# Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

| | |
|---|---|
| Submission author: | Sai Anuroop Kesanapalli |
| Assignment title: | Mid-term report |
| Submission title: | CSCI_535_Midterm_Report.pdf |
| File name: | CSCI_535_Midterm_Report.pdf |
| File size: | 2.2M |
| Page count: | 8 |
| Word count: | 2,848 |
| Character count: | 14,765 |
| Submission date: | 08-Mar-2024 09:41PM (UTC-0800) |
| Submission ID: | 2315856615 |

## A multimodal architecture with shared encoder that uses spectrograms for audio

**Sai Anuroop Kesanapalli, Riya Ranjan, Aashi Goyal, Wilson Tan**
[kesanapa, riyaranj, aashiarv, wtan1167] @usc.edu
University of Southern California

### 1 Problem definition

Multimodal learning aims to create models that process and relate information from multiple modalities. Human communication is multimodal by nature which limits the performance of unimodal models. A shared encoder architecture may be capable of fusing multimodal information while providing better synergy between modalities compared to architectures that use separate encoders.
Multimodal fusion is critical for developing artificial agents that can jointly understand the verbal, non-verbal and contextual cues present in human communication. By aligning multimodal features better, the proposed architectures would be able to implicitly capture these cues that are subtly manifested across modalities in human communication. Furthermore, a shared encoder architecture could lead to improved performance on identifying basic emotions, while allowing the model to identify more complex emotions in social communication such as jealousy or empathy.

### 2 Literature Review

(1) provide architectures that have one encoder tailored per modality. These are specific to voice assistants on smart-watches that utilize accelerometer readings and audio cues. We wish to use a common encoder rather than independent ones. (6) leverage the benefits of complementary information provided by different types of labels and develop three ranking models based on SVM, DNN, and GBDT. This direction is orthogonal to our approach, yet an interesting one to consider since their task is emotion recognition as well. (7) propose one sensor fusion model that is designed for Radar and Lidar data, both of which are visual in nature. Moreover they employ a student-teacher framework. Despite the differences, our work draws inspiration from their sensor fusion pipeline, albeit customized for audio-visual data in our case. (10) propose a method where normalization parameters are exchanged between modes for implicit feature alignment. However they too employ one encoder per modality. Previous works have also leveraged attention mechanisms for fusion. (4) presents a simple modality-agnostic model by using self and cross attention on images and text to learn a common embedding space. Using transformer architectures which utilizes attention mechanisms may also be beneficial for our audio-visual task.

### 3 Data Description

We utilize the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) (2) for our work, offering a rich multimodal experience, integrating audio and video for enhanced emotion analysis. Evaluated by over $2,400$ individuals, CREMA-D includes $7,442$ video clips with performances by $91$ actors, providing a diverse exploration of emotional expression. Within the dataset, each actor presents $12$ sentences, expressing $6$ emotions at different intensity levels. Each video clip is brief, lasting less than $5$ seconds. Importantly, the dataset includes the number of ratings for each emotion, offering valuable insights into the perceived emotional content of the performances.

### 4 Method

We work on a novel audio-visual learning paradigm where audio data is represented as spectrograms, in order for the embeddings to be used with an encoder that is shared between audio and video data. The architecture is visualized in Fig. 1.
Our proposed work is divided into three phases as described below:

- **Video-pipeline**: This is a standard video inference pipeline that shares the same architecture as that of the audio-pipeline except for the spectrogram generation phase, as described next.