set to apply early stopping and choose the best checkpoint and actors 23 and 24 for the final testing of the model. Note that all of the sets are therefore balanced between genders. We also tried a lot of combinations of hyperparameters and only reported the results of the very best models (which might just have been lucky runs). The results are hence not very relevant statistically (the model might just have given good results to this particular choice of actors) and should be taken with a grain of salt. For these reasons, they are only presented in appendix E, mainly as additional information and explanation to some of the choices we made, but should not be used for comparison with other works or to make any stronger claims. In the second part of the project, we wanted to be able to compare our results with some previous studies and to have more confidence about the accuracy we obtained. Following

the work from [21, 29], we decided to use a 5 folds cross-validation scheme, where 4 folds were used for training and 1 for testing, without proper early stopping (we used early stopping on the training set itself, meaning we took the epoch with the lowest training loss, which gave however results very close to the ones we obtained when we just used the last checkpoint). The same folds where used as in the mentioned papers, i.e.:

- **Fold 0**: actors 2, 5, 14, 15, 16
- **Fold 1**: actors 3, 6, 7, 13, 18
- **Fold 2**: actors 10, 11, 12, 19, 20
- **Fold 3**: actors 8, 17, 21, 23, 24
- **Fold 4**: actors 1, 4, 9, 22

We didn't compute a confidence interval for our results, but since they are averaged over 5 folds, the models are less likely to be very lucky or unlucky on all of the folds at the same time (the variance is reduced) and also more robust to different actors, letting us making stronger claims about the results. This training scheme was used to compute the results displayed in tables 6.1, 6.2 and 6.3, where the results on the testing sets are reported. Note that these results were overall lower than the results we obtained without cross-validation and can be better compared with previous works.

In our experiments, we always used label smoothing (with $\epsilon = 0.1$) which is a regularization technique that introduces noise for the labels. In real life, different emotions might be present at the same time as a combination and therefore having only 1 full emotion for each video might be to restrictive. Therefore, label smoothing distributes the probability of the emotion more uniformly between all of the emotions. Mathematically, it is formulated as follows:

$$p_t = 1 - \frac{(c-1)\epsilon}{c} = 1 - 0.0875 = 0.9125 \tag{6.1}$$

$$p_f = \frac{\epsilon}{c} = 0.1/8 = 0.0125 \tag{6.2}$$

Where p_t is the new probability of the true emotion and p_f the new probability for all of the other emotions.

The last important elements concerning the experimental settings are the preprocessing of the data and the chosen hyperparameters. Since they varied quite a lot across experiments, further details about them are provided in appendices B and C.

6.2 Binary Experiments

In order to understand the real impact of the different modalities and their corresponding features on all of the emotions, researchers often rely on confusion matrices, which are matrices showing the true emotion on an axis and the predicted one on the other. Even thought they usually give a pretty good overview of how well the model worked, we realized in some of our experiments that they could suffer quite a lot from the randomness of the seed and the local optimum reached during training. Therefore, two very similar models (or sometimes even the same model trained twice with a different random initialization) could sometimes end up having quite different confusion matrices, even if their overall accuracy is very close.

For this reason, we opted to train models which are specialized for the binary classification of one emotion only (presence versus absence of that particular emotion), which gave us very comparable results across different runs. If the emotion can be classified well, it means that the features are discriminative for that particular emotion. It also gives an overall idea of how well a multi-class model could potentially work and where it might struggle. Moreover, it exhibits how the different modalities could boost each other and what the best combinations could be.

In order to run this experiment on our dataset, we set the label of all of the videos having a different label than the emotion we were classifying as being not of the given emotion. The main problem with this approach is that it makes the dataset very unbalanced as almost (not exactly since neutral has half as much samples) $\frac{7}{8}$ of the samples will be labelled as not having the emotion. This usually makes the model always predict the absence of the emotion and not learning anything useful. In order to resolve this issue, we applied weighting of the two classes, which penalizes more the model to fail to predict the presence of an emotion (predicting its absence when it was present) than its absence (predicting the presence when it was absent) and therefore encourages the model to predict more often the presence of the emotion than it would otherwise. The weight of class c (either the emotion or not the emotion) was computed with equation 6.3 in order to balance the dataset.

$$w_c = \frac{\text{total\_num\_samples}}{\text{num\_classes} \times \text{num\_samples}_c} = \frac{1440}{2 \times \text{num\_samples}_c} = \frac{720}{\text{num\_samples}_c} \tag{6.3}$$

The results obtained with TIM-Net are presented in table 6.1. The macro-F1 score was used as the main metric as it takes into account the data unbalance by averaging over the two classes, which would not be reflected in the accuracy. For the full results including accuracy and F1-score of both classes for each emotion, look at appendix D.

