Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features

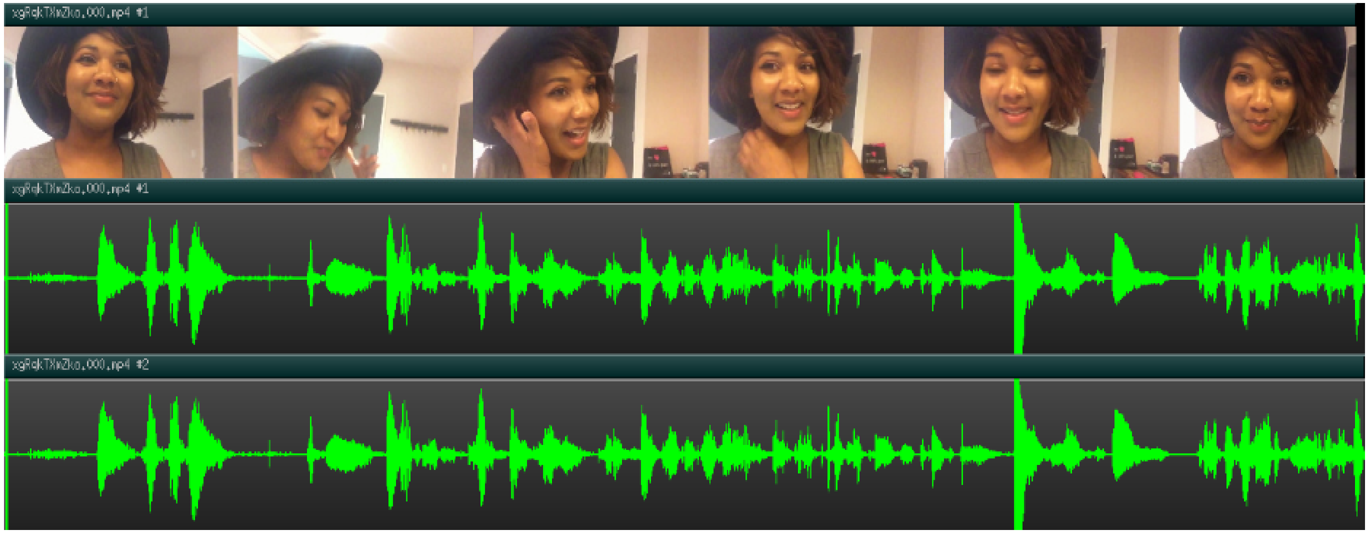
Arulkumar S., Vismay P., Ashish M., Prashanth B., Anurag M.

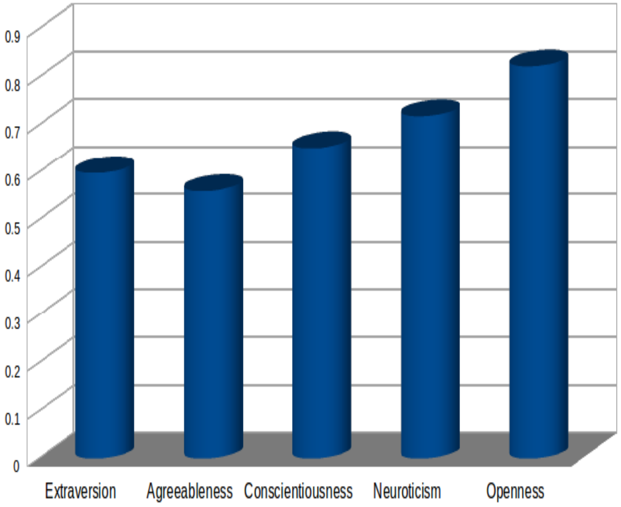
Department of Computer Science and Engineering,
Indian Institute of Technology Madras, India

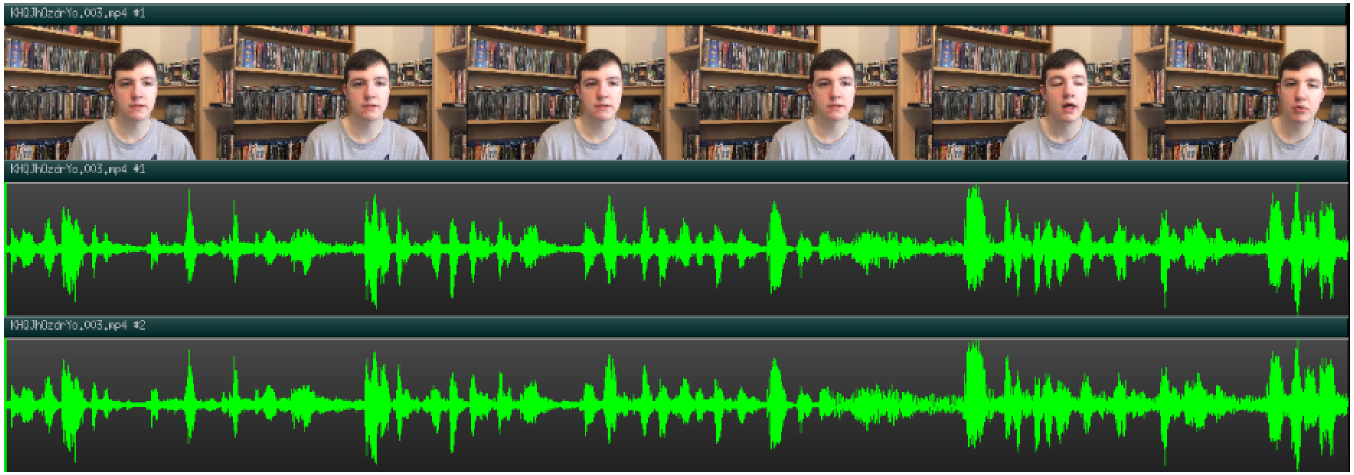
Code: https://github.com/InnovArul/first-impressions

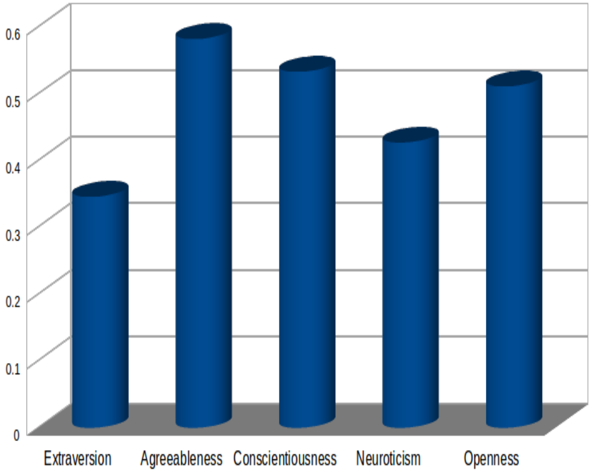
ECCV'16
EUROPEAN CONFERENCE
ON COMPUTER VISION

