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

Arulkumar Subramaniam*, Vismay Patel*, Ashish Mishra,
Prashanth Balasubramanian, Anurag Mittal

Department of Computer Science and Engineering,
Indian Institute of Technology Madras
{aruls, vismay, mishra, bprash, amittal}@cse.iitm.ac.in

Abstract. We propose a novel approach for First Impressions Recognition in terms of the Big Five personality-traits from short videos. The Big Five personality traits is a model to describe human personality using five broad categories: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. We train two bi-modal end-to-end deep neural network architectures using temporally ordered audio and novel stochastic visual features from few frames, without over-fitting. We empirically show that the trained models perform exceptionally well, even after training from a small sub-portions of inputs. Our method is evaluated in ChaLearn LAP 2016 Apparent Personality Analysis (APA) competition using ChaLearn LAP APA2016 dataset and achieved excellent performance.

Keywords: Deep Learning, Bi-modal Neural Networks, First Impressions Analysis, Apparent Personality Analysis

1 Introduction

A "First Impression" is the event when a person encounters another person and forms a mental image about the person [1]. Here the mental image can be based on lot of characteristics such as facial expressions, action, physical appearance, the way of interaction, body language, etc. According to research in Psychology [2], the first impressions are formed even with a limited exposure (as less as 100ms) to unfamiliar faces. Forming a first impression is usually done in terms of Personality-traits recognition. Determining Personality-traits automatically will be helpful in human resourcing, recruitment process. An automatic analysis of Personality-traits will help people to train themselves.

The problem can be represented as in Table 1. A short video with a person's interview is given as input and the output is expected to be 5 fractional values in the range [0, 1] representing Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness. These five are collectively known as the "Big-Five personality-traits".

* Authors contributed equally

There has not been much work in literature for First-impressions recognition, though the researchers have explored Emotion recognition [3–8], a related area in terms of the type of problem and features (hand-crafted features as well as deep features) used. There are many ways, people express their emotions, among which facial expressions are the most useful [3–6]. Cohen et. al [4] used HMM based models to categorize the emotions in a video into six types: (1)happy, (2)angry, (3)surprise, (4)disgust, (5)fear, (6)sad. Their extended work [5] in multilevel HMM performed automatic segmentation and recognition from a continuous signal. Xiaowei Zhao et. al [8] proposed iterative Multi-Output Random Forests for face analysis in images using a combination of three tasks namely Facial landmark detection, Head pose estimation and Facial expression recognition. Deep features have also been used for facial analysis. Javier G. Razuri et. al [9] have extracted features from regions around eyes and mouth for recognizing the human emotions. Their idea was that information related to emotions could be captured by tracking the expressions around eyes and mouth region. The extracted features are then input into a feed-forward neural network trained by back-propagation for classification of emotions.

Although, facial expressions form an important cue, they alone are not sufficient to recognize emotions effectively. Loic et. al [7] used facial expressions, gestures and acoustic analysis of speech based features. In their work, they have used a Bayesian classifier to recognize one of the eight types of emotions (Anger, Despair, Interest, Pleasure, Sadness, Irritation, Joy and Pride). They presented uni-modal (trained separately with all three types of features), bi-modal (combine two modes together) and multi-modal (combine all three modes together). Among all combinations, they observed that multi-modal based classification yielded the best performance.

We propose two end-to-end trained deep learning models that use audio features and face images for recognizing first impressions. In the first model, we propose a Volumetric (3D) convolution based deep neural network for determining personality-traits. 3D convolution was also used by Ji et. al [10], although for the task of action recognition from videos of unconstrained settings. In the second model, we formulate an LSTM(Long Short Term Memory) based deep neural network for learning temporal patterns in the audio and visual features. Both the models concatenate the features extracted from audio and visual data in a later stage. This is in spirit of the observations made in some studies [7] that multi-modal classification yields superior performance.

Our contribution in this paper is two-fold. First, mining temporal patterns in audio and visual features is an important cue for recognizing first impressions effectively. Secondly, such patterns can be mined from a few frames selected in a stochastic manner rather than the complete video, and still predict the first impressions with good accuracy. The proposed methods have been ranked second on the ChaLearn LAP APA2016 challenge(first round) [11].

