# 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.

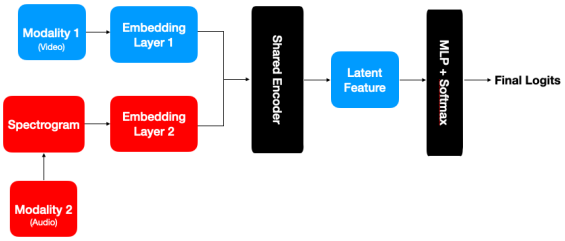


Figure 1: Architecture

- **Audio-pipeline**: In this pipeline audio data first gets converted to spectrograms, which are then passed through an embedding layer to generate embeddings, that get fed to a shared audio-video encoder, and the subsequent latent features are then passed through a fully-connected layer coupled with softmax to generate logits.
- **Bimodal-pipeline**: This is a merger of the aforementioned two pipelines where the audio and video embeddings get fused before being passed on to the shared encoder, and can be categorized under early fusion.

By adopting this unimodal to bimodal development approach, we are tackling the problem in increasing order of complexity.

## 5 Results so far

### 5.1 Experiments

We first implemented unimodal audio and visual pipelines and tested them with three 2D CNNs, namely ResNet18 (5), GoogLeNet (9), and VGG16 (8). For each of these networks, we first established baselines by using their pre-trained versions on the pipelines. Next, we fine-tuned these networks using Mel spectrograms generated from audio files of CREMA-D videos, for the audio pipeline; and using faces cropped from a middle frame out of CREMA-D videos for the video pipeline. Results of fine-tuning were gauged in terms of train / test loss and accuracy. Moreover, we then exchanged the networks with the pipelines, i.e., audio pipeline with video fine-tuned network and vice-versa, for each of the three networks, and checked the performance of both the pipelines.

We also implemented unimodal and multimodal pipelines using 3D CNNs. For data processing, CREMA-D video frames were extracted at ~30 frames per second, converted to grayscale, and resized. The corresponding spectrograms for every video were converted to grayscale and resized, but also vertically divided into evenly sized chunks. They were divided into the same number of frames as their corresponding video such that each chunk of spectrogram would temporally align with each video frame. So far, two models have been tested. I3D (3) followed by average pooling across the logits' spatial and temporal dimensions, and a simple 3D CNN which consisted of 5 convolutional 3D layers followed by 1 final linear classification layer. I3D has only been tested on video. Inputs for I3D were resized to 244x244 (WxH). Simple 3D CNN has been tested on audio, video, and multimodal. For multimodality, the model was trained and tested on both video frames and divided spectrograms. Inputs for simple 3D CNN were resized to 200x100.

For each of the experiments, we used a subset of CREMA-D dataset containing 273 samples from 3 different classes, happy, sad, and angry (91 samples per class). For the multimodal 3D CNN, there was a total of 546 samples because it received both audio and visual data. The 2D experiments used a 7:3 train-test split whereas the 3D experiments used a 9:1 train-test split.

### 5.2 Description of results

Table 1 describes the set of experiments (pipelines) along with their tables containing their results.

### 5.3 Analysis

From Tables 2 and 3, for ResNet18, it is evident that the baselines are dismal, which can be attributed to lack of transfer learning to our specific case of spectrograms and video frames for emotion recognition. Moreover, these tables also suggest that fine-tuning greatly improves the performance of the network for both audio and visual pipelines. However, Table 4 suggests that the networks trained on individual modes do not generalize well to the other mode. Table 5 suggests that the network trained on concatenated spectrograms and faces, which constitutes our multimodal pipeline, performs as good as the individual pipelines.

