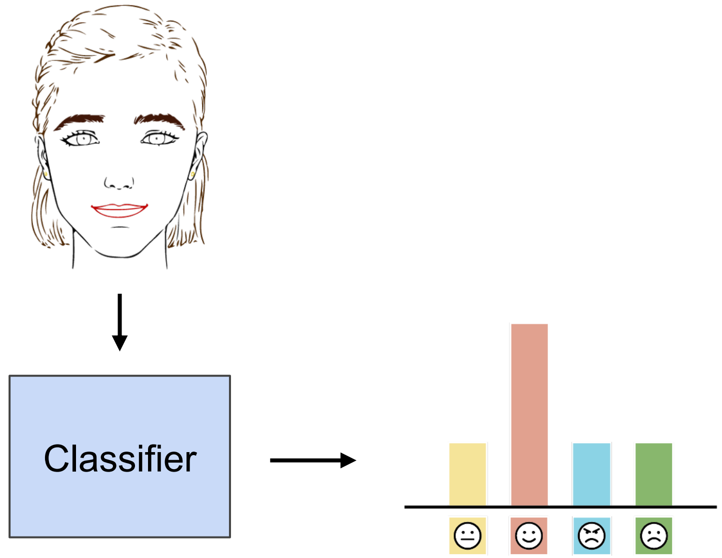
Style Extractor for Facial Expression recognition in the Presence of Speech

Ali N. Salman And Carlos Busso

THE UNIVERSITY OF TEXAS AT DALLAS
UT Dallas MSP Multimodal Signal Processing Laboratory
UTD CRSS

Motivation

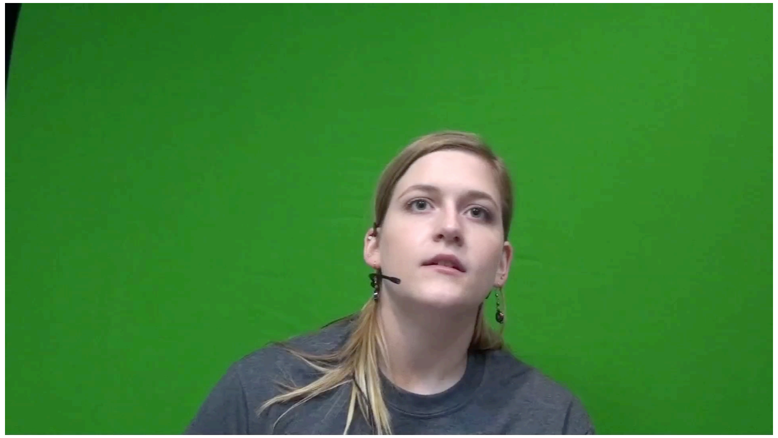
- **Facial Emotion Recognition (FER) is a hard problem**
  - Prediction are not reliable during speech
- **Wide application domains**
  - Human-Computer/Robot Interaction
  - Driver distraction
  - Medical monitoring

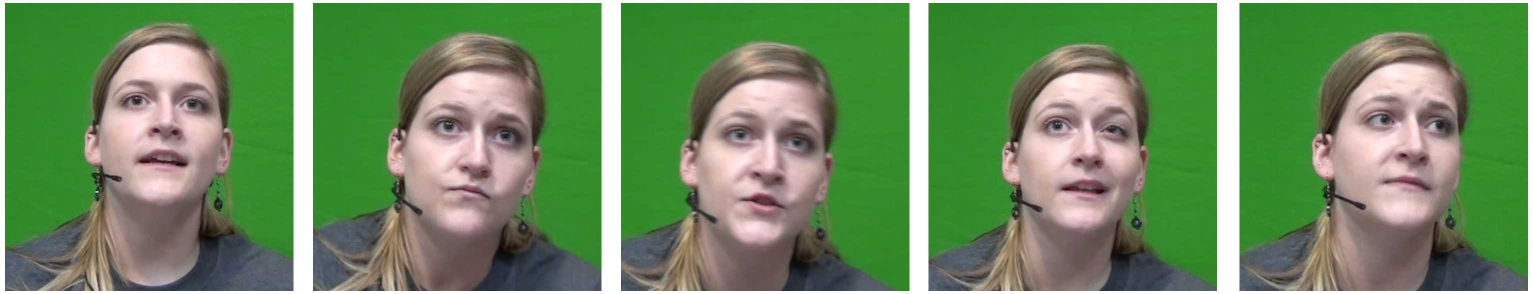


Related Work

- **Dynamic FER**
  - Emotions perceived from isolated frames is different from emotions perceived from watching corresponding video [Salman and Busso 2020]

