IJICT

International Journal of Intelligent Computing and Technology

ISSN: 2457-0249 (Online)

Website: https://www.ijict.com Jan (2022) - Vol.5, Iss.2 Pages:01-14

DEVELOPING A HIRENET WITH CONVOLUTIONAL NEURAL NETWORK FOR PREDICTION OF INTERVIEWED CANDIDATE USING ASYNCHRONOUS VIDEO

A Paulin[1], V Sujitha Angel[2], Dr P Revathi[3]

[1,2]Department of BCA, [3]Assistant Professor, P.G Department of Computer Science,
Holy Cross College (Autonomous), Tiruchirappalli – 620 002, Tamilnadu, India

Article History- Received: June 2021; Published: Jan 2022

Abstract

Many researchers from academic and industry shows interest on analyzing the interviews of right candidate for the particular job in an automatic manner. A vast number of employees can answer to the questions of HR in a grouped way by using the special context of asynchronous video interviews, where deep learning (DL) models uses these monologue videos for prediction process. On the other hand, evolving approaches face some other hurdles, including integrating information from multiple methods and interpreting predictions. In the asynchronous video interviews, the candidate prediction process is explored by using data from live interviews with the help of three models such as facial expression, verbal content and pose of the participant candidate.Furthermore, the architecture we propose is focus-based HireNet model, providing an explanation of the questions, moments, and methods that contribute most to the network's output. In order to visulaize the moments, the other DL models uses the focus centers with specific values and only the particular methods are allowed to predict the interpretation of candidates via comprehensive analysis of social cues' characteristics in these moments.

Keywords: *Job Interviews; Automated Analysis; Video; Questions; Deep Learning; HireNet; Facial Expression; Candidate*

1. INTRODUCTION

Recruiters continue to priorities this. Initially, MONG personal selection procedures, and job interviews were conducted face to face or over the phone, and now job interviews are mostly conducted viavideo recordings or online conferencing systems. Asynchronous video interviews are a growing tool now offered by companies to meet initial assessment requirements for individual testing. The procedure is as follows: Applicants connect to the site using a webcam, smartphone or tablet, record video, and the examiner answers the pre-set questions. In the second stage, candidates will join on the same platform, look at the candidate's answers, evaluate the answers and then decide whether to invite the candidate for a face-to-face interview. These sites speed up application processing time for recruits and provide access to a diverse group of applicants [1,2]. Also, applicants are free to complete the interview at any time.

Recently, researchers have begun to explore methods for skill assignment based on verbal and nonverbal references in self-analysis of social skills and asynchronous video interview systems [3-5]. Recent advances in social signal processing (SSP) have made this investigation easier to obtain a data set (same interview questions, same answer) due to the interview's asymmetry, time and process' structure nature). In many ways, a computer useful in this environment has already proven useful. In order to develop the candidate's social skill, virtual recruiters are implemented and prepare the candidate for face-to-face interviews [6-8]. Therefore, the contribution of the automation process will be useful not only for training applicants but also for analysts and in future, it will help the recruiters for evaluation and selecting the right candidate for the jobs. Therefore, various techniques have been proposed and useful to gain insight into the impact of recruitment outcomes [9-11] social symptoms. This article focuses on recruiting job seekers for direct interviews, but keeping the name "hierarchical" in line with previous studies.

However, little attention has been paid to the study of multimodal indicators. The early and late combinations are presented in the existing multimodal fusion techniques [12]. However, the co-existence of verbal and nonverbal behaviour are not considered by this existing techniques. Since the communication between human-beings is multimodal, multimodal dynamics can have a profound impact on the job interview, and we must investigate and analyze these indicators. Several methods have been suggested to improve performance.

These methods, which mainly focus on decision-making or integration at the articulation level, are now integrated at the word level, showing a lower-level effective unit of instrumental representation [13] this subtle combination is important because this local interaction between styles can help resolve ambiguity or emphasize certain moments. However, these methods are based on the use of silent linguistic transcription, where they are organized and evaluated. Achieving this manual transcription is not always possible, and a practical system will rely on the use of automatic speech recognition (ASR), which can reduce the performance of word-association methods.

Another missing feature for specific multimedia systems is the lack of transparency in the way the integration is conducted and the knowledge gained from these trained models. In addition to the HR context and new legislative regulations (GDPR), interpretation and public understanding are important. One possible way to add a description is to learn about the important moments in a job interview and the ways you contribute to those moments.