Problem setup

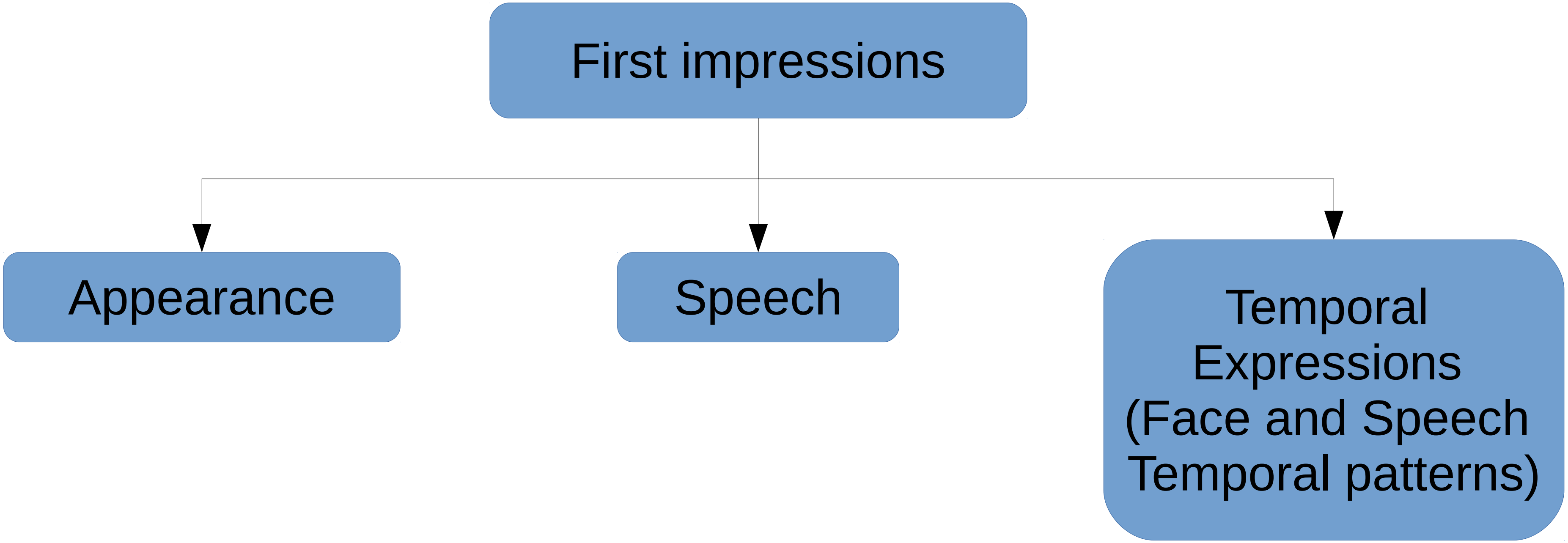








Intuition behind the proposed solution



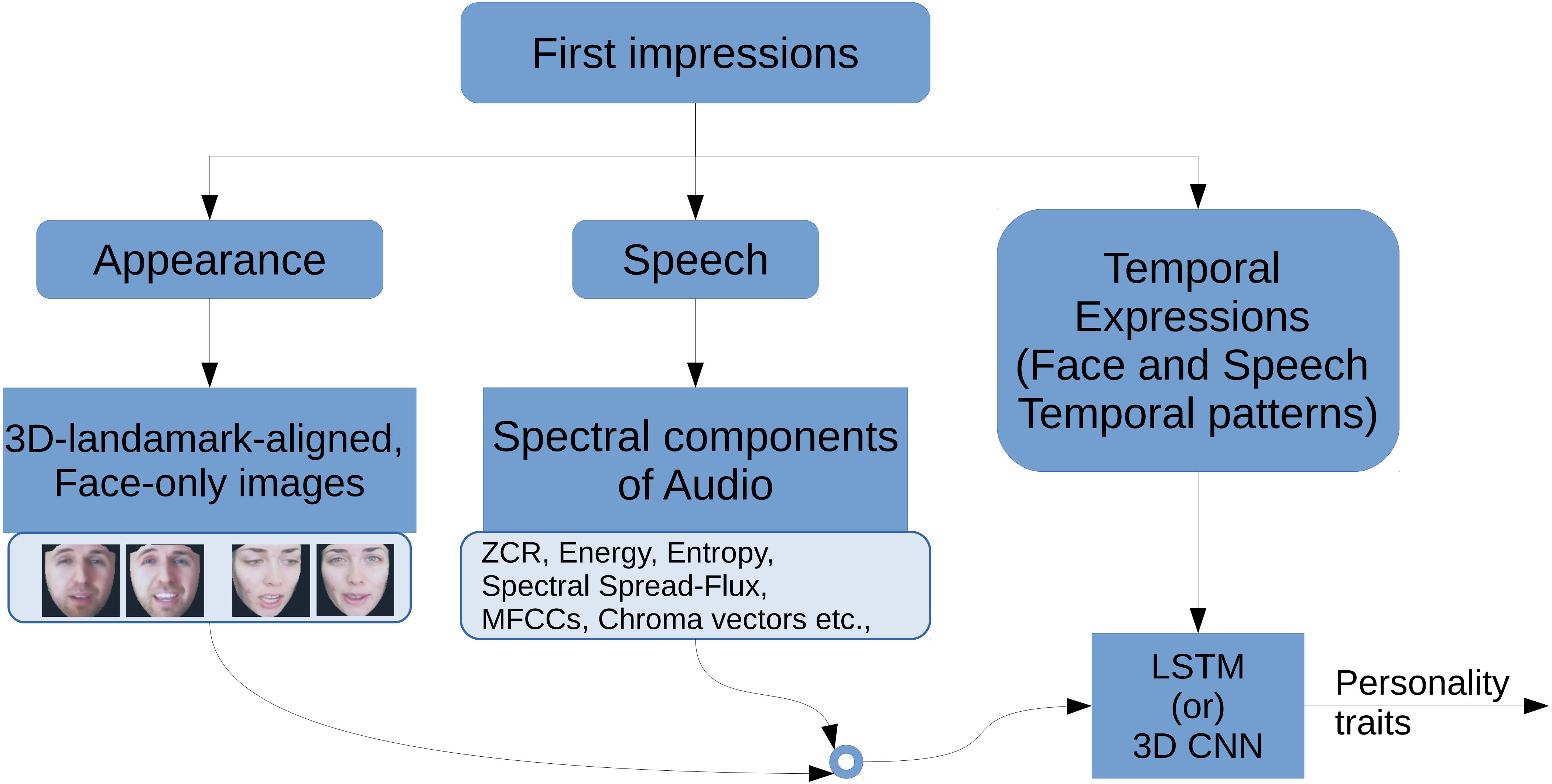
First impressions

Appearance

Speech

Temporal Expressions (Face and Speech Temporal patterns)

Intuition behind the proposed solution



First impressions

Appearance

Speech

Temporal Expressions (Face and Speech Temporal patterns)

3D-landamark-aligned, Face-only images

Spectral components of Audio

ZCR, Energy, Entropy, Spectral Spread-Flux, MFCCs, Chroma vectors etc.,

LSTM (or) 3D CNN

Personality traits

Preprocessing - Audio

- The mean(μ) and standard deviation(σ) of spectral Audio feature attributes

  ZCR, Energy, Spectral properties(Centroid + Spread + Entropy + Rolloff + Flux), Chroma vector + deviation, MFCCs etc., (in total of 34 feature dimensions)

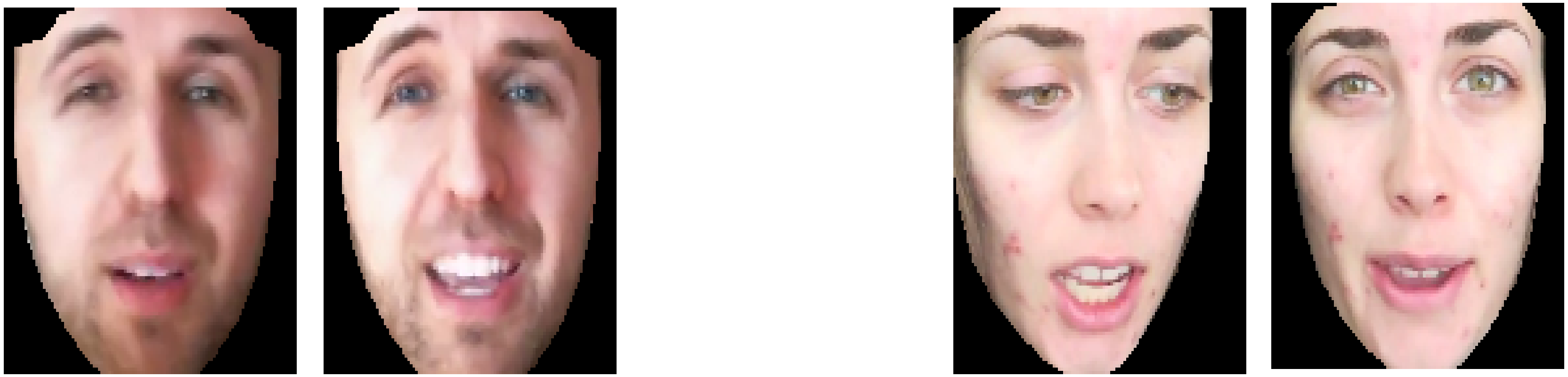
- Total of 68 dimensions (μ and σ for each of 34 feature dimensions)

- Python library[1] 'pyAudioAnalysis' is used for audio feature extraction

[1] https://github.com/tyiannak/pyAudioAnalysis (Theodoros Giannakopoulos)

Preprocessing - Video

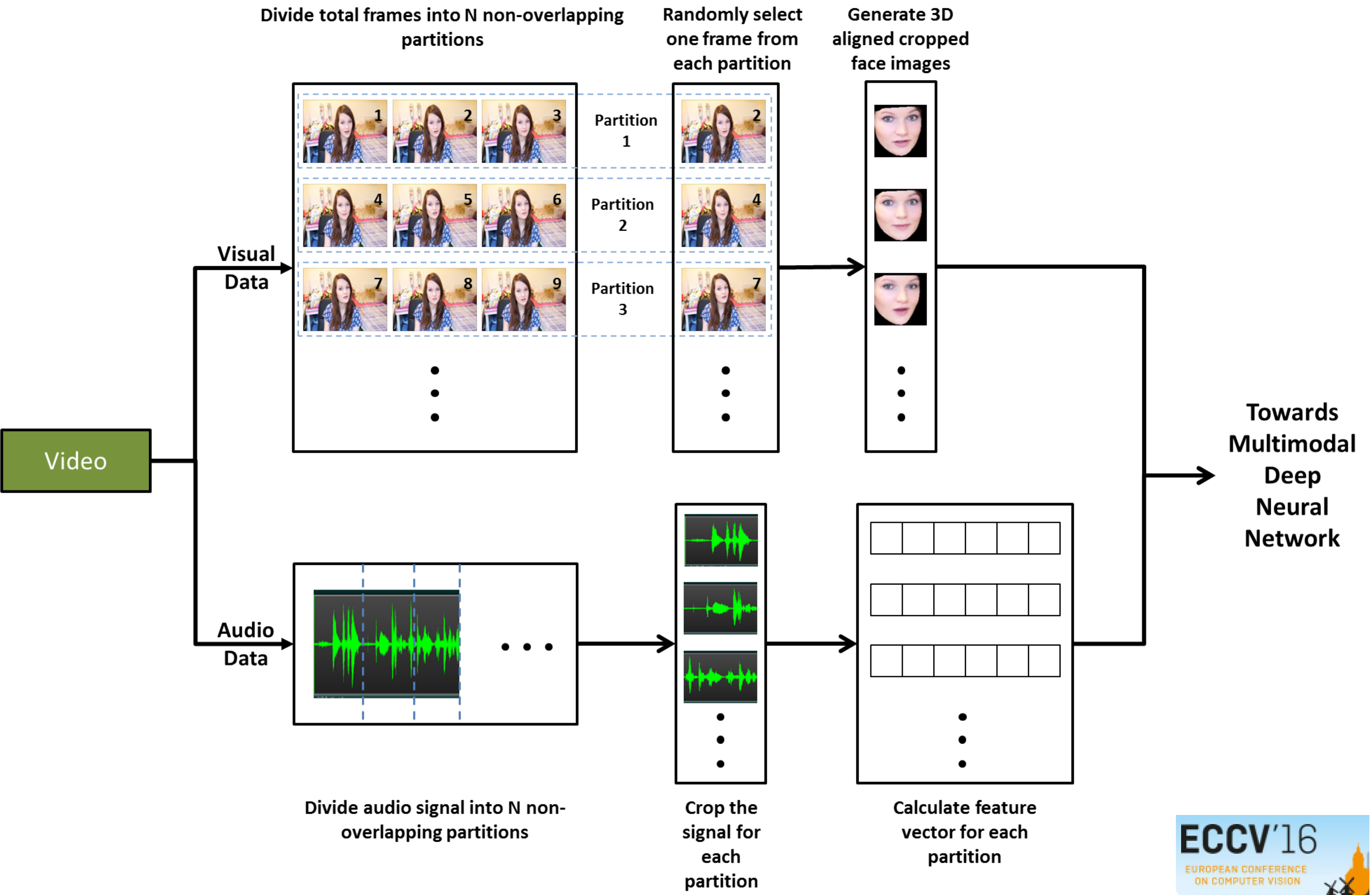
- The 3D-aligned Face is extracted from the frame(s) of the video



- A state-of-the-art open source tool[1] 'OpenFace' is used for Face extraction

[1] https://github.com/TadasBaltrusaitis/OpenFace (Tadas Baltrušaitis)

