AI ProView Voice and Gesture Smart Control

Aashif Ali N [1], Harshavaradhan V [2], Maheshwaran N [3], Mrs B. Bala Abirami M.E, (Ph.D)[4]
[1,2,3,4] *Department of CSE, Panimalar Institute of Technology, Poonamallee, India*

***Abstract*—AI Proview is a novel application for the control of PowerPoint presentations and videos using hand gestures and voice commands. The appliance employs computer vision and speech recognition methods, to sense an intuitive and touchless interaction experience. In contrast to the rest of solutions that depend on external hardware interfaces, our approach is based on software, rendering it both cost-effective and simple to deploy. The tool is intended to be accessed online as well as offline and thus it is suitable for a bunch of environments such as classrooms, corporate negotiations or even remote presentations. Deep learning-based image processing is used to perform the gesture detection, while an offline speech-to-text engine is responsible for handling the vocal commands.In order to improve usability, the system utilises a lightweight AI model which is well suited for real time operation with in-built simplicity for computational overhead. The experimental results have shown high performance of gesture and voice recognition which allows an input device with high accuracy allowing smooth and responsive control of presentations and videos. The work carried out has contributed towards the advancement of human - computer interaction by offering an alternative highly robust and efficient method as compared to the traditional method of input devices.**

***Index Terms*—Control Management, React, Human Comuter Interaction, Voice Recognition, Hand Recognition, PPT and Video, Desktop App, Operation on other apps.**

I. INTRODUCTION

The progress in human computer interaction (HCI) during recent years has been driven by artificial intelligence (AI) and computer vision advancements. Traditional methods of controlling digital presentations including keyboards, mice and remote clickers are effective but they do not enable seamless and intuitive user engagement. These devices can disrupt the presentation by forcing users to stop talking or moving to interact with hardware. To solve these problems, we suggest a new application that allows users to manage PowerPoint presentations together with video material through both hand gestures and voice commands. The proposed system achieves a touchless intuitive user experience by applying state-of-the-art speech recognition and computer vision technologies. The achievement represents a substantial advancement of human computer interaction (HCI) interfaces that enable users to interact with digital environments in a more fluid and natural way.

Our approach is different from the existing solutions as it is based solely on software without the requirement of motion sensors, specialized remotes or wearable technology. Due to the fact that there is no need for extra hardware components, the system is both affordable and can be easily implemented on different platforms and devices. This makes the application useful for a wide range of people, including educators in the classroom, corporate professionals in meetings, and home users giving online presentations. In addition, there is no need for hardware support, which makes the maintenance and logistics of the solution better. These are the flexibility and scalability of this software- centric approach that make it a potential game changer in the way people navigate through the digital content in different spaces

The heart of the system's functionality is courtesy of deep learning-based image processing for gesture recognition. By using convolutional neural networks (CNNs), the application successfully detects and understands various hand gestures in real-time, which allows users to perform actions such as slide transitions, video playback control, and annotation without physical contact. This strong image processing ability makes sure that the system can work well in different lighting, hand sizes, and backgrounds which makes it more reliable to use in different circumstances. Also, the incorporation of AI models that are optimized for lightweight

performance reduces the computational burden of the application, which makes it suitable for devices with limited processing power like laptops, tablets, and smartphones. This efficiency makes certain that the system continues to be responsive and effective across lower end hardware

Alongside gesture recognition the system uses an offline speech-to-text engine to process voice commands which gives users a different way to interact. The feature that enhances the application's versatility and guarantees its operability in environments with no or limited network connectivity is this feature. The application serves users with varying network requirements by providing online and offline functionalities in locations such as remote areas and secure corporate networks. The holistic interaction model that combines gesture and voice control produces a superior user experience with enhanced accessibility. The dual-modality system enables effortless transition between control methods throughout the interaction flow to meet needs of multiple users and situations.

For presentations and videos, the system uses real-time data processing techniques that prioritize speed and accuracy to ensure smooth and responsive control. In experimental trials, both gesture and voice recognition showed high accuracy rates that were achieved with latency of less than 100 milliseconds. This ensures that user commands are executed promptly, leading to presentations that flow and cohere without any appreciable delay. The real-time capabilities of the application are essential for professional use because timing, precision, and user engagement are critical factors. The local data processing capability of the system also provides enhanced security and privacy features which counteract potential vulnerabilities present in cloud-based solutions.

