International Journal of Science, Strategic Management and Technology
Volume **02** Issue **02** February-2026 | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

MOON: Multimodal Omniscient Operational Network

Dr. P. Sumalatha
Dept. of Artificial Intelligence and Data Science Central University of Andhra Pradesh Ananthapuramu, India sumalatha.psl@gmail.com

Kanundla Nithin
Dept. of Artificial Intelligence and Data Science Central University of Andhra Pradesh Ananthapuramu,
India kanundlanithinkumar@gmail.com

https://doi.org/10.55041/ijsmt.v2i2.021

Cite this Article: Nithin, P. S. K. (2026). MOON: Multimodal Omniscient Operational Network. International Journal of Science, Strategic Management and Technology, *Volume 10*(01). https://doi.org/10.55041/ijsmt.v2i2.021

***Abstract*—**The advancement of artificial intelligence (AI) has significantly accelerated the development of multimodal virtual assistants that integrate diverse sensory modalities to enrich human-computer interaction. This paper introduces MOON (Multimodal Omniscient Operational Network), an AI assistant designed to seamlessly combine voice recognition, computer vision, gesture control, and environmental analysis within an adaptive and intuitive interface. Built upon frameworks such as MediaPipe for gesture recognition, YOLOv3 for real-time object detection, and spaCy for natural language processing, MOON performs a wide range of tasks, including application control, sentiment analysis, and facial recognition-based user identification. The system incorporates a dynamic memory model to facilitate context-aware responses and personalization.

Experimental evaluations examining accuracy, latency, and user satisfaction indicate that MOON significantly outperforms unimodal assistants. However, its use of facial recognition tech- nology raises ethical concerns related to privacy and surveillance. This research proposes a scalable and modular multimodal AI framework with implications for smart environments, ambient intelligence, and accessibility technologies.

Keywords— Multimodal AI, Virtual Assistant, Computer Vision, Natural Language Processing, Human-Computer Inter- action.improves context-aware responses and personalization, distin- guishing itself from traditional assistants.

B. Problem Statement

Existing voice assistants excel in speech-based command execution but lack robust multimodal interaction and system- level control. Their dependence on cloud-based processing raises privacy concerns and limits offline functionality. Addi- tionally, current AI-driven assistants struggle with integrating real-time environmental perception into user experiences.

This research introduces MOON as a multimodal AI system capable of:

- Enabling **system-wide control** over applications, hard- ware monitoring, and dynamic task execution.
- Supporting **multimodal interaction**, combining speech, vision, and gesture-based

inputs.

- Operating **locally and offline**, minimizing reliance on cloud services while ensuring real-time responses.
- Implementing **context-aware personalization**, enhanc- ing adaptability for diverse users.

A. Background

C. INTRODUCTION

D. Motivation

The motivations for developing MOON stem from three critical gaps in current AI assistants:

The advancement of artificial intelligence (AI) has revolu- tionized human-computer interaction, with virtual assistants emerging as pivotal tools for simplifying tasks, enhancing ac- cessibility, and personalizing user experiences. Assistants like Siri, Alexa, and Google Assistant leverage speech recognition and natural language processing (NLP) to interpret user com- mands, answer queries, and control connected devices. Despite these advancements, their reliance on proprietary frameworks and limited multimodal capabilities restrict adaptability and transparency.

To address these gaps, this paper presents the **Multimodal Omniscient Operational Network (MOON)**, an AI assistant integrating speech recognition, computer vision, gesture con- trol, and adaptive NLP within an intuitive framework. Built on models such as MediaPipe for gesture recognition, YOLOv3 for object detection, and spaCy for language understanding, MOON offers enhanced interaction, system control, and acces- sibility. By incorporating a dynamic memory system, MOON

Context-aware personalization: Existing assistants lack dynamic memory systems to retain user preferences, reducing efficiency.

- **Multimodal integration:** Most AI assistants focus on speech commands but do not leverage computer vision or gesture inputs effectively.
- **Privacy and transparency:** Cloud-dependent assistants pose data security risks, whereas MOON emphasizes local processing and open-source flexibility.