Modality	Type	Features	Emotion							
			Angry	*Calm*	*Disgust*	*Fear*	*Happy*	*Neutral*	*Sad*	*Surprise*
Audio	Speech	LMS	0.83	0.83	0.85	0.76	0.79	0.70	0.68	0.81
		Wav2Vec2-B	**0.91**	**0.88**	**0.93**	**0.87**	0.85	**0.82**	0.75	**0.88**
Video	Face	ResNet	0.81	0.82	0.84	0.76	**0.90**	0.75	0.75	0.70
		KPD	0.78	0.82	0.85	0.73	0.89	0.79	**0.76**	0.67
		AU	0.77	0.78	0.79	0.67	0.87	0.63	0.62	0.63
	Eyes	ResNet	0.68	0.72	0.77	0.75	0.75	0.68	0.69	0.65
		KPD	0.69	0.69	0.69	0.73	0.73	0.73	0.69	0.59
	Mouth	ResNet	0.80	0.77	0.80	0.67	0.88	0.77	0.63	0.68
	Head Pose	RPY-3D	0.65	0.65	0.56	0.54	0.55	0.55	0.53	0.53

Table 6.1: Macro F1 Score for different emotions, modalities, features and TIM-Net as temporal model. LMS = Log Mel Spectrogram, Wav2Vec2-b = Wav2Vec2-Base, KPD = Key-Point Distances, AU = Action Units, RPY-3D = Roll, Pitch, Yaw and 3D location.

Many interesting discoveries can be highlighted out of this table. First of all, the learnt features manage to give better results than the hand-crafted ones as expected, Wav2Vec being overall the most accurate model. It's interesting to notice that while the audio works better for most of the emotions, video beats it for happiness and sadness. Happiness seems to be mainly provided by the mouth whereas sadness rather from the eyes. The key-points distances always give better results than the AU and very close ones to the ResNet features, showing their great potential. They are especially good for the neutral emotion, probably because the lower number of samples doesn't allow ResNet to learn enough information about neutral characteristics. The mouth seems to provide more emotional information than the eyes in general (except for fear and sadness) as the results obtained are closer to the ones with the full face and the results from the eyes are overall quite poor. Interestingly, some useful signal seems to reside in the head pose, mainly for the classification of anger and calmness even thought this modality achieves way worse results than the other ones as we would have expected.

6.3 Multi-Class Experiments

In this section we highlight our results for the multi-class emotion classification with both unimodal and multimodal models. For the audio we used the large Wav2Vec model as a learnt feature extractor since it is known to give very good results, whereas we used the LMS as an efficient feature since it gave us better results than the MFCCs on our preliminary experiments. We didn't try the autoencoder or FabNet as they were less accurate than ResNet in our previous experiments. Furthermore, we only showed the unimodal results for the head pose, but we tried to combine it with other modalities in some experiments as

Modality → ↓			Audio		
	Type → ↓		Speech		
		Feature → ↓	None	LMS	Wav2Vec-L
	-	None	-	65.6	81.63
Video	Eyes	ResNet	53.18	53.18	82.48
		KPD	52.35	70.30	81.42
	Mouth	ResNet	61.58	59.97	82.04
	Eyes + Mouth	ResNets	63.48	60.98	83.51
	Face	ResNet	71.20	69.58	**83.57**
		KPD	67.13	**76.82**	81.63
		AU	58.63	70.93	81.65
	Head Pose	RPY-3D	33.85	-	-

Table 6.2: Accuracy for multimodal experiments. "None" in the first numerical row and column indicates single-modality cases; other entries are multimodal. LMS = Log Mel Spectrogram, Wav2Vec2-L = Wav2Vec-Large, KPD = Key-Point Distances, AU = Action Units, RPY-3D = Roll, Pitch, Yaw and 3D location.

well without managing to increase their performance (except with the eyes where the results were improved by 1.98% for the KPD and 0.56% for ResNet), so we think that the signal present in that modality is too weak or redundant and hence doesn't help much. Table 6.2 shows the main results of our experiments and table 6.3 compare our best results from audio and video with the two other papers that adopted the same data split and human accuracy.