This paper is organized as follows. In Section 2, we describe the two models in detail and the steps followed to prepare the input data and features for the models. Section 3 describes the novel stochastic method of training and testing

the networks. In Section 4, we discuss the Apparent Personality Analysis 2016: First Impressions Dataset, the evaluation protocol, the implementation details and the experimental results obtained in two phases of the competition. Section 5 concludes the paper providing future direction for the work.

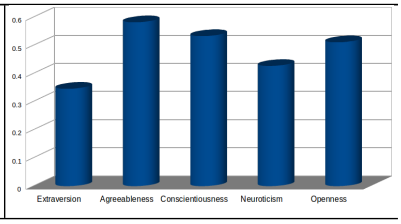
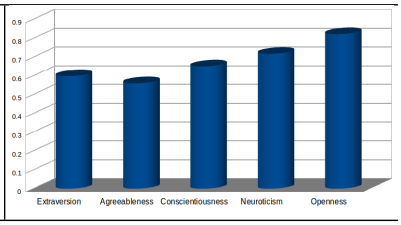
Input	Target
	
	

Table 1: Example of Input and Target. Input is the raw video containing a person's interview & output will be the predicted personality-traits values.

2 Methodology

We propose two bi-modal deep neural network architectures that have two branches, one for encoding audio features and the other for visual features. Inputs to both the audio and visual branches of the model are generated after pre-processing the raw video data. Features extracted from both the branches are fused in a later stage of the model, while the complete network is trained end-to-end. In this section, we describe the pre-processing that was performed on the data and the architecture of models in detail.

2.1 Audio data pre-processing

Given a video, we extract its audio component and split the audio component into N non-overlapping partitions as shown in figure 1. From each individual partition, we extract "mean and standard deviation" of certain properties (table 2) of audio signal. We use an open-source python based audio processing library called pyAudioAnalysis [12,13] for this purpose. The hand-crafted features are of 68 dimensions, which includes the mean and standard deviation of the following attributes:

Attribute Name	**Description**
Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame
Energy	The sum of squares of the signal values, normalized by the respective frame length.

Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
Spectral Centroid	The centre of gravity of the spectrum.
Spectral Spread	The second central moment of the spectrum.
Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Table 2: Audio features extracted using pyAudioAnalysis [14]

2.2 Visual data pre-processing

The visual processing branch of the model takes as input, a set of 'N' 3D aligned segmented face images. We segment the face images to prevent the background from affecting the predictions, which should rather depend only on the features of the face (gaze direction, movements of eye, lips, etc). We use facial landmark detection and tracking to segment the faces. The landmark points are then aligned to fixed locations, which give us segmented face images that have also been aligned. We use an open-sourced C++ library OpenFace [15, 16] for all the visual pre-processing tasks.

2.3 Model Architecture

We propose two models in our work. The models are shown in figure 2a and 2b respectively. We divide each video into N non-overlapping partitions. From each of the N partitions, both audio and visual features are extracted (figure 1) and used as inputs to the models. Here, only the inter-partition variations are learned as temporal patterns, while the intra-partition variations are ignored. We do so, to handle redundancy in consecutive frames especially in high fps videos. As we can see in figures 3 and 4, the audio and visual features from each block are passed through consecutive layers of neural network. Now, in our first model, the temporal patterns across the N sequential partitions are learned using a 3D convolution module. While in the second model, we use an LSTM to learn the temporal patterns across the partitions. The kernel sizes and stride information are available in the figure 2. By empirical analysis, we fixed N as 6.