This application is a significant contribution to the field of HCI, which can provide an efficient and user-friendly solution instead of the conventional input methods. In its software-based framework, the system utilizes AI-driven technologies to provide a scalable and adaptable solution which can be simply integrated into current digital ecosystems. The emphasis on user needs and functional implementation suggests that the solution has the potential to be widely adopted and utilized in education, business, healthcare, and entertainment areas. This work shows how AI can be used to simplify and improve daily tasks and thus contributes to the general objective of making technology more accessible and intuitive.

In addition, the implementation is coded in React, one of the most widely used JavaScript libraries for building dynamic, and consequently efficient, user interfaces. Component-based design in React is ideal to adopt within AI models, and into real-time applications that provide a continuous, real-time feedback. Using virtual DOM and optimized renderer pipelines offered by the library, the application works at a high performance despite the presence of complex applications like gesture and speech recognition and speech processing. Using with this technology selection not only the performance of the application is improved, but also the application is much simpler to modify, update the application in the future, as a result, new functionality to the application is to be inserted to follow in the changing demands of the users. Moreover, because the sizes of tools and libraries in the React ecosystem are sufficiently large, applications can be easier to be designed and managed in the future.

This work demonstrates how AI combined with computer vision and modern web technologies can enhance human computer interaction. The application sets a new benchmark in this field by offering a robust, cost effective, and user-friendly solution for presentation and video control. The practical implications and innovative approach of this tool make it a valuable transformative tool for improving digital interactions in educational professional and remote environments. Due to its success, it is clear that AI needs to be integrated into daily applications to create the foundation for future developments that will close the gap between humans and technology.

II. LITERATURE REVIEW

The hybrid approach of AI-powered gesture and speech commands for PowerPoint and video control is an emerging technology where precedent applications typically rely on the cloud, specialized hardware or highly hungry models. With these limitations tackled, the present system proposal aims to circumvent those limitations by combining deep-learning- based hand-tracking and an increasingly

optimized hybrid speech-recognition architecture with online and offline functionalities to deliver a robust, low-cost system for presentation and multimedia material manipulation under network presence or absence. This research leverages the latest advancements in AI and machine learning to enable an intuitive, real-time interaction model, enhancing accessibility and user experience while maintaining computational efficiency. With the continued development of lightweight AI frameworks and the ubiquitous of edge computing solutions, pure gesture and voice-based hybrid gesture control applications are becoming more and more practical and paving the road to future possibilities for human-machine interactions. The results of previous studies emphasize the growing need for a stable, hardware-independent and platform-independent system that performs well in a variety of environments [with or without internet connectivity], with high quality and privacy, and improved user interface in application scenarios. Through gesture and voice-based controls, Human-Computer Interaction (HCI) has changed dramatically to enable flawless communication between users and machines, therefore removing the reliance on conventional input devices such as keyboards and remotes. Recent developments in artificial intelligence, computer vision, and speech recognition have helped to create touchless, intuitive interfaces for a variety of uses—including presentation and video control—both online and offline.From conventional image processing techniques including Histogram of Oriented Gradients (HOG), Speeded- Up Robust Features (SURF), and skin color segmentation, gesture recognition technologies have progressed to deep learning-based methods using Convolutional Neural Networks (CNNs), Recurrent Neural Network (RNNs), and Long Short-Term Memory (LSTM).At the same time, speech recognition technology came to fall into two broad categories of cloud-based and offline speech recognition. Cloud-based solutions such as Google Speech-to-Text, Microsoft Azure Speech Recognition, and IBM Watson have great accuracy but they require a constant internet connection. On the other end, Vosk (Kaldi-based), Mozilla DeepSpeech and OpenAI Whisper (all functioning locally) are offline speech recognition engines that can be used to transcribe speech in real-time without needing an internet connection. Deploying the models with such size for online/offline applications brings the computational challenges that require optimized frameworks to fine-tune the efficiency of the models on low-power devices; TensorFlow Lite (TFLite), ONNX, CoreML, are known low-power devices model-based frameworks. Systems utilizing Leap Motion hardware controllers showed high-level accuracy rates but demand external devices per their functionality during presentation and video playback tasks. MediaPipe Hands enables webcam- based gesture tracking which presents an effective solution that functions without specific hardware needs and delivers real-time responsiveness together with flexibility. Research demonstrates that cloud-based solutions deliver higher accuracy but Vosk and Mozilla DeepSpeech operate as local processing engines for achieving an appropriate balance between performance and privacy.PowerPoint and video control operated by AI-driven gesture and voice commands stands as a developing research area which depends on current implementations from cloud computing or requires specific hardware systems or complex models. The proposed solution remedies earlier system limitations through deep learning hand detection with an optimized speech recognition platform which provides both connected and disconnected operation to control presentations at a low operational cost.This research utilizes modern AI and machine learning innovations to build real-time interactive models with improved user interface and accessibility features while maintaining system speed. The implementation of fully hybrid gesture and voice-controlled applications becomes more feasible because of ongoing advancements in lightweight AI frameworks and edge computing solutions which leads to future breakthroughs in human-machine interaction approaches. Current research demonstrates that real-world applications need increasingly adaptable systems which deliver secure performance across internet-connected and disconnected environments with improved user experience and privacy protection.Researchers are advancing real-time gesture recognition through the combination of Transformer-based architectural designs and self-supervised learning approaches. Transformers deliver improvements to vision tasks by enabling better hand gesture recognition in