Paper	**Audio Acc. (%)**	**Video Acc. (%)**	**multimodal Acc. (%)**	**multimodal Improvement (%)**
Human accuracy [6]	62	**72**	80	8
[21]	76.58	57.08	80.08	3.5
[29]	**81.82**	62.13	**86.70**	4.88
Our (best features)	81.63	71.20	83.57	1.94
Our (efficient features)	65.6	67.13	76.82	**9.69**

Table 6.3: Comparison of our framework with previous works using the same evaluation scheme and human accuracy. Best and second best models are highlighted for each modality. Last column shows the improvement of the multimodal model with regard to its best unimodal counterpart.

From these tables, we can confirm that learnt features work better and give an overall higher accuracy. For the unimodal models, Wav2Vec features surpassed

all of the other ones, followed by the ResNet features from the video and the key-point distances. The action units performed surprisingly very poorly, which might mean that OpenFace wasn't able to extract them perfectly since there should normally exist a one-to-one mapping between them and the emotions. The eyes didn't perform extremely well and are a weaker source of information compared to the mouth. The head pose managed to give results significantly better than randomness, but still not good enough to add information to the other modalities (except for the eyes which are quite weak as well). On the multimodal front, the combination of the bigger learnt feature extractors gave the best results, but not a huge improvement over their unimodal counterparts. This is probably due to the fact that the features overfitted to the unimodal setting and hence could not be perfectly combined by the multimodal model. This theory seems coherent when we look at the combination of learnt and handcrafted features, one for each modality, which never worked well. The explanation could be that the learnt feature extractors were overfitting, giving features which were way easier to use than the hand-crafted ones from the other modalities and so the model learnt to classify the emotion solely from them. On the other hand, the efficient features seemed to have a great symbiosis, giving the best improvement of almost 10% over the unimodal models. It might be argued that one of the reason is that there is more room for improvement for them than for the learnt features where the unimodal models already achieved pretty good results, but we think that the fact that our multimodal model can learn jointly from both sources is one of the reason for these great improvement in comparison to an ensemble of models as proposed in the two other papers. From an HMD perspective, it is also interesting to see that the eyes key-point distances give a boost of 4.7% to the audio LMS, as both can be very easily computed and therefore could be directly used on device. As a final conclusion on these results, it is nice to point out the fact that our best model, even if it isn't state-of-the-art, is more accurate than humans are, especially from the audio, showing the great potential of AI to detect emotions. Even our model using more efficient features obtain results which are comparable to human accuracy which is encouraging for future researches on the topic.

Chapter 7

Conclusion

This final chapter will first be used to give an overview of the main findings from this project. Then we will conclude by giving some ideas for future researches on the topic.

7.1 Overview

In this work, we focused on the emotion classification from multimodal sources and we tried to understand which modality was most discriminative for different emotions. We started by looking at what had previously be done and by searching for the right dataset. As not much studies had been conducted on the impact of partial parts of the face and on head-pose and therefore no dataset seemed to fit exactly our requirements, we opted to focus on RAVDESS and to extract ourselves the modalities and different features out of the full video. We then used the full face to partially reproduce some results from the literature and achieved results which were quite close using some of their techniques (CNN-14 and Wav2Vec fine-tuning, frame level classification with aggregation at video level with max-pooling). Then, we wanted to better understand the impact of the different modalities on each emotion. Since our focus was on HMD, we decided to compute some efficient features which could be used in real-time and some more heavy ones in order to make better comparisons between their performance. We used these different features for the binary classification of the emotions, which gave us many useful insights. The main finding was that video and audio were more adapted to learn different emotions (anger and surprise by audio, happiness by mouth and sadness by video, mainly from eyes), highlighting the intrinsic advantage of using a multimodal approach. Finally, we aimed to compare the accuracy of unimodal and multimodal models. To create a unimodal benchmark, we decided to use TIM-Net as our temporal model on top of our different features. Taking inspiration from it, we developed our own multimodal model called Multi-Net, which learns discriminative features from different modalities with a temporal model, concatenate them and fuse them with a final MLP. Using this model, we ran many experiments combining the different features together. Our best model

used the learnt features from Wav2Vec and ResNet and achieved an accuracy of 83.57%, surpassing human accuracy on the task. However, it didn't improve the unimodal results as much as we would have expected and didn't achieve SOTA results, probably due to the fact that the feature extractors partially overfitted to the unimodal setting, making it hard for the model to process the two modalities together. Nevertheless, Mutli-Net showed more promise when it was used with efficient features, improving the unimodal counterpart by almost 10% with full face key-points and by almost 5% with the eyes key-points when they were combined with the log mel spectrogram. In summary, the main contributions of our work are the following:

- We partially reproduced some experiments from the literature, establishing a strong benchmark for different modalities.
- We analysed the impact of many modalities and different types of features, some lightweight and some learnt, on the detection of 8 basic emotions, showing that different modalities were better adapted for different emotions.
- We developed our own end-to-end multimodal model inspired by TIM-Net, moving away from conventional ensemble-based approaches. It was proven to work well for the fusion of efficient features, boosting their individual performance by almost 10%. When used on top of learnt features, it even managed to beat humans by reaching an accuracy of 83.57%.