Data selection for the model



Divide total frames into N non-overlapping partitions

Randomly select one frame from each partition

Generate 3D aligned cropped face images

Visual Data

Partition 1

Partition 2

Partition 3

Video

Towards Multimodal Deep Neural Network

Audio Data

Divide audio signal into N non-overlapping partitions

Crop the signal for each partition

Calculate feature vector for each partition

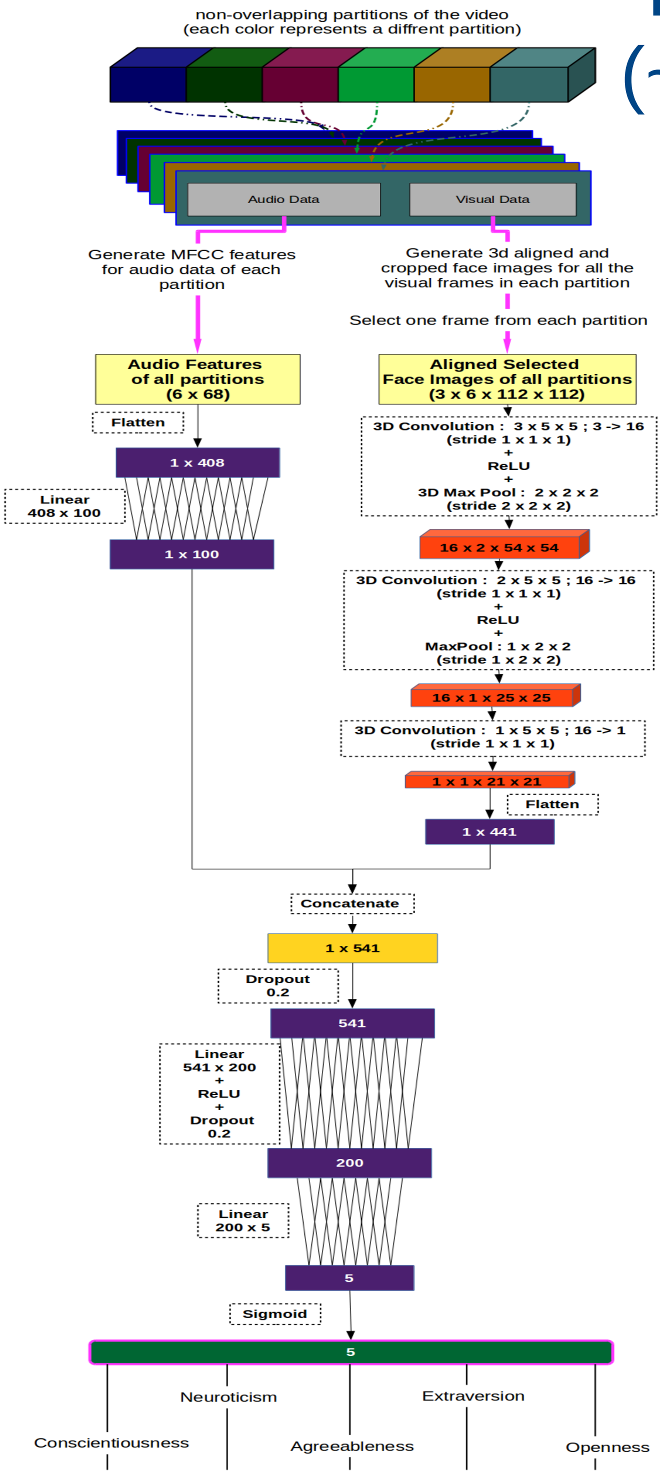
Stochastic feature selection

- Keeping N = 6

  (split the Audio and Video into non-overlapping 6 partitions)

Audio	Visual
68 dimensional feature vector for each of 6 partitions = 6 x 68 feature vectors	For each of 6 non-overlapping partitions, single randomly selected image of 3 x 112 x 112. (= 6 x 3 x 112 x 112) Typically, video length = ~15 seconds 30 frames / second = ~450 frames in total = ~75 frames / partitions = 75^6 combinations of selecting frames **(helps in increasing data points & avoids overfitting)**

Bi-Modal 3D CNN model (~0.17 million parameters)



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

Visual Data

Generate MFCC features for audio data of each partition

Generate 3d aligned and cropped face images for all the visual frames in each partition

Select one frame from each partition

Audio Features of all partitions (6 x 68)

Aligned Selected Face Images of all partitions (3 x 6 x 112 x 112)

Flatten

1 x 408

Linear 408 x 100

1 x 100

3D Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1 x 1) + ReLU + 3D Max Pool : 2 x 2 x 2 (stride 2 x 2 x 2)

16 x 2 x 54 x 54

3D Convolution : 2 x 5 x 5 ; 16 -> 16 (stride 1 x 1 x 1) + ReLU + MaxPool : 1 x 2 x 2 (stride 1 x 2 x 2)

16 x 1 x 25 x 25

3D Convolution : 1 x 5 x 5 ; 16 -> 1 (stride 1 x 1 x 1)

1 x 1 x 21 x 21

Flatten

1 x 441

Concatenate

1 x 541

Dropout 0.2

541

Linear 541 x 200 + ReLU + Dropout 0.2

200

Linear 200 x 5

5

Sigmoid

5

Neuroticism

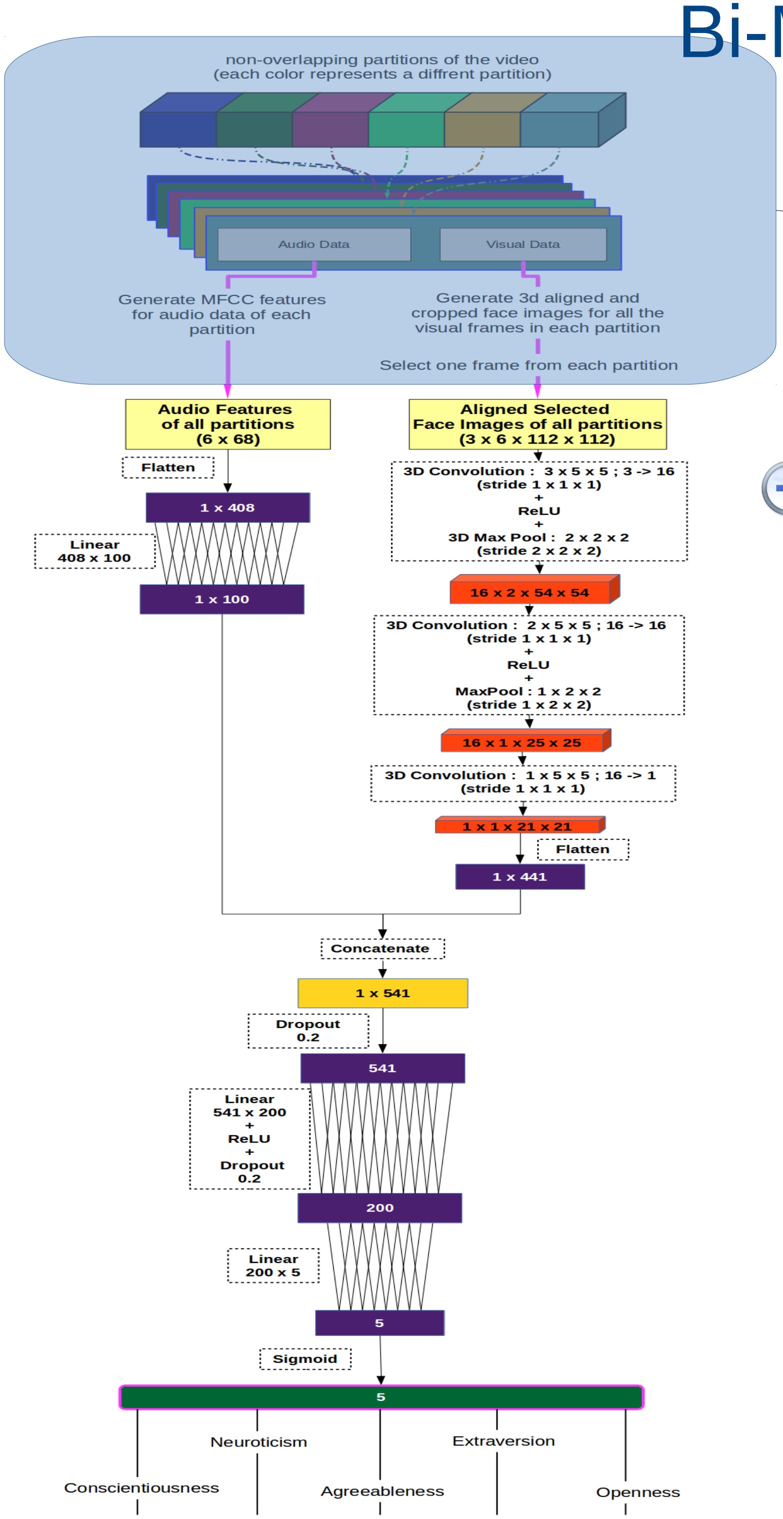
Extraversion

Conscientiousness

Agreeableness

Openness

Bi-Modal 3D CNN model



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

Visual Data

Generate MFCC features for audio data of each partition

Generate 3d aligned and cropped face images for all the visual frames in each partition

Select one frame from each partition

Audio Features of all partitions (6 x 68)

Flatten

1 x 408

Linear 408 x 100

1 x 100

Aligned Selected Face Images of all partitions (3 x 6 x 112 x 112)

3D Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1 x 1) + ReLU + 3D Max Pool : 2 x 2 x 2 (stride 2 x 2 x 2)

16 x 2 x 54 x 54

3D Convolution : 2 x 5 x 5 ; 16 -> 16 (stride 1 x 1 x 1) + ReLU + MaxPool : 1 x 2 x 2 (stride 1 x 2 x 2)

16 x 1 x 25 x 25

3D Convolution : 1 x 5 x 5 ; 16 -> 1 (stride 1 x 1 x 1)