GoogLeNet, similar to ResNet18, demonstrates substantial performance improvement in emotion recognition on spectrograms and video frames after fine-tuning (Table 2 and 3). However, like the findings in Table 4, it struggles to generalize this knowledge to the other modality, performing poorly when tested on unseen data (e.g., video-trained

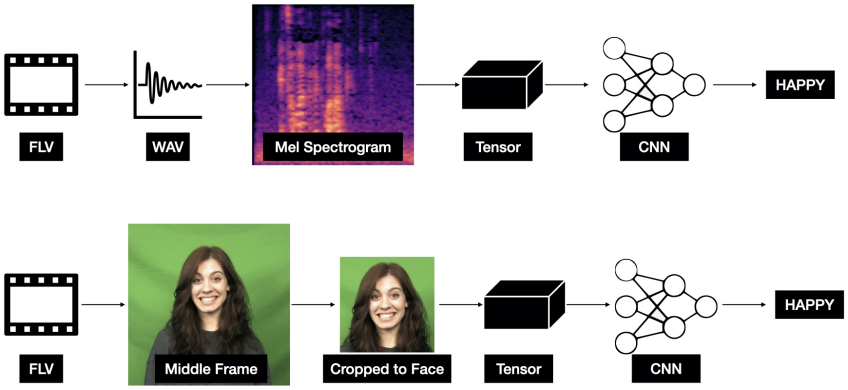


Figure 2: Unimodal audio and video pipelines with 2D CNN

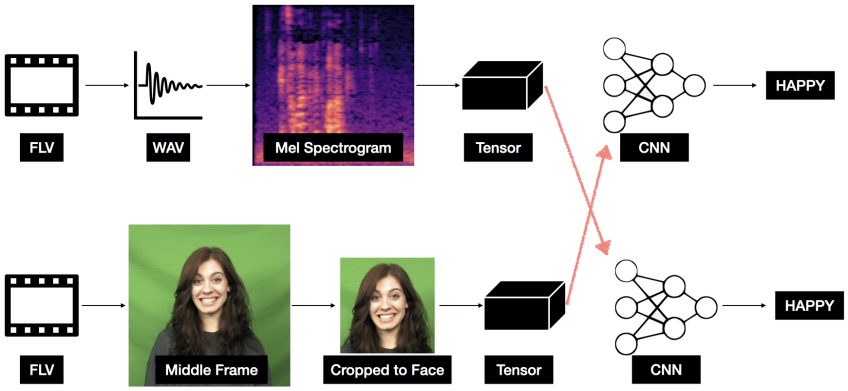


Figure 3: Unimodal audio and video pipelines with 2D CNN crossed

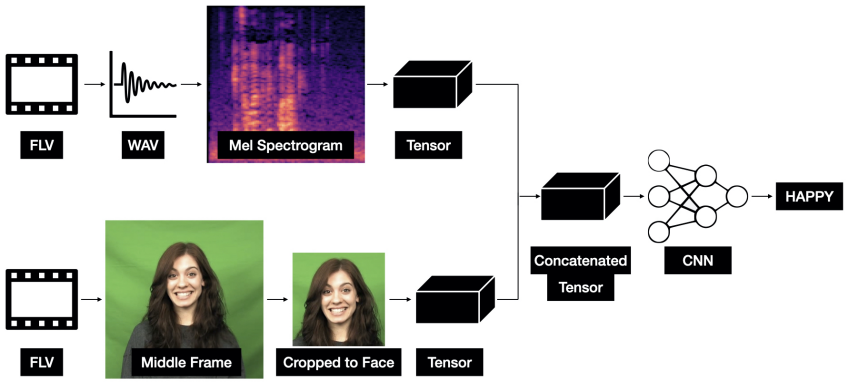


Figure 4: Multimodal audio and video pipelines with 2D CNN

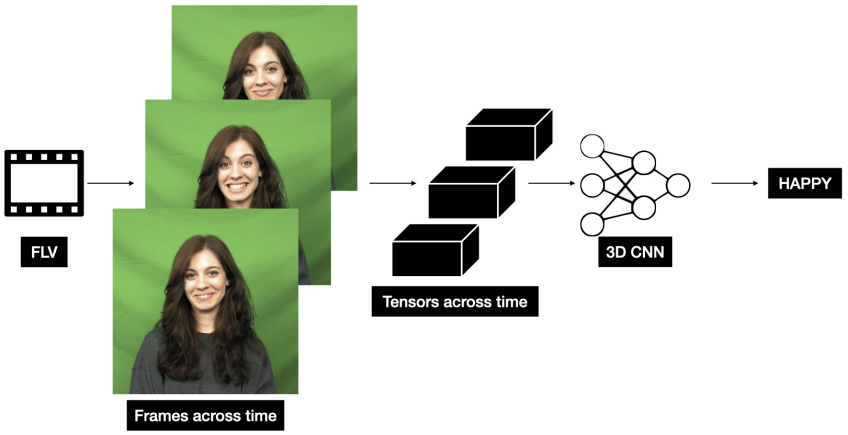


Figure 5: Unimodal video pipeline with 3D CNN

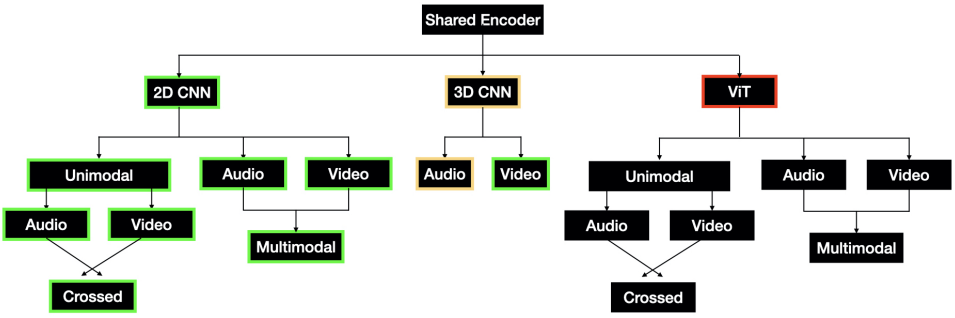


Figure 6: Project Status

model tested on spectrograms). This limitation in generalizability between audio and visual data is evident throughout the provided information. Interestingly, as shown in Table 5, combining both modalities achieves competitive results in GoogLeNet, suggesting potential for exploring multimodal approaches for improved emotion recognition.

VGG16 gave similar trends in emotion recognition tasks using spectrograms and video frames, as the above two CNN models. It exhibited poor baseline performances. Fine-tuning significantly enhanced its effectiveness across both modalities as seen in Tables 2 and 3. Moreover, the model struggled to perform well on unseen data, i.e., cross performance is dismal (Table 4), similar to the other two CNN models. A multimodal pipeline combining spectrograms and faces achieved comparable results to individual modalities, as seen in Table 5, suggesting that integrating audio and visual data through a multimodal approach can effectively improve emotion recognition performance.
We note with caution here that the preliminary results obtained are not final in the sense that we have not yet optimized the training for best parameters of batch size, learning rate, optimizer, and number of epochs. These preliminary results show that the models are ready to be trained on the full CREMA-D dataset, but should not be interpreted for anything else. Based on the small subset of CREMA-D used for the initial experiments and the 7:3 / 9:1 train-test split ratio used in training the 2D / 3D CNNs, the current results are not representative of their true performance because of two reasons. Firstly, there is not enough data in the subset to effectively train and test the models which resulted in high variability in accuracy scores for both models. Secondly, the subset only contains 3 out of 6 classes in the full dataset, meaning that the models are unvalidated on half of the classes.

## 6 Plans to finish the project

Figure 6 outlines the scope of our project, and the status of various parts of it. We have successfully implemented the unimodal and multimodal audio and visual pipelines with 2D CNN, and 3D CNN. We propose to work on the target items listed below in the next phase of our project. Kindly note that some of these targets may be ambitious given the time frame.