7.2 Future Work

In this final section, we will discus some ideas for future work on the subject.

- **Improving generalization:** our model was only trained and tested on the RAVDESS database, giving very few guarantees about its performance in real-life settings when the data comes from another distribution. Also, it was only trained on two different English sentences, thus we cannot attest that it would generalize to other languages or different sentences in English. Finally both modalities were always present, which is not the case in real-life settings where the person wearing the HMD device might not talk for a while and therefore our current model would probably fail in such scenarios. Using modality dropout as proposed in [41] might be a solution to overcome this issue. In general, we didn't pay a lot of attention to generalization in our study but it would definitely be an important step to conduct next, especially if we would like to use the model on device.
- **Better data labelling with clustering:** it is still not extremely clear why the current models fail to predict the emotions correctly, but it might

somehow be linked to the way the actors are playing the emotions which might not always be optimal. In order to see if some of the videos are played poorly, it would be interesting to cluster the dataset by emotions and analyze the outliers. By removing them from the training set, the model might be able to predict the emotions more accurately and it would give more insights about why it might struggle on some samples more than on others.

- **Per person training while wearing the device:** usually the models used on device are trained beforehand and deployed on device to work with new users. However, it would be interesting to be able to add a component to the architecture which learns specific features about the user gradually while it wears it in order to improve the overall accuracy. It is still not extremely clear how this could be best implemented in practice, but the idea is worth thinking about.

- **Make the model run in real-time on device:** using the efficient features, it should theoretically be possible to make the model run on device. However, mainly due to time constraint, we were not able to demonstrate it. Hence it would be interesting to develop a demo on device, proving the feasibility of the approach.

Bibliography

[1] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[3] C. Hjortsjö, *Man's Face and Mimic Language*. Studentlitteratur, 1969. [Online]. Available: https://books.google.ch/books?id=BakQAQAAIAAJ

[4] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[5] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System: Facial action coding system : the manual : on CD-ROM*, ser. Facial Action Coding System. Research Nexus, 2002. [Online]. Available: https://books.google.ch/books?id=wphFzwEACAAJ

[6] 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, 2018.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[8] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[9] M. J. Lyons, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets (ivc special issue)," *arXiv preprint arXiv:2009.05938*, 2020.

[10] M. J. Lyons, "" excavating ai" re-excavated: Debunking a fallacious account of the jaffe dataset," *arXiv preprint arXiv:2107.13998*, 2021.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer, 2013, pp. 117–124.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[13] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[15] C. Eroglu Erdem, C. Turan, and Z. Aydin, "Baum-2: A multilingual audio-visual affective face database," *Multimedia tools and applications*, vol. 74, no. 18, pp. 7429–7459, 2015.

[16] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 907–929, 2019.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[18] K. Kaur and P. Singh, "Trends in speech emotion recognition: a comprehensive survey," *Multimedia Tools and Applications*, pp. 1–45, 2023.

[19] K. Venkataramanan and H. R. Rajamohan, "Emotion recognition from speech," *arXiv preprint arXiv:1912.10458*, 2019.

[20] J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on ravdess dataset using transfer learning," *Sensors*, vol. 21, no. 22, p. 7665, 2021.

[22] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[23] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[24] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective computing*, vol. 13, no. 2, pp. 992–1004, 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[27] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[28] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[29] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," *Applied Sciences*, vol. 12, no. 1, p. 327, 2021.

[30] N. Chanthaphan, K. Uchimura, T. Satonaka, and T. Makioka, "Facial emotion recognition based on facial motion stream generated by kinect," in *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2015, pp. 117–124.

[31] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[32] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[33] C. Turan, K.-M. Lam, and X. He, "Facial expression recognition with emotion-based feature fusion," in *2015 Asia-Pacific Signal and Information Processing*

Association Annual Summit and Conference (APSIPA). IEEE, 2015, pp. 1–6.

[34] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Computing and Applications*, pp. 1–18, 2021.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] O. Wiles, A. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," *arXiv preprint arXiv:1808.06882*, 2018.

[38] J.-R. Chang, Y.-S. Chen, and W.-C. Chiu, "Learning facial representations from the cycle-consistency of face," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9680–9689.

[39] H. Yong, J. Lee, and J. Choi, "Emotion recognition in gamers wearing head-mounted display," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 1251–1252.

[40] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[41] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 2822–2828.

[42] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 4, 2021, pp. 3510–3519.

[43] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, 2022.

[44] M. S. Nagawade and V. R. Ratnaparkhe, "Musical instrument identification using mfcc," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2017, pp. 2198–2202.

[45] N. Sheng, Y. Cai, C. Zhan, C. Qiu, Y. Cui, and X. Gao, "3d facial expression recognition using distance features and lbp features based on automatically detected keypoints," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2016, pp. 396–401.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[48] "Wav2vec 2.0: A framework for self-supervised learning of speech representations," https://neurosys.com/blog/wav2vec-2-0-framework, accessed: 2023-10-11.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[50] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[51] "Facs - facial action coding system," https://www.cs.cmu.edu/~face/facs.htm, accessed: 2023-10-09.

Appendix A

OpenFace's Action Units

In this appendix, we give a complete overview of all the action units extracted with OpenFace that were used in our experiments. All AUs were present with both binary and continuous values except AU28 which was only in binary format. The table and the images are taken from [51].

AU	Description	Facial Muscle	Example
1	Inner Brow Raiser	Frontalis, pars medialis	
2	Outer Brow Raiser	Frontalis, pars lateralis	
4	Brow Lowerer	Corrugator supercilii, Depressor supercilii	
5	Upper Lid Raiser	Levator palpebrae superioris	
6	Cheek Raiser	Orbicularis oculi, pars orbitalis	
7	Lid Tightener	Orbicularis oculi, pars palpebralis	

AU	Description	Facial Muscle	Example
9	Nose Wrinkler	Levator labii superioris alaquae nasi	
10	Upper Lip Raiser	Levator labii superioris	
12	Lip Corner Puller	Zygomaticus major	
14	Dimpler	Buccinator	
15	Lip Corner Depressor	Depressor anguli oris	
17	Chin Raiser	Mentalis	
20	Lip stretcher	Risorius	
23	Lip Tightener	Orbicularis oris	

AU	Description	Facial Muscle	Example
25	Lips part	Depressor Labii, Relaxation of Mentalis, Orbicularis Oris	
26	Jaw Drop	Masseter, relaxed Temporalis and internal Pterygoid	
28	Lip Suck	Orbicularis oris	
45	Blink	Relaxation of Levator palpebrae superioris; Orbicularis oculi, pars palpebralis	-

Table A.1: Action Units detected by OpenFace used in our experiments.

Appendix B

Data Preparation

In this appendix, we outline the steps taken to prepare each modality for our experiments, ensuring the reproducibility of our results. This section is copied from our paper called "The Role of Facial and Speech Features in Emotion Classification" which was just submitted before the writing of this document started and is currently under review and hence wasn't added to the Bibliography.

- *ResNet:* ResNet served as the feature extractor for eyes, mouth, and face in our experiments. All images were resized to 224x224 pixels. For facial images, each frame was first center-cropped to 720x720 pixels to remove the borders, which did not contain facial information. These cropped frames were then scaled from 0.0 to 1.0 and normalized using mean=[0.485, 0.456, 0.406] and standard deviation (std)=[0.229, 0.224, 0.225], following the pre-training procedures outlined in PyTorch's documentation. In unimodal and binary experiments where fine-tuning was performed, we selected 20 equally spaced samples from the full video. For multimodal experiments, we employed the weights derived from the unimodal experiments, froze the ResNet extractor, and utilized the 88 middle frames of each video. We experimented with using 88 center frames for the unimodal experiments (with a frozen ResNet fine-tuned using 20 equally spaced samples from the video) to ensure that the superiority of our multimodal results was not merely due to the increased number of frames. However, this approach yielded slightly inferior results (0.3-1% reduced accuracy), and therefore, these findings are not reported in our tables.
- *Video key-points:* For video key-points, we used OpenFace to extract 40 3D landmarks from the initial 68, as some landmarks were closely positioned and didn't offer substantial additional information, hence were omitted to simplify the feature space. We calculated the pairwise distance between each landmark, constructing a 780-dimensional feature vector for each of the 88 middle frames in the video. These distances were then normalized against those from the most neutral frame—identified as the frame with the lowest activated action units—from the first repetition of a neutral video where the actor vocalized the same sentence. This process ensures a consistent and comparative basis for analysis across different frames and videos.

