Cutting-Edge Methods in Visage Expression Recognition and Ocular Tracking for Analyzing Remote Video Interviews

Avni Soni[1]

[1]*Computer Science Engineering, Sage University, Indore*

---***---

Abstract - This paper focuses to augment an in the flesh image and video processor, capacitated with an artificial intelligence (AI) arbitrator that can speculate a job interviewee's behavioral adeptness according to the facial dictions or expressions. We put forward a sentiment analysis model using machine learning and CNN for the reinforcement of decision-making in the job interview process. This is bought up by histogram of oriented gradients and support vector machine (HOG-SVM) with the addition of convolutional neural network (CNN) recognition in real-time video recorded interviews. The goal of applying this technology, is to develop a method that could automatically decode a candidate's behavior by his or her facial language (or micro expressions) based on the behavioral ecology view of facial displays (BECV) which is different from the classical perspective of recognizing emotional states. This paper dispenses a trailblazing technique of determining the performance of a candidate in a video interview. To do this, a delineation of the analysis of sentiments and eye tracking technique study is intricated in which the results can be processed on a single screen to select the right person for the hire.

Key Words: Artificial Intelligence(AI), Histogram of Oriented Gradients(HOG), Support Vector Machine(SVM), Convolutional Neural Network(CNN), Behavioral Ecology View of Facial Displays, Eye Tracking Technique, AVI(Automatic Visual Inspection).

1.INTRODUCTION

At recent time, online video interviews have been progressively used in the employment process which came into account after the occurrence of pandemic situation, since the recruitment market was rapidly diminishing due to the continuous spread of COVID-19[1]. The motive of the paper is to automate the rank of a video-interview based on the facial directions, like head movement and eye contact of the candidates, where the emerging technologies such as AI, big data and machine learning could assist in making the route facile. In the context of video-recorded job interview process, it is an arduous task to detect an object or to recognize an image. There are 9 types of interview[2]: structured interview, unstructured interview, stress interview, one to one interview, panel interview, telephonic interview, video interview, depth interview, open call interview, and exit interview. Image Recognition has application in the various field of computer vision, some of which are facial recognition, emotion detection, image restoration, robotics, sentiment analysis, biometric systems, self-driving cars, and many more[3]. Hence while making hiring decisions, interviewers derive a variation of information from the candidates, such as work experience, technical skills, answers to their questions, and other soft skills. For both the interviewers and interviewees, the job interview process is a high-stakes task seeking to achieve both hiring success and maximize person-job fit[4]. So, distant from what a candidate mentions in an interview, the body movement, attentiveness, eye pupil and face direction are an requisite role of how well a person communicate. Therefore, it is tremendously important to understand the role of nonverbal communication in a video interview. The nonverbal communication includes eye contact, facial movement, posture, personality traits and hand gestures that helps in the interpretation of candidate's answers[5]. Therefore, this can help to predict how a job candidate will perform or behave at a specific job for which he or she is applying [6]. Image processing is a form of signal processing wherefore the input is an image, such as pictures or frames of video; the output of this can be either an image or a set of characteristics or parameters those are related to the image. Video processing is a peculiar case of signal processing, in which the input and output signals are video files or video streams[7]. There are different approaches for developing competency models in human resource management[8]: the job based approach, the future-based approach, the person-based approach, and the value-based approach. Job-based competency defines what should be tendered for a specific role and is commonly adopted with a static context with specific job duties and requirements. Future based competency defines innovative approaches to assessment emerging to meet the demands of any dynamic field. Person-based competency defines which generic distinctive attributes are interpretated into behavioral patterns that can holdup the human capital supremacy in an organization.

The paper aims at predicting the candidate's facial and eye movements as suitable or not suitable for the interview[5]. The paper accompanies an approach of ascertaining the facial and eye regulation with the assistance of Image Processing primarily, followed by training the Convolutional Neural Networks and Recurrent Neural Networks for multi-layered classification. In this paper, Image Processing serves as a preliminary treatment method which emphasizes the area of interest, in this instance, the candidate's face, eyes, and pupils are the focus during the interview. This aids in supplying the necessary attributes to the classifiers, CNN, and RNN. CNN models are designed to assess their performance on image recognition and detection datasets. This paper introduces an innovative method for evaluating a candidate's performance in a video interview. The candidate's confidence and attentiveness, or lack thereof, are assessed based on the direction of their face and eyes(Fig-1).