1 x 1 x 21 x 21

Flatten

1 x 441

Concatenate

1 x 541

Dropout 0.2

541

Linear 541 x 200 + ReLU + Dropout 0.2

200

Linear 200 x 5

5

Sigmoid

5

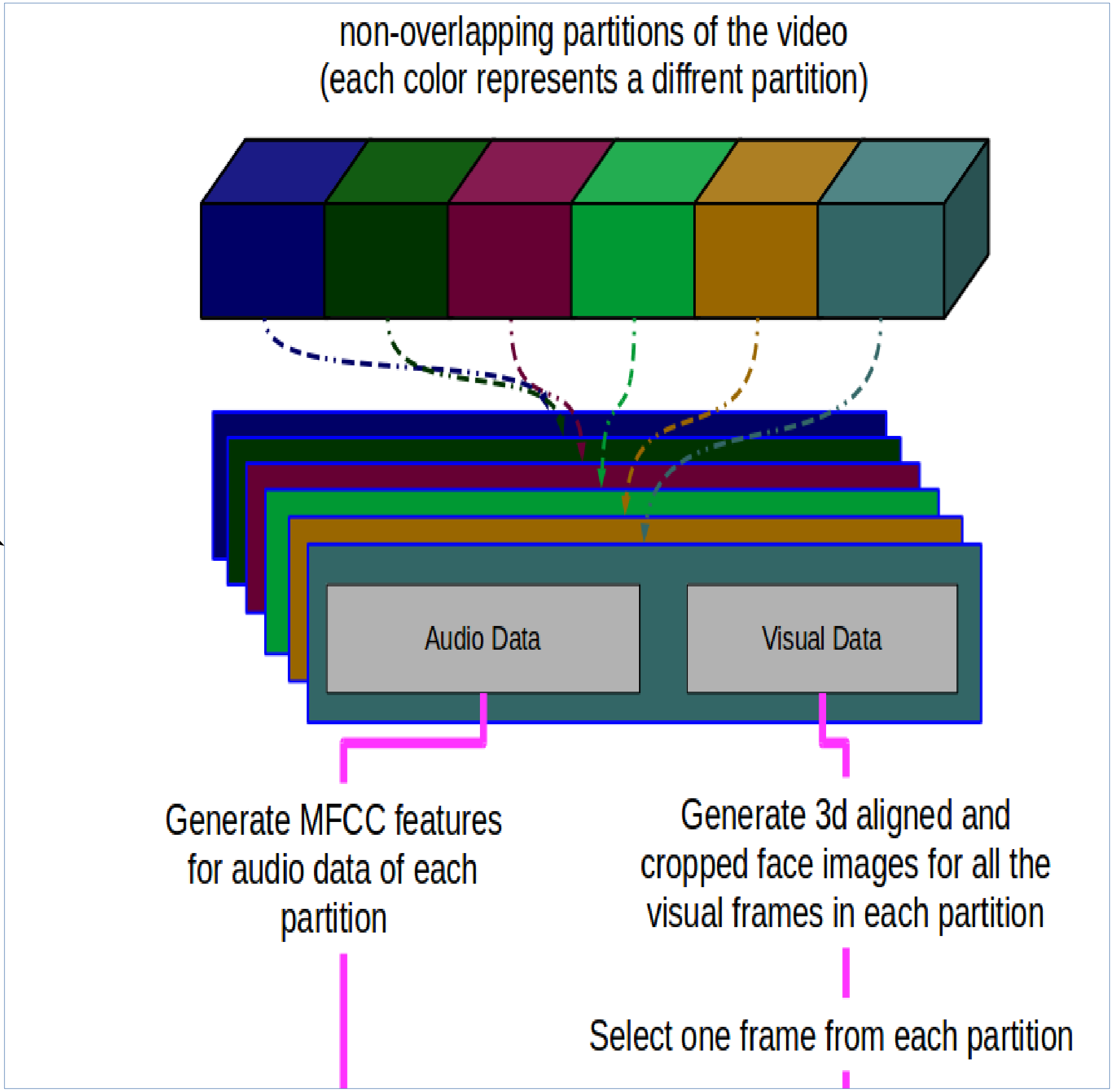
Neuroticism

Extraversion

Conscientiousness

Agreeableness

Openness



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

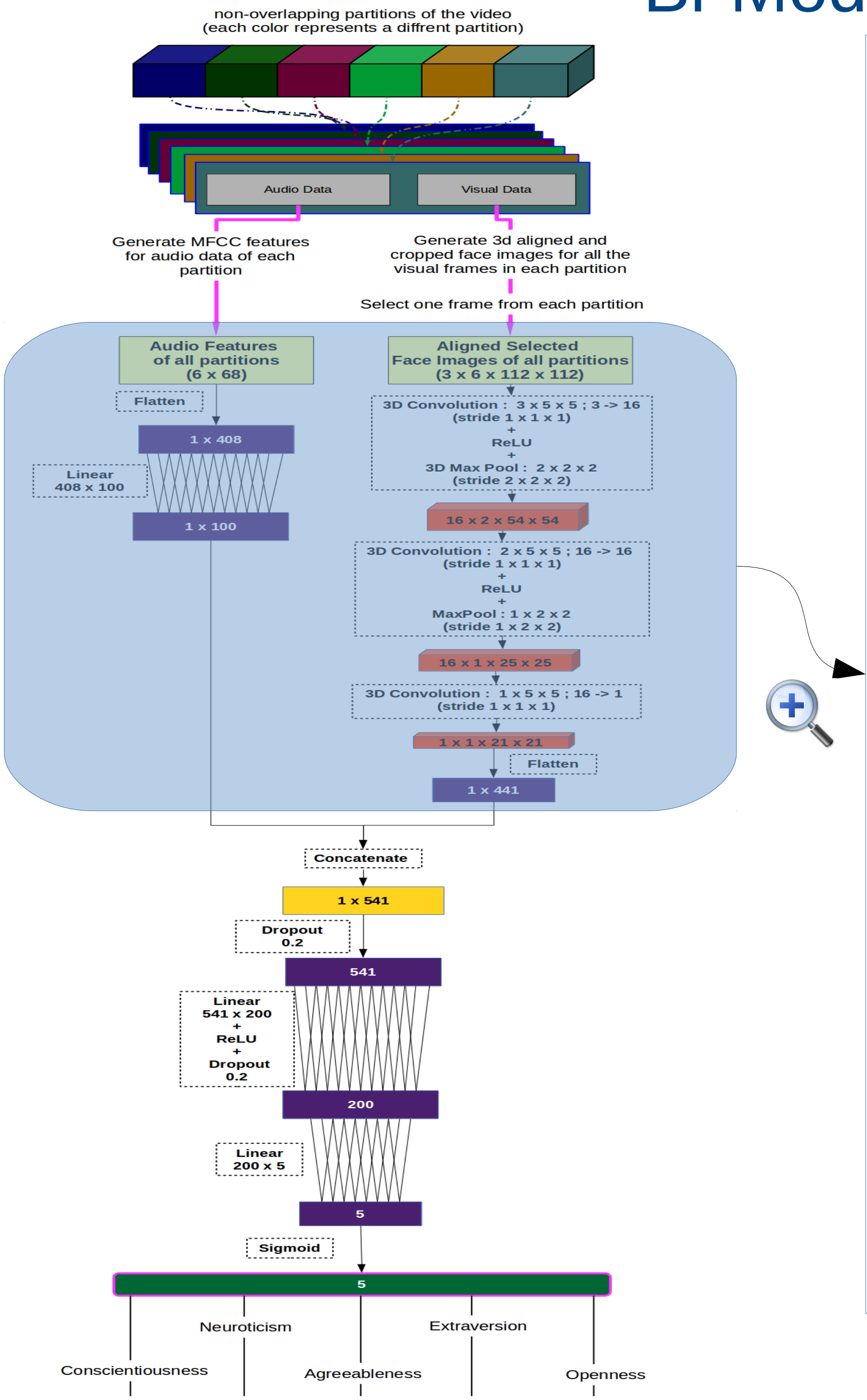
Visual Data

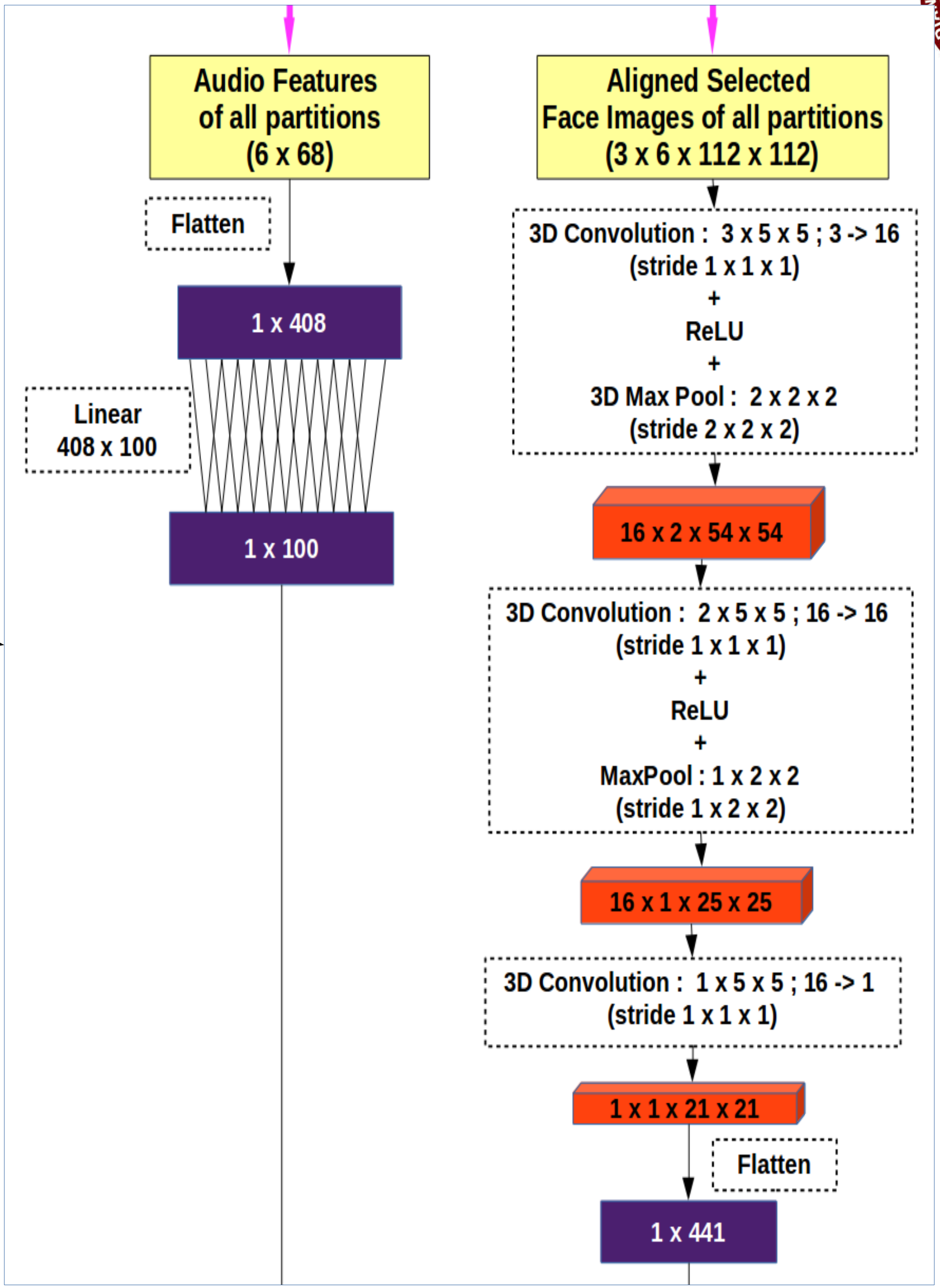
Generate MFCC features for audio data of each partition

