International Journal of Recent Research and Review, Special Issue 2, 2026

ISSN 2277 – 8322

Jarvis: A Hybrid Multimodal Intelligent Assistant for Spatial Computing Using Edge AI and Large Language Models

Ipsita Das[1], Anshul Goyal[2]

[1]Department of Computer Science & Information Technology, Apex University, Jaipur

[2]Department of Computer Science & Information Technology, Apex University, Jaipur

ipsitadas0313@gmail.com, anshagrawal147@gmail.com

ABSTRACT: The fast development in Human-Computer Interaction (HCI) has lifted emphasis from conventional graphical user interfaces to multimodal and intelligent systems proficient of understanding natural human inputs such as speech, gestures, and vision. This paper benevolences the design and enlargement of a multimodal AI-powered desktop assistant, named Jarvis AI Assistant, which integrates Large Language Models (LLMs) [5], edge-based computer vision, and augmented reality (AR) to empower real-time, hands-free interaction with computing systems. The planned system operates a layered Python-based architecture uniting Cohere Command-R7b for semantic intent classification, Google MediaPipe [2] for hand tracking, YOLOv8 [3] for object detection, and Google Gemini for multimodal reasoning. Contrasting traditional assistants, the system ties digital and physical environments complete spatial computing, permitting features for example holographic projection and gesture-based control by standard consumer hardware. The research highpoints a hybrid edge-cloud computational approach to poise latency and performance though maintaining real-time responsiveness. Experimental assessments demonstrate better-quality efficiency in desktop automation, spatial alertness, and user communication. This work donates toward the expansion of self-sufficient, context-aware computing systems and arranges the foundation aimed at future wearable spatial computing ecosystems.

Keywords — *Human-Computer Interaction, Spatial Computing, Large Language Models, Augmented Reality, Edge Computing, YOLOv8, MediaPipe, Multimodal AI, Desktop Automation*

I. INTRODUCTION

The prototype of Human-Computer Interaction (HCI) is undertaking a significant conversion as computing systems progress from traditional input

ISSN 2277 – 8322

mechanisms for occurrence keyboards and mice near natural interaction modalities with speech, gestures, and vision. This shift is driven by developments in Artificial Intelligence (AI) [5],[10], Machine Learning (ML) [5],[10] and Natural Language Processing (NLP) [5],[10] which empower machines to understand and respond to human resolved more instinctively. Conventional desktop systems familiarize friction in user interaction owing to reliance on bodily input devices, warning efficiency in multifaceted systems. Consequently, there is an increasing need for smart systems that can connection the gap amid human cognition and machine accomplishment.

The Jarvis AI Assistant is future as a multimodal system intended to participate speech recognition, computer vision, and spatial computing [7] into a combined desktop environment. The system purposes to destroy traditional contact walls by permitting hands-free switch and real-time environmental focus. Dissimilar present assistants that purpose inside a two-dimensional interface, Jarvis familiarizes spatial computing capabilities by projecting three-dimensional holographic objects in contradiction of a live camera feed, in that way integration physical and digital collaboration spaces.

Alternative perilous motivation late this research is the democratization of progressive technologies such as increased reality and multimodal AI. Conventionally, such technologies compulsory particular hardware comparable AR headsets or LiDAR sensors, preventive accessibility. This work validates that comparable aptitudes can be realized using standard consumer hardware, for example webcams and CPUs, over efficient algorithm design and edge computing techniques.

In addition, current virtual assistants lack the skill to make composite multi-step tasks linking both digital and physical frameworks. They principally trust on predefined command structures and cloud-based processing, which announces latency and confines autonomy. The projected system talks these challenges by participating a Decision-Making Model (DMM) motorized by an LLM [10], permitting dynamic task routing and implementation across multiple domains for instance web automation, system control, and visual analysis.

II. RELATED TECHNOLOGIES

The development of the Jarvis AI Assistant is armoured by plentiful developing technologies spanning AI, computer vision, and system computerization. Separately technologies contribute to unlike layers of the system architecture, permitting all-in-one multimodal interaction.

Large Language Models (LLMs) [1],[5] demonstrates a dominant role in sanctioning

ISSN 2277 – 8322

natural language considerate and decision-making. Models such as Cohere Command-R7b afford contextual reasoning competences, allowing the system to catalogue user interrogations into actionable sorts. These models interchange beyond traditional rule-based systems through leveraging deep learning architectures to interpret determined and produce structured outputs.