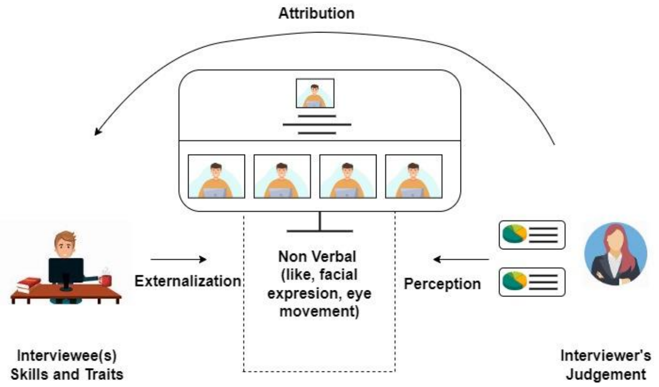
Attribution
Externalization
Non Verbal (like, facial expresion, eye movement)
Perception
Interviewee(s) Skills and Traits
Interviewer's Judgement

Fig -1: The Interview process judgement with regards to the Interviewee's skills and traits

2. Literature Review

Human detection: Nguyen (2013) identified humans using contour-based local motion features. This study consists of two main sections. The first section involves template generation, where training data is used to create templates of the entire human body and identify key points by assigning weights to each point. The second section is the testing phase, which employs a sliding window technique to extract candidate regions and applies Canny Edge Detection to detect edges within these regions[9]. In the context of Canny Edge Detection, the process involves four steps: Noise Reduction utilizing a Gaussian filter, calculating the intensity gradient of the images, Non-maximum Suppression, and Hysteresis Thresholding. The Gaussian filter is used to blur the images by applying a kernel or filter.

Facial Action Coding System (FACS): Created by Ekman and Friesen, has traditionally been used by psychology researchers to manually code facial expressions. Over the past decade, advancements in machine learning and computer vision have enabled the automatic execution of high-quality FACS coding[10].

Emotion Analysis: Automatic Speech Emotion Recognition (SER) is a technology designed to identify the emotions of speakers from both recorded and live speech. It relies on acoustic features such as energy distribution, pitch patterns, spectral properties, speaking rate, and the spoken content. Numerous studies have systematically reviewed the collective research in this field. Implementing Speech Emotion Analysis (SEA) in interviews with Machine Learning (ML) can reveal important insights into a candidate's emotional condition. This is particularly valuable for positions that demand emotional intelligence, stress management, and effective communication skills. The key components of Speech Emotion Analysis are[11]: Pre-processing: Cleaning and preparing the audio data for analysis. Feature Extraction: Identifying key features from the speech, such as prosody, pitch and rhythm. Classification: Employing machine learning algorithms to systematically categorize passions based on derived features.

Face Detection: It is a technology implemented in computing to ascertain the positions and dimensions of human faces in digital pictorial representation. It emphasizes on recognizing physiognomy while disregarding other elements like constructions, evergreens, and figures or various objects. Face detection is regarded as a specialized form of object-class detection. In this process, the positions, and dimensions of all articles within an image that fall under a specific category are pinpointed. Face detection shall be seen as a broader concept compared to face localization[12]. In face localization, the objective is to conclude the positions and sizes of a predetermined number of faces. In contrast, face detection does not have this prior information. Initial face-detection algorithms were designed to identify frontal human faces, while more recent algorithms aim to tackle the more complex challenge of detecting faces via multiple angles. This involves detecting faces that are either rotated around the axis from the face to the viewer (in-plane rotation), or rotated around the vertical or horizontal axis(out-of-plane rotation), or both[13].

