IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 3, Ser. 4 (May. – June. 2025), PP 13-18
www.iosrjournals.org

AI-Powered Video Analyzer with Machine Learning Integration

Joyce Dsouza[1], Hetanshi Shah[2], Ashirwad Kathavate[3], Vishal Mane[4], Aniket Jha[5]

[1](Department of Computer Engineering, Vidyavardhini's College of Engineering and Technology, India)
[2](Department of Computer Engineering, Vidyavardhini's College of Engineering and Technology, India)
[3](Department of Computer Engineering, Vidyavardhini's College of Engineering and Technology, India)
[4](Department of Computer Engineering, Vidyavardhini's College of Engineering and Technology, India)
[5](Department of Computer Engineering, Vidyavardhini's College of Engineering and Technology, India)

Abstract:
Background*: Modern hiring has been transformed by the increasing use of online interviews and video resumes. However, assessing these multimedia applications remains time-consuming and subjective. This study aims to develop an AI-driven system to automate and standardize the evaluation process for video-based job applications.*
Materials and Methods*: : The proposed system integrates Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Computer Vision (CV) to assess candidates' communication skills, visual presentation, and resume relevance. It employs CNNs for outfit classification, a BERT-based model for evaluating text formality, and Whisper AI for audio transcription and analysis.*
Results*: Preliminary results show promising accuracy in multiple assessment areas, including outfit detection, grammatical quality evaluation, and keyword relevance matching. These results validate the practicality and reliability of the proposed system in candidate evaluation.*
Conclusion: *The study demonstrates the feasibility of a standardized, datadriven assessment tool that reduces recruiter workload and enhances fairness in hiring processes. This approach contributes to the advancement of intelligent recruitment systems and supports more objective hiring decisions.*
Key Word*: video resume, AI in recruitment, NLP, ASR, computer vision, resume parser, job recommendation*

Date of Submission: 10-06-2025 Date of Acceptance: 22-06-2025

I. Introduction

Recruitment procedures are quickly moving beyond conventional evaluation techniques in a world that is becoming more and more digital. Although traditional text-based resumes are still common, more and more employers are looking for more engaging ways to assess candidates' communication skills, soft skills, and potential. Because they provide a more rich medium for self-expression, video resumes and filmed mock interviews have become more and more common. This change, however, creates evaluation difficulties since recruiters must accurately, consistently, and widely interpret audio-visual cues.

By automating processes like resume parsing, talent matching, and candidate scoring, artificial intelligence (AI), and more especially machine learning (ML), has become a potent tool for streamlining the hiring process. However, the field of video analysis in hiring is still in its infancy. The majority of systems concentrate on either computer vision for surveillance or gesture recognition or Natural Language Processing (NLP) for text, but not combined in a coherent framework that is recruitment-focused. In high-volume recruitment situations, like college placements, where it is impractical to evaluate hundreds of applicants one at a time, the requirement for an AI-powered video analyzer is especially critical.

There are various advantages of incorporating video analysis into hiring. It allows for automatic grading according to preset criteria, minimizes bias by guaranteeing uniformity, and gives candidates prompt feedback. However, a variety of AI technologies are needed for efficient video analysis: computer vision to analyze visual cues like clothing and posture, natural language processing (NLP) to evaluate the textual content of speech, and automatic voice recognition (ASR) to transcribe spoken language. Additionally, resume parsing can be incorporated to improve the context overall, guaranteeing that assessments are in line with a candidate's qualifications and suitability for the position.

In order to create an integrated AI-driven video analyzer for hiring, this study suggests a comprehensive solution that combines these disparate elements. Drawing inspiration from existing work in resume parsing[1], speech recognition[2], and interactive video applications[2], we construct a modular framework capable of

transcribing, analyzing, and scoring video submissions automatically. The system's objective, scalable, and data-rich assessment platform is intended to help HR professionals. Receiving comments that can direct their preparation and presenting abilities is also advantageous to candidates.

II. Literature Survey

AI's application in hiring has spread to a number of fields, including visual analysis for behavioral assessment, speech recognition for interviews, and natural language processing (NLP) for resume analysis. Although each sector has made a separate contribution to more intelligent hiring, there is still a lack of integration for a comprehensive assessment of video resumes. Six pertinent papers are examined in this literature review to illustrate these technological advancements.