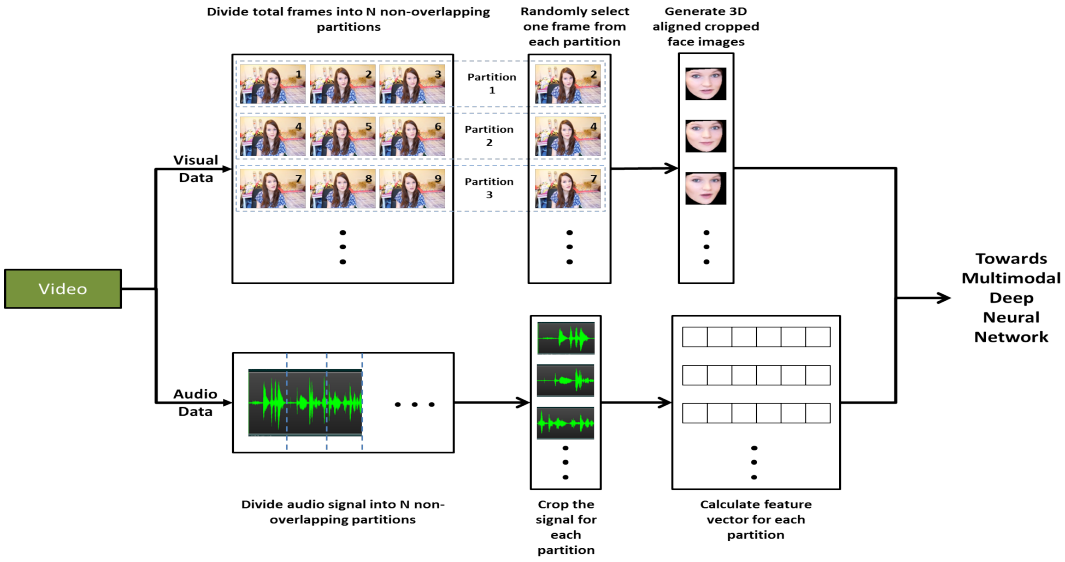
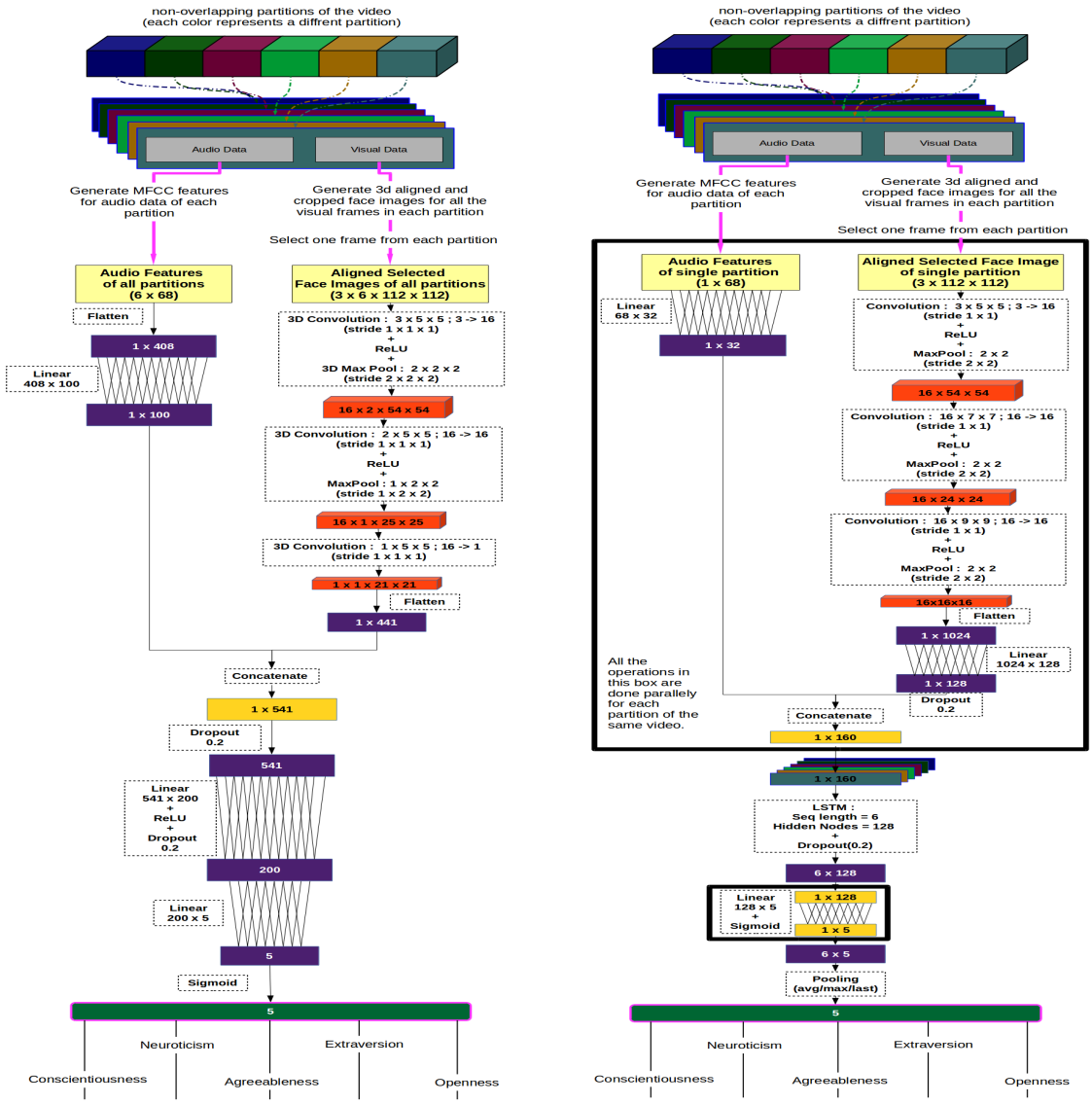