2. INTERVIEW THE CANDIDATE FOR JOBS

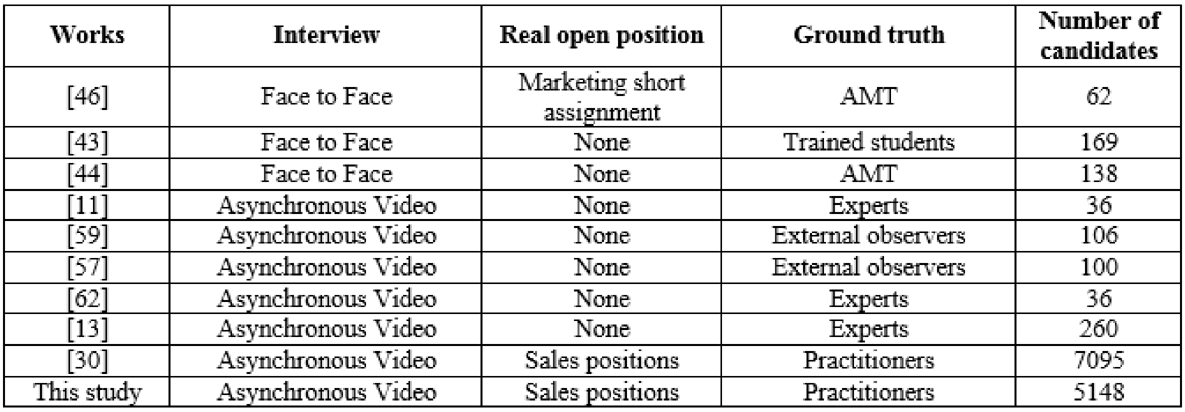
2.1 Effect of different cues of verbal and non-verbal speech

Visual and aesthetic indicators of job seekers are the first records of recruiters [14]. Recruiters seek to infer various characteristics related to job situation from verbal content and non-verbal references expressed by staff [15]. In this regard, the researchers sought to highlight the relationship between nonverbal behaviors and constructs in job interviews. For example, candidates' personalities [16], attempts at fraud [17], anxiety [18] or, more broadly, performance in job interviews [19] were examined. The influence of verbal references on nonverbal references is minimal in the literature [20] and [21]. Finally, given the diversity of human communication, we need to examine the multimedia symptoms that occur in job interviews. DeGrootetGooty [16] explores the channel's own contributions to subjective understanding (audio-only and video-audio-video-only). Multimodal specific indicators have been used successfully in previous studies, such as speech vibration [10]. However, since the contextual significance or deception of a job interview is fertile ground for multimedia strategies such as [22] there is another way to identify influential multimedia indicators. This paper takes a step in this direction by designing a model that takes into account multimedia cues and highlights important multimedia indicators in the video interview.

2.2 Identify the correct candidate for jobs using databases for automatic analysis

Pre-tasks for automated job interview analysis differ in the group used to train and evaluate systems. This group differs in several points, namely: the type of interview (face-to-face or asynchronous), interview settings (actual or simulated location), the form of the designations (trained, expert or non-expert), and body size. Indeed, interview systems and the type of factors that influence employee behavior during the interview [23] Second, since the task is a complex label, it is important to know who labeled the data. Finally, the number of candidates affects credibility and style. Figure 1 contains a summary of job interviews used in previous jobs. Note that only the other two databases are stored on the original system [11].

Table 1 Summary of job interview databases. AMT stands for Amazon Mechanical Turk



Works	Interview	Real open position	Ground truth	Number of candidates
[46]	Face to Face	Marketing short assignment	AMT	62
[43]	Face to Face	None	Trained students	169
[44]	Face to Face	None	AMT	138
[11]	Asynchronous Video	None	Experts	36
[59]	Asynchronous Video	None	External observers	106
[57]	Asynchronous Video	None	External observers	100
[62]	Asynchronous Video	None	Experts	36
[13]	Asynchronous Video	None	Experts	260
[30]	Asynchronous Video	Sales positions	Practitioners	7095
This study	Asynchronous Video	Sales positions	Practitioners	5148

In addition to the current study database, we have previously compiled a comparable database of candidates [24]. We had to continue with this second package for two reasons: First, for legal and privacy reasons, many videos from the first database are no longer available due to expiration dates.

Second, the timestamp of the spoken words was not included in the previous database, which could not be combined with low-level methods. In summary, the database used in this study includes 5,148 actual asynchronous video interviews of candidates for sales positions, which are evaluated by practitioners.

2.3 Video job interviews based on Automatic Analysis

Recent developments have greatly reduced the time it takes to manually encode behavioral observations. Auto-encoders for audio [25] or visual feedback [26] are now available. In addition, advances in automated speech recognition enable researchers to obtain automated transcripts of a candidate's verbal content. Since asynchronous job interviews are videos, in order to create a rating model, the features (verbal, audio, and video content) must be distinguished from each other. The vocal notes are mainly related to prose features (eg, fundamental frequency, intensity, slope-frequency barrier coefficients), and speaking activity (eg pauses, silences, lowercase) [11]. Features derived from facial expressions (eg units of facial movement, head rotation and position, visual orientation) create highly isolated visual cues. To describe oral content, researchers use lexical statistics (eg, word count, number of unique words), dictionaries (language test word count), title type, word bags or inclusion of words or recent documents [11].

Once the features are extracted based on the law, the standard representation of the axes should be considered to train the classification or regression model. The most common approach is to simplify the time aspect by reducing the amount of time using statistical functions (eg mean, constant variance, etc.). Another well-known method is the representation of pockets of visual, audio or visual-acoustic words [11]. The idea is to treat each frame shot as a word from a specific dictionary. This dictionary or code book is created by assembly. Then each frame is associated with a specific code word. Finally, the interview can be described as a symbolic word map. Given the consistent representation of the job interview, a classification or regression model is applied. The most commonly used mechanisms are logistic gradient control (lasso or hillside), random forest and support vector machines. However, these models do not take into account the internal and intervening dynamics.

To fill the gap, we propose a CNN model with Hire networkfor the automatic analysis of asynchronous video job interviews inspired by state-of-the-art multimodal analysis of socio-emotional behaviour.

2.4 Automatic Analysis by different approaches

Multimedia is an open issue in advanced applications such as visual query answer, action recognition, video description generation or video summary [27]. Since human communication is multimodal, consideration of multimodal representation in video job interviews for automated analysis would increase the efficiency of such a system. Multimodal fusion is a robust system that enables efficient exploitation of filling systems and potential resolution of noise at each sample entry. In the field of video robotic job interviews, some attempts have been made to create a multimedia system. For example, early (set of features) and late (last set) were explored and the hire rating was good for the task. Interestingly, visual system integration reduces performance compared to audio-text integration [12]. Creating a multimedia representation through feature engineering involves another approach (eg staring while speaking) implementing features (eg bag of visual and audible words).

In addition to early and late integration, as well as automated job interview analysis, several models have been proposed to combine methods. Therefore, multimodal integration through neural networks has become of great importance in recent years, especially in the field of emotion recognition. Models based on memory cells [28], focus algorithms [29], adapter structures [31], modulated word embedding or group representation [32] show increased performance in the emotional prediction process. However, the design of these structures largely depends on the structure of the project being interpreted. For example, sentiment databases such as MOUD [33], MOSI [34] or IEMOCAP [35] are categorized by discrimination level (as opposed to the whole video naming), allowing for sentiment dynamics to be modeled. Contrary to popular belief, tasks tend to take at least one minute (at least a minute) to evaluate.

As a result, there is a special need to propose another model suitable for the computed structure. Also, unlike existing multimedia sentiment databases, verbal content is not copied manually, but retrieved automatically. While most algorithms combine methods at the word level, automatic speech recognition failures can be very harmful. In this article, we propose a neural network approach to synchronize multimedia information at regular intervals, considering the peculiarities of asynchronous job interviews when dealing with ASR scripts.

2.5 The importance of machine learning in job interview

Simultaneous translation is an important factor in building trust between users and machine learning systems, especially when applied to critical applications such as health, justice, or human resources. Opinions on transparency and interpretation are not yet clear, and in our particular case, the impact of social references on hiring decisions is still being studied [34]. To this end, we model our work by rating Yang and others, respectively. (2019) [37], the basic fact of the interpretation is unknown. The authors classify the description into two dimensions: i) Descriptive: Does the model explain the whole or does it explain each prediction of events separately and 2) Interpretation method: is it built into the model or is it given somehow after Hook Previous methods for automated analysis of job interviews are described models with an intuitive description in general. In contrast to these models, the use of neural networks comes at the expense of extreme transparency. Learning by imitation or the use of local classifiers. Among these methods, methods focused on improving performance and interpretation have recently become popular. The focus mechanisms inherent in the model allow us to visualize an event by visualizing the basic elements of the conclusion. In this way, we design our model using multiple focus mechanisms to identify the key moments of the interview, the key pair of questions and answers, and the main pattern during the multimodal fusion.

