# Multimodal Emotion Recognition

Petros Mitseas[1]

[1]National Centre for Scientific Research "Demokritos"

June 23, 2022

## Abstract

In this assignment we examine the task of recognizing emotion from video and audio, using a combination of machine learning models. The multimodal predictions are used as input to a meta-classifier, to produce the final output. We test the performance on the RAVNESS dataset, and show that the combined model demonstrates improved performance over each separate model.

**Keywords**
SVM, CNN, multimodal, emotion, ensemble

## 1 Introduction

Recognizing human emotion is a challenging problem that involves many different scientific domains. Applications of this task can be found in education, product sales, customer support and product user experience, to name a few. Recent progress in artificial intelligence, deep learning and sensor technology makes it possible to combine audio, visual and physiological readings into models that can classify the emotional state of a person.

For classification based on extracted audio features, simple machine learning models such as Hidden Markov Models and Support Vector Machines can be used [1]. There are also approaches based on Convolutional Neural Networks that use raw features like the spectogram of the audio signal [2]. Latest approaches include fine-tuning a pretrained Wav2Vec[3] model, originally trained on large amounts of unlabeled audio data to perform this specific task [4]. Multimodal solutions have also been applied to this problem. Panagiotis Tzirakis et al. [5] used two separate deep CNNs for extracting features from audio and visual signals. The output was combined using a two-layer LSTM. In another research, Cristina Luna-Jiménez et al [6] also used two discrete networks for extracting the posteriors, which were combined using a support vector classifier, claiming an accuracy of 80%.

## 2 Methodology

Our model is based on combining multiple weak classifiers (base models) using a meta-classifier to achieve improved results.

### 2.1 Audio

The first base model is a Support Vector Classifier (SVM) which is trained on audio samples from a portion of the train set.

#### 2.1.1 Feature Extraction

The feature vector used as input to the model, consists of 48 features derived from time and frequency domain. Most common time domain features include Energy (i.e. the sum of squared values normalized by the window length), Zero Crossing Rate (the number of sign changes during the window) and the Energy Entropy which measures the changes in the signal's energy. In the frequency domain, we perform an FFT transform for a short fixed window of time and calculate statistical features of the spectrum (harmonics, center of gravity, moments). Other features include the most important Mel-frequency cepstral coefficients, and the Chroma Vector.

To calculate the features that enter the classifier, the audio is split into short fixed-size frames of 50 milliseconds. At each frame we calculate the above features. We reference these as short term features. We then calculate the mean and standard deviation of these features over a larger frame (1 second long), to obtain the mid-term features. The final step is to perform a long term average of the mid-term features, and normalize, to obtain a single vector that describes the full audio signal. This is the feature vector that will be used by the classifier.

In practice we make use of the pyAudioAnalysis [7] library to extract the short and mid term features from raw audio.

### 2.1.2 Classifier

For classifying the audio features we used a Support Vector Machine model. The goal of SVMs is to find the hyperplane that best separates the data points of different classes (i.e the distance between the closest point of each class to the hyperplane is maximized). The algorithm takes advantage of the fact that the data could be linearly separable in higher dimensions, while using the kernel trick to simplify computations. The advantages of SVMs include the ability to work with many dimensions on small datasets, while requiring little tuning. For our task, sample efficiency was a critical factor since the dataset consisted of only a few hundred samples.

## 2.2 Video

The second base model is a pretrained Convolutional Neural Network used for facial emotion recognition.

### 2.2.1 Model description

Convolutional neural networks (CNNs) are wildly used for various image recognition tasks with great success. The basic components of a CNN are:

- Convolutional layers: At this layer, a number of filters with learned parameters is applied to the NxN image.

- Pooling layers: Used to reduce the frame size by averaging (or taking the max value) of neighbour elements.

A sequence of convolutional and pooling layers is used, and the last frame is flattened and used as input for a standard MLP (multilayer perceptron).

The model that we used follows the VGG-19 [8] architecture of 19 layers: 16 convolution layers, 3 Fully connected layer, 5 MaxPool layers and 1 SoftMax layer. The input is a 48 by 48 image.

The model was pretrained on the FER-2013 dataset. The training set consists of 28,709 samples of 48x48 pixel grayscale facial images. The images are labeled based on the facial expressions, into one of seven categories: angry, disgust, fear, happy, sad, surprise, neutral.

### 2.2.2 Predictions

In order to make a prediction from raw video, we perform the following:

- The video is processed frame by frame

- The frame is scaled to lower resolution to 48 by 48 resolution

- We apply OpenCV's cascade classifier [9] to identify the regions of the image, that contain faces

- The image is cropped to the region of interest

- The image is then forward-passed to the VGG model to get the prediction.

- Finally we calculate the average of all frame predictions and normalize the vector.

## 2.3 Ensemble

Ensemble models combine many different weak classifiers to get the final prediction. In our case we used the stacking approach, where we combine the two classifiers using a random forest meta-classifier. The inputs to this model are the output predictions of the two base models. The model is trained on the rest of the training set.

## 2.4 Dataset

For training and testing our model we used the RAVDESS [10] dataset. The dataset contains 24 actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. The model was trained on

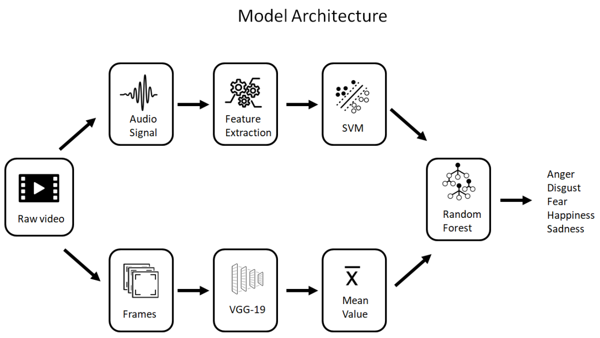

Figure 1: Visualization of the individual components of the combined model.