Fig. 1: Data pre-processing pipeline, where the face aligned images are extracted from image frames and spectral audio features are extracted from audio data.

Volumetric (3D) convolution model: Our first model is inspired from the work of Ji et. al [10]. The architecture is shown in figure 2a and the pipeline is demonstrated in figure 3. The visual data processing branch learns the change in facial expressions from face aligned images using 3D convolution. At first, the 6 face aligned temporally ordered images of size $3 \times 112 \times 112$ are passed through a 3D convolution layer, followed by a ReLU and a 3D max-pooling layer. The 3D convolution as well as max-pooling are done in a volume comprised of X, Y and t dimensions. The resulting feature maps are in-turn passed through a second set of similar layers of 3D convolution, ReLU and 3D max-pooling but with different kernel sizes (refer to figure 2a for details about parameters). This is followed by another layer of 3D convolution, which result in a single feature map of size $1 \times 21 \times 21$ which is flattened to a 441 dimensional feature vector. Simultaneously, the audio-data processing branch gets a 6×68 dimensional feature vector which is reduced to a 100 dimensional vector using a fully connected layer. The feature vectors from audio and visual branches are concatenated and yields a 541 (100 from audio + 441 from visual data) dimensional feature vector, which is then input to a fully connected (FC) layer of 200 nodes and a ReLU layer, followed by another FC layer of 5 nodes which has the activation function as sigmoid. These 5 nodes represent the predicted values of the Big-Five Personality traits.

LSTM based model: We designed our second model to learn the task based on temporal relationship within the input. The architecture and pipeline of the model are shown in figure 2b and figure 4 respectively. We propose LSTM units to capture the temporal patterns of the input data to predict the personality traits. Each aligned face image is passed through a series of spatial convolution, ReLU and spatial max-pooling layers of varying kernel sizes (refer to figure 2b for details about parameters). The generated feature maps are flattened to get



(a) Bi-modal Volumetric Convolutional Neural Network architecture

(b) Bi-modal LSTM Neural Network architecture

Fig. 2: Model Architecture Diagram

1024 dimensional feature vector and it is connected to a fully connected layer of 128 nodes. Simultaneously, the audio-data (6 feature vectors of 68 dimension) is passed through a 32-node fully connected layer and reduced to 32-dimension. After these steps, the output feature vectors from audio and visual data processing branches are concatenated to yield 6 feature vectors of 160 dimension (32 dim of audio + 128 dim of visual data for each 6 partition) which are still maintained in temporal order. The extracted temporally ordered 6 feature vectors are then passed through an LSTM with output dimension of 128. The LSTM takes 6×160 dimensional input and outputs a sequence of 6 128-dimensional feature vector. The LSTM generates output for each time step and then, each output is passed through 5 dimensional fully-connected layer with sigmoid activation function. Thus, we get 6 outputs of predicted 5 personality traits. For each personality

trait, we average the predicted value, output by all 6 LSTM output units. Thus we get a single prediction value for each of the Big Five personality traits.