No Audio



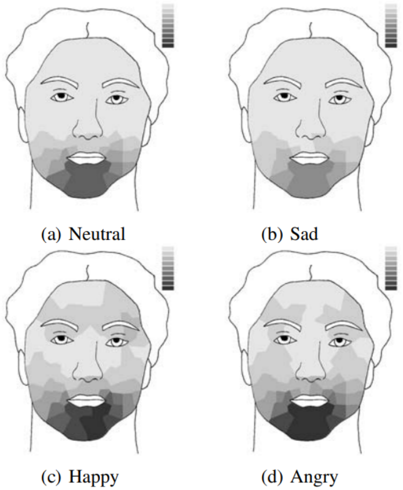


Label	Set	Precision	Recall	F1-Score
Happiness	Video/Video	0.91	0.84	0.87
	Video/Frame	0.67	0.97	0.79
Anger	Video/Video	0.73	0.67	0.70
	Video/Frame	0.55	0.14	0.22
Sadness	Video/Video	0.77	0.79	0.78
	Video/Frame	0.66	0.57	0.61
Neutral	Video/Video	0.72	0.72	0.72
	Video/Frame	0.54	0.77	0.63
Average	Video/Video	0.78	0.76	0.77
	Video/Frame	0.61	0.61	0.56

Table: Compares the perceptual evaluation between videos (different annotators) or videos compared to frames.

Related Work

- **Facial Regions**
  - Lower facial regions are greatly affected by speech articulation
    - Lower regions contain valuable features for emotion classification
    - Some emotions are better perceived in the lower regions [Hoffmann et. al. 2013, Busso and Narayanan 2006]
  - Lower facial regions contain both emotional and lexical information
    - Separating emotion facial features from speech articulations is challenging



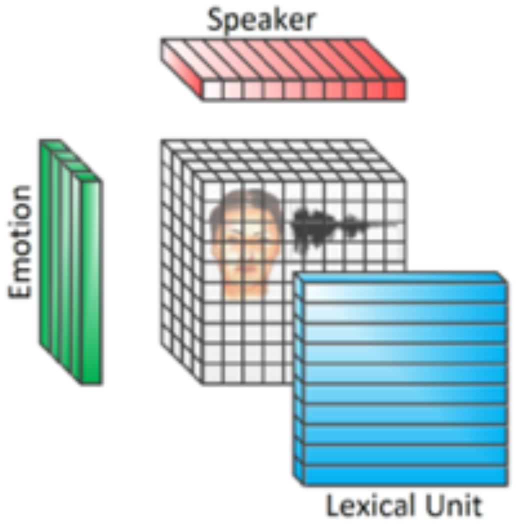
Activeness of different facial regions during speech

Related Work

- **Lexical Dependent FER**
  - A phone or viseme dependent classifier can be used to increase the reliability of emotion recognition [Mariooryad and Busso 2013, Kim and Provost 2015]
  - Another approach is to treat the lower and upper area and use a phoneme dependent classifier on the lower region [Kim and Provost 2019]
  - Transcriptions can be costly (manual), unreliable (ASR), or not feasible to attain (no audio) in real world application
- **Blind-Lexical FER**
  - Separating the facial region into many area can improve accuracy [Kim and Provost 2015]
  - Using an asymmetric bilinear factorization model to extract emotion information without knowing the phonetic labels [Mariooryad and Busso 2015].