Decrypt and analyze facial expressions: Research on human facial expressions follows two main approaches[14]: basic emotion theory (BET) and the behavioral ecology view of facial displays (BECV). BET, often referred to as the 'common view' posits that human internal feelings and emotions are expressed through facial articulations at both massive and small scales. Facial expressions are often described as: "Our visage is an intricate, highly specialized component of our anatomy" – indeed, it represents one of the most sophisticated signaling systems at our disposal. It comprises over 40 structurally and functionally distinct muscles, each capable of being activated independently[15]. Scientifically proven methodology suggests about the nerve that forms expressions on the human face(Fig- 2).

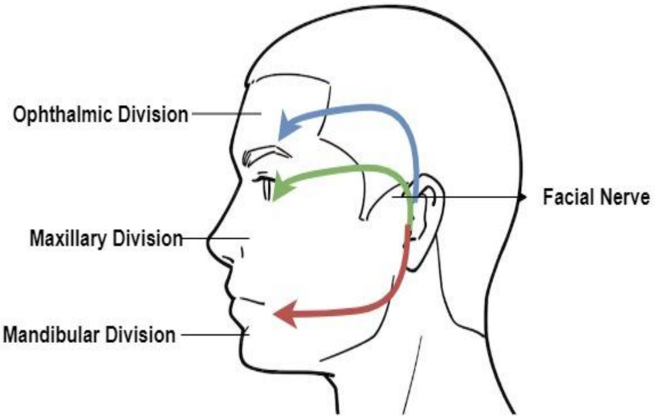

Fig -2: Facial Nerve with different divisions

The below table(Table-1) demonstrates difference between Emotions, Moods, Feelings and Affect.

Table -1: A table describing difference between Emotion, Feelings and Moods

Aspect	Description
Emotions	Acute, ephemeral experiences precipitated by distinct events or stimuli.
Feelings	Subjective interpretations of emotional action schemas, propelled by conscious cognition and introspection. Emotions can exist independently of feelings, yet feelings cannot manifest without the presence of emotions.
Moods	Subtle, intrinsic, subjective states that are typically less intense than emotions and endure significantly longer. Profoundly shaped by personality traits and characteristics. Deliberate facial expressions can simulate emotional effects.
Affect	A broad term that includes emotions, feelings, and moods.

3. Methodology

Facial Expression Detection: Facial expression detection in AI and ML is a fascinating application, especially in the context of interviews. This technology, often referred to as Facial Expression Recognition (FER), involves using machine learning algorithms to analyze and elucidate human facial expressions from images or video feeds[16]. Facial detection refers to identifying the region(s) of interest (ROIs) of the human face in the frames of huge image sequences. FER can be divided into few major processes: **Emotion Analysis**: FER systems can detect and classify emotions such as happiness, sadness, anger, surprise, and more. This can enable interviewers to discern a candidate's emotional disposition throughout the interview. **Behavioral Insights:** By analyzing facial expressions, these systems can provide insights into a candidate's behavior and reactions. This can be instrumental in evaluating attributes such as self-assurance, integrity, and attentiveness. **Deception Detection:** Some leading-edge systems combine facial expression analysis with other biometric data, such as pulse rate, to detect signs of deception. This can assist interviewers in discerning the veracity of a candidate's responses. **Competency Prediction:** AI models can predict a candidate's competencies based on their facial expressions and other non-verbal cues. This can facilitate the assessment of whether a candidate embodies the requisite skills and attributes for the position. **Bias Reduction:** Programmed systems can mitigate human bias in the interview process by furnishing objective insights into a candidate's emotional and behavioral responses.