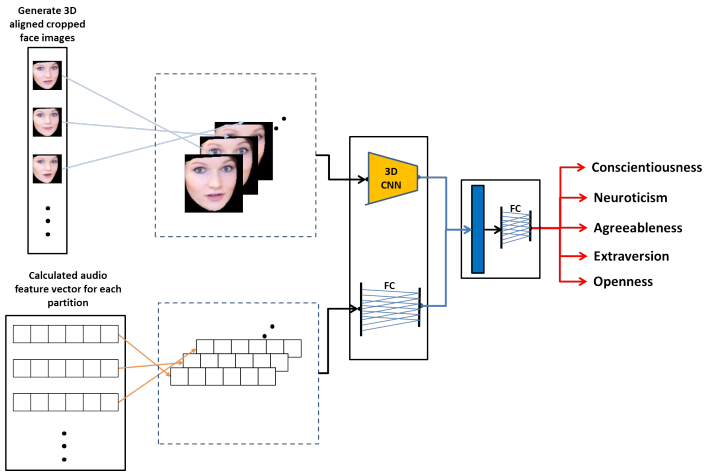


Fig. 3: Pipeline of 3D-Convolution model

3 Stochastic Training and Testing

According to Psychology research [2], it is observed that first impressions of unfamiliar faces can be formed even with exposure times as small as 100-ms. Their results suggest that predictions made with a 100-ms exposure correlated highly with judgments made in the absence of time constraints, suggesting that small exposure times were sufficient for participants to form an impression. On similar lines, we also hypothesize that deep models can learn effective representations for recognizing first impressions from a few randomly selected frames.

3.1 Stochastic Training

Training of the two proposed models is carried out using Stochastic Gradient Optimization (SGD) method. The parameters used for SGD are: learning rate = 0.05, weight decay = $5 \times e^{-4}$, momentum = 0.9, batch size = 128, learning rate decay = $1 \times e^{-4}$.

As mentioned earlier (figure 1), each raw video file is split into non-overlapping 6 partitions and the audio as well as visual features are extracted from each partition individually. We propose to train the models by using a combined feature set such that we take single face aligned image from each partition, as well as the pre-processed audio features from each partition. Particularly, in video data, since we are only using 1 frame from whole partition, there are multiple combinations of frames from each partition possible for training. Consider there are N partitions & F frames per partition and we intend to take a single frame from each partition, hence F^N combinations of frames are possible per video. We

assume N as 6 and typically, F is in the range of ~75 (considering 30 fps and each video of 15 seconds). Training the model with 75^6 combinations of frames is an overkill. Empirically, we found that training only on several hundreds of combinations (typically ~ 500) is enough for the model to generalize for whole dataset.

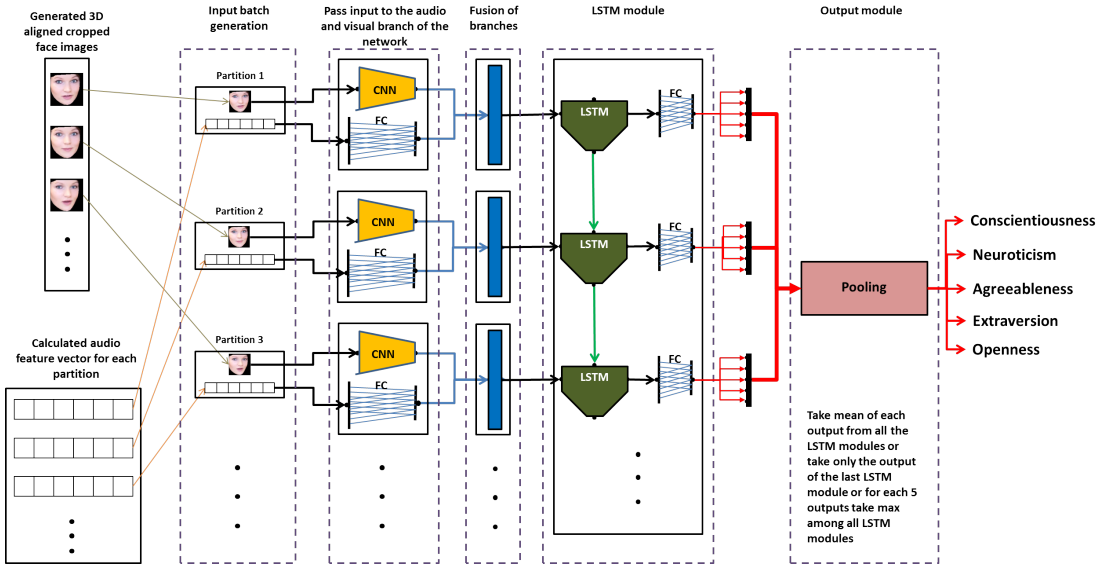


Fig. 4: Pipeline of LSTM model

Going with the above explanation, the 6 input frames (single frame from each partition) for model training is selected randomly by keeping temporal ordering in mind. At every epoch, the random selection will yield new input combination for each video. This stochastic way of training produces new sample at every epoch and "regularizes" the learning effectively, thus increasing the generalization of the model.