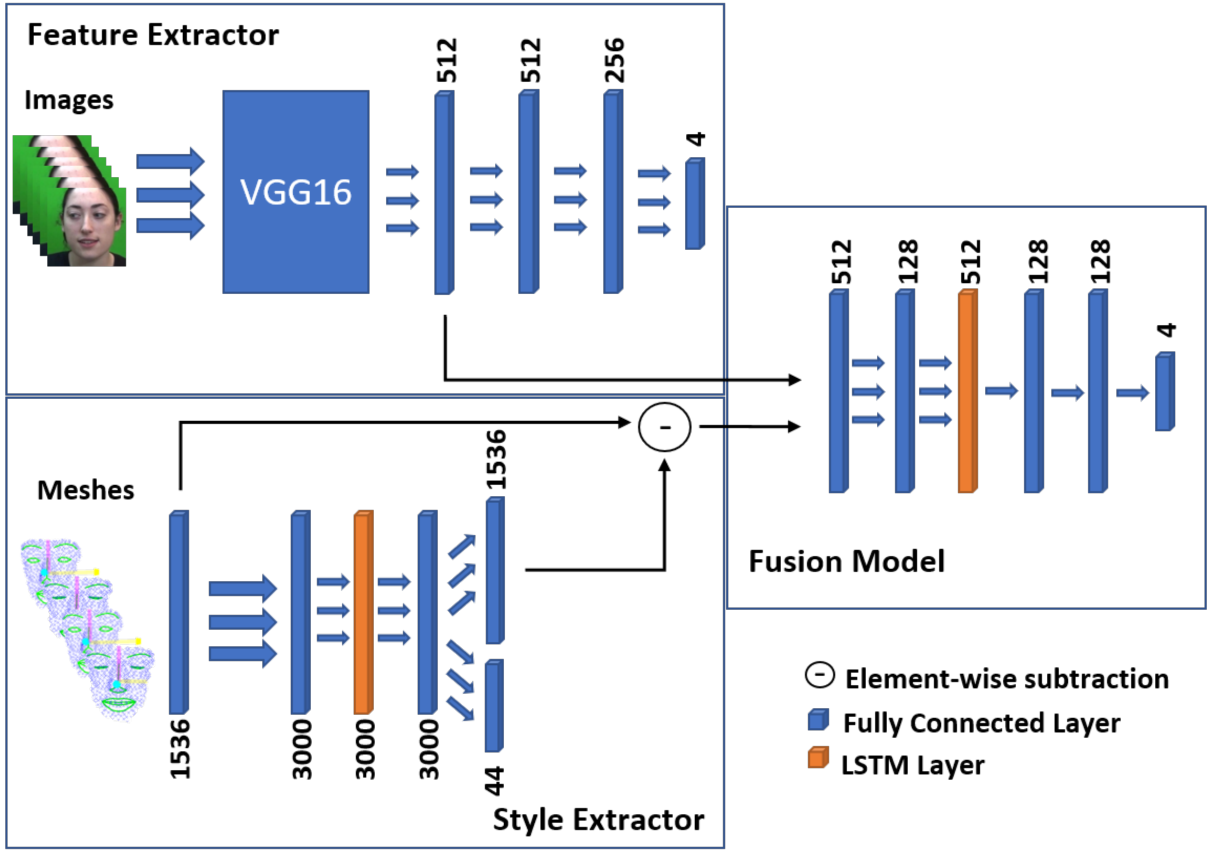
Goal

- **Dynamic FER**
  - Static FER have restriction in real-world application
  - Aggregating features collected from static FER systems is not enough to capture temporal information
- **Blind-Lexical compensation**
  - Transcriptions can be costly and might not be available during real world application
  - The use of transcription during training is valid
  - Separate the emotional and lexical attribute
- **End-to-End Image FER**
  - Using just image sequences as input and not relying on special hardware to capture (i.e., motion capture, depth sensing)



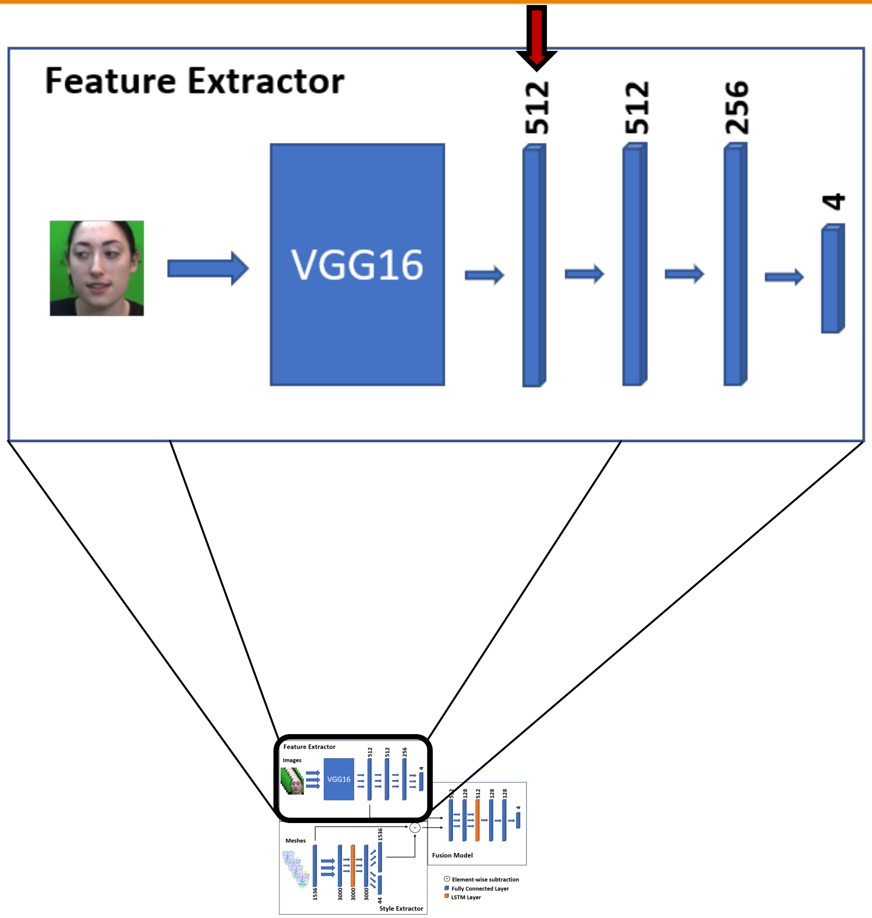
Proposed Model

- **Feature Extraction**
  - Extracts facial features using a CNN model
- **Style Extractor**
  - Separates the emotional facial information (i.e., style) from the lexical facial information (i.e., content)
- **Fusion Model**
  - Combines the features extractor and style extractor features to predict the emotion