Generate 3d aligned and cropped face images for all the visual frames in each partition

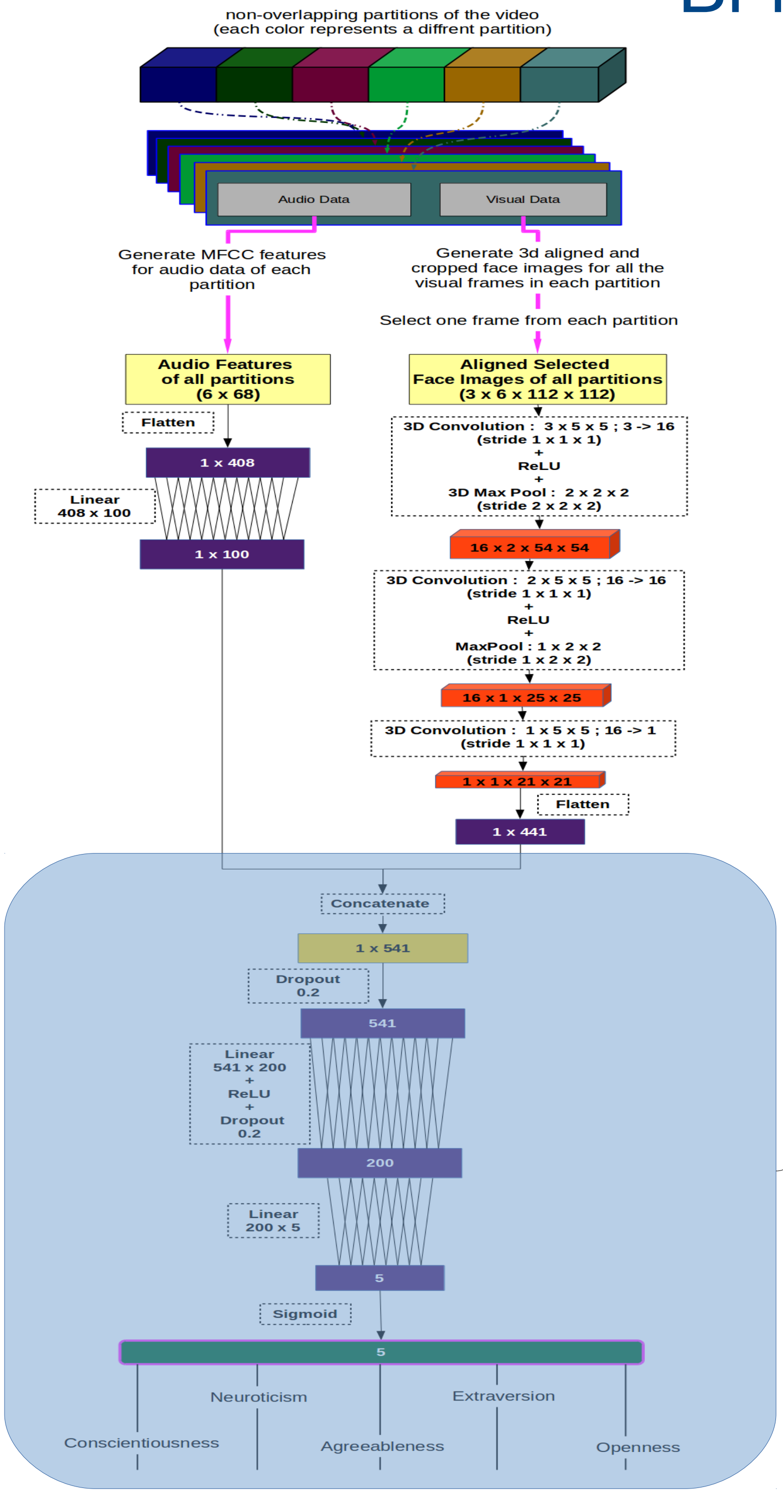
Select one frame from each partition

Bi-Modal 3D CNN model


non-overlapping partitions of the video
(each color represents a diffrent partition)
Audio Data
Visual Data
Generate MFCC features for audio data of each partition
Generate 3d aligned and cropped face images for all the visual frames in each partition
Select one frame from each partition
Audio Features of all partitions (6 x 68)
Flatten
1 x 408
Linear 408 x 100
1 x 100
Aligned Selected Face Images of all partitions (3 x 6 x 112 x 112)
3D Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1 x 1) + ReLU + 3D Max Pool : 2 x 2 x 2 (stride 2 x 2 x 2)
16 x 2 x 54 x 54
3D Convolution : 2 x 5 x 5 ; 16 -> 16 (stride 1 x 1 x 1) + ReLU + MaxPool : 1 x 2 x 2 (stride 1 x 2 x 2)
16 x 1 x 25 x 25
3D Convolution : 1 x 5 x 5 ; 16 -> 1 (stride 1 x 1 x 1)
1 x 1 x 21 x 21
Flatten
1 x 441
Concatenate
1 x 541
Dropout 0.2
541
Linear 541 x 200 + ReLU + Dropout 0.2
200
Linear 200 x 5
5
Sigmoid
5
Neuroticism
Extraversion
Conscientiousness
Agreeableness
Openness


Audio Features of all partitions (6 x 68)
Flatten
1 x 408
Linear 408 x 100
1 x 100
Aligned Selected Face Images of all partitions (3 x 6 x 112 x 112)
3D Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1 x 1) + ReLU + 3D Max Pool : 2 x 2 x 2 (stride 2 x 2 x 2)
16 x 2 x 54 x 54
3D Convolution : 2 x 5 x 5 ; 16 -> 16 (stride 1 x 1 x 1) + ReLU + MaxPool : 1 x 2 x 2 (stride 1 x 2 x 2)
16 x 1 x 25 x 25
3D Convolution : 1 x 5 x 5 ; 16 -> 1 (stride 1 x 1 x 1)
1 x 1 x 21 x 21
Flatten
1 x 441

Bi-Modal 3D CNN model



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

Visual Data

Generate MFCC features
for audio data of each
partition

Generate 3d aligned and
cropped face images for all the
visual frames in each partition

Select one frame from each partition

Audio Features
of all partitions
(6 x 68)

Aligned Selected
Face Images of all partitions
(3 x 6 x 112 x 112)

Flatten

1 x 408

Linear
408 x 100

1 x 100

3D Convolution : 3 x 5 x 5 ; 3 -> 16
(stride 1 x 1 x 1)
+
ReLU
+
3D Max Pool : 2 x 2 x 2
(stride 2 x 2 x 2)

16 x 2 x 54 x 54

3D Convolution : 2 x 5 x 5 ; 16 -> 16
(stride 1 x 1 x 1)
+
ReLU
+
MaxPool : 1 x 2 x 2
(stride 1 x 2 x 2)

16 x 1 x 25 x 25

3D Convolution : 1 x 5 x 5 ; 16 -> 1
(stride 1 x 1 x 1)

