Multimodal Emotion Recognition

Petros Mitseas

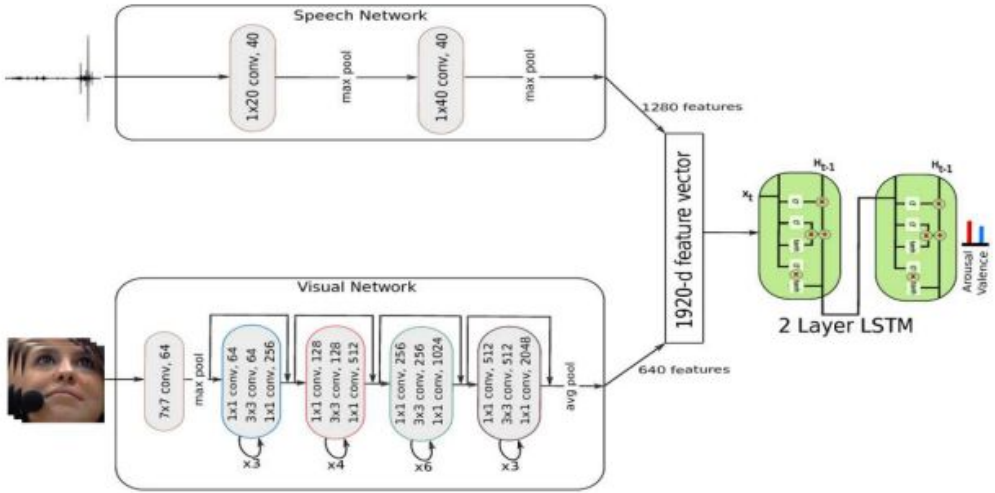
Outline

- Introduction
- Challenges
- Model Description
- Experiments
- Demo
- Conclusion

Introduction

Human emotion recognition: A challenging task

Audio, Visual and Text modalities



Example of a complex multimodal classifier using CNNs and LSTMs

Challenges

- Not many publicly available datasets, that combine all required modalities
  - We used the RAVDESS dataset
- Complex task that requires a large number of diverse data to train deep models, yet gathering these data is a difficult process
- Great diversity among people (gender, language, etc.)
- Handling missing modalities
- Visual models require a clear image of the speaker's face, which is often infeasible in real life
- Since we use 2 + 1 models, training becomes time and data consuming

Model Description

Rules for picking our architecture

1. Choose reasonable models that fit our data (and deadline) - start simple
2. Leverage pretrained models
3. Don't aim for the highest base model performance

Model Description

Audio

- Split audio signal to short windows of 50 msec.
- For each window calculate the short-term features: Zero-crossing rate, energy, spectral features, chroma, mfcc, …
- Average the short-term features over larger 1 sec windows to get mid-term features
- Finally average the mid-term feature to get the representation for the whole signal
- The features are passed to an SVM model