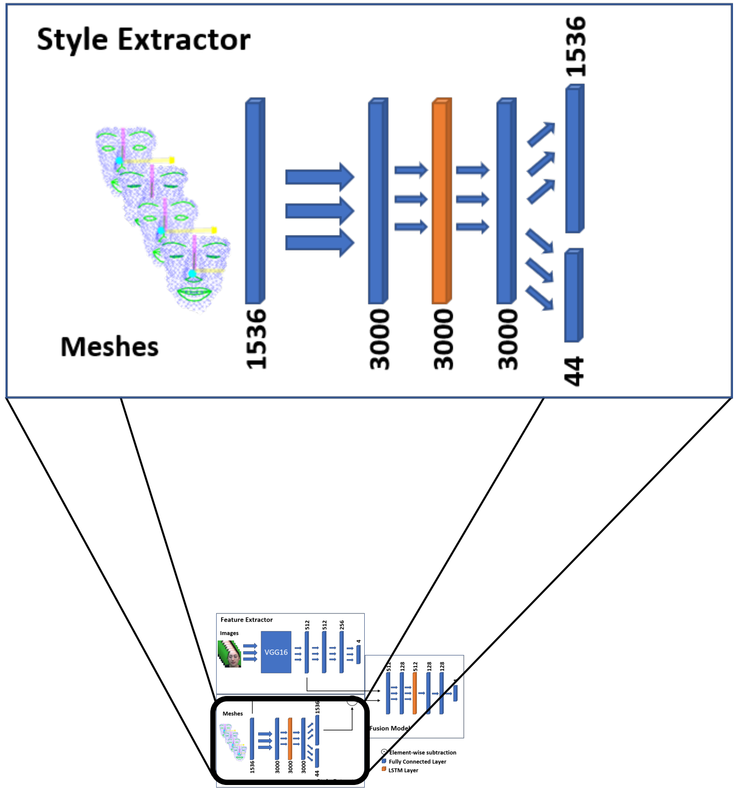
Feature Extraction

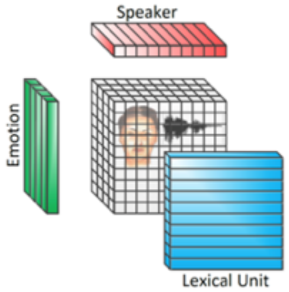
- **VGG16 architecture**
  - Initialize model using VGG-Face weigths [Parkhi et. al. 2015]
  - Train the model for the emotional classes using categorical cross-entropy.
- **Static Features**
  - The first fully connected layer (red arrow) to represent the features



Style Extraction

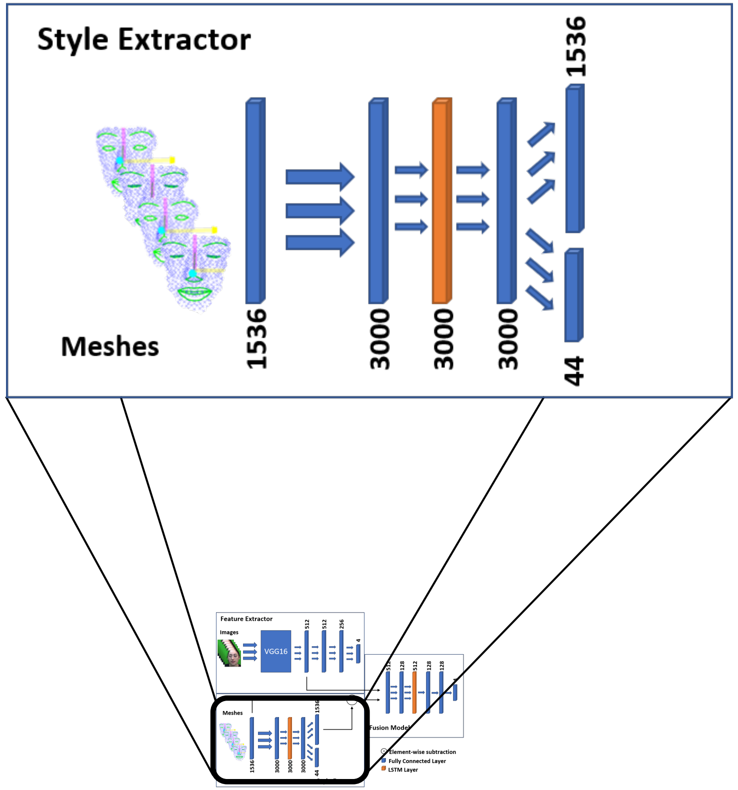
- **Model**
  - FC (blue) and LSTM (orange) model to transform the input sequence from emotional to neutral
  - The model also predicts the phoneme for each
  - The model takes a facial mesh as input and normalize the emotional features
  - We use the difference between the input mesh and output mesh to represent the style
  - Additionally, we predict the phoneme for each mesh to assist in learning phoneme dependent features