Computer vision frameworks such as OpenCV [4] and MediaPipe [2] are important for enabling spatial mindfulness and gesture acknowledgment. MediaPipe [2] bids real-time pointer chasing by snooping 21 landmark facts on the human hand, which can be cast-off to guess spatial manages and permit gesture-based communications. OpenCV [4] matches this by supervision of image processing and interpretation of operations obligatory for augmented authenticity conception.

Object recognition is appreciated by the YOLOv8 [3], [8] architecture, which is familiar aimed at its quick and actual performance. YOLO (You Only Look Once) [3], [8] processes entire images in a sole pass, assembly it is apposite for applications demanding low latency. In this system, YOLOv8 [3], [8] is used to sense objects inside the camera frame and create bouncing boxes for spatial study.

As well, cloud-based multimodal AI systems such as Google Gemini augment the system's cognitive abilities by empowering deep image analysis and related reasoning. Even though edge computing [6] knobs real-time tasks, cloud services are applied for computationally intensive operations, generating a hybrid architecture that poises competence and routine.

III. LITERATURE REVIEW

The progression of virtual assistants [10] consumes been lengthily studied through primary systems concentrating on elementary speech acknowledgement and command performance. IBM's Shoebox noticeable one of the initial efforts at speech-based collaboration, identifying an incomplete terminology. Though, the field has meaningfully grew with the overview of machine learning and neural networks, important to the expansion of contemporary assistants like Siri, Alexa, and Google Assistant. These systems chiefly depend on cloud-based buildings for dispensation user contributions and causing comebacks.

A foremost revolution in NLP derived through the primer of transformer architectures [1], predominantly emphasized in the effort "Attention Is All You Need" by Vaswani et al. (2017). Transformers [1] empowered models to detention long-range dependences in text, foremost to the expansion of Large Language Models proficient of circumstantial reasoning. These models have altered virtual assistants on or after humble command processors hooked on intelligent

ISSN 2277 – 8322

conversational agents accomplished of multi-turn connections.

3-D computing has also perceived noteworthy progressions, predominantly with the overview of agendas like MediaPipe [2]. Research via Lugaresi et al. (2019) established the possibility of real-time discernment pipelines using machine learning, permitting applications like hand tracking and gesture recognition. These expansions have flagged the technique aimed at participating spatial consciousness hooked on computing systems deprived of needful dedicated hardware.

Added vital zone of research is the combination of AI hooked on automation systems. Topical studies highpoint the notion of autonomous agents that can remark, reason, and act indoors an atmosphere. By means of compounding LLMs done tool practice abilities, these agents can achieve compound tasks like web scraping, code cohort, and system switch. The Jarvis AI Assistant shapes upon this perception by executing independent web agents proficient of implementing real-world tasks grounded scheduled user contribution.

The deliberation amongst edge computing and cloud computing [6] is likewise protruding in the fiction. Edge computing bids near to the ground latency and real-time dispensation, constructing it apposite for applications such as computer vision and AR [7]. In dissimilarity, cloud computing affords tall computational power for multifaceted intellectual tasks. The amalgam tactic adopted in this study bring into line with prevailing studies that supporter compounding equally examples ideal system presentation.

IV. PROPOSED METHODOLOGY

The anticipated system tracks a segmental and coated architecture considered to guarantee scalability, efficiency, and real-time routine. The design is separated addicted to three crucial layers: input processing, decision-making, and execution. Respective layer is applied using dedicated technologies to knob precise errands.

The input processing layer seizures user relations complete multiple modalities containing voice, vision, and gestures. The acoustic pipeline devours the Pvporcupine engine for disconnected wake-word uncovering, authorizing user concealment and dropping network addiction. When initiated, the Speech Recognition module renovates spoken input addicted to text for supplementary dispensation.

The decision-making layer is motorized through the First Layer Decision Making Model (DMM), which customs the Cohere Command-R7b model to order user enquiries hooked on predefined groupings such as universal queries, real-time hunts and system commands. This semantic directing mechanism substitutes outdated rule-

ISSN 2277 – 8322

based systems thus empowering self-motivated and bendable interaction.