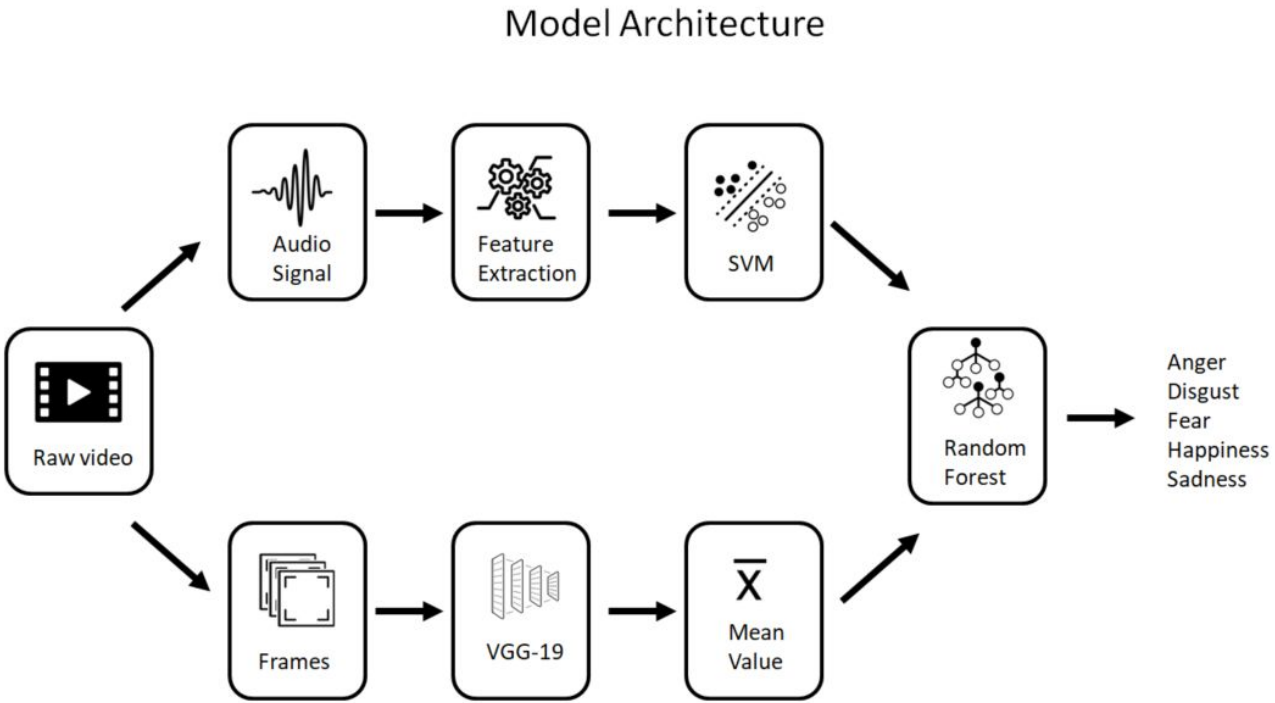
Visual

- Detect faces in frame using a pretrained classifier from OpenCV
- Used a pretrained VGG-19 model on the FER Dataset (48x48 images of seven emotions)
- Forward pass each frame to the model
- Average the output probabilities to get the representation for the whole video

Ensemble

- Combine the output of the two models into a random forest
- Train on different part of the dataset

Model Description



Experiments

- Training and testing on the RAVDESS dataset
- 24 actors: 8 for training the audio model, 12 for training the combined model and 4 for testing

Time for the results:

F1: **37.4** Acc: **36.8**	F1: **23.3** Acc: **23.7**
Audio model	Visual model

Experiments

- Training and testing on the RAVDESS dataset
- 24 actors: 8 for training the audio model, 12 for training the combined model and 4 for testing

Time for the results:

F1: **37.4**
Acc: **36.8**

Audio model

F1: **23.3**
Acc: **23.7**

Visual model

That did not go as planned…

Experiments

What exactly happened?

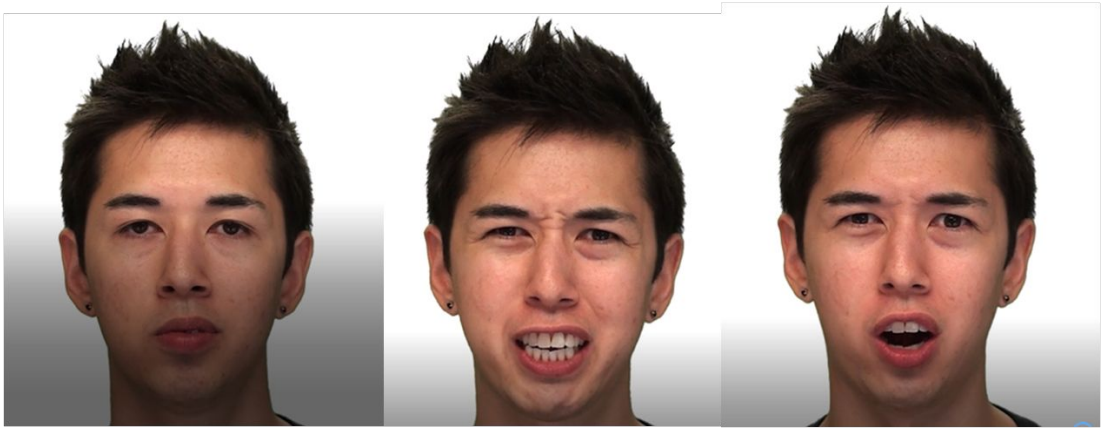
Let's examine the confusion matrix of the **visual** model (anger, disgust, fear, happiness, **sadness**, surprise, **neutral**)

```
0  0  0  0 19  0 13
0  3  2  0 16  0 11
0  0  0  0 20  0 12
0  0  1 15  5  1 10
0  0  0  0 20  0 12
0  0  0  0  0  0  0
0  0  0  0  0  0  0
```

The model is heavily **biased**. And of course it is!