Style Extraction

- **Data**
  - Manually align emotional and neutral videos that contain the same lexical contents but different emotions
    - Alignment at the phone level
  - Z-Face [Jeni et. al. 2015] to extract the 3D facial mesh
  - Use the 3D mesh of the aligned pairs (emotional to neutral) to train the model

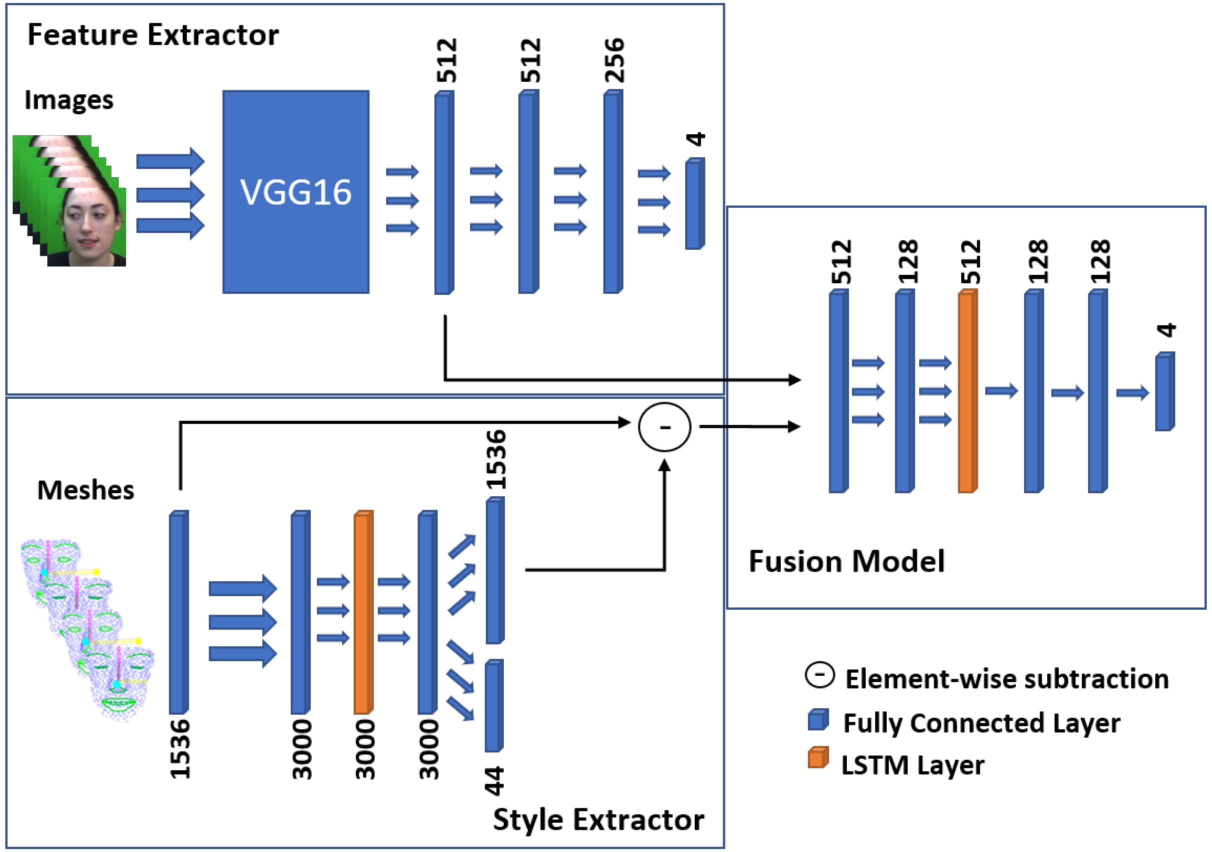


Neutral Emotional

No Audio

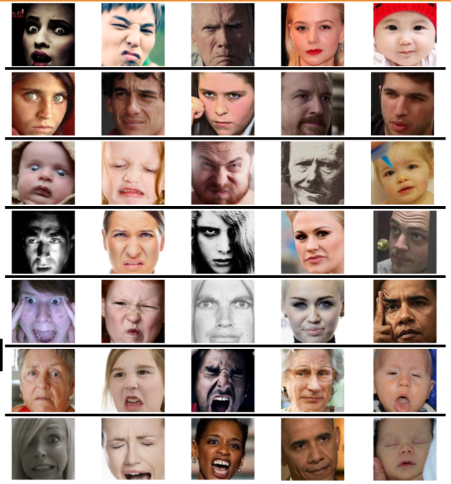
Fusion Model

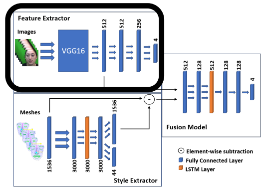
- **Feature Extraction**
  - Extracts facial features using a CNN model
- **Style Extractor**
  - Separates the emotional facial information (i.e., style) from the lexical facial information (i.e., content)
- **Fusion Model**
  - Combines the features extractor and style extractor features to predict the emotion