– *Eyes key-points:* For eyes key-points, we adopted an approach similar to the one used for video key-points. We extracted 12 specific key-points within each eye, designated as eye_lmk_i in OpenFace, and computed the pairwise distances within each eye individually. Additionally, pairwise distances between 10 chosen key-points located on both eyes and the eyebrows were concatenated to the initial distances to incorporate inter-eye values. These distances were normalized using the same procedure previously described for video key-points. The average gaze angles of both eyes across two axes were then calculated and appended to the data, resulting in a 179-dimensional vector.
– *Action units (AU):* OpenFace supplies both binary action units (either activated or not, for 18 AUs) and a continuous intensity measure (ranging from 0 to 5, for 17 AUs). As these two sets of values are generated through different models, there might be inconsistencies in their correspondence. To address this, we integrated both types of values. The continuous intensities were first rescaled to a range between 0 and 1. Following this, we concatenated the rescaled intensities with the binary action units to form a combined 35-dimensional vector for each of the 88 frames.
– *Audio:* We applied straightforward preprocessing to the audio. This involved extracting a segment of 140,800 units in length from the center of the audio, which is equivalent to the duration of 88 frames or approximately 2.9 seconds. The stereo signal was then converted into a mono signal by averaging the two channels. In the case of the Wav2Vec experiments, the audio signal was resampled from 48KHz down to 16KHz and normalized by subtracting the mean and dividing by the standard deviation. For the LMS computation, we used 24 coefficients, a window length of 4800, a hop length of 1600, and a maximum frequency of 17kHz giving a 88x24 dimensional feature.

Appendix C

Hyperparameters

In this appendix, we offer a comprehensive overview of the architecture's hyperparameters and provide detailed training information necessary for reproducing our results. The complete set of hyperparameters, along with specific training details are shown in Table C.1. This section is copied from our paper called "The Role of Facial and Speech Features in Emotion Classification" which was just submitted before the writing of this document started and is currently under review and hence wasn't added to the Bibliography.

In our experiments, the MutiNet network utilized three layers within its temporal block, compared to the original TIM-Net's two, and employed a kernel of size 2 for 1D convolutions. We experimented with both 32 and 64 convolution channels, with 64 yielding superior results in most cases (the exceptions being the multi-class eyes ResNet and LMS + Eyes KPD experiments, reported with 32 channels only). The dilation factors used for multi-scale feature extraction were always powers of two, consistent with the original paper. When scales equal n, it indicates the use of n different temporal blocks with dilation factors ranging from 2^0 to 2^{n-1}. These factors were selected to ensure the largest temporal block dilation factor was smaller than the total feature temporal length (hence for 88 frames, we chose 7 so that $2^6 = 64 < 88$).

Model	Experiment	Learning Rate	Batch Size	Epochs	Scales
Wav2Vec2-L Wav2Vec2-B	both	0.00005	16	100	8
LMS	both	0.001	64	100	7
Video ResNet Mouth ResNet Eyes ResNet	both	0.0002	8	60	5
Face KPD Eyes KPD	multi-class	0.0005	64	100	7
Face KPD Eyes KPD	binary	0.0002	64	100	7
AU	both	0.0002	64	40	7
Head Pose	both	0.0002	64	100	8
Multimodal with Wav2Vec-L + ...	multi-class	0.0001	64	100	7
Multimodal with LMS + ...	multi-class	0.002	64	100	7

Table C.1: Hyperparameters used in our Experiments. Each experiment type is categorized as binary, multi-class, or both; "both" is used when identical hyperparameters were applied to both experiment types. "Scales" denotes the count of distinct scales at which each modality's features were extracted in the temporal model before being fused by the dynamic fusion module.

Appendix D

Full Binary Experiments Results

In this appendix, we give the full results of our binary experiments with accuracy and F1-score for both the presence and absence of the emotion

Modality	**Features**	**Accuracy (%)**	**F1-score angry**	**F1-score not angry**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	91.27	0.72	0.95	0.83	0.92
	Wav2Vec (base)	**95.92**	**0.84**	**0.98**	**0.91**	**0.96**
Video	ResNet	90.08	0.69	0.94	0.81	0.91
	Key-points Distance	88.50	0.64	0.93	0.78	0.89
	Action Units	86.40	0.62	0.92	0.77	0.88
Eyes	ResNet	81.80	0.47	0.89	0.68	0.83
	Key-points Distance	82.63	0.48	0.90	0.69	0.84
Mouth	ResNet	89.56	0.66	0.94	0.80	0.90
Head Pose	Head Pose	77.50	0.45	0.86	0.65	0.80