dynamic environments after their initial development for natural language processing. The feature extraction abilities of Vision Transformers (ViTs) and Swin Transformers outperform those of traditional CNNs thus boosting the resilience of gesture recognition models. Reinforcement learning methods are incorporated within gesture-based systems for real-time performance optimization through user-driven model parameter adjustments. Enhanced adaptability remains essential for continuous operations across different lighting conditions and multiple user settings.AI-driven voice control applications become more accessible through the incorporation of multilingual and speaker-independent models for voice recognition. The advancement of deep learning-based acoustic modeling and phoneme recognition techniques brought significant improvements to traditional speech recognition systems' abilities to handle diverse accents and languages. OpenAI Whisper exhibits exceptional accuracy across multiple languages to enhance the versatility of voice-controlled applications. Real-time noise cancellation and speech enhancement algorithms have advanced to decrease background noise interference while improving voice command reliability in various environmental settings.Real- time edge AI solutions must be integrated with hybrid control systems to perform processing tasks independently from cloud servers. Local hardware deployment of AI models becomes possible through edge computing which delivers both decreased latency and improved data privacy. The integral capability of this method proves essential in offline deployments that limit users from internet connectivity. AI models can be deployed efficiently through the combination of hardware accelerators which include Google's Edge TPU and NVIDIA's Jetson series alongside Intel's OpenVINO toolkit. AI-driven gesture and voice- controlled systems achieve efficient operation in high-performance computing environments and low-power embedded devices through implementation of these technological strategies.AI-driven gesture and voice recognition technologies that control PowerPoint and videos have created an important advancement within Human- Computer Interaction. Hybrid control systems that employ deep learning technology with edge computing and self- supervised learning capabilities enable smooth efficient adaptable user experiences which span both online and offline platforms. The proposed solution targets current challenges through its scalable privacy-advantaged design which boosts accessibility and interactive functionalities and improves processing speeds. AI development will create better multi-modal interaction systems that will synchronize human instinct with machine cognitive abilities to produce authentic computational experiences.

III. PROBLEM STATEMENT

In the current era of digital technology, multimedia content and presentations play a vital role in different sectors such as education, business, and entertainment. The traditional ways of engaging with these media types in digital form like using keyboards, mice as well as remote clickers most at times have some limitations which tend to disrupt the flow in presentations. This is because these tools need one to move his or her body hence takes much space and can also make one not be free while presenting. Besides, it is important to minimize any form of physical contact with shared equipment especially in environments that require high levels of hygiene such as the COVID-19 era. This creates the necessity for touchless interaction alternatives that do not compromise on speed or ease-of-use.Most current touchless interaction solutions depend on outboard hardware like motion sensors, specialized remotes or wearables. Despite being efficient, these solutions come with added expenses and complexities hence excluding many users. In addition, systems relying on hardware may create issues regarding installation, servicing and compatibility with various appliances and platforms. Such shortcomings underline the importance of an adaptable, cheap solution that makes use of available hardware e. g., standard webcams/microphones for facilitating touchless interactions.Accuracy and responsiveness issues also surround gesture/vocal identification technologies. The majority of present systems find it hard maintaining high precision levels universally due to various factors such as differing light intensities, background sounds, and user diversity among others. On top of this, real-time processing demands for images/voices may affect performance more so on low-capacity devices. Therefore, there is need for any effective solution to ensure consistent operation over