- Unimodal and multimodal pipeline with 3D CNN - I3D requires more testing on unimodal audio and multimodal data. Also, an MLP can be used as a final fully connected layer for the I3D model instead of average pooling across temporal and spatial dimensions to get classifications. We will also try fusion techniques for multimodality. Moving the 3D CNN work to HPC will also be important, especially when we scale to the full CREMA-D dataset. It may also be worthwhile looking at other 3D models or transformers.
- Unimodal and multimodal pipelines with ViT - We plan to replace the 2D CNN in the pipelines that we built with that of a Vision Transformer. We expect to face challenges in re-writing the audio and image pre-processing part of our pipelines to make it suitable for Vision Transformer.
- Improve the existing CNN-based pipelines for better accuracy by experimenting with different values for the hyper-parameters.
- Scaling to full CREMA-D dataset - We propose to extend our work to full CREMA-D dataset that contains 6 emotion classes and 7,442 emotion clips. This will require us to make slight modifications to the preprocessing and training code to include the new classes.

| Experiment | Table |
|---|---|
| Unimodal audio pipeline with 2D CNN | 2 |
| Unimodal video pipeline with 2D CNN | 3 |
| Unimodal audio and video pipelines with 2D CNN crossed | 4 |
| Multimodal audio and video pipelines with 2D CNN | 5 |
| Unimodal video pipeline with 3D CNN | 6 |
| Unimodal audio pipeline with 3D CNN | 7 |
| Multimodal audio and video pipeline with 3D CNN | 8 |
| Training setup for unimodal and multimodal pipelines with 2D CNN | 9 |
| Training setup for unimodal and multimodal pipelines with 3D CNN | 10 |

Table 1: Description of experiments and tables containing their results

## 7 Codebase

Code-base hosted on GitHub (private) repository - https://github.com/ksanu1998/multimodal_course_project. Our experiments are available as `.ipynb` notebooks and `.py` scripts accompanied with README files and can be reproduced. Please contact any of the team members for access and information. I3D PyTorch implementation was taken from https://github.com/piergiaj/pytorch-i3d/tree/master.

## 8 Contributions

- Anuroop
  1. Unimodal audio and video pipelines with 2D CNN - ResNet18
  2. Multimodal pipeline with 2D CNN - ResNet18
  3. Midterm presentation deck and report
- Riya
  1. Unimodal audio and video pipelines with 2D CNN - GoogLeNet
  2. Multimodal pipeline with 2D CNN - GoogLeNet
  3. Fine tune GoogLeNet model
- Aashi
  1. Unimodal audio and video pipelines with 2D CNN - VGG16
  2. Mulitomodal pipeline with 2D CNN - VGG16
  3. Trying out different combinations of neural networks and optimizers
- Wilson
  1. Unimodal and multimodal 3D CNN pipelines
  2. 3D data processing (video to frames and dividing spectrograms)
  3. Midterm report

All team members are actively involved in and contributing towards the project. Furthermore, everyone contributed to proof-reading both the presentation deck and the report.

## References


[1] Buddi, S. S., Sarawgi, U. O., Heeramun, T., Sawnhey, K., Yanosik, E., Rathinam, S., and Adya, S. Efficient multimodal neural networks for trigger-less voice assistants, 2023.

[2] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing 5*, 4 (2014), 377–390.

[3] Carreira, J., and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[4] Dodds, E., Culpepper, J., Herdade, S., Zhang, Y., and Boakye, K. Modality-agnostic attention fusion for visual search with text feedback, 2020.

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

[6] Lei, Y., and Cao, H. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing 14*, 4 (2023), 2954–2969.

[7] Li, Y.-J., Park, J., O'Toole, M., and Kitani, K. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 918–927.