In the context of using AI and ML for the interview process, feature extraction plays a crucial role in identifying key facial features or data attributes. This process is essential for facial expression recognition, where feature detection and feature extraction are the primary activities involved in analyzing a candidate's facial expressions. By accurately extracting these features, the system can provide valuable insights into the candidate's emotional state and behavioral responses during the interview[17]. Discriminative Response Map Fitting (DRMF) is a frequently employed holistic texture-based model adept at detecting facial features from various expressions in video clips. This sophisticated technique involves annotating 68 landmark points on the face[18](Fig-3), which serve as critical inputs for recognizing and analyzing patterns of facial expressions. Each image of the candidate's face is divided into small

DOI: 10.55041/IJSREM37620

spatial regions, and the gradients and orientations are meticulously calculated for every pixel. This method, known as the Histogram of Oriented Gradients (HOG)[19], is employed to extract essential features, enabling the system to analyze and interpret the candidate's facial expressions accurately. These attributes facilitate the training of machine learning algorithms, notably linear support vector machines (SVMs), which are capable of executing classification tasks utilizing a nonlinear decision boundary formulated through a polynomial kernel[20].

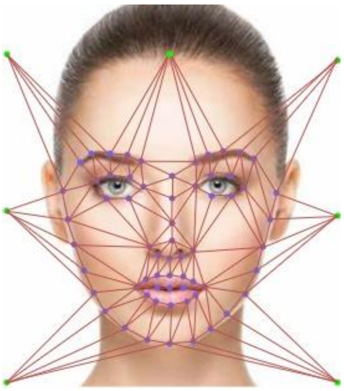

Fig -3: Facial Landmark Points

Eye Tracking: Eye tracking is a technique used to measure where and how long a person looks at various points in their visual field. The eye tracker uses machine learning and image processing to analyze the camera feed and calculate data points.(Fig-4). Eye tracking involves the use of sensors, cameras to film the movement and position of the eyeballs. The key metrics typically recorded include[21]: **Fixations:** Locations where the gaze remains stationary. **Saccades:** Swift transitions between points of fixation. **Pupil Dilation:** Changes in pupil size, often linked to cognitive load or emotional state. **Blink Rate:** Frequency of blinking, which can indicate stress or fatigue. Macro-expressions, which are closely linked to emotions, can be observed and categorized into six universal basic emotions of a person: happiness, fear, sadness, surprise, anger, and disgust"[22]. Micro-expressions occur involuntarily and last only a brief moment, making them difficult to detect with the naked eye. The key difference between macro and micro-expressions is their duration, with micro-expressions lasting less than 1/5 of a second[23]. During an asynchronous video interview, Ocular tracking software can monitor a candidate's visual acuity to assess their engagement and honesty. For instance, consistent eye contact with the camera might point to confidence and attentiveness, while frequent darting of the eyes might suggest nervousness or distraction. This data can be analyzed alongside the candidate's verbal responses to provide a more comprehensive evaluation of their suitability for the role. Eye tracking can also help identify moments when a candidate is reading from notes or a script, which might be important for roles requiring spontaneous thinking and communication skills.

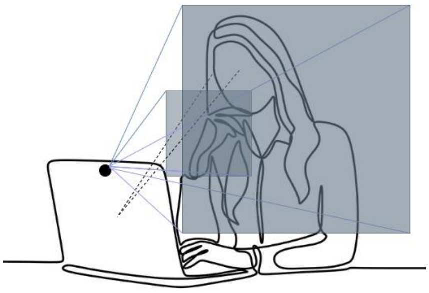

Fig -4: Eye Tracking Technique

Eye accessing cues refer to the movements of the eyes that indicate which representational system (visual, auditory, or kinesthetic) a person is using to think. According to research, there are six primary categories of eye accessing cues[24]: **Visual Constructed(Vc) :** When someone's eyes shift upward to the left, it suggests they are visualizing images they haven't previously encountered. **Visual Remembered(Vr)**: When someone's eyes shift upward to the right, it suggests they are remembering images they have previously seen. **Auditory Constructed(Ac)**: When someone's eyes move horizontally to the left, it suggests they are imagining sounds they have not heard before. **Auditory Remembered(Ar)**: When someone's eyes move horizontally to the right, it suggests they are remembering sounds they have previously heard. **Kinesthetic(K)**: When someone's eyes move down to the left, it suggests they are accessing their feelings, tastes, or smells. **Auditory Digital(Ad):** When someone's eyes shift downward to the right, it suggests they are engaging in internal dialogue or self-talk.