1 x 1 x 21 x 21

Flatten

1 x 441

Concatenate

1 x 541

Dropout
0.2

541

Linear
541 x 200
+
ReLU
+
Dropout
0.2

200

Linear
200 x 5

5

Sigmoid

5

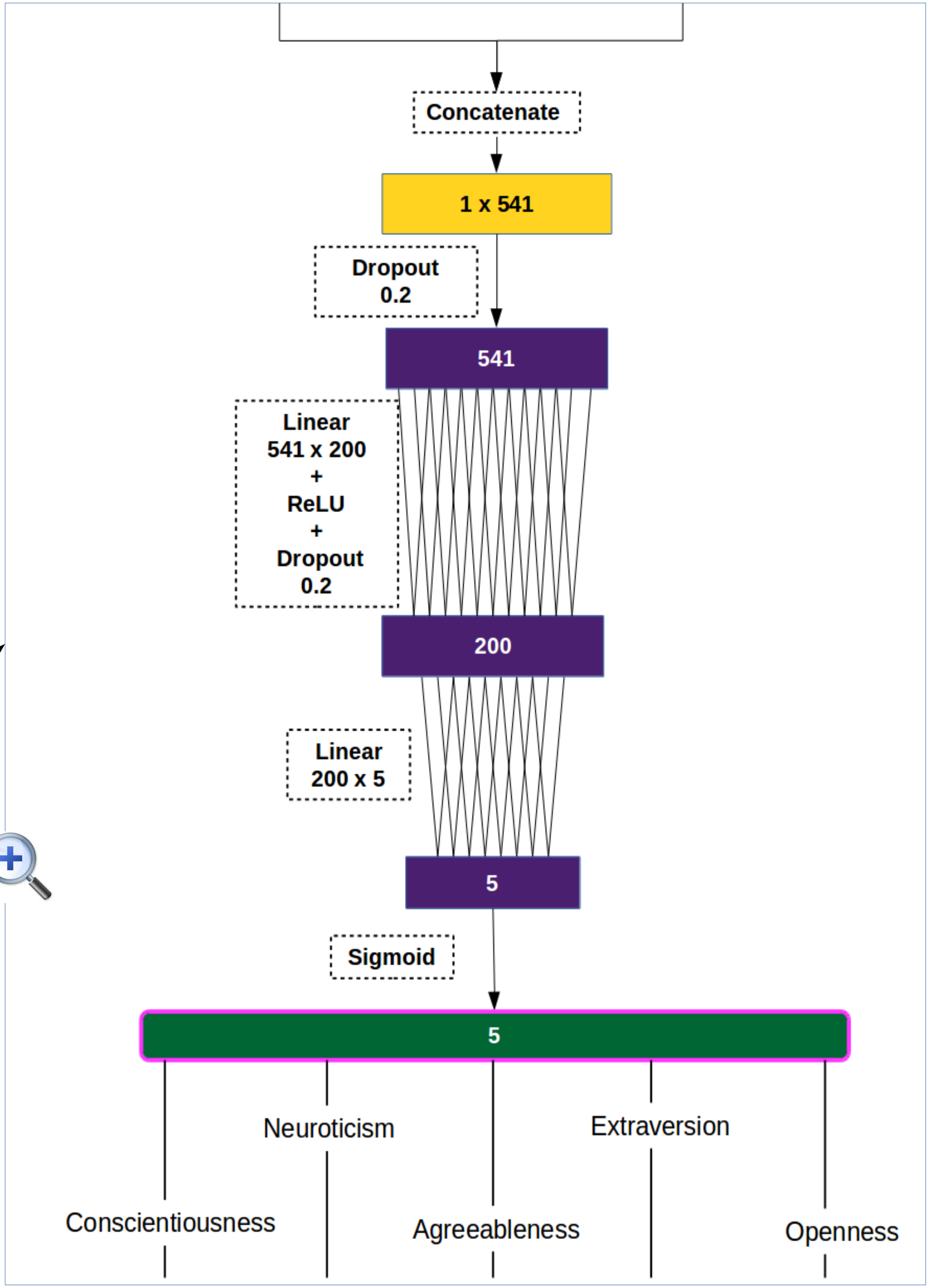
Neuroticism

Extraversion

Conscientiousness

Agreeableness

Openness



Concatenate

1 x 541

Dropout
0.2

541

Linear
541 x 200
+
ReLU
+
Dropout
0.2

200

Linear
200 x 5

5

Sigmoid

5

Neuroticism

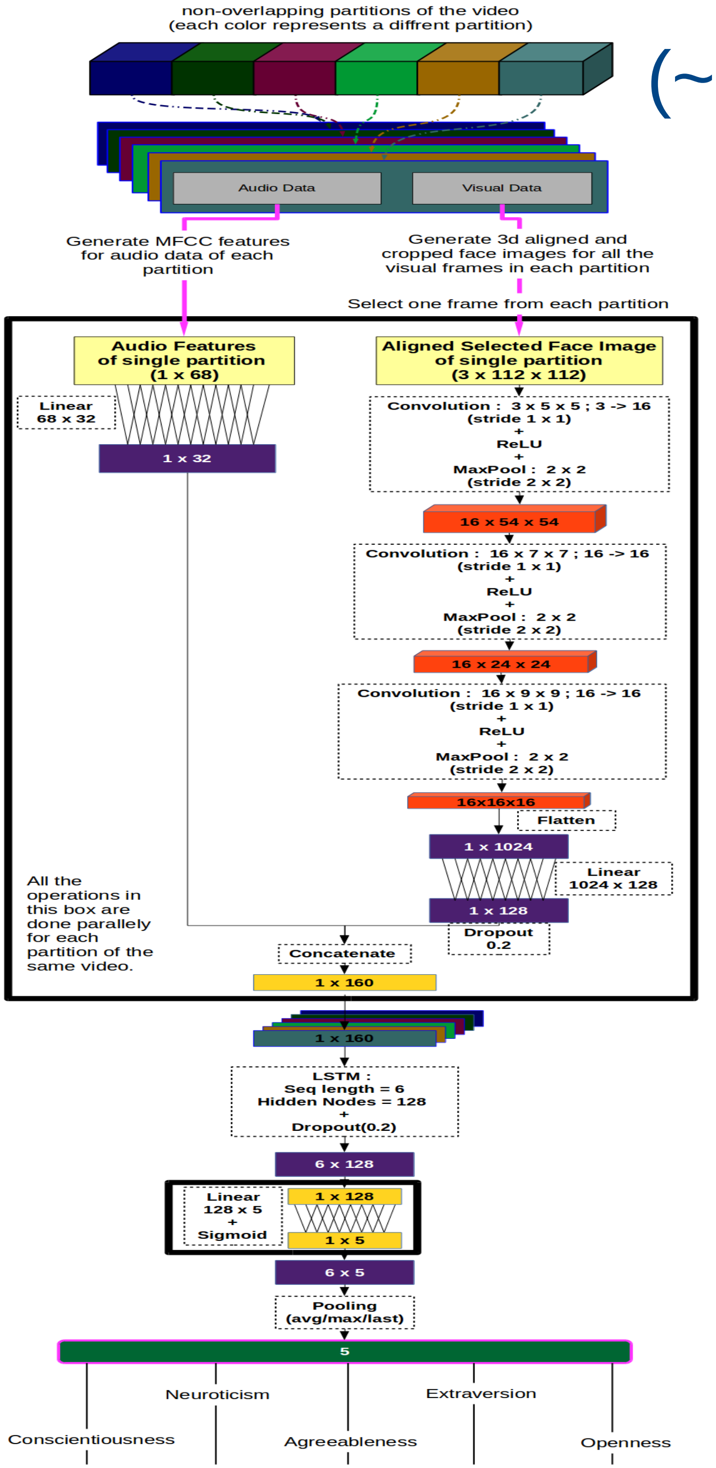
Extraversion

Conscientiousness

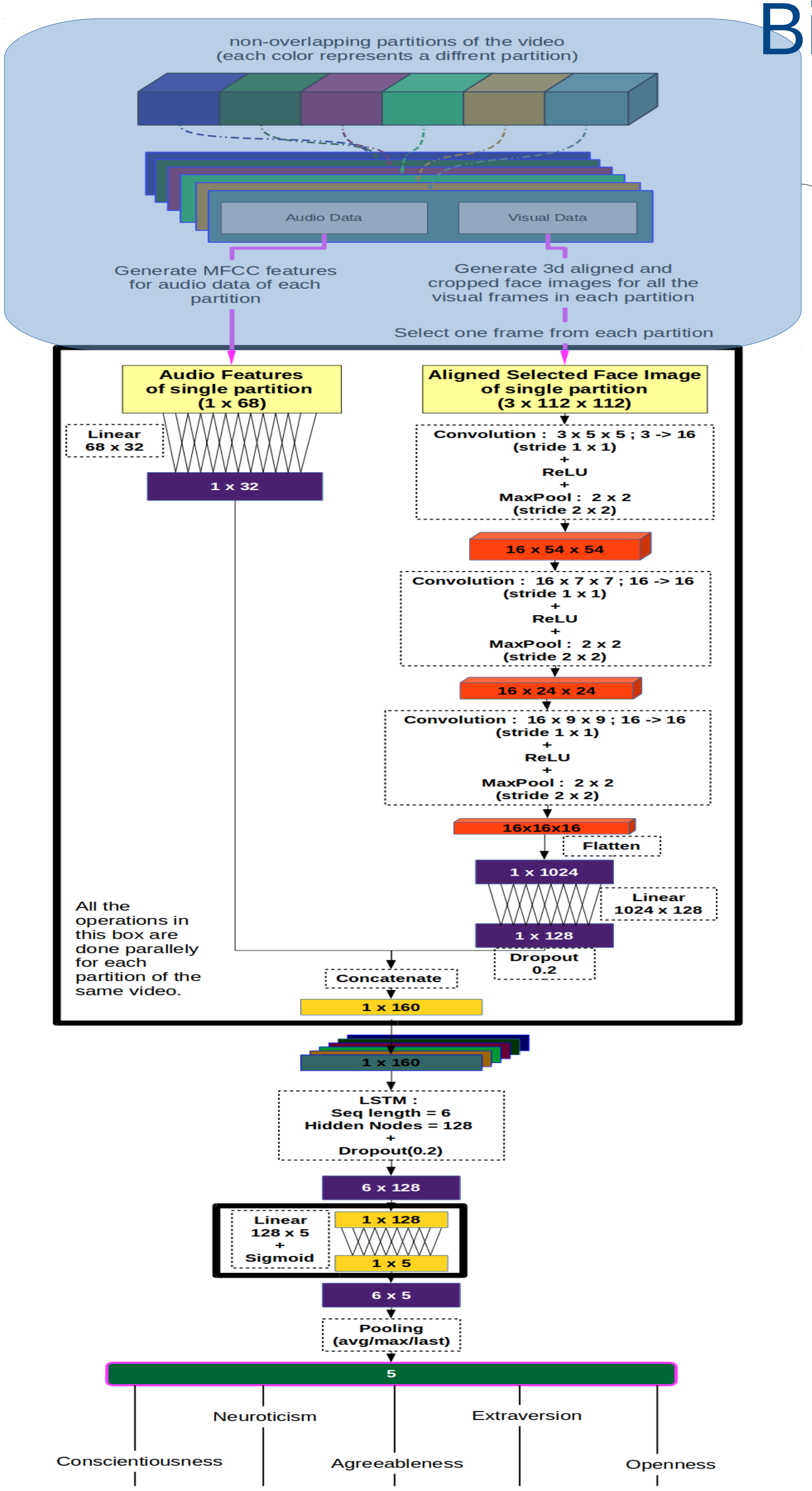
Agreeableness

Openness

Bi-Modal LSTM model (~0.32 million parameters)


non-overlapping partitions of the video
(each color represents a diffrent partition)
Audio Data
Visual Data
Generate MFCC features for audio data of each partition
Generate 3d aligned and cropped face images for all the visual frames in each partition
Select one frame from each partition
Audio Features of single partition (1 x 68)
Linear 68 x 32
1 x 32
Aligned Selected Face Image of single partition (3 x 112 x 112)
Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16 x 54 x 54
Convolution : 16 x 7 x 7 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16 x 24 x 24
Convolution : 16 x 9 x 9 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16x16x16
Flatten
1 x 1024
Linear 1024 x 128
1 x 128
Dropout 0.2
All the operations in this box are done parallely for each partition of the same video.
Concatenate
1 x 160
1 x 160
LSTM : Seq length = 6 Hidden Nodes = 128 + Dropout(0.2)
6 x 128
Linear 128 x 5 + Sigmoid
1 x 128
1 x 5
6 x 5
Pooling (avg/max/last)
5
Neuroticism
Extraversion
Conscientiousness
Agreeableness
Openness

Bi-Modal LSTM model



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

Visual Data

Generate MFCC features for audio data of each partition

Generate 3d aligned and cropped face images for all the visual frames in each partition

Select one frame from each partition

Audio Features of single partition (1 x 68)

Linear 68 x 32

1 x 32

Aligned Selected Face Image of single partition (3 x 112 x 112)

Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)

16 x 54 x 54

Convolution : 16 x 7 x 7 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)

16 x 24 x 24

Convolution : 16 x 9 x 9 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)

16x16x16

Flatten

1 x 1024

Linear 1024 x 128

1 x 128

Dropout 0.2

All the operations in this box are done parallely for each partition of the same video.