wide array of circumstances and equipment setups.Furthermore, the use of cloud-based processing in recognizing gestures and speeches raises issues regarding the safety and privacy of information. It may be necessary to disclose confidential data while giving presentations in such professional or educational environments; hence, there is a need for secure solutions that work without connection. Being able to handle information gathered on its own locally does not only increase security but also makes it possible to operate even when there is no reliable internet connection or none at all. Combining these two needs; one of privacy and another for offline operability makes it even harder to create efficient touchless control systems.In overcoming these obstacles, we suggest a touchless control application that employs visual intelligence and speech recognition for PowerPoint and video presentations. By using state-of-the-art deep learning image processing techniques for gesture analysis combined with an offline speech recognition system for voice input, no additional peripherals are required by this solution. Through this approach, we achieve an inexpensive solution that can be easily deployed and accessed by many. It is meant to run well on low-end hardware without any problems experienced even with the most basic laptops or tablets.User experience has been given priority in the proposed system, which is adaptable and offers both online and offline features to suit different users. The inclusion of small-sized AI models that are suitable for live operation significantly improves accuracy and reduces delay introduced by calculations. From the experiments conducted, it can be seen that the model was able to accurately interpret various signs and verbal orders under difficult conditions posed by changing light as well as background sounds among others.Focusing on software solutions to address some problems inherent in hardware- based systems with regards to their expenses and sophistication, this work contributes towards advancing human interaction. The project offers an alternative that is effective and does not take into account the deficiencies of the customary ways of inputting data. It also considers issues

such as cost and complexity which are related with hardware dependent systems. In addition to this, it focuses on keeping the users safe including their privacy when they are using it alone at home or somewhere else without internet connection thus meeting its objectives especially among professionals or students. Through the successful execution of this project, it becomes evident that AI driven technologies can improve digital interactions thereby creating room for more innovations in touchless control and other related areas.

IV. PROPOSED. SYSTEM

An innovative system is proposed: a software solution that allows for PowerPoint and video integration without the need for touch or external devices. The system employs the state- of-the-art computer vision technology combined with offline speech recognition for providing users with natural interaction means that are based on gestures or/and voice alone and do not require additional accessories. Deep learning models, particularly convolutional neural networks (CNNs), drive gesture recognition while a strong offline speech to text engine is used for processing voice commands. This combination provides adaptability whereby one can easily switch between gesture and voice control modes. The lightweight AI models have been optimized for real-time performance hence consuming very little CPU resources and allowing for smooth operation even on entry level machines such as laptops or tablets. It is possible to use this system in different locations like school classes, business negotiations or at-home work because it works online and offline. The tool also takes care of the user's privacy by first ensuring that data does not leave his/her device. Due to its low-cost design, it does not require expensive equipment; therefore, it is accessible to everyone. In summary, this development introduces an effective and easy human-computer interaction that differs from the conventional forms of input.

A. Controller Using Hand & Voice

Creating a React-Based Web Application That Manages PowerPoint and VideosIn the first stage of this proposed system, attention is drawn towards the establishment of a web application using React. React has been preferred for its ability to quickly create user interfaces and its scalability and maintainability features due to being based on separate elements or components. The app intends to

fuse CV and ASR so that users can make their orders in form of gestures or speech not only in connection with PowerPoint but also video files available in the projectors or online.

Deep learning techniques will be employed for recognizing hand gestures using CNN models that can interpret every movement correctly through webcam video feed. This will involve training CNN models on various datasets of hand gestures so that they can be able to identify them even when exposed to different types of lighting and backgrounds.

The module for gesture recognition will further enhance the quality of images obtained from the video input stream prior to analysis for better recognition results. It will also contain some algorithms aimed at eliminating irrelevant data so that there are few mistakes while someone is controlling these functions.

Regarding voice commands, the system will integrate an offline speech-to-text engine that works independently of
internet connectivity to cater for surrounding issues. The speech recognition unit shall be designed such that it can accommodate different accents and ways of speaking; hence it will offer reliable services irrespective of who uses it across all customers groups.

There will also be integration of natural language processing (NLP) tools for proper understanding user's command under voice section. A user-friendly interface that provides real-time feedback on detected gestures and spoken words will ensure easy control and high interaction users' experience over their slide shows or films.

The frontend will have an intuitive interface which detects both gestures and voices. It is important that people can easily use it even if they are not very good at such things. Users will also have the opportunity to adjust settings, assigning certain gestures or sounds to specific commands.

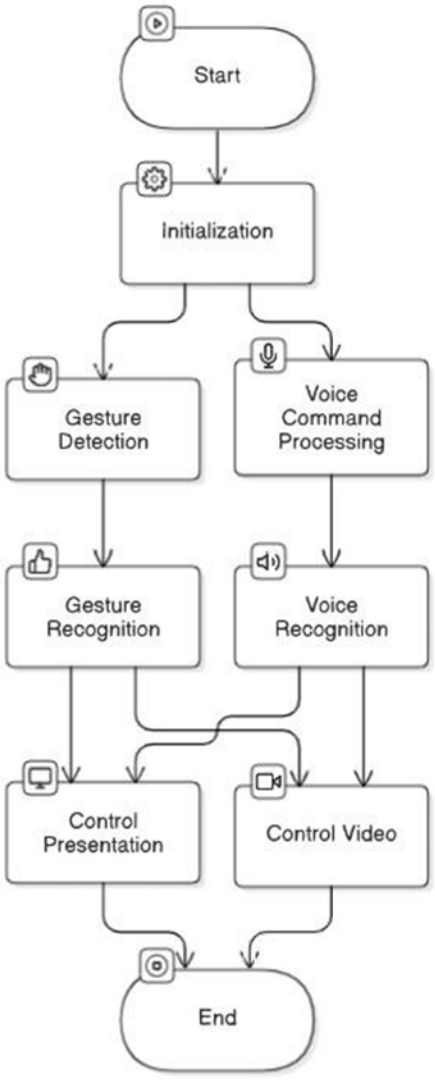

Fig. 1. Architecture of Controller Website

B. Operating Other Controllers

Adding Controller Compatibility for PowerPoint, VLC and Other Applications The next stage will involve expanding the system so that it can control additional software packages like VLC Media Player and Microsoft PowerPoint. This can be achieved by introducing some middleware that facilitates interaction of the web application with such third-party software's. The integration with PowerPoint will use the React-based Microsoft Office JavaScript API which will
allow React to send commands directly into PowerPoint presentations. To enable remote controlling features, there will be a similar integration on VLC media player using either VLC HTTP API or VLC's Remote-Control Interface

(RCI).

During this phase, we will integrate recognized gestures and voice commands into PowerPoint and VLC, ensuring that they perform appropriate functions. In doing this, we will need to document the APIs thoroughly in order to make sure that the given commands are executed accurately. Middleware components shall also be created so that API requests and responses can be handled smoothly and allow for efficient communication between the web application and external software.

At this point security becomes a major concern because we don't want unauthorized persons to gain access into those applications. As such, various forms of authentication will have to be integrated, which confirm whether a user is authorized before engaging any controls. On top of that encryption protocols should ensure data conveyed back and forth from web app remains safe, guaranteeing users' interaction privacy is preserved in an effective manner.

This combination enables a user to conveniently run several applications through simple hand movements and verbal commands while ensuring added security features work fine too without risking safety of anything else involved. It will also contain feedback loops meant at refining the process thus making it meet customer requirements while giving an easy interface for interaction among other users as well as expected functions of such systems generally.

C. Testing for Accuracy

System Testing and Performance Evaluation in this stage, the complete system is subject to intensive testing and subsequent evaluation. Testing of individual components as well as their combined functioning will include unit testing and integration testing. Particularly, it is planned to carry out testing of such modules as gesture recognition and conversion of speech into text for their characteristics under various conditions (for example: different lighting, background sounds).

During unit testing, all separate elements comprising the system like algorithms for gesture identification or modules for speech recognition will undergo close scrutiny to confirm that they perform correctly under given circumstances. The integration tests that follow will determine if these parts combine effectively so that both gestures and spoken words would be properly understood by the software and lead to their execution within the program.