AffectNet Database

- **AffectNet [Mollahosseini et. al. 2019]**
  - Collected from the internet using major search engines
    - 1250 emotional keywords in 6 different languages
  - Over 1 million images
    - Around 440 thousand are manually annotated with seven discrete emotional
    - Valence and arousal annotation (not used in this study)
    - 425x425 average resolution
  - We consider 4 classes (happiness, anger, sadness, and neutral state)
    - Downsample to 24,882 images per class (training set)
      - Random split 80/20 for training/validation
    - Validation set as testing set
  - This dataset is used to train the feature extractor





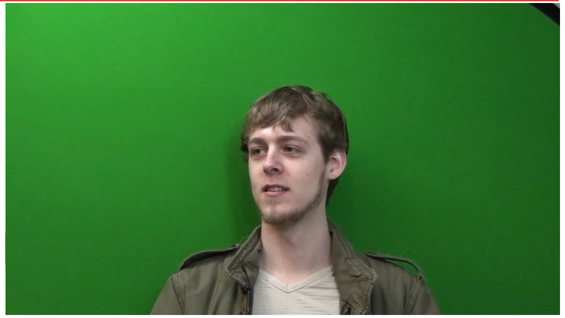
Neutral	75374
Happy	134915
Sad	25959
Surprise	14590
Fear	6878
Disgust	4303
Anger	25382
Contempt	4250
None	33588
Uncertain	12145
Non-Face	82915
Total	420299

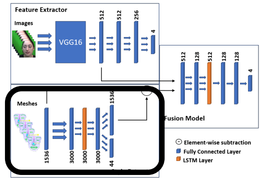
Number of images for each discrete label

MSP-IMPROV Dataset

- **MSP-IMPROV [Busso et. al. 2017]**
  - Multimodal emotional database
    - 12 subjects (six males, six females)
    - 1,440 x 1,080 resolution
    - Same sentences are spoken with different target emotions
      - Improvisations are used before/after to the target sentences to capture naturalistic data
    - Target sentences are manually annotated in different modalities
      - 652 speaking turns
      - We only consider video-only annotations (happiness, anger, sadness, and neutra state)
  - This dataset is used to train the style extractor

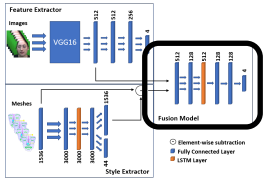
Dyadic Setting   High Quality Recordings





CREMA-D Dataset

- **CREMA-D [Cao et. al. 2014]**
  - Multimodal emotional dataset
    - 91 subjects (six males, six females)
    - 960 x 720 resolution
    - Same sentences are spoken with different target emotions
    - Target sentences are manually annotated in different modalities
      - 7,442 annotated clips
      - We only consider 5,093 video-only labeled clips (happiness, anger, sadness, and neutral state)
      - 81/4/7 actors for train/validate/test
  - After training the style/feature extractor models we use this dataset to train the fusion model



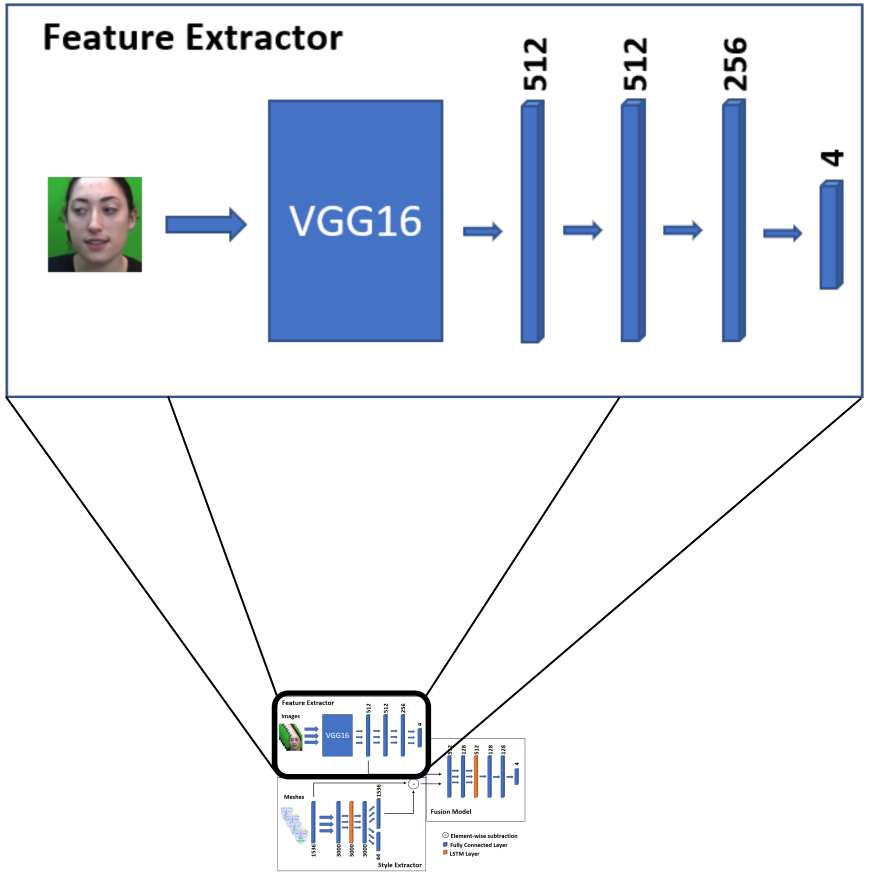
https://github.com/CheyneyComputerScience/CREMA-D