Table D.1: Binary results on RAVDESS for angry emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score calm**	**F1-score not calm**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	92.15	0.71	0.96	0.83	0.92
	Wav2Vec (base)	**94.57**	**0.80**	**0.97**	**0.88**	**0.95**
Video	ResNet	90.92	0.69	0.95	0.82	0.91
	Key-points Distance	90.08	0.70	0.94	0.82	0.91
	Action Units	87.92	0.64	0.93	0.78	0.89
Eyes	ResNet	85.48	0.53	0.92	0.72	0.86
	Key-points Distance	82.23	0.49	0.89	0.69	0.84
Mouth	ResNet	86.57	0.62	0.92	0.77	0.88
Head Pose	Head Pose	77.18	0.44	0.86	0.65	0.80

Table D.2: Binary results on RAVDESS for calm emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score disgust**	**F1-score not disgust**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	92.95	0.75	0.96	0.85	0.93
	Wav2Vec (base)	**96.70**	**0.87**	**0.98**	**0.93**	**0.97**
Video	ResNet	91.52	0.74	0.95	0.84	0.92
	Key-points Distance	92.23	0.74	0.95	0.85	0.93
	Action Units	88.38	0.65	0.93	0.79	0.89
Eyes	ResNet	87.07	0.61	0.92	0.77	0.88
	Key-points Distance	80.87	0.49	0.89	0.69	0.83
Mouth	ResNet	88.54	0.67	0.93	0.80	0.89
Head Pose	Head Pose	65.77	0.34	0.77	0.56	0.71

Table D.3: Binary results on RAVDESS for disgust emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score fear**	**F1-score not fear**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	86.95	0.59	0.92	0.76	0.88
	Wav2Vec (base)	**94.42**	**0.78**	**0.97**	**0.87**	**0.94**
Video	ResNet	88.47	0.59	0.93	0.76	0.89
	Key-points Distance	86.45	0.54	0.92	0.73	0.87
	Action Units	77.75	0.49	0.86	0.67	0.81
Eyes	ResNet	88.10	0.57	0.93	0.75	0.88
	Key-points Distance	85.83	0.54	0.92	0.73	0.87
Mouth	ResNet	86.14	0.41	0.92	0.67	0.85
Head Pose	Head Pose	65.53	0.30	0.77	0.54	0.71

Table D.4: Binary results on RAVDESS for fear emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score happy**	**F1-score not happy**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	89.38	0.65	0.94	0.79	0.90
	Wav2Vec (base)	93.23	0.74	0.96	0.85	0.93
Video	ResNet	**95.28**	**0.83**	**0.97**	**0.90**	**0.95**
	Key-points Distance	94.70	0.81	**0.97**	0.89	**0.95**
	Action Units	93.57	0.78	0.96	0.87	0.94
Eyes	ResNet	87.38	0.58	0.92	0.75	0.88
	Key-points Distance	86.48	0.54	0.92	0.73	0.87
Mouth	ResNet	94.28	0.80	**0.97**	0.88	0.94
Head Pose	Head Pose	66.97	0.32	0.78	0.55	0.72

Table D.5: Binary results on RAVDESS for happy emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score neutral**	**F1-score not neutral**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	89.13	0.45	0.94	0.70	0.91
	Wav2Vec (base)	**94.88**	**0.66**	**0.97**	**0.82**	**0.95**
Video	ResNet	92.35	0.54	0.96	0.75	0.93
	Key-points Distance	92.18	0.62	0.96	0.79	0.93
	Action Units	80.25	0.38	0.89	0.63	0.85
Eyes	ResNet	92.33	0.41	0.96	0.68	0.92
	Key-points Distance	89.68	0.52	0.94	0.73	0.91
Mouth	ResNet	92.90	0.58	0.96	0.77	0.94
Head Pose	Head Pose	70.60	0.28	0.82	0.55	0.78

Table D.6: Binary results on RAVDESS for neutral emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score sad**	**F1-score not sad**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	82.62	0.47	0.90	0.68	0.84
	Wav2Vec (base)	**88.13**	0.57	**0.93**	0.75	**0.88**
Video	ResNet	87.58	0.57	**0.93**	0.75	**0.88**
	Key-points Distance	87.28	**0.60**	0.92	**0.76**	**0.88**
	Action Units	74.13	0.40	0.84	0.62	0.78
Eyes	ResNet	84.15	0.47	0.91	0.69	0.85
	Key-points Distance	83.35	0.47	0.90	0.69	0.84
Mouth	ResNet	80.05	0.38	0.88	0.63	0.82
Head Pose	Head Pose	64.08	0.29	0.76	0.53	0.70

Table D.7: Binary results on RAVDESS for sad emotion with different features types.