3. PROPOSED SYSTEM

In order to predict the legitimacy of employees based on multimedia records of interviews, questions, and job names, the new model is developed and it is briefly explained in this section.

3.1 Description of Dataset

3.1.1 Collection of Data and Pre-processing as Filtering

In cooperation with the HR Industrial Estate, 5148 joint French video transcripts of video conferencing were received. Videos are completely different from what can be created in the query team. Various technologies such as smartphone, tablets, webcam are used to record the videos of candidate. Often there are non-existent parking and low-quality equipment. Given these real conditions, full learning of the word generator will fail if one of the methods chooses the feature. One example is when face units (AUs) are difficult to detect in poor lighting. If some of the answers are removed, especially 1) if the independent speech recognition is less than 85%, 2) the open face is mistaken for 20% more face overall. Finally, candidates who do not have at least one wrestling answer can be removed. These autoclave job descriptions are similar to a specific type of job case, i.e. sales levels. The database contains more than 450 meters of different sales levels selected from the job titles of educators and HR professionals. Even if applicants agree to use their live accounts, videos will not be advertised outside the scope of this inquiry as they are subject to more personal restrictions.

3.1.2 Labelling of Data

When watching the video, the recruiter's opinions are used to provide the posters at interview level. The different types of comments such as criteria for pre-set, shortlist, "like" button, etc are provided for labelling the data.Applicants with positive feedback will be considered eligible and applicants with negative or neutral feedback will be considered eligible. The specific rating scale area is the inferior curve (AUC), which is the mean F1 score for the rental and non-lease category. It is suitable for binary classification and has been used in previous studies [36]. We divide the dataset into training packages, AUC-based ultra-labor selection evaluation and test package for final evaluation of each sample. These subgroups make up 70%, 15% and 15% of the total database, respectively.

3.2. Sequence of Input data and extracting the features

At regular intervals, effective representation of input data is identified by combining the methods, where frame representation is developed for every filter response methods. Here, the various frequencies for every system is controlled by the proposed model and Figure 2 shows the features sequence.

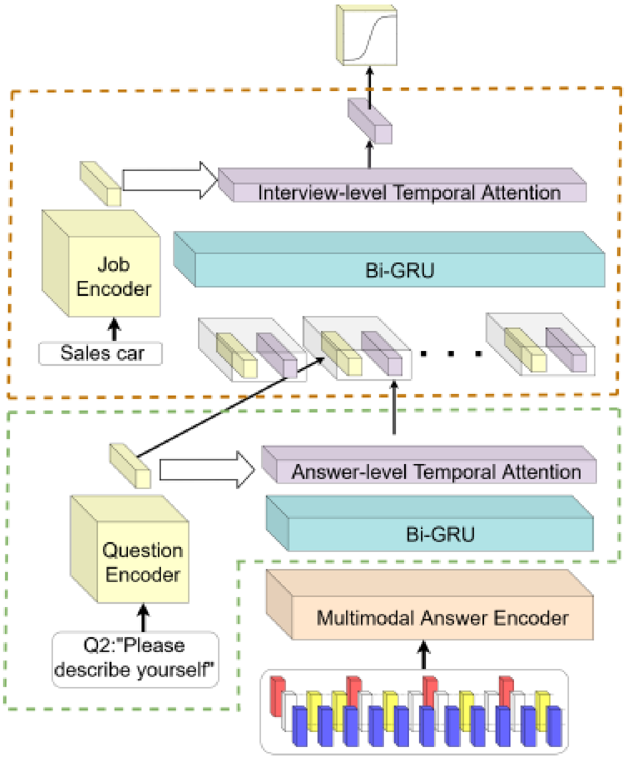

Interview-level Temporal Attention
Job Encoder
Bi-GRU
Sales car
Answer-level Temporal Attention
Bi-GRU
Question Encoder
Q2:"Please describe yourself"
Multimodal Answer Encoder

Fig 2 Proposed HireNet Architecture.

Appearance: We extract visual features at the frame level with Openface, an advanced visual behavior analysis software that gives multiple meaningful dimensions to each frame. For each frame, the difference between head's position and rotation is chosen, the intensity of the units of work, and the direction in which the velocity and scale appear in vector 52. We decided to soften the values because different videos have different frames. The time window is 0.3 eV.

General Notes:In order to extract the community computing, the eGeMAPS feature set is selected in this research study. Thanks to OpenSmile, we can extract 23 vectors with a frequency of hundred Hz. We loosen these values with a 0.1 second time frame.

Verbal content: To receive transcripts from the oral content, the study uses automatic speech recognition in interviews and timestamp each written word. Then the verbal content of the text is converted into a series of contextual embed vectors. Using the pre-trained French version of

RoBERTa "CamemBERT" published by huggingFace, we obtain 768 volumes per word in the oral version.

In order to publish the oral content, the order of language is changed. In other words, an array of placeholder is developed as language system order in this study, where the word's last moments is excluded, which is represented in Figure 2. Therefore, the placeholder is replaced with BERT representation for being spoken.

3.3 Classification using CNN

This is followed by combining and feeding the spark-based CNN classifier for facial emotion classification with the deep spatial information created by CNN model and dynamic characteristics attained by proposed Hire Network. This section defines the proposed CNN's layer definitions.

3.3.1. CNN classification

It consists of different sorts of layers, such as a convolution layer, a architecture layerlayer, pooling layers, and output levels with fully connected output layers.

3.3.2. Convolutional Layer

In CNN construction, the initial layer is always a Convolutional Layer. A CNN accepts M×N×1 as an input layer. A two-dimensional image with single layers has a two-dimensional size of M×N. This filter has the same depth as the input image and is convolved with the image. As a result, the input image is convolved with this curve or form, resulting in the final image. During convolution, the shape that closely approaches the curve in the input image and is signified by the filter ends up with higher values. Equation can be used to represent a convolution process (1).

$$s(t) = (x^{*}w)(t) \qquad (1)$$

3.3.3. Pooling layer:

To minimize the size of the data, a pooling layer is used. It includes dividing the matrix data into segments and replacing each segment with a single value.

3.3.4. Fully Connected layer

Dimensional changes are made in a fully linked layer in order to accommodate the network layer architecture. Each dimension of input and output of a completely connected layer is connected to each other.

3.3.5. Softmax layer

When the Softmax function is called, input from preceding layers is translated into a possibility for the classes that total to 1. As a result, this layer plays a significant part in the anticipated output, as it is the class which has the largest possibility for the given data input. It is given as follows:

$$f_j(z) = \frac{e^z j}{\sum_k e^z k} \tag{2}$$

Where, $e^z j$ is represented as a standard exponential function for output vector and k is a number of classes in the CNN classifier. Images may be classified using deep neural networks. These networks are pre-trained to categories different images, but they may be modified to meet our classification problematic via transfer learning by modifying essential parameters. For all the networks, hyper-parameter training has been maintained. Multiple epochs (maximum 25) were used to divide data into segments. As the name suggests, mini-batch size is the sum of samples that are used to update a model's parameters. Training mini-batch size was retained at 7 and initial learning rate was fixed at 0.0001.

4. RESULTS AND DISCUSSION

To estimate confidence intervals, we run five times more neural network experiments than random startups that have an impact on performance. All of our testing was done on an Nvidia K80 GPU. The training time for the specific parameters of the multimedia architecture is about 8 hours.

4.1. Evaluation Parameters

As quantitative measurements in this investigation, we employed precision (Acc.) and the F 1 score. We also employed the average $Mean_{Acc}$ depending on the major diagonal of the standardized M_{norm} confusion matrix, for the performing results to be evaluated, as was the case in [33]. These measurements are derived accordingly.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

$$F_1 = 2\frac{Precision.Recall}{Precision+Recall} \tag{4}$$

$$Mean_{Acc} = \frac{\sum_{i=1}^{n} g_{i,j}}{n} \tag{5}$$

$$Std_{Acc} = \sqrt{\frac{\sum_{i=1}^{n}(g_{i,j}-Mean_{Acc})^2}{n}} \tag{6}$$

Where $g_{i,j} \in diag(M_{norm})$is the ith diagonal value of the normalized confusion matrix M_{norm}, $n$is the size of M_{norm}, and *TP*, *TN*, *FP*, and *FN*, respectively, are true positive, false positive, true negative, and false negative.

4.2. Performance of Proposed model by considering training data as 20% and testing data as 80%

In this section, the proposed HireNet-CNN is tested and compared with various pre-trained network of DNN and CNN by considering 80% of testing data and 20% of training data. The results are provided in Table 1.

Table 1 Comparative analysis of various pretrained network with proposed HireNet in terms of Accuracy and F1-score for 20% and 80% of testing data

Classification Model	Acc (%)	F_1 (%)	$Mean_{Acc} \pm Std$
ConvNet-DNN	51.70	46.17	46.51± 34.38
DenseNet-DNN	52.22	48.26	47.33 ± 31.73
ResNet-DNN	54.05	50.78	48.98 ± 32.28
ConvNet-CNN	55.87	52.76	51.21 29.87
DenseNet-CNN	56.14	54.61	52.35 ± 25.53
Proposed HireNet-CNN	**58.66**	**58.50**	**56.25 ± 15.63**

In the analysis of accuracy, the proposed HireNet with CNN achieved only 58.66%, where the other networks such as ConvNet, DenseNet and ResNet with CNN and DNN achieved nearly 51% to 55% of accuracy. The reason for poor performance is that the data distribution is less in training set and more in testing set. As like accuracy, the proposed model achieved only 58.50% of F1-score, where ConvNet with DNN and CNN achieved nearly 48% to 50% of F1-score. Table 2 shows the performance of proposed model with existing techniques by training data as 40% and testing data as 60%.

Table 2 Comparative analysis of various pretrained network with proposed HireNet in terms of Accuracy and F1-score for 40% to 60% of testing data

Classification Model	Acc (%)	F_1 (%)	$Mean_{Acc} \pm Std$
ConvNet-DNN	56.26	56.38	56.23 ± 11.18
DenseNet-DNN	61.51	61.50	61.51 ± 10.40
ResNet-DNN	61.57	61.46	61.57 ± 10.79
ConvNet-CNN	81.23	81.79	77.08 ± 08.10
DenseNet-CNN	83.64	83.81	76.96 ± 11.12
Proposed HireNet-CNN	**87.22**	**87.38**	**82.45**± 09.20

When the training data is increased and testing data is decreased, the performance of the proposed model achieved better performance than existing techniques in terms of accuracy and F1-score. Figure 3 and 4 shows the comparison of proposed technique with existing techniques in terms of both parameters.

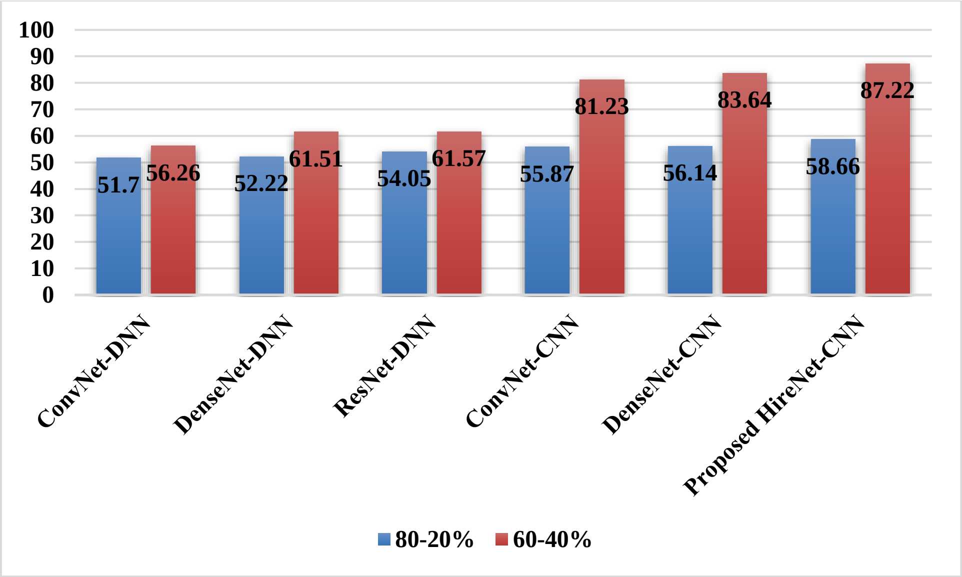


Fig 3 Graphical Representation of Proposed Model with existing techniques in terms of Accuracy

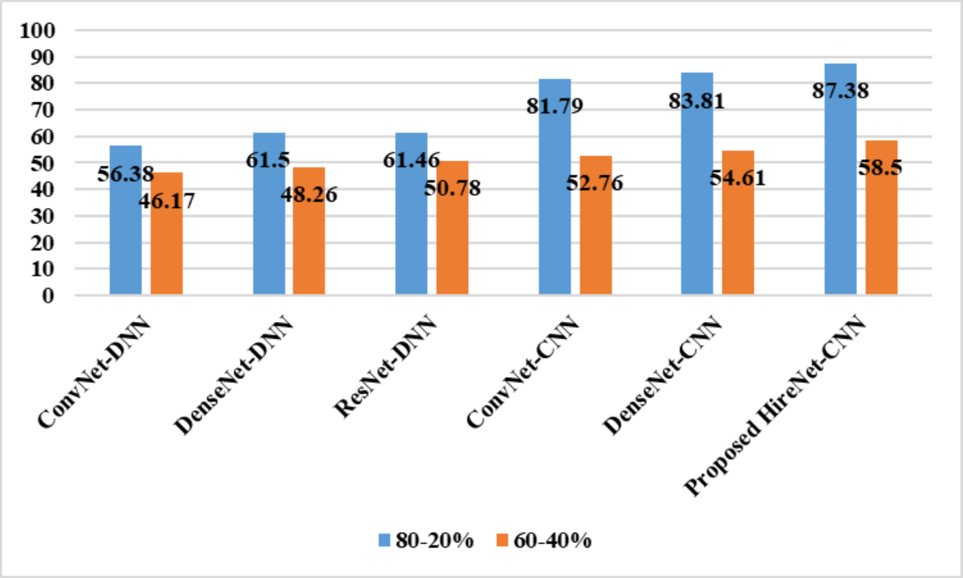


Figure 4: Graphical Representation of Proposed Model with existing techniques in terms of F1 measure

5. CONCLUSION

In the candidate selection model, a significant changes leads a development of new technologies on human resources. The new tool called self-categorized video interview provides a significant impact on marketed and developed tools. Therefore, the attention of scientific community is attracted by these solutions, which is highly studied in the contexts. The practitioners use the functionality and tools validity for selecting the right candidate, where industry didn't publish their functionality. In the context of asynchronous video interviews, a new pertained network called HireNet with CNN model is designed for predicting the right candidates of appropriate jobs from the various applicants. This model, designed to help recruits decide using a hierarchical focus neural network called HireNet, expands on earlier tasks. When creating HireNet, we design a model based on a number of practical guidelines: audio, multimedia interpretation, and ASR through an appropriate fusion mechanism. This last feature is achieved by using focus centers. We recommend and evaluate two different temporal attention functions, and use the Focus method to combine the method of CNN and HireNet. Our experiments show that the proposed HireNet with CNN perform better than existing pre-trained networks of CNN and DNN.

REFERENCES

1. Schmitt, N., &Ott-Holland, C. (2012). Theoretical and practical issues: Research needs. In The Oxford Handbook of Personnel Assessment and Selection.
2. Torres, Edwin N., and Cynthia Mejia. "Asynchronous video interviews in the hospitality industry: Considerations for virtual employee selection." International Journal of Hospitality Management 61 (2017): 4-13.
3. Chen, Lei, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. "Automated video interview judgment on a large-sized corpus collected online." In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 504-509. IEEE, 2017.
4. Lo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-claude Martin, and ChloClavel. HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. Proceedings of the AAAI Conference on Artificial Intelligence, 33:573–581, 7 2019.
5. Sowmya Rasipuram and Dinesh BabuJayagopi. Automatic assessment of communication skill in interview-based interactions. Multimedia Tools and Applications, 77(14):18709–18739, 2018.
6. Mohammed Ehsan Hoque, MatthieuCourgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. MACH. In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13, page 697, New York, New York, USA, 2013. ACM Press.
7. Keith Anderson, Elisabeth Andre, T Baur, Sara Bernardini, ´M Chollet, E Chryssafidou, I. Damian, C Ennis, A Egges, P Gebhard, H Jones, M Ochs, C Pelachaud, Kaka Porayska-Pomsta,P Rizzo, and Nicolas Sabouret. The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews. Pages476–491. 2013.
8. Patrick Gebhard, Tanja Schneeberger, Elisabeth Andr, Tobias Baur, Ionut Damian, Gregor Mehlmann, Cornelius Knig, and Markus Langer. Serious games for training social skills in job interviews. IEEE Transactions on Games, 11(4):340–351, 2019.
9. Leo Hemamou, Ghazi Felhi, Jean-claude Martin, and Chloe Clavel. Slices of Attention in Asynchronous Video Job Interviews. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–7. IEEE, 9 2019.
10. Iftekhar Naim, Md. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated Analysis and Prediction of Job Interview Performance. IEEE Transactions on Affective Computing, 9(2):191–204, 4 2018.
11. Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. IEEE Transactions on Multimedia, 16(4):1018–1031, 6 2014.
12. Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. Automated video interview judgment on a large-sized corpus collected online. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 504–509. IEEE, 10 2017.
13. Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 1:2225–2235, 2018
14. Robert Gifford, Cheuk Fan Ng, and Margaret Wilkinson. Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. Journal of Applied Psychology, 70(4):729–736, 1985.
15. Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Cross-modal Learning by Hallucinating Missing Modalities in RGB-D Vision. Elsevier Inc., 2019.