Concatenate

1 x 160

1 x 160

LSTM : Seq length = 6 Hidden Nodes = 128 + Dropout(0.2)

6 x 128

Linear 128 x 5 + Sigmoid

1 x 128

1 x 5

6 x 5

Pooling (avg/max/last)

5

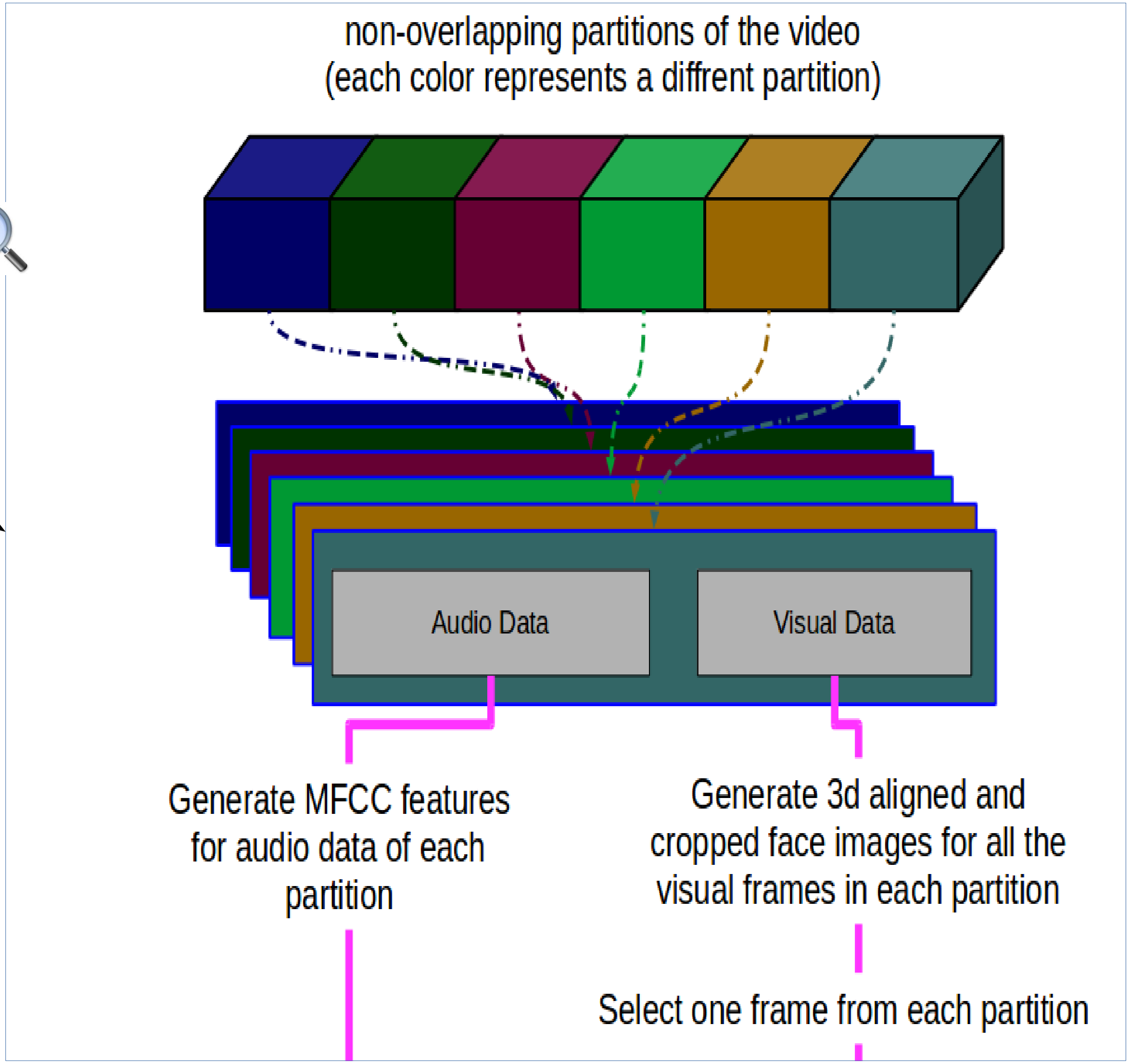
Neuroticism

Extraversion

Conscientiousness

Agreeableness

Openness



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

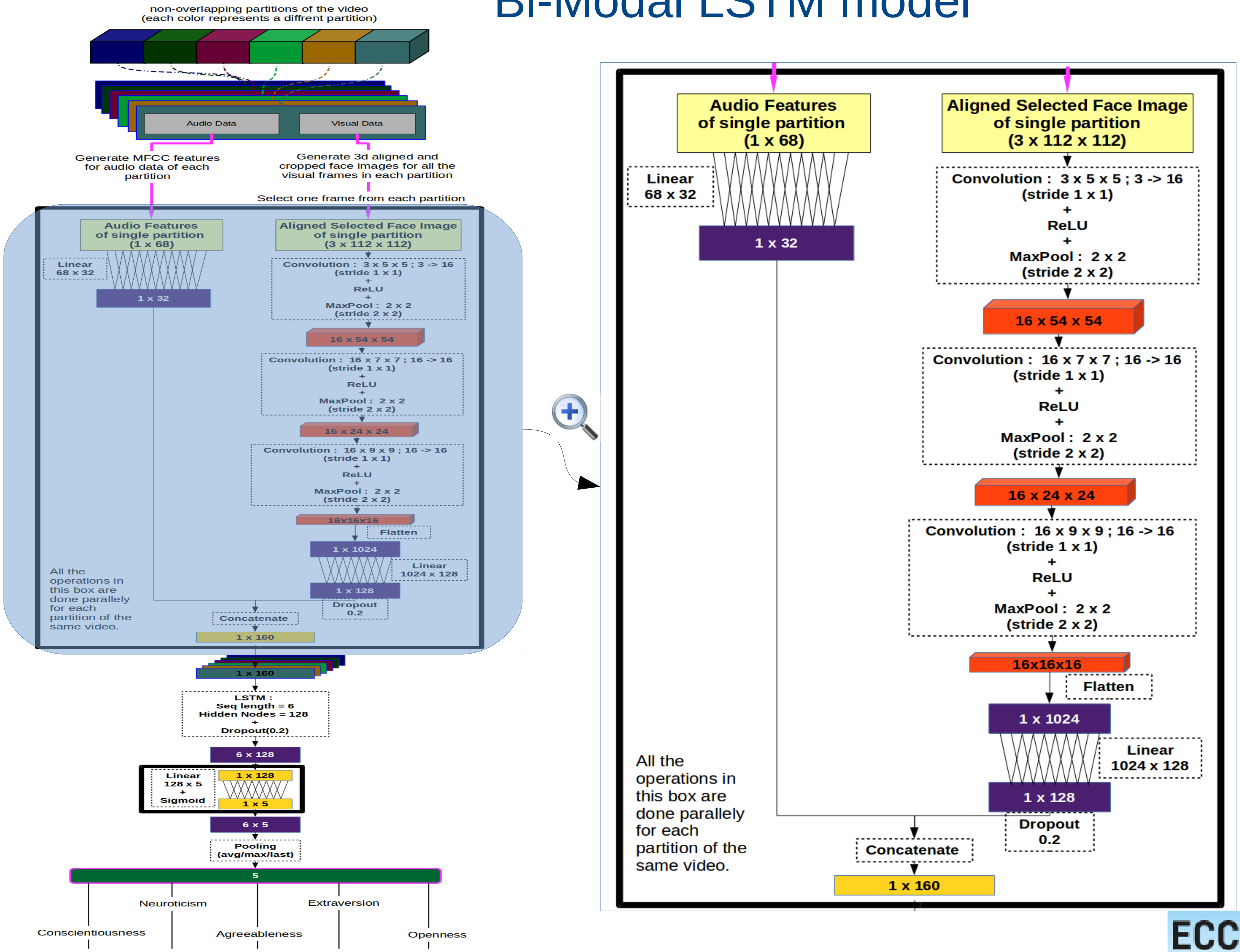
Visual Data

Generate MFCC features for audio data of each partition

Generate 3d aligned and cropped face images for all the visual frames in each partition