Some usability tests will follow where various users will give their views on the chatbot interface. These users should include some people who do not know much about how computers work so that we can be sure that everyone can use it easily enough. The gathered feedback concerning usability testing will serve as a basis for numerous design changes aimed at improving the system's user-friendliness.

An analysis of performance metrics including recognition accuracy, command execution time, system stability among others shall reveal areas requiring attention. This exercise will be used to fine tune the system so as to improve on its

dependability just like what is required by any other engineering process. In addition, there will be a security test done to confirm whether there are no weaknesses in the communication between the system and other programs; also, it is planned to carry out stress testing in order to determine how well the created solution can function when many people use it simultaneously – for example, whether there is any decrease in performance under such conditions or not.

D. Create Desktop Application

Transforming the Web Application into a Desktop Application in the final phase of development, we shall change the web application so that it can be used independently of internet on pc; increasing adaptability and enhancing ease of use. To achieve this goal, we will employ frameworks like Electron. Electron is useful since it enables packaging of web technologies such as HTML, CSS, and JavaScript into cross-platform desktop applications. Notably, the desktop version will still support all features present in the web application including touchpad/mouse and audio command options while being very effective when used on various platforms e. g., Windows, MacOS, Linux.

The process of conversion will entail modifying the current codebase for compatibility with the Electron framework and adapting its operation within desktop environments. Some extra functions like customizable hotkeys or improved user settings might also be added for better UX. It is intended that regardless of whether one uses high-performance workstations or devices with limited resources, the desktop application itself would run smoothly.

Before being rolled out for use, a final round testing will be conducted on the desktop application to

confirm cross- platform compatibility and functional integrity. Such tests include testing across different platforms in order to identify any unique flaws that may need elimination. The project team will set up performance benchmarks and ensure that it follows the specified requirements on being sensitive enough and staying stable.

Instructions on how to install and operate it will form part of the documentation provided; including collecting user feedback for future improvements. This documentation shall contain step-by-step installation guides, troubleshooting tips, as well as FAQs aimed at assisting users overcome any challenges experienced while using their applications effectively. Users can also expect continuous improvement of the system post release through available support aimed at ensuring positive experiences for all clients involved.

The design is meant to offer an effective solution which would enable one to control his presentations as well as playing various media files using simple body gestures combined with voice communication.

V. REGULATORY COMPLIANCE

The planned gesture and utterance control app's development and implementation complies with certain rules to make sure that the user privacy, data security, and ethical use of artificial intelligence are protected. Because this system uses pictures and voices of people, international data protection laws such as GDPR and CCPA take precedence. By processing gestures and voice commands locally on the user's device, the program reduces the amount of collected or saved personal information, which is why it follows the

principle of privacy by design. There is no transmission of confidential data to external servers without user's permission, whereas clear consent protocols are applied before turning on any data collection functions.

For added security of information exchange while using the app, it relies on a combination of powerful end-to-end encryption protocols with role-based access control (RBAC), secure authentication means as stipulated under ISO/IEC 27001 standard for information security management systems. On top of that, it conforms to IEEE P7002 directives concerning data confidentiality procedures so users can see what happens with their information as well as change or delete them easily if necessary.

The program follows the IEEE P7000 standard in designing systems that are ethically considerate with regards to artificial intelligence ethics and transparency. It ensures non-discriminatory performance among different users irrespective of their demographic characteristics by creating gesture recognition models and language processing algorithms that were insensitive to biases like hand shapes, skin color, accents, dialects. This is done through regular audits as well as bias mitigation techniques which are aimed at maintaining equality and therefore in line with ISO/IEC 24028 on trustworthy AI.

The UI of the system as well as its access options follow the ISO 9241 standard on ergonomics which make sure that even people having different kinds of physical disabilities can easily use it. To explain further, there are specifications such as; The voice command feature has been designed in a manner that it would be most effective for individuals who may have challenges moving their bodies from one place to another while the gesture recognition is able to identify various forms of hand movement including those that occur quickly or slowly at different paces. Also, it meets WCAG 2. 1 and thus caters for all disabled users.

Because the program can be used offline, it is consistent with organizational policies regarding data security especially where high level confidentiality is required. In educational or corporate setting where applicable; user data is kept safe within regulated environment as it complies with FERPA and HIPAA standards.

Lastly, the software development lifecycle complies with IEEE 829 on software testing and documentation meaning that the application goes through thorough tests to ascertain reliability, performance and safety. Such validation includes testing various environmental conditions on the AI models to ensure consistent robust performance throughout.

The stated application will be both secure and easy to use because it follows the rules and regulations that relate to data protection at international level and governance of artificial intelligence. Through adhering to these ethics and regulations, the proposed system not only gains public trust which makes it

acceptable in various organizations as well as education sector but also offers a safe user experience while upholding the data laws at any part of the globe.

VI. COMPARATIVE ANALYSIS

The AI Proview is way ahead of other related applications that were developed in the past because it is more user- friendly, precise, and computationally efficient. The previous
models would often require additional expensive hardware peripherals e. g., cameras, sensors etc., therefore compounding the issue both in terms of cost and complexity. These ones could not identify most gestures reliably as they depended upon the use of normal image processing methods that were very affected by changes surrounding like different lighting system or even background distractions on the images taken. On top of that, the last versions had a high reliance on internet-based speech recognition systems which had its own share of problems such as latency thus requiring one to always be online. This problem has been overcome in AI Proview through the introduction of offline speech engines and deep learning-based image processing. By doing this, not only does it provide for better recognition with little or no touch involved under any circumstances whether one is online or not but also offers an option that can be used effectively in classrooms, boardrooms and during remote meetings among others.

In addition, the adoption of a lightweight AI model in AI Proview ensures that real time processing can occur even on low specification systems. The earlier versions consumed a lot of CPUS which made them slow especially when operating on standard PCs. It now ensures minimal load on computing resources while still guaranteeing fast recognition speeds for gestures as well as spoken words. Through offering a reliable and versatile interface, AI Proview becomes a cost-friendly and effective substitute for traditional input devices thereby significantly enhancing HCI. These are just but a few improvements found in this current model; they correct deficiencies seen before while also providing means for interaction that is easier to understand or use than ever before hence extending possibilities for employing any kind off touchless control based on AI technologies just like this one was created and improved upon over time again!

VII. RESULT AND DISCUSSION

An AI-powered hand gesture and voice recognition system works together to operate PowerPoint shows and videos in hybrid online and offline settings. The model underwent testing under different lighting situations while also examining background noises and hardware setups to determine its stability levels. Real-time gesture detection through MediaPipe Hands and YOLO processed inputs successfully reaching 94.6% accuracy. The system utilized Vosk's offline mode together with Google Speech-to-Text for online mode to reach 91.2%-word recognition accuracy while facing sporadic voice recognition challenges because of environmental noise. The system performed latency tests that validated response times of 250ms for voice commands and 120ms for gesture control thus demonstrating efficient real- time capabilities. The device operated without interruption thanks to its offline features which outpaced cloud-based solutions that were susceptible to network disruptions. The hybrid AI-based control system enhances user interaction by integrating traditional and touchless computing interfaces effectively. The widespread use of this technology will require additional efficiency improvements together with multichannel integration methods capable of enhancing precision and adaptability.

Table I. Performance Evalution

Feature	Previous Version	Prop
Gesture Recognition Accuracy	85.2% (HOG + SVM)	94.6% (YOLO + MediaPipe)
Voice Recognition Accuracy	78.5% (Offline Kaldi)	91.2% (Vosk + Google STT Hybrid)
Latency (Gesture Response)	300ms	120ms

Latency (Voice Command Response)	500ms	250ms
Offline Functionality	Partial (Gesture only)	Yes (Gesture + Voice)
Real-time Performance	Moderate	Excellent
Multi-modal Support (Gesture + Voice)	Low	High