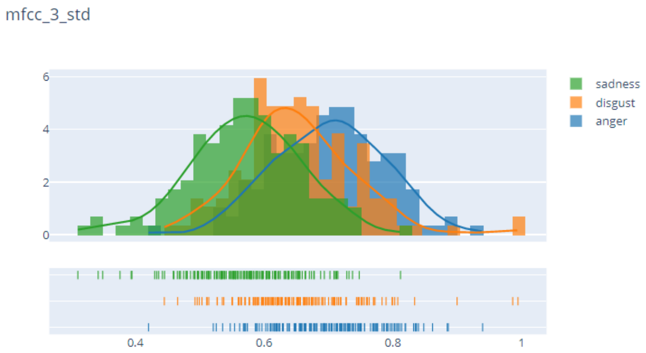

Figure 2: The standard deviation of the third mfcc coefficient.

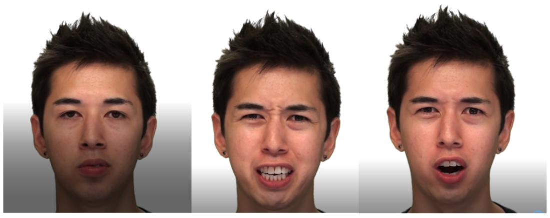

Figure 3: Different frames of the same clip, labeled 'neutral', 'sadness' and 'surprise'. The actual clip represents anger.

speech samples which include calm, happy, sad, angry, fearful, surprise, and disgust expressions.

Samples from 8 actors were used to train the audio model (64 samples per class), with a 0.9 split train/test ratio. Another set of 12 actors' samples were used to train the meta-classifier. During this phase, the base models are only used to extract the features for the meta-classifier, and therefore are not modified. Finally the last 4 actors' samples (32 per class) were used as the test set. In this way we eliminate correlations between the train and test set, since the latter consists of unseen actors.

## 3 Experimental Results

The audio model demonstrated a 37.4 f1 score / 36.8 accuracy on the test set. The confusion matrix is shown in table 1. The classes that demonstrate the best performance are 'anger' and 'disgust' with score 0.44 and 0.42, while the worst performance happens with class 'happiness' and 'sadness', with a score of 0.31. Examining the histograms of each individual feature, grouped by the respective label, it is shown that neither one demonstrates any strong discriminating ability. For the classes 'angry', 'disgust', 'sadness' we visualize the standard deviation of the third mfcc coefficient, in figure 2.

The CNN model demonstrated a poor performance of 23.3 f1 score / 23.7 accuracy. Examining the confusion matrix, we can observe that the model is heavily biased towards the neutral class. In order to explain the outcome we need to revisit the prediction process. Each video frame is forward-fed into the model and the final prediction is calculated as the average of all frames. The issue with this approach is that at the start and end of the video, most frames are labeled as neutral. Only an handful of frames actually capture the true expression of the actor. Therefore the final vector is biased towards neutral. Furthermore, an actor may display several expressions that are not clear and may resemble other emotions. An example can be found in figure 2 where the label of the video is 'anger'.

The combined model showed increased performance over the two separate base models, with an f1 score of 41.0 and 44.3 accuracy. Observing the confusion matrices of the combined and base models, the former seems to perform better on anger, happiness and sadness classes, but lot worse in the fear and disgust.

## Conclusions

In this assignment we proposed a combined architecture of audio and visual models to predict emotion from video. We showed that the introduction of a meta-classifier on top of the base models, demonstrated an increase in per-

| - | anger | fear | disgust | happiness | sadness |
|---|---|---|---|---|---|
| anger | 23 | 0 | 2 | 0 | 7 |
| fear | 8 | 4 | 15 | 0 | 5 |
| disgust | 4 | 11 | 2 | 0 | 15 |
| happiness | 2 | 2 | 4 | 13 | 11 |
| sadness | 1 | 2 | 0 | 0 | 29 |

Table 1: Confusion matrix for the combined model, on the test set

| - | anger | fear | disgust | happiness | sadness |
|---|---|---|---|---|---|
| anger | 11 | 6 | 3 | 9 | 3 |
| fear | 3 | 13 | 5 | 7 | 4 |
| disgust | 0 | 8 | 13 | 3 | 8 |
| happiness | 3 | 4 | 2 | 11 | 12 |
| sadness | 1 | 6 | 6 | 8 | 11 |

Table 2: Confusion matrix for the audio model, on the test set

formance over each individual model. Further modifications can be made to improve the proposed architecture, such as:

- Using an LSTM on features extracted from the CNN, to perform classification over the sequence of frames, in order to improve the video model performance.
- Swapping the SVM for a Wav2Vec model
- Introduce a third modality, that is the transcript of the video, and use SOTA NLP models to analyze sentiment.
- Introduce a bigger dataset with greater diversity

## References

[1] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *2005 International Conference on Machine Learning and Cybernetics*, volume 8, pages 4898–4901 Vol. 8, 2005.

[2] Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017. Advances in Cognitive Engineering Using Neural Networks.

[3] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019.

[4] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *CoRR*, abs/2104.03502, 2021.

[5] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Bjorn W. Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, dec 2017.

[6] Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M. Montero, and Fernando Fernández-Martínez. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22), 2021.

[7] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[9] Vandna Singh, Vinod Shokeen, and Bhupendra Singh. Face detection by haar cascade classifier with simple and complex backgrounds images using opencv implementation. 2013.

[10] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of

Emotional Speech and Song (RAVDESS), April 2018. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak.