IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

An Intelligent Architecture for Recruitment Process Using Machine Learning

Jiso K. Joy[1], Sreedev S.B.[2], Vishnu A.K.[3] , Rejimoan R[4*]

[1,2,3,4]Dept. of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, APJ Abdul Kalam Technological University, Trivandrum, India

[*]*Corresponding Author: rejimoan@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i8.1115 | Available online at: www.ijcseonline.org

Accepted: 10/Aug/2019, Published: 31/Aug/2019

***Abstract*—**Recruitment process has become one of the laying foundations for the development of an organization. All organizations are looking for the perfect candidate to build their enterprises. Finding the right candidate for the right job is becoming more and more difficult. Recruiter and other HR professionals that don't use innovative recruiting strategies are often unable to find job candidates that are suitable for the job. To find the right candidates, recruiters have to have a well-planned and developed recruiting and hiring strategies. Machine learning is emerging as a strategy to help employers more efficiently conduct talent sourcing and recruitment. Traditional recruiting process requires lot of time and effort along with various costs that comes with it for filtering out the candidate. This paper will propose an automated interview system which uses machine learning to gauge the candidates based on the emotions expressed in the interview process and thus find the right person for the right job.

***Keywords*—**Machine Learning, Neural Network , Recruitment, Emotion, Speech

I. INTRODUCTION

The future of the company relies on the shoulder of the employees, the company's future depends on the work done by various employees. We need the right person for the job. Most of the time this a major problem faced by the company they have to do a lot of filtering to find the suitable person which requires valuable time. So, recruitment holds a crucial role in the development of company all around the world. A candidate-interviewer interaction is susceptible to many categories of judgment and subjectivity. Such subjectivity makes it hard to determine whether candidate's personality is a good fit for the job. Identifying what a candidate is trying to say is out of our hands because of the multiple layers of language interpretation, cognitive biases, and context that lie in between. AI can measure candidate's facial expressions to capture their moods and further assess their personality traits. An automated interview system will reduce time taken by replacing tasks like resume reviews and phone screens with an on-demand video interview with this recruiters and managers can evaluate more candidates on their own time. Machine learning has been used to analyze the resume and papers have been presented to analyze the video input to filter out candidates but little work has been done on combining the audio and video input. In this paper we propose a method to gauge the skills of a candidate by analyzing the video and audio of an online interview of the candidate.

The remaining sections of this paper are organized as follows: Section II contain the related work done on this field, Section III contain proposed architecture of the model in details, Section IV contain the implementation of the model containing four subsections describing different phases of the model, Section V describes results obtained from various research cases in details and Section VI concludes research work with future directions.

II. RELATED WORK

Humans detect and interpret faces and facial expressions in a scene with little or no effort. Still, development of an automated system that accomplishes this task is rather difficult. There are several related problems: detection of an image segment as a face, extraction of the facial expression information, and classification of the expression.

Zhiding Yu et al. proposed a deep CNN model by combining the face detection and classification modules with a collection of multiple deep CNN's. Each CNN model was pretrained on a larger dataset provided by the Facial Expression [1].

Ali et al. proposed a deep neural network architecture that consists of two convolution layers each followed by max pooling. The network takes recorded face image as the input and classifies it into one of the six basic emotions. This neural network was trained on facial expression databases, viz. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013 [2].

From the above researches we observe that a neural network with a combination of multiple convolutional layers followed by some other layers is a best architecture for face emotion recognition. Furthermore, that FER2013 dataset is best available dataset for pre-training of an emotion recognition system from kaggle.

Most researchers believe that prosody continuous features such as pitch and energy convey much of the emotional content of an utterance. According to the studies performed by Williams and Stevens, the arousal state of the speaker (high activation versus low activation) affects the overall energy, energy distribution across the frequency spectrum and the frequency and duration of pauses of speech signal. However, there are contradictory reports on the effect of emotions on prosodic features. For example, while Murray and Arnott indicate that a high speaking rate is associated with the emotion of anger, Oster and Risberg have an opposite conclusion.

Coming to cepstral features, Cepstral-based features can be derived from the corresponding linear features as in the case of linear predictor cepstral coefficients (LPCC) and cepstral-based OSALPC. In 'A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process', it was shown that features based on cepstral analysis such as LPCC, OSALPCC, and Mel-frequency cepstral coefficients(MFCC) clearly outperform the performance of the linear-based features of LPC and OSALPC, in detecting emotions in speech signal [3].

MFCC extraction is used to find the features in the audio file. These features can then be analyzed to detect the emotion in the tone. To properly study the MFCC work done by A. Charisma and M. R. Hidayat was referred in which feature was extracted from speech signal using Mel Frequency Cepstral Coefficients (MFCC) which was further used in the speaker verification system [4].

III. ARCHITECTURE

The visual and audio module is separated from the video. The video module is passed through Haar classifier for identifying face which is then fed to the Convolution Neural Network. The features are extracted from Audio module using Mel-frequency cepstral coefficients which is later fed into Neural Network. The corresponding audio and video analysis are done and the final result is aggregated to rank the candidates.

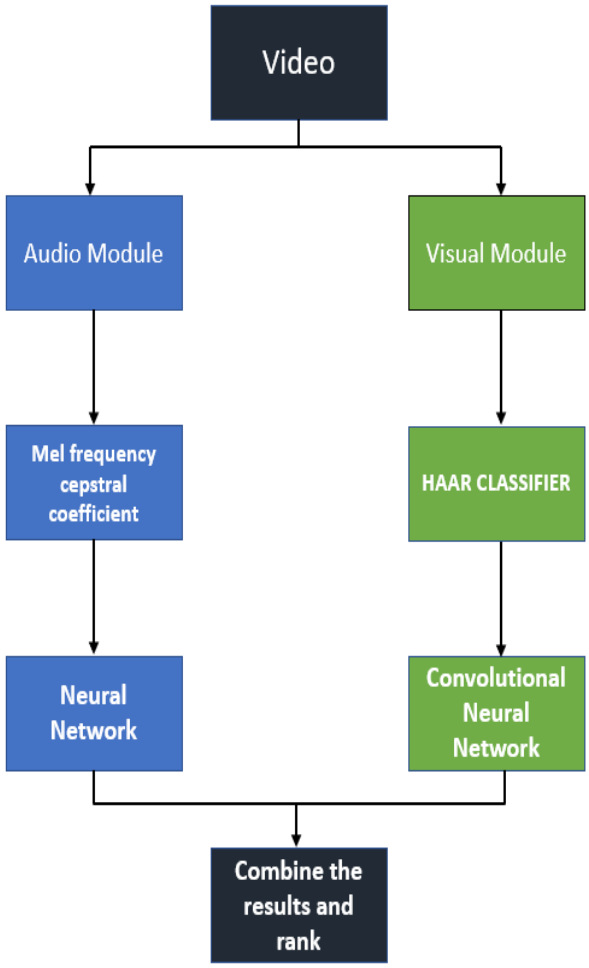

Figure 1. Data Flow Diagram

The Haar classifier identifies the face in the video and a corresponding set of frames are taken to be fed into the pre-trained Convolution Neural Network to identify the seven basic facial emotions. The Convolution Neural Network is trained using Fer2013 dataset. The features from audio module is extracted using Mel-frequency cepstral coefficients [4] which is fed into a pre-trained Neural Network to identify seven basic emotions in voice and is trained with RAVDESS dataset.

IV. IMPLEMENTATION

IV.I Facial Emotion Recognition

The Haar classifier identifies the face in the video and a corresponding 100 frames are taken to be fed into the pre-trained convolutional neural network consisting of 8

convolutional layers, 4 max pooling layers with a modified version of sigmoid activation function in dense layer and the activation function for the convolutional layers being Rectified Linear Unit [5]. The Convolutional neural network is trained using fer2013 dataset. The structure of the proposed network is given in Table 1.

Table 1. Various layers of the neural network used in the visual module

Layer (type)	Output Shape	Parameters
Convolution Layer1	(None, 32, 48, 48)	320
Convolution Layer2	(None, 32, 48, 48)	9248
Max Pooling Layer1	(None, 32, 24, 24)	0
Convolution Layer3	(None, 64, 24, 24)	18496
Convolution Layer4	(None, 64, 24, 24)	36928
Max Pooling Layer2	(None, 64, 12, 12)	0
Convolution Layer5	(None, 96, 12, 12)	55392
Convolution Layer6	(None, 96, 10, 10)	83040
Max Pooling Layer3	(None, 96, 5, 5)	0
Convolution Layer7	(None, 128, 5, 5)	110720
Convolution Layer8	(None, 128, 3, 3)	147584
Max Pooling Layer4	(None, 128, 1, 1)	0
Flatten Layer	(None, 128)	0
Dense Layer	(None, 64)	8256
Dropout Layer1	(None, 64)	0
Dense Layer	(None, 7)	455

The first Convolutional layer takes 48x48 pixel image as the input because the dataset we trained the CNN on is in that form. We transform the input image into that form and feed it to the CNN. The first layer uses the padding to prevent the output from reducing in size. As we can see from the Table I the output of the first layer is 48x48, this is because of the padding. A sliding window of 3x3 is used in the convolutional layer to detect the minute features in the image. Small windows are used to detect complex features [6].

In the last fully connected layer 128 neurons are used. The output of the last layer is 7 neurons which give the probability of 7 emotions in the image.

IV.II Voice Emotion Recognition

The features from audio module is extracted using Mel frequency cepstral coefficients which is fed into a pre-trained neural network consisting of two convolutional layers. The activation function for the convolutional layers is Rectified Linear Unit with the last activation layer having a Softmax function. The optimizer used for the model is RMSprop [7].

The neural network is trained using RAVDESS dataset. The structure of the proposed network is given in Table 2.

Table 2: Various layers of the neural network used in audio module

Layer (type)	Output Shape	Parameters
Convolution Layer1	(None, 40, 128)	768
Activation Layer1	(None, 40, 128)	0
Dropout Layer1	(None, 40, 128)	0
MaxPooling Layer	(None, 5, 128)	0
Convolution Layer2	(None, 5, 128)	82048
Activation Layer2	(None, 5, 128)	0
Dropout Layer2	(None, 5, 128)	0
Flatten Layer	(None, 640)	0
Dense	(None, 8)	5128
Activation Layer3	(None, 8)	0

IV.III Combining Both Results

The results which are the probability values of seven emotions from video and voice respectively are combined based on score level fusion method.

IV.IV Normalization

Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1. Using the (1) the given data was normalized

$$v1 = \frac{v - min}{max - min} \quad (1)$$

where v_1 and v is the new and old value respectively while max and min are the maximum and minimum values respectively.

IV.V Score Level Fusion Method

The simple weighted fusion was used as a score fusion strategy in this paper, and the fusion scores (S) is calculated as in (2).

$$S = w_1 S_1 + w_2 S_2 \quad (2)$$

where S_1 and S_2 a[illegible] score and voice emotion recognition score, w_1 and w_2 are their weights, S is the fusion score. Here S_1 is the probability vector of the seven emotions from the face and S_2 is the probability vector of the seven emotions from the voice. The weights w_1 and w_2 are varied over the range [0, 1], such that the constraint $w_1 + w_2 = 1$ is satisfied.

V. RESULT

The convolution neural network for detecting emotions from face was able to achieve an accuracy of 87.6% while the

neural network for detecting the emotions from voice was able to achieve an accuracy of 91.2\%.

An interview consists of N questions. The video module for each question is separately recorded. After recording, the facial and voice emotion for a single question is predicted. The predicted value for both face and voice were normalized based on (1).

Table 3. Different weights used and observed accuracy

Weighted sum (Face,Voice)	Accuracy
(0.9,0.1)	78.9
(0.8,0.2)	73.52
(0.7,0.3)	82.6
(0.6,0.4)	81.23
(0.5,0.5)	80.25
(0.4,0.6)	72.2
(0.3,0.7)	71.56
(0.2,0.8)	66.23
(0.1,0.9)	62.3

The two features are then combined based on (2). The weights for score level fusion is experimentally calculated based on several test cases. There are seven emotions for each emotion five videos were used as test cases. Therefore, there was total 35 test cases. Table 3 shows the accuracy for each value of weights taken for these 35 test cases. From Table 3 it is clear that weight value 0.7 and 0.3 for face and voice respectively gave best accuracy. [8]

The seven emotions for a single question are classified into positive affect (Happy, Neutral, Surprise) and negative effect (Fear, Angry, Disgust, Sad). The positive effect is the set of emotions which produce positive emotional experience while negative effect is the set of emotions which produce negative emotional experience in humans.

Table 4. Different emotions and corresponding weights assigned

Emotion	Weights
Happy	0.8
Surprise	0.15
Neutral	0.05
Sad	-0.1
Disgust	-0.2
Fear	-0.3
Angry	-0.4

The Table 4 shows different emotions and their weights. Positive emotions are emotions that we typically find pleasurable to experience. It can be described as pleasant or desirable situational responses, distinct from pleasurable sensation and are pleasant responses to our environment (or our own internal dialogue) that are more complex and targeted than simple sensations. Negative emotions can be defined as an unpleasant or unhappy emotion which is evoked in individuals to express a negative effect towards an event or person. [9]

Positive emotions have been shown to impact the brain in the following ways:

- They can increase our performance on a cognitive task by lifting our spirits without distracting us like negative emotions do.
- Positive emotions can trigger the reward pathways in the brain, contributing to lower levels of a stress hormone and greater well-being.
- Positive emotions may help us broaden our horizons and widen our brain's scope of focus.