Experiments

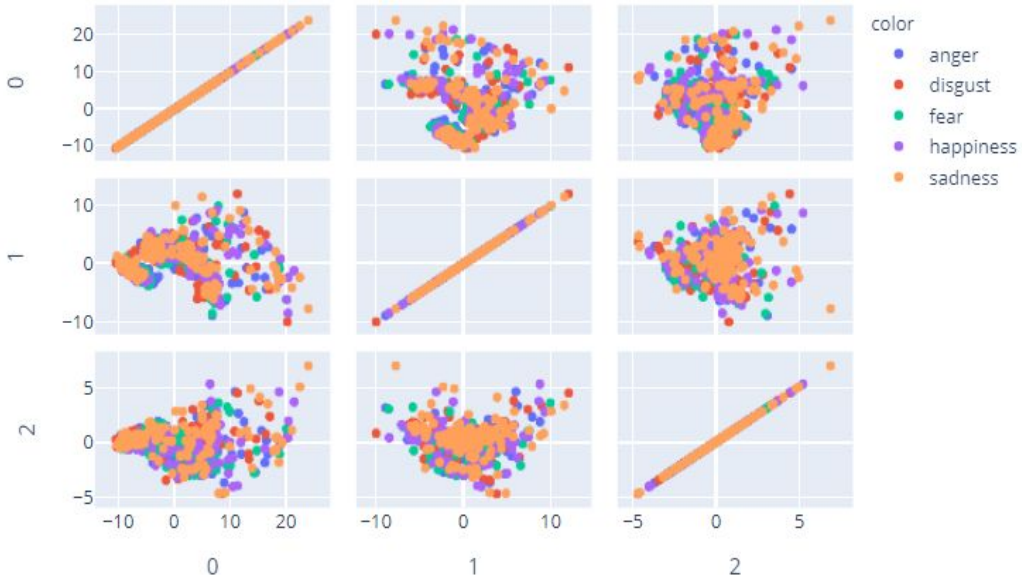
- If we examine each frame of the video individually, most frames appear mostly as **neutral** expressions. Only in a small fraction of the video, do the frames resemble the person's actual emotion.
- Furthermore an emotion may be represented by a variety of expressions, that would individually give incorrect predictions



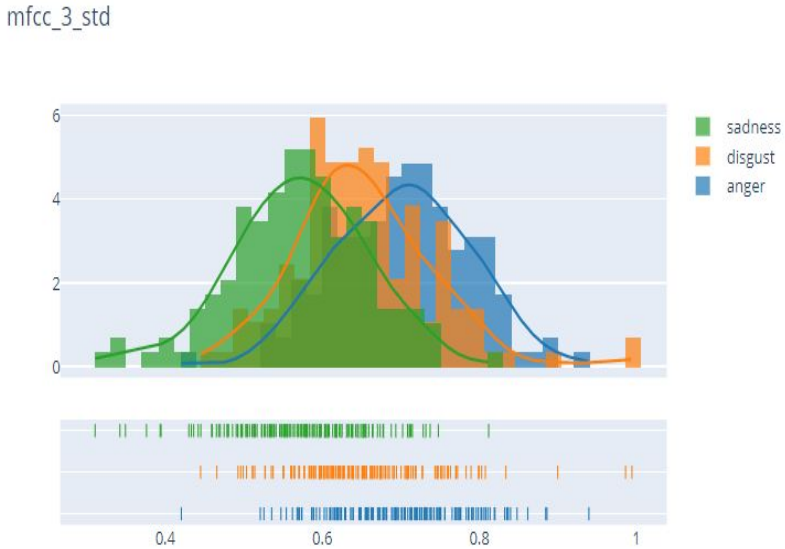
What is this person overall emotion? sadness, surprise?

Experiments

Audio features are not easily separable



PCA visualization of the 3 most important dimensions



The most discriminative feature

Experiments

Performance of the combined model

- F1: **41.0**
- Acc: **44.3**
- Confusion matrix:

23	0	2	0	7	0	0
8	4	15	0	5	0	0
4	11	2	0	15	0	0
2	2	4	13	11	0	0
1	2	0	0	29	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

precision:	0.60	0.21	0.08	1.	0.43
recall:	0.71	0.12	0.06	0.40	0.90
fscore:	**0.65**	0.15	0.07	**0.57**	**0.58**

Conclusion

- 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
- Train an end-to-end deep model to capture cross-modality relations
- Introduce another modality, the transcript of the video
- Bigger dataset + bigger models + bigger hardware

Thank you