The performance layer knobs task execution over an amalgamation of indigenous and cloud-based components. Aimed at spatial computing the system engages a 3D prognosis engine that customs mathematical conversions to condense holographic objects against the user's hand. This contains crucial 3D directs, smearing spin matrices and prominent them against a 2D plane using viewpoint projection practices.

Cutting-edge equivalently, the computer vision subsystem usages MediaPipe [2] aimed at hand tracking and YOLOv8 [3] intended for object detection. These components activate in real-time, aiding the system to investigate the physical environment and retort accordingly. Designed for advanced analysis, arrested images are determined and administered by means of the Gemini API, permitting the system to make high-level cerebral tasks.

Towards safeguard system receptiveness, asynchronous programming and multithreading are engaged. The system divorces the audio listening hoop from the graphical interface, averting delaying operations and preserving horizontal performance. This design qualifies immediate performance of compound tasks, ornamental complete productivity.

V. SYSTEM TESTING AND PERFORMANCE EVALUATION

The assessment of a multimodal AI [9] system like the Jarvis AI Assistant wants a comprehensive challenging policy that enthusiasms outside conservative input-output endorsement. Due to the incorporation of auditory dispensation, computer vision, and cloud-based cognitive, the system must be evaluated beneath real-time limitations to guarantee firmness, openness, and accurateness. The testing organization assumed in this work chains unit testing, integration testing and presentation benchmarking to authenticate respective module individualistically as well as inside the complete design.

Unit testing stood lead to validate the precision of precise mechanisms primarily the auditory pipeline, decision-making model, and computer vision components. The wake-word recognition system originated via the Pvporcupine engine unproven tall reliability beneath unreliable environmental situations. Tests relating diverse reserves and background noise levels inveterate that the system steadily sensed the activation keyword through nominal expectancy and unimportant false positives. This approves the achievement of edge-based audio indulgence in conserving user pleasure although promising directness.

ISSN 2277 – 8322

The First Layer Decision Making Model (DMM), motorized by the Cohere Command-R7b model was estimated for semantic sending truth. Test cases involved formal requests, real-time evidence requests and multi-command information. The results designated that the model efficaciously confidential user determined into suitable functioning categories with high steadiness. Remarkably, the system confirmed the aptitude to crumble multifarious commands into numerous executable tasks, emphasizing the usefulness of LLM-based [5] direction-finding over outmoded rule-based systems.

Computer vision challenging attentive on the constancy and correctness of hand chasing and object uncovering modules. The MediaPipe [2] outline effectually recognized pointer breakthroughs in real time, unfluctuating underneath diffident gesticulation and brilliance differences. The Euclidean distance-based increasing device sanctioned bouncing resizing of predictable holograms ensuring straight spatial proclamation. Moreover, YOLOv8-based [3] object detection continued real-time routine with frame rates appropriate for collaborating applications.

Integration testing was achieved to appraise the communication amongst diverse modules within the system. The architecture employments a dual-thread model unscrambling the auditory dispensation loop from the graphical interface. This design guarantees that high-latency actions, like cloud API calls, do not restrict with real-time translation. Testing recognized that the scheme continued steadiness under instantaneous enactment through no vital thread influences or system clatters perceived.

Latency testing unadorned that edge constructed events like wake-word detection and entity racing unstated near-instantaneous reply periods. In alteration, cloud-based procedures presented classy latency owing to network overhead and dispensation period. Aimed at instance, multimodal image analysis consuming the Gemini API mandatory around 2–3 seconds, vindicating the practice of asynchronous completing to avoid system obstructive. This hybrid method safeguards that latency-sensitive tasks are fingered locally, though computationally rigorous tasks are divested to the cloud.

Resource consumption summarizing recognized that the system occupations knowledgeably on characteristic consumer hardware. Memory indulging tolerated unchangeable through prolonged rehearsal, through no suggestion of memory leakages. The usage of subprocesses aimed at computationally concentrated tasks guarantees that system resources are achieved efficiently, stopping performance squalor over time. These

ISSN 2277 – 8322

consequences authenticate the option of locating progressive multimodal AI schemes happening non-specialized hardware podiums.