3.2 Testing

Testing the model also faces the same issue of exponential combination of frames per video. Empirically, we choose to use only a random subset (10 combinations) from total possible combinations and use the average of 10 evaluations as the Personality-traits recognition results. The validation and test results suggest that the model and evaluation method performs significantly better than the other submissions and the LSTM model stood at second place in the Final evaluation phase of competition.

4 Experiments and Results

In this section, we first briefly describe about the dataset and the evaluation protocol from our experiments. Then we provide the implementation details for our method and discuss the results.

4.1 Dataset: Apparent Personality Analysis (APA) - First impressions

In our validation experiment, we use the ChaLearn LAP 2016 APA dataset provided by the challenge organizers [11]. This dataset has 6000 videos for training with ground truth Personality-traits, 2000 videos for validation without ground truth (performance is revealed on submission of predictions) and 2000 videos for test (Ground truth is not available until the competition is finished). Each video is of length 15 seconds and generally has 30 frames/second. The ground truth consists of fractional scores in the range between 0 to 1 for each of Big-Five Personality traits : Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness.

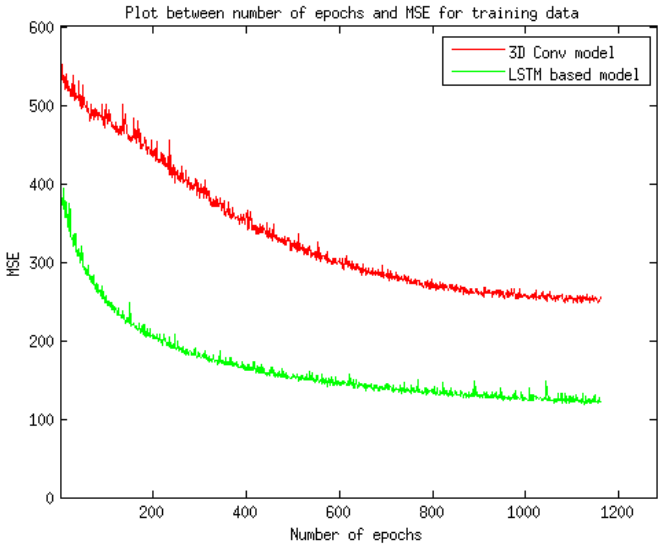


Fig. 5: Number of Epochs vs. Mean Squared Error (MSE) for individual models during training phase

4.2 Evaluation Protocol

The evaluation is done in terms of Mean Average Accuracy.
The individual personality traits Average Accuracy is calculated as,

$$\text{Average Accuracy}_j = \frac{1}{N}\sum_{i=1}^{N}(1 - |Target_{ij} - y_{ij}|) \tag{1}$$

where j = 1...5, N is the number of total videos, $Target_{ij}$ is the ground truth value for i^{th} video and j^{th} personality-trait, y_{ij} is the predicted value for i^{th} video and j^{th} personality-trait.

The Mean Average Accuracy between the predictions and the ground truth personality-traits values:

$$\text{Mean Average Accuracy} = \frac{1}{m}\sum_{j=1}^{m}(\text{Average accuracy}_j) \tag{2}$$

where m = 5 (the number of Personality Traits).

Note, that the maximum value of the Mean Average Accuracy, as well as Average Accuracy is equal to 1, which represents the best result and the minimum is equal to 0 representing the worst match.

4.3 Implementation details

Both of the deep learning models are implemented using Torch [17] scientific computing framework. The training of 3D convolution based model takes 30 seconds per epoch and LSTM based model takes 3 minutes per epoch on a GeForce GTX Titan Black graphics card. The training of each individual model is done for up-to whole 1 day. We used only the ChaLearn LAP 2016 APA dataset [11] for training. The comparison of mean squared error(MSE) of both models during training is shown in figure 5. The source code files of both the training and final proposed prediction method are available in github [1] repository.