Modality	**Features**	**Accuracy (%)**	**F1-score surprise**	**F1-score not surprise**	**macro avg**	**weighted avg**
Audio	Log Mel Spectrogram	91.20	0.67	0.95	0.81	0.91
	Wav2Vec (base)	**94.41**	**0.80**	**0.97**	**0.88**	**0.95**
Video	ResNet	84.33	0.49	0.91	0.70	0.85
	Key-points Distance	81.72	0.45	0.89	0.67	0.83
	Action Units	75.53	0.42	0.85	0.63	0.79
Eyes	ResNet	82.05	0.40	0.89	0.65	0.83
	Key-points Distance	78.55	0.32	0.87	0.59	0.80
Mouth	ResNet	84.37	0.45	0.91	0.68	0.85
Head Pose	Head Pose	63.08	0.32	0.75	0.53	0.69

Table D.8: Binary results on RAVDESS for surprise emotion with different features types.

Appendix E

Preliminary Multi-Class Results

In this appendix, we give an overview of some of the results we obtained in our preliminary surveys, where we always used the 2 last actors in the testing set and the 2 before in the validation set. As mentioned earlier, these numbers are mainly there as a complement to give an overview of a few more models we tried out, but should not be used as a comparison to previous works as they lack statistical significance and were sometimes computed using different preprocessing steps, making them harder to compare. A simple summary of some meaningful experiments and the best results on both validation and test set are given. Note that many more experiments were tried out, but we only reported the ones that seemed most meaningful.

Modality	**Features**	**Model**	**Test acc. (%)**	**Validation acc. (%)**
Eyes	FabNet	TIM-Net on 20 equidistant frames, extractor unfrozen	54	50
Eyes	ResNet	TIM-Net on 20 equidistant frames, extractor unfrozen	53	58
Eyes	Autoencoder	TIM-Net on 88 middle frames, no fine-tuning of feature extractor, pretrained on Magic Leap and RAVDESS eyes	43	53
Mouth	FabNet	TIM-Net on 20 equidistant frames, extractor unfrozen	48	60
Mouth	ResNet	TIM-Net on 20 equidistant frames, extractor unfrozen	53	67
Video	-	Max-Pooling on 88 first frames classified with ResNet fine-tuned on each frame with parent label	52	61
Video	-	Max-Pooling on 88 first frames prediction on groups of 30 frames separated by 5 frames classified with own CNN trained on each frame with parent label	61	62

Modality	**Features**	**Model**	**Test acc. (%)**	**Validation acc. (%)**
Video	FabNet	TIM-Net on 20 equidistant frames, extractor unfrozen	68	64
Video	ResNet	TIM-Net on 20 equidistant frames, extractor unfrozen	71	74
Video	KPD	TIM-Net on 88 equidistant frames, using 40 points and neutral frame of same video normalization	55	66
Video	KPD	TIM-Net on 88 equidistant frames, using 40 points and neutral frame of neutral video normalization	67	73
Video	KPD	TIM-Net on 88 equidistant frames, using 18 points and neutral frame of same video normalization	60	62
Video	KPD	TIM-Net on 88 equidistant frames, using 18 points and neutral frame of neutral video normalization	64	69
Video	KPD	TIM-Net on 88 equidistant frames, using 18 points and path normalization	63	61
Video	AU	TIM-Net on 88 equidistant frames, using AU regression values	53	68
Video	AU	TIM-Net on 88 equidistant frames, using AU binary values	60	64
Video	AU	TIM-Net on 88 equidistant frames, using AU both regression and binary values	63	66
Audio	Spectrogram	CNN-14 from PANN fine-tuning on 3 seconds of audio	71	76
Audio	LMS	TIM-Net on LMS with 32 coefficients, 2048 window length, 512 hop length max frequency of 24000 and 2.083 seconds of audio	70	76
Audio	MFCCs	TIM-Net on MFCCs with 39 coefficients, 2048 window length, 512 hop length max frequency of 24000 and 2.083 seconds of audio	67	66
Audio	Wav2Vec-b	TIM-Net on Wav2Vec features using 2.93 seconds of audio	84	84
Audio, Video	Wav2Vec-b, ResNet	Multi-Net on 2.93 seconds of video using late fusion	88	93
Audio, Video	Wav2Vec-b, ResNet	Multi-Net on 2.93 seconds of video using early concatenation fusion	85	93
Audio, Video	Wav2Vec-b, ResNet	Multi-Net on 2.93 seconds of video using early no concatenation fusion	86	93

Table E.1: Preliminary results used to guide our choices of models.