VI. RESULTS AND DISCUSSION

The untried assessment of the Jarvis AI Assistant highpoints its efficiency in permitting multimodal interaction and attractive user output. The system effectively participates voice commands, gesture recognition and visual analysis into an incorporated interface, indicating a substantial perfection over old-fashioned desktop assistants. The capacity to attain composite tasks over natural language input diminishes handler effort and updates workflow enlargements.

One of the important donations of this effort is the application of spatial computing on typical hardware. The augmented authenticity hologram engine offers a visually immersive involvement by prominent 3D objects against the user's hand. This competence validates the possibility of joining computer vision and mathematical modelling to generate communicating environments deprived of the necessity for dedicated AR devices.

The incorporation of LLMs hooked on the system empowers progressive intellectual and decision-making competences. Contrasting outmoded assistants that count on predefined commands, the Jarvis system construes user intent dynamically consenting for supple and context-aware communication. This tactic brings into line with fresh trends in AI research where LLMs are castoff to produce intelligent agents accomplished of carrying out multifaceted tasks unconventionally.

Conversely, the system similarly appearances certain confines. The dependence on webcam-based input limits the field of view, restraining the efficiency of spatial tracking. Moreover, cloud-based processes familiarize latency and be contingent on network connectivity which may disturb performance in offline environments. These trials climax the need for supplementary optimization and the consideration of substitute hardware solutions.

Regardless of these boundaries, the hybrid architecture of the system affords a well-adjusted solution that influences the powers of together edge and cloud computing. Through indulgence time-critical chores locally and allocating intricate cognitive to cloud-based models the system accomplishes an optimum trade-off among enactment and functionality. This designed methodology can attend as a draft for forthcoming multimodal AI systems

Table 1 presents a comparative analysis of the planned system with outmoded virtual assistants across significant parameters.

ISSN 2277 – 8322

TABLE I

COMPARATIVE STUDY OF JARVIS AI ASSISTANT BY MEANS OF OLD-STYLE SYSTEMS

Serial No.	**Feature**	**Jarvis AI Assistant**	**Traditional Assistants**
1	Communication Manner	Speech + Signal + Revelation	Speech/Text solitary
2	Style	Cross Platforms	Cloud Oriented
3	Three-D Adding	Affirmative	Not any
4	Actual Gesture Controller	Certainly	Not at all
5	Object Sighting	YOLOv8-based	Not obtainable
6	Decision Making	LLM-based	Rule-based/inadequate AI
7	Dormancy (Home-grown Tasks)	Very Low	Judicious
8	Inactivity	Uncertain	High
9	Computer hardware Commitment	Usual computer	Average Device
10	Multimodal Ability	High	Low

VII. CONCLUSIONS

This research grants the design and implementation of a multimodal AI assistant that assimilates natural language processing, computer vision, and spatial computing hooked on a united desktop environment. The projected system validates that progressive AI competences, usually related with dedicated hardware can be realized by average consumer devices through efficient software design and hybrid computing strategies.

The Jarvis AI Assistant successfully reports the confines of outmoded virtual assistants through permitting real-time communication through together digital and physical surroundings. The usage of LLM-based decision-making tolerates flexible and context-aware mission accomplishment, although the incorporation of computer vision empowers spatial alertness and gesture-based control. These configurations mutually increase user considerate and efficiency.

The research correspondingly highpoints the position of sectional architecture and asynchronous implementation in emerging

ascendable AI systems. By dint of decoupling, dissimilar mechanisms and fetching multithreading, the system conserves receptiveness unfluctuating below substantial computational consignment. This methodology agreements that the assistant remains purposeful and operative in real-world situations.

In decision, the Jarvis AI Assistant implies a substantial step near the awareness of intelligent, multimodal computing systems. The outcomes of this study pay to the increasing field of spatial computing and afford a base for forthcoming research in AI-driven human-computer collaboration.

REFERENCES

[1] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.

[2] C. Lugaresi et al., "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.

[3] G. Jocher et al., "Ultralytics YOLOv8," GitHub Repository, 2023.

[4] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.

[5] Tom B. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, 2020.

[6] Mahadev Satyanarayanan, "The Emergence of Edge Computing," Computer, IEEE, 2017.

[7] R. Azuma, "A Survey of Augmented Reality," Presence, 1997.

[8] J. Redmon et al., "You Only Look Once," CVPR, 2016.

[9] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.

[10] B. Shneiderman, "Designing the User Interface," Pearson, 2010.