4.4 Development phase

In the development phase of the APA2016 competition [11], only the training set ground truths were released and the methods were evaluated online by submitting the predictions on the validation videos to a server. The best performance of our models during development phase is shown in Table 3.

4.5 Test phase

In the test phase of the APA2016 competition [11], the testing videos were released. The testing ground truths were kept secret and the teams were invited to submit their results on the testing videos. The organizers announced the final ranking after the test phase. The results are summarized in Table 4. The proposed LSTM model secured the second place in the leader-board and shown in bold font.

4.6 Results and Discussion

The performance of CNN (3D convolution) based model and LSTM model can be seen from learning phase evaluation shown in table 3:

[1] refer https://github.com/InnovArul/first-impressions for more information

	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

Table 3: Evaluation during learning phase on ChaLearn LAP 2016 APA : First Impressions challenge

The test phase leader-board standings is shown in the table 4.

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

Table 4: Leaderboard of Test-phase on ChaLearn LAP 2016 APA : First Impressions challenge. our entry is with **bold**

As we noticed from the table 3, during learning phase, LSTM based model performs superior to 3D convolution based model. It maybe due to the fact that, LSTM is able to learn better temporal relationships than 3D convolution based approach. Also, the audio-features were not used to define temporal relationship in 3D convolution based model (only 3D face aligned images are used), but LSTM model used both audio and visual features to learn the temporal correspondences, which could have made it perform better. Because of these reasons, we chose LSTM model to be used for test phase: Our method secured second place in ChaLearn LAP 2016: APA challenge [11].

5 Conclusions and Future Works

In this work, we proposed two deep neural network based models that use audio and visual features for the task of First Impressions Recognition. These networks mine the temporal patterns that exist in a sequence of frames. It was also shown that such sequences can be small and selected in a stochastic manner respecting the temporal order. The proposed methods have been shown to yield excellent performance on the ChaLearn LAP APA2016 Challenge [11]. As deep neural networks are known for their representation and feature extracting ability, they

can be used to learn the optimal representations without having to pre-process the data. Appearance and Pose features can also be explored to see if they improve the performance given by the proposed audio and visual features.

References

1. Wikipedia. `https://en.wikipedia.org/wiki/First_impression_(psychology)` Definition of psychological term First impression.
2. Willis, J., Todorov, A.: First impressions making up your mind after a 100-ms exposure to a face. Psychological science **17**(7) (2006) 592–598
3. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal processing magazine **18**(1) (2001) 32–80
4. Cohen, I., Garg, A., Huang, T.S., et al.: Emotion recognition from facial expressions using multilevel hmm. In: Neural information processing systems. Volume 2., Citeseer (2000)
5. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and image understanding **91**(1) (2003) 160–187
6. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE (2013) 3687–3691
7. L.Kessous, G.Castellano, G.: Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces **3**(1) (2010) 33–48
8. Zhao, X., Kim, T.K., Luo, W.: Unified face analysis by iterative multi-output random forests. (2014)
9. Razuri, J.G., Sundgren, D., Rahmani, R., Moran Cardenas, A.: Automatic emotion recognition through facial expression analysis in merged images based on an artificial neural network. (2013)
10. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1) (2013) 221–231
11. Lopez, V.P., Chen, B., Places, A., Oliu, M., Corneanu, C., Baro, X., Escalante, H.J., Guyon, I., Escalera, S.: Chalearn lap 2016: First round challenge on first impressions - dataset and results. chalearn looking at people workshop on apparent personality analysis. In: ECCV Workshop proceedings. (2016)
12. Giannakopoulos, T.: pyaudioanalysis: An open-source python library for audio signal analysis. PloS one **10**(12) (2015) e0144610
13. Giannakopoulos, T.: pyaudioanalysis. `https://github.com/tyiannak/pyAudioAnalysis` an open Python library that provides a wide range of audio-related functionalities.
14. Giannakopoulos, T.: pyaudioanalysis. `https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction` Features extracted using pyAudioAnalysis.
15. Baltru, T., Robinson, P., Morency, L.P., et al.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2016) 1–10

16. Baltru, T., Robinson, P., Morency, L.P., et al.: Openface. `https://github.com/TadasBaltrusaitis/OpenFace` a state-of-the art open source tool intended for facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation.
17. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A matlab-like environment for machine learning. In: BigLearn, NIPS Workshop. Number EPFL-CONF-192376 (2011)