V.I Candidate Score

Different positive weight values are assigned for emotions in positive affect like w1, w2, w3 and different negative weight values are assigned for emotions in negative affect like -w4, -w5, -w6, -w7. The values chosen should be in such a way that the sum of positive weights should be 1 and the sum of negative weights should be -1.

The value of each emotion is multiplied with the corresponding weights. Let p and n be the weighted sum of positive and negative affect respectively. The sum of p and n for each candidate is recorded. Based on this candidate score we can choose the best candidate for the job.

VI. CONCLUSION AND FUTURE SCOPE

By using the proposed method, we are able to come up with a solution to reduce the effort and time put into initial interview in any recruitment process. Here we are analysing the seven different emotions from visual and audio extract and using which we are able to reach an aggregate result. The emotions are analysed based on the questions issued by the interviewer. By classifying the emotions into group of positive and negative we are able to rank and from which induce how suitable the person is for the job thus initiating an automated approach for recruitment. Since the candidates are classified based solely on their emotions there is always a chance that candidate could fake an emotion. But since we are also taking the emotions from the voice it will be much difficult for the candidate to fake the emotions from the voice. The proposed model can be modified to analyse the presence of depression in a person, analysing the emotions of the driver to check weather he/she is able to drive in the emotional state they are

in. It can also be used to adjust the game based on the gamer's emotions or in case of E-learning adjust the subject difficulty to suite the learner.

REFERENCES

[1] Z. Yu and C. Zhang, "*Image based Static Facial Expression Recognition with MultipleDeep Network Learning*",Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15, **2015**.

[2] A. Mollahosseini, D. Chan and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks" ,*IEEE Winter Conference on Applications of Computer Vision (WACV)*, **Lake Placid,** NY, pp. **1-10**, **2016.**

[3] S. E. Bou-Ghazale and J. H. L. Hansen, *"A comparative study of traditional and newly proposed features for recognition of speech under stress*", IEEE Transactionson Speech and Audio Processing, vol. **8**, no. **4**, pp. **429-442**,**2015**

[4] A. Charisma, M. R. Hidayat and Y. B. Zainal, "*Speaker recognition using mel-frequency cepstrum coefficients and sum square error"*,3rd International Conferenceon Wireless and Telematics (ICWT), Palembang, pp. **160-163**, **2017**

[5] M. Pantie and L. Rothkrantz, "Automatic analysis of facial expressions: the state of the art", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. **22**, no. **12**, pp. **1424-1445**, **2000**.

[6] S. Albawi, T. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network", *2017 International Conference on Engineering and Technology (ICET)*, **2017**.

[7] P. Dasgupta, "Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing", *International Journal of Computer Trends and Technology*, vol. **52**, no. **1**, pp. **1-3**, **2017.**

[8] A. Mardin, T. Anwar, B. Anwer, "*Image Compression: Combination of Discrete Transformation and Matrix Reduction*", International Journal of Computer Sciences and Engineering, Vol.**5**, Issue.**1**, pp.**1-6**, **2017.**

[9] Malathi Sriram, L. Gandhi, "*Exploring the dynamica virtus of Machine Learning (ML) in Human Resource Management - A Critical Analysis of IT industry*", International Journal of Computer Sciences and Engineering, Vol.**5**, Issue.**12**, pp.**173-180**, **2017**.

Authors Profile

Mr. Rejimoan R pursed Bachelor of Technology in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala in 2019. His main areas of interest includes image processing, data mining, competitive programming.

Mr. Jiso K Joy pursued Bachelor of Technology in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala in 2019. His main areas of interest includes image processing, data mining, competitive programming.

Mr. Sreedev S B pursued Bachelor of Technology in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala in 2019. His main areas of interest includes machine learning, machine learning, web development.

Mr. Vishnu A K pursued Bachelor of Technology in Computer Science and Engineering from APJ Abdul Kalam Technological University, Kerala in 2019. His main areas of interest includes machine learning, emotional analysis, audio analysis.