Use case for Eye Tracking in Interview Process: Objective Assessment of Engagement: Eye tracking software can monitor a candidate in order to assess their level of engagement through eye movements during the interview. For instance, consistent eye contact with the camera might indicate attentiveness and confidence, while frequent shifts in gaze could suggest distraction or nervousness[25]. This data provides an additional layer of insight beyond verbal responses. **Detection of Reading from Notes:** In roles requiring spontaneous thinking and communication skills, it's crucial to know if a candidate is reading from notes. Eye tracking can identify patterns that suggest reading, helping interviewers assess the candidate's ability to think on their feet. **Emotional Analysis:** Eye movements can also be linked to emotional states[26]. For example, rapid blinking or pupil dilation might indicate stress or excitement. This can help interviewers understand how candidates react under pressure, which is valuable for high-stress roles. **Consistency and Honesty:** By analyzing eye movement patterns, interviewers can detect inconsistencies in a candidate's responses. For example, avoiding eye contact when answering certain questions might suggest discomfort or dishonesty. **Improved Candidate Experience:** For candidates, knowing that their engagement and honesty are being objectively assessed can encourage more genuine and focused responses. This can lead to a more fair and transparent interview process.

Data collection and Preparation: (Fig-5)

Selecting Participants: In the realm of asynchronous video recorded interviews, AVI can be used to analyze candidates' facial expressions, eye movements, and other visual cues to assess their responses. This can aid in evaluating non-verbal communication skills and ensuring a more comprehensive assessment process. AVI technology leverages ML algorithms to enhance the exactitude and proficiency of visual inspections, which are crucial in various industries, such as manufacturing and pharmaceuticals. Hence, based on this, candidates will be selected and trained to confer interviews. Participants will be asked to respond to a series of core competency interview questions[27], which will determine their suitability for the role. These questions are crafted to elicit responses that will serve as a dataset for a machine learning model. Interviewees will be notified that their answers will guide the hiring manager's decision and that their facial expressions will be analyzed by an AI algorithm to assess their competencies.

Data Analysis and Visualization: Machine learning algorithms derive their intelligence from data. It is imperative to provide them with the appropriate data tailored to the distinct challenge at hand. The more meticulous and disciplined you are in managing this data, the more consistent and superior the outcomes you are likely to achieve. The process of preparing data for a machine learning algorithm can be encapsulated in three pivotal steps[28]. Step1: Select Data, Step 2: Preprocess Data, Step 3: Transform data. The three prevalent steps in data preprocessing include formatting, cleansing, and sampling[29]. **Formatting**: The statistics chosen might not be in a format conducive to the needs. For instance, it could reside in a relational database when you require it in a sequential file, or it might be in a specialized format when you need it in a database or a text file. **Cleaning**: Data cleaning involves the elimination or correction of missing data. The instances where data is incomplete and lacks the necessary information to address the problem at hand might exhibit. Such instances may need to be discarded. **Sampling**: An excess of selected data might not be necessary for your work. Handling more data can lead to significantly longer algorithm run times and increased computational and memory-related aspects in the demands. Instead, you can opt for a smaller, stratified sample of the data, which can expedite the process of engaging in the exploration and prototyping of solutions prior to implementation, tackling the entire dataset.

Model Selection: Model selection necessitates the decision of a sole machine learning model derived from a pool of potential models tailored for a definite training dataset. This methodology can be put into practice into various types of as well as to models of the same type but with different hyperparameters (like various kernels used in a SVM). Ultimately, model opting is about identifying the terminal model that best remediates the concern at hand. Once a model is selected, it can be assessed to convey its anticipated overall performance[30]. Thus, a 'good enough' model can have various meanings depending on your project, such as: a model that satisfies the stipulations and limitations set by project collaborators, demonstrates adequate skill given the awarded temporal and material assets, performs better than uncomplicated models, shows superior performance when contrasted with some differently tested models, is competitive with the state-of-the-art[31].

Estimating Accuracy and Performance: Estimating the exactness and capability of a model in the interview process involves evaluating how well the model predicts candidate success based on historical data, ensuring it effectively identifies top talent while minimizing biases and errors. Once the model is deployed, monitoring its performance is essential to maintaining the quality of your ML system. Calculating metrics such as accuracy, precision, recall, or F1-score requires labels[32]. To evaluate precision, divide the number of genuine positive predictions by the addition of genuine positive and spurious positive predictions[33].

Precision = True Positives / (True Positives + False Positives)

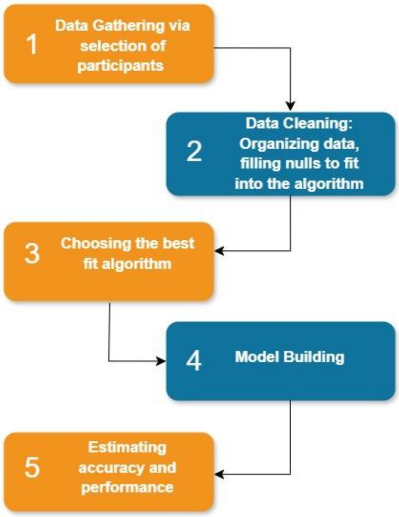
1 Data Gathering via selection of participants
2 Data Cleaning: Organizing data, filling nulls to fit into the algorithm
3 Choosing the best fit algorithm
4 Model Building
5 Estimating accuracy and performance

Fig -5: Machine Learning Process Flow

4. CONCLUSIONS

In this study, we try to whip up a prototype system that rocks a HOG-SVM detection and feature extraction method for facial expressions, paired with a CNN classification setup. Snagging the interviewees' facial expressions using an AVI platform, then deep learning will show its magic to match these expressions with the competency scores rated by their supervisors. So, our study rolled out a cutting-edge HR selection and assessment method that can automatically predict behavioural competencies from facial expressions in asynchronous video interviews. This could totally replace human interviewers and traditional competency assessments during the candidate screening stage, offering higher predictive accuracy and slashing selection costs.

REFERENCES

1. Youel Park and Chang-Bae Ko" Proposal for AI Video Interview Using Image Data Analysis" International Journal Internet, Broadcasting and Communication Vol.14 No.2 212 218 (2022).
2. Suen, H., Hung, K., Lin, C.: TensorFlow-based automatic person ality recognition used in asynchronous video interviews. IEEE Access 7, 61018–61023 (2019).
3. Yann LeCun, Yoshua Bengio, Geoffery Hinton, "Deep Learning", Nature, Volume 521, pp. 436-444, Macmillan Publishers, May 2015.
4. Lei Chen, Su-Youn Yoon and Chee Wee Leong "An Initial Analysis of Structured Video Interviews by Using Multimodal Emotion Detection" General— Multimedia Information Systems ERM4HCI'14, November 16, 2014.
5. Nishank Singhal, Neetika Singhal and Srishti "Comparing CNN and RNN for Prediction of Judgement in Video Interview Based on Facial Gestures" 5th International Conference on Signal Processing and Integrated Networks 2018.
6. Hofrichter, D.A., Spencer, L.M.: Competencies: the right foundation the right foundation for effective human resources management. Compens. Benefts Rev. 28, 21–26 (1996).
7. Byeong-Ho KANG "A Review on Image and Video processing" International Journal of Multimedia and Ubiquitous Engineering Vol. 2, No. 2, April, 2007.
8. Cardy, R.L., Selvarajan, T.T.: Competencies: alternative frameworks for competitive advantage. Bus. Horiz. 49, 235–245 (2006).
9. Yi Yang, Deva Ramanan **:** Articulated Human Detection with Flexible Mixtures of Parts. December 2013, IEEE Transactions on Pattern Analysis and Machine Intelligence 35(12):2878-90.
10. Crivelli, C., Fridlund, A.J.: Facial displays are tools for social infuence. Trends Cogn. Sci. 22, 388–399 (2018).
11. Paul Viola, Michael J. Jones : Robust Real-Time Face Detection, Published: May 2004 Volume 57.
12. Jay Prakash Maurya, Mr.Akhilesh A. Waoo, DR. P.S Patheja, DR Sanjay Sharma : A Survey on Face Recognition Techniques: Computer Engineering and Intelligent Systems ISSN 2222-1719.
13. Llerena, M.: Desarrollo de una metodología basada en la programación neurolingüística utilizando software educativo para mejorar el proceso enseñanza-aprendizaje. Msc tesis. Escuela Superior politécnica de Chimborazo, Ecuador (2016).
14. Shen, X.B., Wu, Q., Fu, X.I.: Efects of the duration of expressions on the recognition of microexpressions. J. Zhejiang Univ. Sci. B 13, 221–230 (2012).
15. Takalkar, M., Xu, M., Wu, Q., Chaczko, Z.: A survey: facial micro-expression recognition. Multimed. Tools Appl. 77, 19301– 19325 (2018).
16. Asthana A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map ftting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3444–3451. IEEE, Portland, OR, USA (2013).
17. Merget D., Rock, M., Rigoll, G.: Robust facial landmark detection via a fully-convolutional local-global context network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 781–790. IEEE, Salt Lake City, UT, USA (2018).
18. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893. IEEE, San Diego, CA, USA (2005).