Results - Feature Extractor

- **Feature Extractor**
  - Trained on a subset of the AffectNet database
  - Down sampled to match the minimum number of samples in a class
  - Results are reported on the validation set, which we use as our testing set

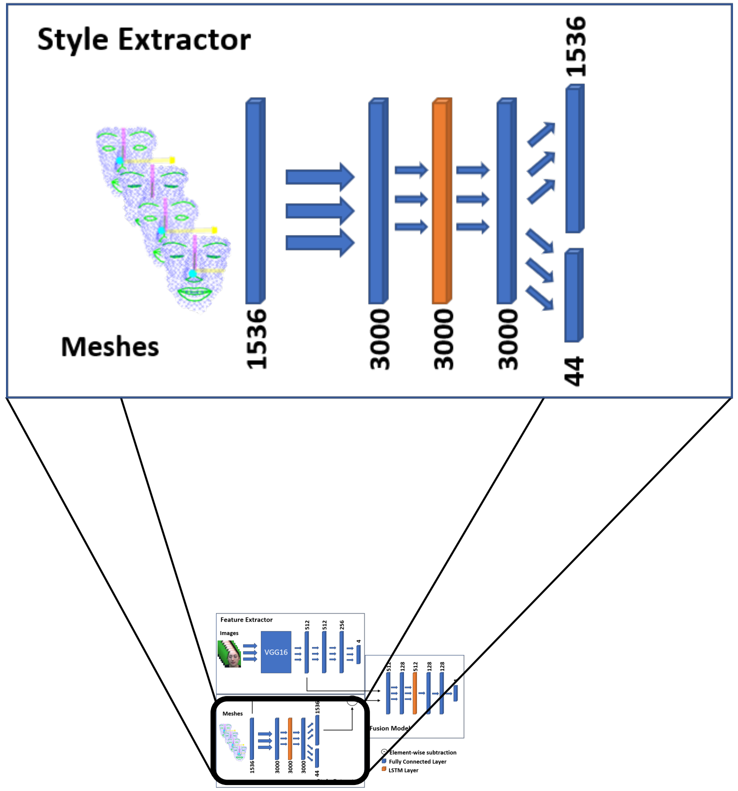
Emotion	Precision [%]	Recall [%]	F1-score [%]
Happiness	89.8	91.0	90.5
Anger	76.7	71.2	73.9
Sadness	75.8	71.6	73.7
Neutral	63.7	70.1	67.0
Average	76.5	76.2	76.3

Performance of the static FER system in the feature extractor model. The reported values are on the AffectNet corpus.



Results - Style Extractor

- **Style Extractor**
  - We expect that the mesh from the style extractor looks more neutral than the original input
  - We trained a vanilla 3D mesh classifier on MSP-IMPROV
    - model achieved 60% F1-score
  - Testing the 3D mesh emotion classifier on CREMA-D
    - Around 4% of the original meshes are classified as neutral
    - 31% of the normalized meshes are classified as neutral



Emotion\Mesh	Original	Normalized
Happiness	15,886	3,332
Anger	191,309	122,075
Sadness	332,825	260,350
Neutral	25,060	179,323

Results - Proposed Model

UT Dallas MSP Multimodal Signal Processing Laboratory

- **Proposed Model**
  - To assess the effectiveness of the proposed approach we train two models
    - With the Style Extractor (Model [A])
    - Without the Style Extractor (Model [B])
  - The model with the Style Extractor performs 7% better (absolute)
  - The Style Extractor helps with generalization
    - Similar performance on train set (A vs B)
    - Smaller gap between train and validation