Select one frame from each partition

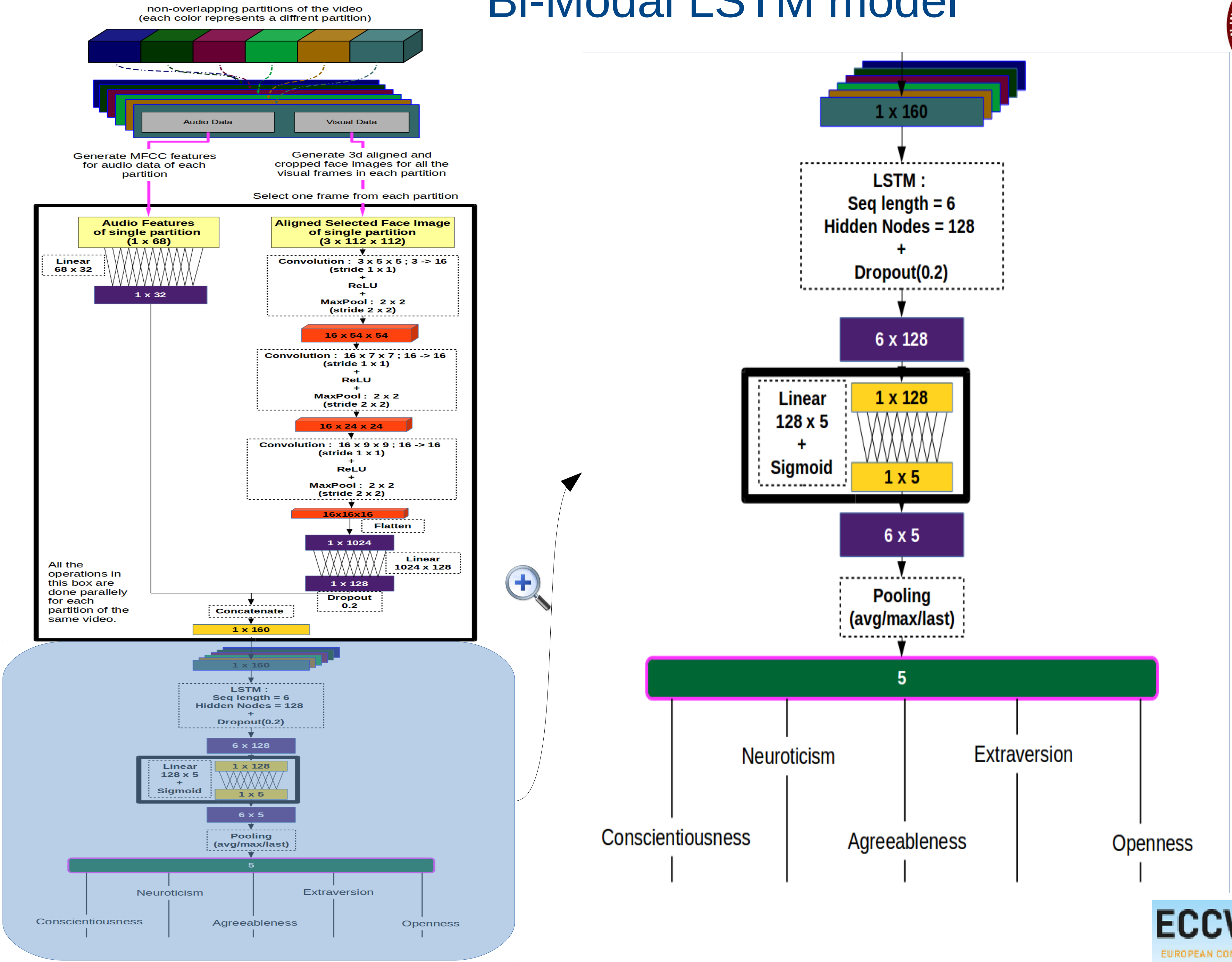
Bi-Modal LSTM model



non-overlapping partitions of the video
(each color represents a diffrent partition)
Audio Data
Visual Data
Generate MFCC features for audio data of each partition
Generate 3d aligned and cropped face images for all the visual frames in each partition
Select one frame from each partition
Audio Features of single partition (1 x 68)
Aligned Selected Face Image of single partition (3 x 112 x 112)
Linear 68 x 32
1 x 32
Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16 x 54 x 54
Convolution : 16 x 7 x 7 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16 x 24 x 24
Convolution : 16 x 9 x 9 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16x16x16
Flatten
1 x 1024
Linear 1024 x 128
1 x 128
Dropout 0.2
Concatenate
1 x 160
All the operations in this box are done parallely for each partition of the same video.
1 x 160
LSTM : Seq length = 6 Hidden Nodes = 128 + Dropout(0.2)
6 x 128
Linear 128 x 5 + Sigmoid
1 x 128
1 x 5
6 x 5
Pooling (avg/max/last)
5
Conscientiousness
Neuroticism
Agreeableness
Extraversion
Openness
Audio Features of single partition (1 x 68)
Aligned Selected Face Image of single partition (3 x 112 x 112)
Linear 68 x 32
1 x 32
Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16 x 54 x 54
Convolution : 16 x 7 x 7 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16 x 24 x 24
Convolution : 16 x 9 x 9 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)
16x16x16
Flatten
1 x 1024
Linear 1024 x 128
1 x 128
Dropout 0.2
Concatenate
1 x 160
All the operations in this box are done parallely for each partition of the same video.

ECCV'16
EUROPEAN CONFERENCE ON COMPUTER VISION

Bi-Modal LSTM model



non-overlapping partitions of the video
(each color represents a diffrent partition)

Audio Data

Visual Data

Generate MFCC features for audio data of each partition

Generate 3d aligned and cropped face images for all the visual frames in each partition

Select one frame from each partition

Audio Features of single partition (1 x 68)

Linear 68 x 32

1 x 32

Aligned Selected Face Image of single partition (3 x 112 x 112)

Convolution : 3 x 5 x 5 ; 3 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)

16 x 54 x 54

Convolution : 16 x 7 x 7 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)

16 x 24 x 24

Convolution : 16 x 9 x 9 ; 16 -> 16 (stride 1 x 1) + ReLU + MaxPool : 2 x 2 (stride 2 x 2)

16x16x16

Flatten

1 x 1024

Linear 1024 x 128

1 x 128

Dropout 0.2

Concatenate

1 x 160

All the operations in this box are done parallely for each partition of the same video.

1 x 160

LSTM : Seq length = 6 Hidden Nodes = 128 + Dropout(0.2)

6 x 128

Linear 128 x 5 + Sigmoid

1 x 128

1 x 5

6 x 5

Pooling (avg/max/last)

5

Neuroticism

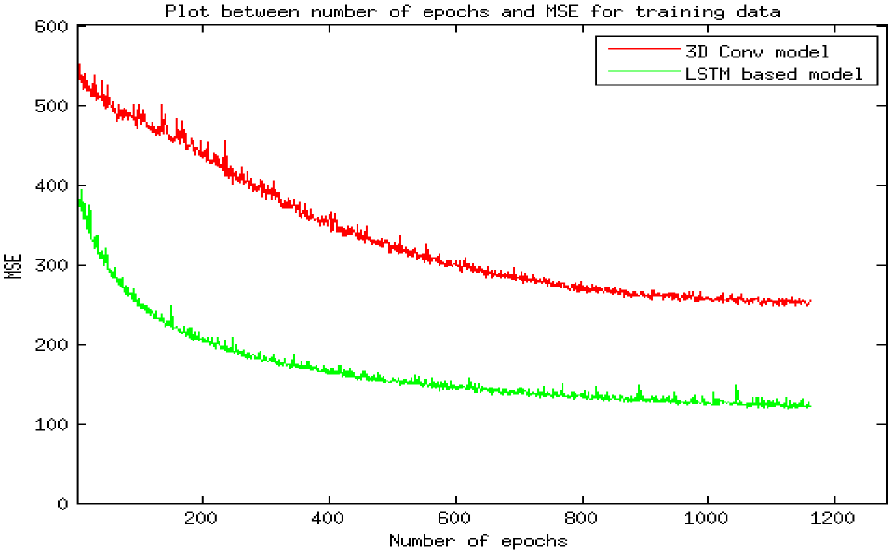
Extraversion

Conscientiousness

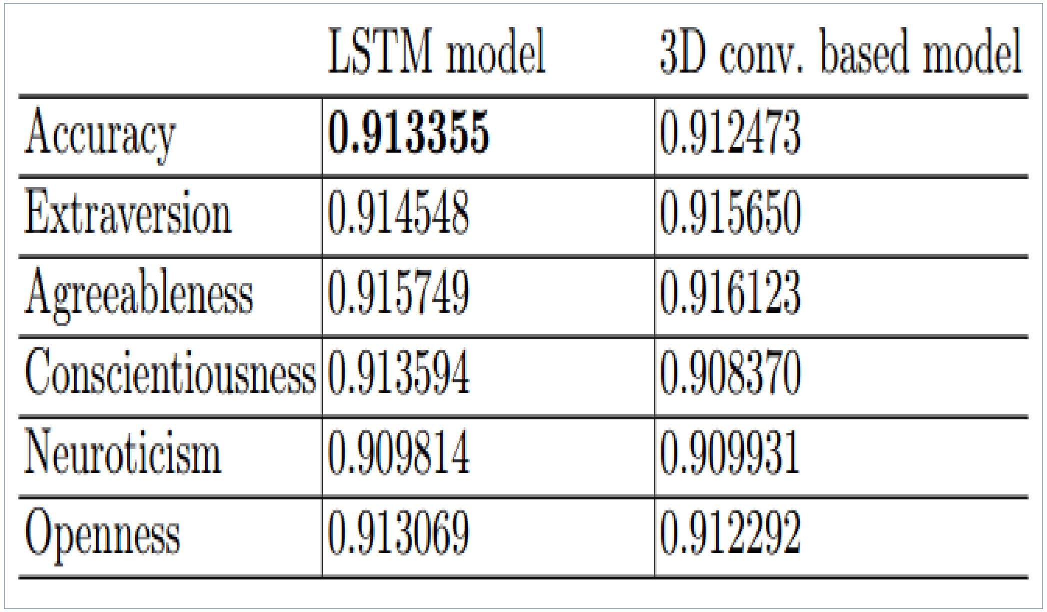
Agreeableness

Openness

Results

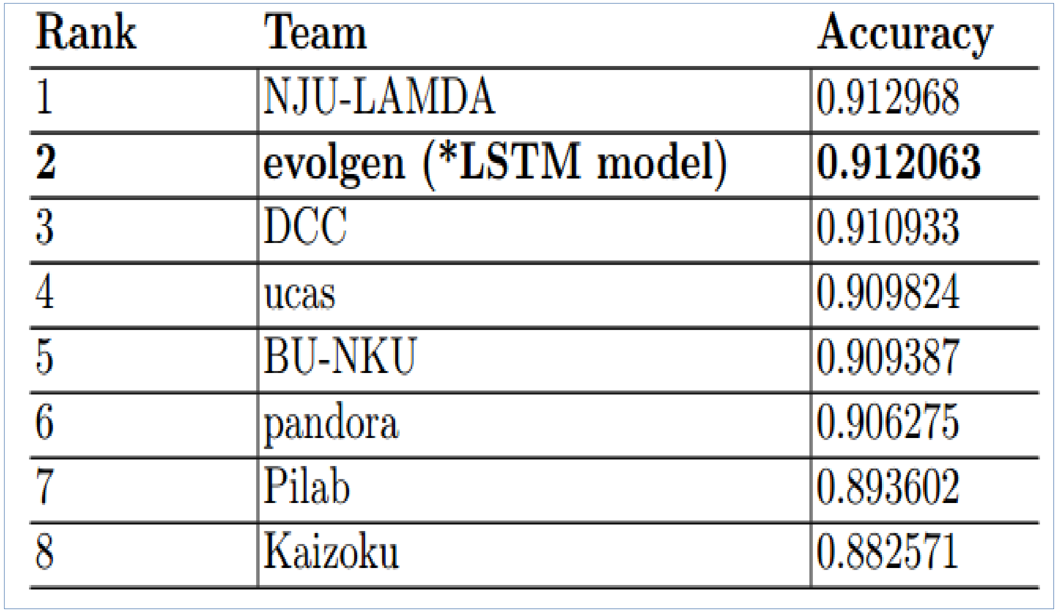


Validation phase:



	LSTM model	3D conv. based model
Accuracy	**0.913355**	0.912473
Extraversion	0.914548	0.915650
Agreeableness	0.915749	0.916123
Conscientiousness	0.913594	0.908370
Neuroticism	0.909814	0.909931
Openness	0.913069	0.912292

Test phase:



Rank	Team	Accuracy
1	NJU-LAMDA	0.912968
2	**evolgen (*LSTM model)**	**0.912063**
3	DCC	0.910933
4	ucas	0.909824
5	BU-NKU	0.909387
6	pandora	0.906275
7	Pilab	0.893602
8	Kaizoku	0.882571

Possible future directions

- Add linguistic feature descriptors along with Audio and Visual features (using speech recognition)?
- Eliminate preprocessing
  - of video frames (i.e., to include Background cues)
  - of Audio frames (i.e., extract features directly from Audio using CNN-like setup)