19. Carcagnì, P., Del Coco, M., Leo, M., Distante, C.: Facial expression recognition and histograms of oriented gradients: a comprehensive study.
20. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to decep tion. Psych. 32, 88–106 (1969).
21. Cuve, Helio Clemente, Jelka Stojanov, Xavier Roberts-Gaal, Caroline Catmur, and Geoffrey Bird. 2022. Validation of Gazepoint low-cost eye-tracking and psychophysiology bundle. *Behavior Research Methods* 54: 1027–49..
22. Dalmaso, Mario, Luigi Castelli, and Giovanni Galfano. 2020. Social modulators of gaze-mediated orienting of attention: A review. *Psychonomic Bulletin & Review* 27: 833–55
23. Norambuena, B. K., Lettura, E. F., and Villegas, C. M. (2019). Sentiment analysis and opinion mining ap plied to scientific paper reviews. Intell. Data Anal., 23(1):191–214.
24. Su, Y.S., Lin, C.L., Chen, S.Y., Lai, C.F.: Bibliometric study of social network analysis literature. Libr. Hi Tech. 38, 420–433 (2019).
25. T. DeGroot and J. Gooty. Can nonverbal cues be used to make meaningful personality attributions in employment interviews? Journal of Business and Psychology, 24(2):179–192, 2009.
26. A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma Combining Audio, Video and Lexical Indicators of Affect in Spontaneous Conversation via Particle Filtering Multimodal Interaction (ICMI), International Conference on, pages 485–492. ACM, 2012.
27. D. M. Gavrila, J. Giebel, and S. Munder. Vision-based pedes trian detection: the protector+ system. Proc. of the IEEE Intelligent Vehicles Symposium, Parma, Italy, 2004.
28. Sun- Jeong Jeong and Hwi-Eun Nam, "Status and Implications of Private-led Human Resources Development Program in New Technology Field," Journal of the Korea Academia-Industrial cooperation Society, Vol. 23, No. 2, pp. 322-330, Feb 2022.
29. T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in Proc. Eur. Conf. Mach. Learn., vol. 98, 1998, pp. 137142.
30. Ping-Hsien Lin and Tong-Yee Lee, "A Fast 2D Shape Interpolation Technique", O. Gervasi et al. (Eds.): ICCSA 2005, LNCS 3480.
31. R. A. Popa, F. H. Li, and N. Zeldovich, “An ideal-security protocol for order-preserving encoding,” in Proceedings of the 2013 IEEE Symposium on Security and Privacy (SP’13). IEEE, pp. 463–477, 2013.
32. Jianming, Chujie Chen, Zequan Liang: Automated Scoring of Asynchronous Interview Videos Based on Multi-modal Window Consistency Fusion, 2024.
33. A System for MANET to detect selfish nodes using NS2 SD Padiya, R Pandit, S Patel Int. J. Eng. Sci. Innov. Technol 1 (2), 25-30