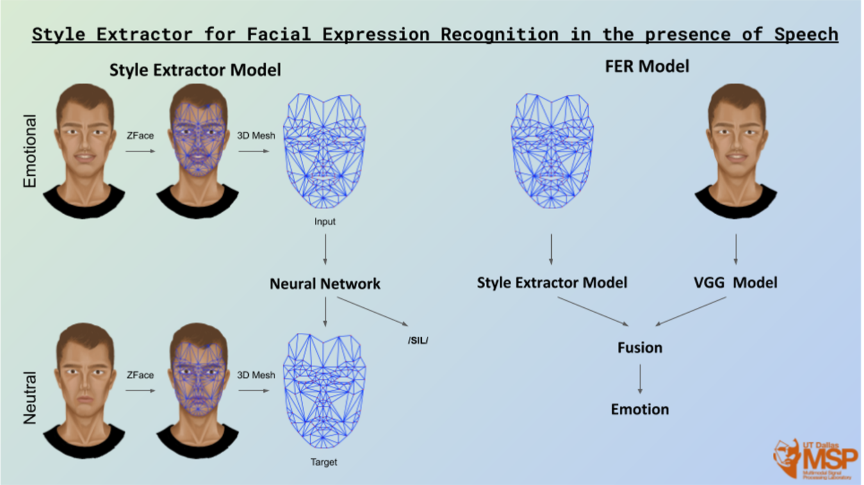
Emotion	Precision		Recall		F1-score	
Model	A [%]	B [%]	A [%]	B [%]	A [%]	B [%]
Happiness	87.8	81.1	83.0	83.5	85.3 ←	82.3
Anger	89.2	51.0	50.9	65.0	64.8 ←	57.1
Sadness	78.6	83.0	60.5	52.3	68.4 ←	64.1
Neutral	68.8	65.0	89.9	65.0	78.0 ←	65.0
Average	81.1	70.0	71.0	66.4	74.1 ←	67.1

Performance of the proposed FER system for videos on the test set of the CREMA-D corpus.

Conclusion

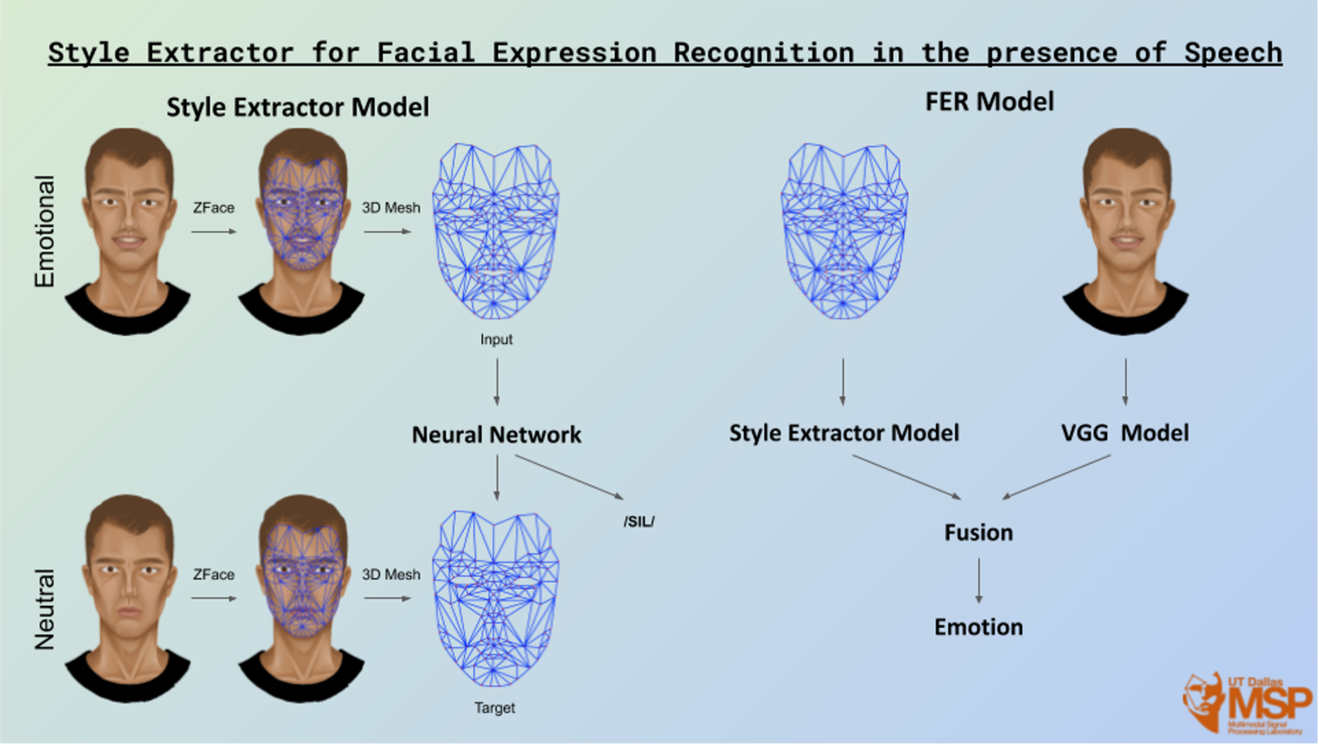
- **Proposed Approach**
  - FER system that does not require transcription during inference
  - Style extractor that extracts the emotional features, not speech articulations

- **Future Research**
  - Find ways to align/pair data for training
  - Improve the feature extractor by extracting spatial-temporal features
  - Improve the style extractor by using images instead of 3D mesh



Thank you

- **This work was funded by NEC Foundation and NSF under Grant IIS-1718944**



NEC \Orchestrating a brighter world

Our Research: msp.utdallas.edu