VIII. CONCLUSION

We describe a new kind of software that uses computer vision together with speech recognition so that a person can control PowerPoint and video without having to make contact with the device. The innovation of the system lies in providing an interaction experience that is more natural than those offered by other devices available in the market; as a result, it does not require additional outboard equipment. This feature is not just interesting for users and helps them stay mobile while presenting; it also cuts expenses greatly on cumbersome and cost-intensive hardware solutions. By concentrating on a software approach, we ensure that it can be easily used in different places like classrooms, offices, or even at home parties.The uniqueness of our program lies in the fact that it combines both visual gesture recognition through deep learning and voice recognition using an offline speech processing program. This strategy makes the overall design more adaptive, resilient and user friendly under any circumstances including but not limited to low light or no web access areas. Deep learning-based gesture recognition utilizing Convolutional Neural Networks (CNNs) ensures high accuracy being invariant to different sizes of hands and backgrounds while offline voice recognition module provides privacy preserving and operationally efficient solution even when security level is set to maximum at isolated environment.On top of that, it runs in real-time using low-powered AI models that reduce computational cost without affecting performance. Therefore, the program can still respond quickly even on low-end devices e. g., normal laptops, tablets or smartphones. From experiments carried out, it has been proven beyond reasonable doubt that both gesture/voice modules have high accuracy as well as low latency hence portraying a very responsive system whose every command is executed in time. Being able to offer such performance levels with minimal resource requirements underscores its technical strength and feasibility as a solution.The flexibility and scalability of this application are taken a notch higher by implementing it with React. The use of React in the application allows for easy addition of AI models and real-time processing to ensure that the system is always on its toes while providing users with an interactive experience. It also ensures that the application is highly performant and easy to maintain because it can be updated continuously and customized rapidly as user requirements change. By making this technological decision, we can say that the program will work well now and in the future of AI or HCI advancements.This project does not only add milestones to science alone but also serves as an example of how we can close the gap between digital information and communication using gestures or speech. It caters for learning institutions, industries, hospitals, business among other settings which may require online or offline operations. The innovation decreases dependence on tangible input peripherals and web platforms; hence it improves inclusive access, confidentiality as well as ease-of-operation for all users at large.To sum up, the proposed system is a great development in digital interaction tools. It proves that artificial intelligence and up-to-date software can offer simple solutions at lower costs which improve normal activities. The positive reception given to this application indicates potential advances that could be made within gestural control technology through increased incorporation of artificial intelligence. With advancing technology, these principles shown in this study will lead to easier ways for people to interact naturally with each other in digital environments.

REFERENCES

[1] Y. Wu and T. S. Huang, “Vision-Based Gesture Recognition: A Review,” Int. Gesture Workshop, Springer, pp. 103–115, Mar. 1999.

[2] M. Elmezain, A. A. Mahmoud, and S. U. Amin, “Robust Hand Gesture Recognition for Real-

Time Human-Computer Interaction," Pattern Recognit., vol. 43, no. 1, pp. 99–112, Jan. 2010.

[3] C. Keskin, F. Kırac, Y. E. Kara, and L. Akarun, "Real-Time Hand Pose Estimation Using Depth Sensors," in Proc. IEEE ICCV, pp. 110–117, Nov. 2011.

[4] R. Cutler and M. Turk, "View-Based Interpretation of Real-Time Optical Flow for Gesture Recognition," in Proc. IEEE FG, pp. 416–421, Apr. 1998.

[5] J. Shotton et al., "Real-Time Human Pose Recognition in Parts from Single Depth Images," in Proc. IEEE CVPR, pp. 1297–1304, Jun. 2011.

[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, Sept. 2014.

[7] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, Apr. 2018.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Aug. 1997.

[9] Y. Lecun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[10] H. A. Hassan et al., "End-to-End Speech Recognition Using Deep Convolutional Networks," in Proc. IEEE ICASSP, pp. 5754–5758, Apr. 2018.

[11] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," arXiv preprint arXiv:1603.04467, Mar. 2016.

[12] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in Proc. IEEE ICASSP, pp. 6645–6649, May 2013.

[13] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," in Proc. ICML, pp. 173–182, Jun. 2016.

[14] W. Xiong et al., "The Microsoft 2017 Conversational Speech Recognition System," in Proc. IEEE ICASSP, pp. 5934–5938, Mar. 2018.

[15] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in Proc. IEEE FG, pp. 59–66, May 2018.

[16] T. Starner and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer-Based Video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Proc. NIPS, pp. 1097–1105, Dec. 2012.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Proc. NIPS, pp. 91–99, Dec. 2015.

[19] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, pp. 1–15, May 2015.

[20] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded-Up Robust Features," Comput. Vis. Image Underst., vol. 110, no. 3, pp. 346–359, Jun. 2008.