[8] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

| Type | 2D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| Baseline | ResNet18 | - | - | 1.0953 | 0.3415 |
| FT | ResNet18 | 0.5624 | 0.9895 | 0.6576 | **0.8902** |
| Baseline | GoogLeNet | - | - | 1.0993 | 0.2805 |
| FT | GoogLeNet | 0.6148 | 0.9319 | 0.8268 | 0.6829 |
| Baseline | VGG16 | - | - | 1.1027 | 0.3537 |
| FT | VGG16 | 0.6657 | 0.8848 | 0.7831 | 0.7683 |

Table 2: Unimodal Audio with 2D CNN

| Type | 2D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| Baseline | ResNet18 | - | - | 1.0967 | 0.3659 |
| FT | ResNet18 | 0.5809 | 0.9686 | 0.7736 | **0.7805** |
| Baseline | GoogLeNet | - | - | 1.0987 | 0.3171 |
| FT | GoogLeNet | 0.6992 | 0.8429 | 0.9076 | 0.6463 |
| Baseline | VGG16 | - | - | 1.0919 | 0.3902 |
| FT | VGG16 | 0.5927 | 0.9579 | 0.8393 | 0.7073 |

Table 3: Unimodal Video with 2D CNN

| Type | 2D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| Audio | ResNet18 (Video) | - | - | 1.2287 | 0.3171 |
| Video | ResNet18 (Audio) | - | - | 1.2611 | 0.2561 |
| Audio | GoogLeNet (Video) | - | - | 1.2038 | 0.3293 |
| Video | GoogLeNet (Audio) | - | - | 1.1606 | 0.3902 |
| Audio | VGG16 (Video) | - | - | 1.1726 | 0.3415 |
| Video | VGG16 (Audio) | - | - | 1.2138 | 0.3049 |

Table 4: Unimodal crossed with 2D CNN

| Type | 2D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| Baseline | ResNet18 | - | - | 1.0839 | 0.3780 |
| FT | ResNet18 | 0.5515 | 1.0000 | 0.7124 | **0.8415** |
| Baseline | GoogLeNet | - | - | 1.0987 | 0.3171 |
| FT | GoogLeNet | 0.6793 | 0.8639 | 0.8967 | 0.6585 |
| Baseline | VGG16 | - | - | 1.1029 | 0.2683 |
| FT | VGG16 | 0.5942 | 0.9579 | 0.9061 | 0.6220 |

Table 5: Multimodal with 2D CNN

| Type | 3D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| FT | Simple3D CNN | 0.3320 | 0.8490 | 0.4763 | **0.8214** |
| FT | I3D | 0.6406 | 0.7061 | 0.8172 | 0.7143 |

Table 6: Unimodal video with 3D CNN

| Type | 3D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| FT | Simple3D CNN | 0.2632 | 0.9102 | 0.3176 | **0.9286** |

Table 7: Unimodal audio with 3D CNN

| Type | 3D CNN | Train Loss | Train Acc. | Test Loss | Test Acc. |
|---|---|---|---|---|---|
| FT | Simple3D CNN | 0.2239 | 0.9124 | 1.0055 | **0.8000** |

Table 8: Multimodal with 3D CNN

| Parameter | ResNet18, GoogLeNet, VGG16 |
|---|---|
| # samples (train + test) | 273 (191 + 82) |
| Batch size | 32 |
| Learning rate | 0.001 (0.0001 for VGG16) |
| Optimizer | Adam |
| Loss | Cross Entropy |
| # train epochs | 50 |
| GPU | T4 (via Colab) |

Table 9: Training setup for unimodal and multimodal pipelines with 2D CNN

| Parameter | Simple3D(Video), Simple3D(Audio), Simple3D(Multi) | I3D(Video) |
|---|---|---|
| # samples (train + test) | 273(245+28), 273(245+28), 546(491+55) | 273(245+28) |
| Batch size | 8 | 2 |
| Learning rate | 0.001 | 0.001 |
| Optimizer | Adam | Adam |
| Loss | Cross Entropy | Cross Entropy |
| # train epochs | 23, 9, 33 | 32 |
| GPU | T4 | T4 |

Table 10: Training setup for unimodal and multimodal pipelines with 3D CNN

[9] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[10] YIN, Y., XU, J., ZU, T., AND SOLEYMANI, M. X-norm: Exchanging normalization parameters for bimodal fusion. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (2022), pp. 605–614.