16. Timothy Degroot and Janaki Gooty. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? Journal of Business and Psychology, 24(2):179–192, 2009.
17. Leann Schneider, Deborah M. Powell, and Nicolas Roulin. Cues to deception in the employment interview. International Journal of Selection and Assessment, 23(2):182–190, 2015.
18. Amanda R. Feiler and Deborah M. Powell. Behavioral Expression of Job Interview Anxiety. Journal of Business and Psychology, 31(1):155–171, 2016.
19. Ray J. Forbes and Paul R. Jackson. Nonverbal behaviour and the outcome of selection interviews. Journal of Occupational Psychology, 53(1):65–72, 1980.
20. Ryan Miller, Brianne Gayfer, and Deborah Powell. Influence of Vocal and Verbal Cues on Ratings of Interview Anxiety and Interview Performance. Personnel Assessment and Decisions, 4(2), 2018.
21. SkandaMuralidhar, Laurent Nguyen, and Daniel Gatica-Perez. Words Worth: Verbal Content and Hirability Impressions in YouTube Video Resumes. In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 322–327, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics.
22. Anne Kathrin Buehl, Klaus G. Melchers, Therese Macan, and Jana Kuhnel. Tell Me Sweet Little Lies: How Does Faking in ¨ Interviews Affect Interview Scores and Interview Validity? Journal of Business and Psychology, 34(1), 2019.
23. Richard A. Posthuma, Frederick P. Morgeson, and Michael A. Campion. Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. Personnel Psychology, 55(1):1–81, 3 2002.
24. Lo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-claude Martin, and ChloClavel. HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. Proceedings of the AAAI Conference on Artificial Intelligence, 33:573–581, 7 2019.
25. Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, 7(2):190–202, 2016.
26. TadasBaltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, pages 59–66, 2018.
27. ShrutiPalaskar, JindichLibovicky, SpandanaGella, and Florian Metze. Multimodal Abstractive Summarization for How2 Videos In Proceedings of the 57th Annual Meeting of the Associationfor Computational Linguistics, pages 6587–6596, Stroudsburg, PA,USA, 2019. Association for Computational Linguistics.
28. Amir Zadeh, SoujanyaPoria, Paul Pu Liang, Erik Cambria,NavonilMazumder, and Louis Philippe Morency. Memory fusion network for multi-view sequential learning. 32nd AAAIConference on Artificial Intelligence, AAAI 2018, pages 5634–5641,2018.
29. Amir Zadeh, Paul Pu Liang, SoujanyaPoria, PrateekVij, Erik Cambria, and Louis-Philippe Morency. Multi-attention Recurrent Network for Human Communication Comprehension. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018, 2 2018.
30. Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and RuslanSalakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6558–6569, Stroudsburg, PA, USA, 6 2019. Association for Computational Linguistics.
31. Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. Proceedings of the AAAI Conference on Artificial Intelligence, 33:7216–7223, 2019.
32. Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and BarnabsPoczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. Proceedings of the AAAI Conference on Artificial Intelligence, 33:6892–6899, 7 2019.
33. Rada Mihalcea and Louis-philippeMorency. Utterance-Level Multimodal Sentiment Analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 973–982, 2013.
34. Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-PhilippeMorency. MOSI: Multimodal Corpus of Sentiment Intensity andSubjectivity Analysis in Online Opinion Videos. Proceedings ofthe 56th Annual Meeting of the Association for ComputationalLinguistics, 2016.

35. Carlos Busso, MurtazaBulut, Chi Chun Lee, Abe Kazemzadeh,Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, andShrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadicmotion capture database. Language Resources and Evaluation,42(4):335–359, 2008.
36. Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. Automated video interview judgment on a large-sized corpus collected online. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 504–509. IEEE, 10 2017.
37. Fan Yang, Mengnan Du, and Xia Hu. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. 7 2019.

How to cite this article:

A Paulin, V Sujitha Angel, Dr P Revathi, “Developing a HireNet with Convolutional Neural Network for Prediction of Interviewed Candidate using Asynchronous Video”, International Journal of Intelligent Computing and Technology (IJICT), Vol.5, Iss.2, pp.01-14, 2022