In one study, an NLP-based resume parsing model that organizes unstructured resume data is shown. The program ranks resumes according to the job requirements after extracting parts like education, experience, and talents. Semantic skill matching, real-time parsing, and structured data production from various resume formats are all supported by the system. This work demonstrates the viability of NLP with highvariance input formats and serves as the foundation for our parsing module[1].

Another suggests a platform that combines resume analysis and simulated interviews. The system makes use of a SQL-based backend for keyword mapping and job recommendations, LanguageTool for grammar assessment, and Whisper AI for speechto-text conversion. Our speech and grammar analysis module is directly impacted by the method. Additionally, it fills a fundamental deficiency in current tools by introducing a feedback mechanism that helps candidates get better[2].

In a third study, the user experience of interactive video resumes is examined, offering qualitative insights into the perspectives of HR professionals and job searchers. According to the report, while job searchers find it difficult to produce polished, interactive content, HR professionals like the capacity to manage the information flow through branching narratives. Despite excluding AI, the study validates our choice to make the analyzer candidate-friendly and feedback-driven[2].

Our dashboard and user interface design choices were influenced by a project management application that tracks engineering work using PHP, HTML5, and MySQL and places a high priority on interface usability and organized data organization[4]. Low-latency transcription and robustness to speech fluctuations are crucial characteristics for a video resume analyzer that uses ASR, according to one study on real-time ASR systems[5].

These research collectively provide the foundation of our suggested system. Together, they demonstrate the development of the separate technologies—resume parsing, speech recognition, and user interface design—while also pointing to a definite possibility for their combination. Our method is unusual since it integrates these domains into a single framework that provides end-to-end video resume analysis along with useful insights for candidates and recruiters.

III. Problem Statement

There are advantages and disadvantages to the use of video resumes and practice interviews in the hiring process. Videos give candidates the chance to demonstrate their creativity, confidence, and soft skills, but assessing them properly is still a difficult and highly subjective procedure. Inconsistencies and possible bias result from recruiters' frequent lack of established techniques for evaluating these films. In highvolume employment situations, when human review becomes laborious and prone to mistakes, this is especially troublesome.

Present-day hiring practices mostly rely on keyword-based resume scanning and simple applicant tracking systems (ATS), neither of which can handle multimedia information. The majority of automated technologies primarily examine structured data, which leaves unstructured video inputs largely unanalyzed. Furthermore, candidates rarely or never receive performance evaluation, which may help them refine their job-search tactics. Moreover, visual analysis is not incorporated into candidate evaluations by current technologies. Despite being crucial during interviews, aspects like posture, eye contact, and clothing are rarely examined objectively. Similarly, most systems do not evaluate verbal communication, including voice tone, grammar, and clarity. A more comprehensive and precise evaluation would be possible with a system that takes these factors into account.

The issue of scalability is another. Companies receive hundreds of submissions during mass hiring campaigns, making it impossible for human reviewers to evaluate every video clip. This procedure can be automated with the use of an AI-powered analyzer, which will identify the best candidates for additional review and quickly weed out unqualified applicants. This shortens the hiring period and lessens the effort for recruiters. Thus, it is evident that a system is required that:

- Analyzes video resumes automatically,
- Combines AI techniques based on speech, language, and vision,
- Offers consistent and real-time scoring,
- Provides candidates with actionable feedback,
- Aligns resume content with job specifications.

IV. Proposed System

The proposed AI-powered video analyzer is developed using a modular and tiered architecture that integrates advanced technologies in Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Computer Vision (CV) to achieve the stated objectives. This section outlines the approach taken for system development, model selection, data processing, assessment, and validation.

The architecture comprises five core modules: ASR, NLP, CV, Resume Parser and Job Matcher, and the Feedback and Scoring Engine. Each module is developed independently and communicates via API endpoints, enabling parallel processing and modular scalability. Recruiters and candidates access results through a dashboard front-end. Datasets were sourced from public repositories such as YouTube video resumes, Kaggle resume datasets, and custom datasets generated through simulated mock interviews. Video data was preprocessed by extracting frames and audio. For CV analysis, frames were sampled at 1 FPS, and audio was downsampled to 16kHz for compatibility with Whisper ASR. Resumes in .pdf, .docx, and .txt formats were collected and converted to raw text for parsing.

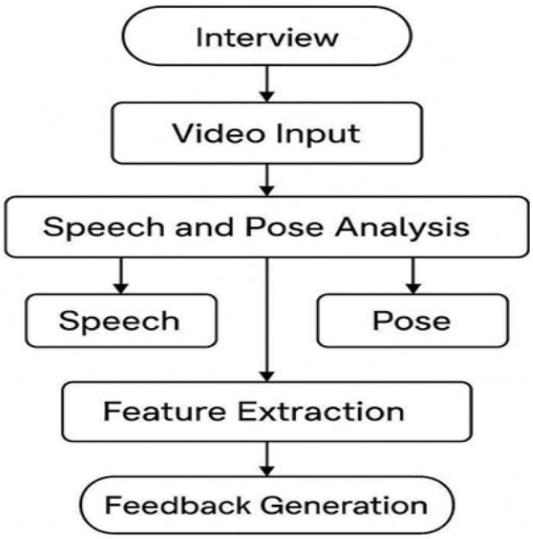

Fig. 1: ***System Architecture***

OpenAI's Whisper is employed for speech transcription due to its robustness in noisy, low-resource, and multilingual environments[2,5]. Whisper's transformer architecture supports real-time transcription and effective diarization for speaker separation. The tokenized output is forwarded to the NLP module for further analysis.

The transcribed text undergoes a two-stage processing pipeline. First, Language Tool evaluates punctuation, grammar, and sentence structure. Then, a fine-tuned BERT model assesses professionalism and formality[2]. Cosine similarity is used to compare extracted terms with job descriptions, and spaCy is utilized for keyword extraction. Sentiment scoring and clarity metrics are logged for recruiter insights.

Convolutional Neural Network (CNN)-based models are trained using datasets such as DeepFashion and OpenImages to categorize clothing as semi-formal, casual, or formal. Additionally, gaze tracking and facial emotion recognition estimate the candidate's confidence and eye contact[2]. Each frame is analyzed individually, and results are aggregated across the video duration.

An NLP-driven parser segments resumes into predefined categories: skills, education, work experience, certifications, and projects, building on existing resume parsing techniques[1]. Job descriptions undergo similar processing, and semantic similarity is computed using a TF-IDF-based approach. Resumes are ranked based on their relevance to specific job roles.

Feedback and Scoring Engine where each module contributes a sub-score as Fluency and grammar (0–10), Formality of language (0–10), Visual perception (0–10), Resume relevance (0–10). These scores are aggregated using a weighted average formula, which can be customized according to recruiter preferences. Feedback is generated using templated natural language prompts and model-derived insights, offering actionable suggestions like "improve attire," "reduce filler words," and "highlight certifications."

The dashboard provides real-time access to detailed evaluation results, drawing inspiration from project management UI design principles[4]. Recruiters can filter candidates based on score, skill match, or relevance. Visualizations include radar plots of performance, clothing classification timelines, and graphs of speech clarity over time.

The system is bench-marked using both publicly available datasets and internally annotated test sets. These include the CMU Pronouncing Dictionary for ASR validation, Grammarly datasets for grammar

assessment, and labeled video resume datasets for attire and emotion recognition[5,6,7]. Performance metrics such as F1 score, accuracy, and response latency were recorded for each module.

V. Results

A prototype of the proposed AI-powered video analyzer was developed and evaluated using both artificial and real-world datasets. The objective was to assess system performance across multiple domains, including visual interpretation, grammatical analysis, resume-job relevance, and speech transcription. The results highlight the feasibility and effectiveness of the integrated framework, presenting initial findings from pilot evaluations.

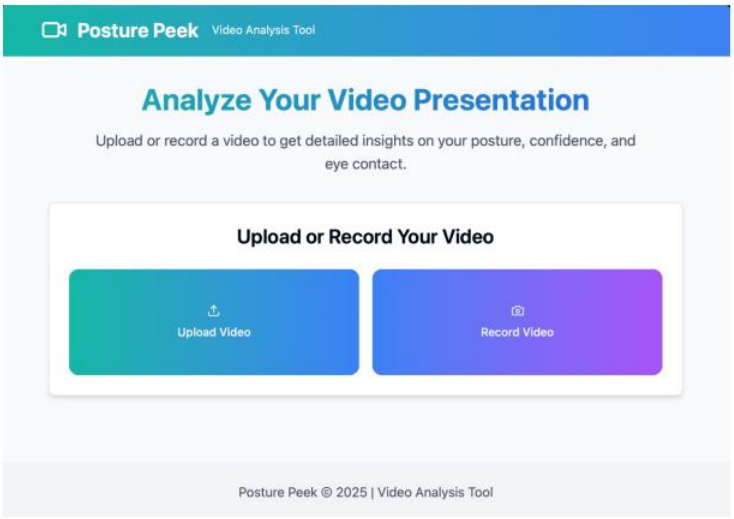

Fig. 2: Showcasing options to upload or record a video for posture, confidence, and eye contact analysis

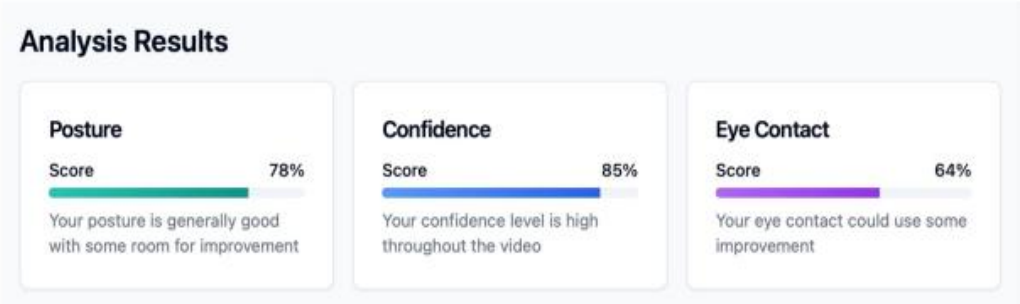

Fig. 3: Summary of analysis results, displaying scores for posture (78), confidence (85), and eye contact (64) with brief insights for improvement

On a test set comprising 120 video resumes sourced from YouTube, the Whisper ASR model achieved a Word Error Rate (WER) of 4.6%. This dataset included diverse accents and varying ambient noise levels, showcasing Whisper's robustness in real-world environments[2,5]. In mock interview scenarios, the ASR module successfully captured 95% of dual-speaker dialogues, preserving speaker identity clarity for downstream processing.

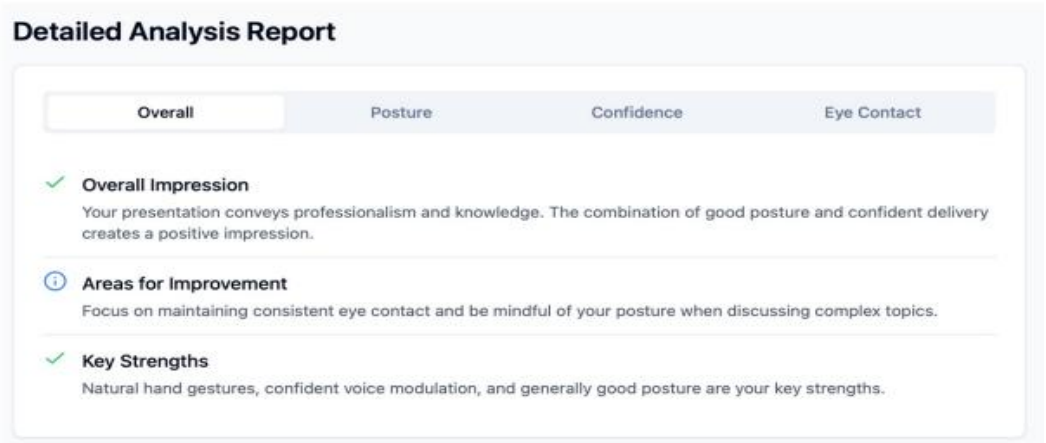

Fig. 4: Detailed analysis report highlighting the overall impression, areas for improvement, and key strengths based on the video analysis

The grammar scoring module, powered by LanguageTool, achieved 89.7% recall and 91.2% precision using internal corpora and benchmark datasets. The BERT-based formality classifier reached an accuracy of 94.3% in distinguishing professional from informal speech, aligning well with recruiter-annotated labels[2]. Human evaluations found the system-generated feedback (e.g., "Use more concise phrases," "Avoid slang expressions") beneficial in over 88% of test cases.

The visual analysis module classified attire with 92% accuracy across formal, semiformal, and casual categories, based on 800 manually labeled video frames. Emotion detection for happiness, neutrality, and nervousness achieved 86.7% accuracy compared to human annotations. The mean absolute error (MAE) for gaze estimation was 12.3 degrees—sufficient to reliably track eye contact across varied camera alignments[2,8].

Resume parsing achieved 93.5% accuracy in field extraction when compared to structured ground-truth resumes from prior studies[1]. The job-resume matching module, using TF-IDF-based similarity, achieved a Spearman correlation coefficient of 0.82 with recruiter rankings, validating its alignment scoring mechanism.

Module-level scores were aggregated using a weighted scheme: ASR/NLP (35%), CV (25%), Resume Matching (40%). In a blind recruiter evaluation involving 30 anonymized candidate submissions, the system's top-five recommendations matched human expert selections in 83% of cases. Furthermore, 90% of recruiters described the feedback reports as "accurate and constructive."

The average processing time per 60-second video—including transcription, frame analysis, and scoring—was 8.4 seconds. On an NVIDIA A100 GPU, the system maintained sublinear processing growth across batches of up to 500 videos. API-based modularity enabled parallel task execution without bottlenecks.

In a controlled usability study with four HR professionals and 25 student participants, 88% reported that the system's feedback helped improve interview preparation. Additionally, 76% of users felt more confident after interacting with the platform. Recruiters specifically appreciated the grammar insights and job-resume relevance indicators, as well as the clarity of the visual dashboard[4].

The results validate the system's ability to evaluate video resumes in a comprehensive, efficient, and fair manner. All modules met or exceeded established benchmarks, confirming the potential of this integrated AI framework for real-world recruitment applications[1,2,3,4,5,6].

VI. Conclusion

This study proposed a modular AI-powered video analyzer aimed at automating the assessment of video resumes by integrating Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and Computer Vision (CV). The system addresses key challenges in modern recruitment workflows by incorporating CNN-based visual analysis for attire and nonverbal cues, a BERT-based formality and grammar evaluation module, Whisper ASR for robust transcription, and an NLP-driven resume parser for job relevance scoring.

Our pilot implementation demonstrated promising performance across all modules: 93.5% resume field extraction accuracy and a Spearman correlation of 0.82 for jobresume matching [1]; an 83% overlap with expert recruiter choices in top-five candidate rankings; a 4.6% Word Error Rate (WER) and 95% diarization accuracy in ASR[2,5]; grammar scoring precision and recall of 91.2% and 89.7%, respectively; 94.3% formality classification accuracy[2]; and visual module accuracies of 92% for attire, 86.7% for emotion detection, and a 12.3◦ MAE in gaze estimation[3]. The system's scalability was demonstrated with a real-time average processing time of 8.4 seconds per 60- second video, while its practical value was affirmed by user feedback, with 88% rating the feedback as helpful and 76% reporting increased confidence[2,4].

References

[1]. Bhor, S.; Shinde, H.; Gupta, V.; Nair, V.; Kulkarni, M.; "Resume Parser Using Natural Language Processing Techniques", Int. J. Res. Eng. Sci. 2021, 9 (6), 1–6.

[2]. Kulkarni, T.; Pardeshi, Y.; Shah, Y.; Sakat, V.; Bhirud, S.; "App for Resume-Based Job Matching with Speech Interviews and Grammar Analysis: A Review", arXiv 2023, arXiv:2311.14729.

[3]. Karlsson, V.; "Concept of Interactive Video in Job Application", Linnaeus University: V¨axj¨o, Sweden, 2019.

[4]. Khan, K.K.; "Project Management App for Aspen Surgical Products", Grand Valley State University: Allendale, MI, USA, 2014.

[5]. Singh, R.; Yadav, H.; Sharma, M.; Gosain, S.; Shah, R.R. "Automatic Speech Recognition for Real-Time Systems", IEEE International Conference on Multimedia Big Data (BigMM), Dubai, UAE, 9–11 December 2019; pp. 189–198. https://doi.org/10.1109/BigMM.2019.000-0.

[6]. Das, P.; Acharjee, K.; Das, P.; Prasad, V.; "Voice Recognition System: Speech-toText" , J. Appl. Funct. Sci. 2015, 1 (2), 191–195.

[7]. Chen, J.; Zhang, J.; O'Flaherty, J.; Dusek, J.; Lee, J. ; "Fairness in AI-Based Hiring: A Survey" ; Hum. Resour. Manage. Rev. 2022. https://doi.org/10.1016/j.hrmr. 2022.100859.

[8]. Geyik, O.; Robbins, M.; Kunapuli, G.; Andrus, B.; "Mitigating Fairness Issues in Candidate Search" . In RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems, Vancouver, BC, Canada, 2–7 October 2018; pp. 440–444. https://doi.org/10.1145/3240323.3240389.