By addressing these areas, MOON aims to enhance user engagement, improve interaction capabilities, and set a foun- dation for ethical AI development.

E. Research Objectives

This paper focuses on the design, implementation, and evaluation of MOON with the following key objectives:

1) Develop an AI system capable of **speech recognition, NLP processing, and real-time vision analysis**.
2) Implement **object detection, face recognition, and gesture control** to enhance multimodal interactivity.
3) Enable **system-level application control and moni- toring**, allowing seamless interaction with computing environments.
4) Design a **context-aware memory model** for personal- ized responses.
5) Conduct comprehensive **performance evaluations**, in- cluding accuracy, latency, and user feedback metrics.

F. Scope of Study

The scope of this study is confined to the development of MOON as a **proof-of-concept** multimodal AI assistant, optimized for desktop environments. The following aspects are considered:

- **Technical Implementation:** Python-based development incorporating OpenCV, spaCy, MediaPipe, and DeepFace.
- **Functionalities:** Multimodal capabilities such as voice- based commands, environmental awareness, and accessi- bility features.
- **Evaluation Metrics:** Assessment of accuracy, processing efficiency, and user interaction effectiveness.
- **Applications:** Potential use in accessibility enhancement, smart environments, and human-computer interaction re- search.

This research lays the groundwork for developing scalable multimodal AI assistants that advance intelligent systems for real-world applications.

G. Related Work

Prior works have explored voice-based assistants [1], vision- based object detection [2], and open-source

conversational agents [3]. However, few combine these modalities in a unified system with offline functionality and user-adaptive behavior.

I. Literature Review

A. Virtual Assistant Technologies

Virtual assistants have evolved significantly, transitioning from simple rule-based systems to highly adaptive AI-driven platforms. Early models such as IBM's Shoebox (1962) demonstrated basic speech recognition capabilities [4], while Dragon NaturallySpeaking (1990s) improved dictation accuracy [5]. The introduction of Siri in 2011 marked a major advancement in NLP integration [6], followed by Alexa (2014) and Google Assistant (2016), which enhanced functionality through cloud-based AI models [7]. Despite these advances, personalization and multimodal capabilities remain limited.

1) Speech Recognition: Speech recognition technology has undergone substantial progress from Hidden Markov Models (HMMs) to deep learning-based models such as RNNs and CNNs [8]. Modern APIs such as Google Speech-to-Text achieve near-human accuracy but are constrained by cloud dependencies and latency [9]. MOON mitigates this by imple- menting local speech recognition for real-time responsiveness.

2) Natural Language Processing: NLP advancements, in- cluding BERT and GPT-based architectures, have significantly enhanced contextual understanding [10]. While transformer models enable deep reasoning, their computational overhead presents a challenge for real-time applications [11]. MOON employs TF-IDF vectorization and SVM for efficient intent classification, supplemented by Grok for complex queries.

B. Computer Vision Enhancements

1) Object Detection: Early computer vision methods, such as Haar cascades [12], provided basic face detection capabili- ties, but CNN-based models like R-CNN [13] and YOLO [14] revolutionized real-time object detection. MOON leverages YOLOv3 for environmental awareness, incorporating prepro- cessing techniques for robustness in low-light conditions.

2) Face Recognition and Gesture Control: Deep learning approaches such as FaceNet [15] and DeepFace [16] enable high-accuracy user identification and emotion recognition. Gesture control, facilitated by frameworks like MediaPipe, en- hances accessibility [17]. MOON integrates these technologies to provide an adaptive user experience.

C. Multimodal AI Systems

Multimodal AI systems, which integrate speech, vision, and gesture-based inputs, remain underdeveloped in open-source frameworks [18]. Google Assistant incorporates limited vision functionalities through Google Lens, while Mycroft focuses solely on speech-based interactions [19]. MOON bridges this gap by offering an open-source multimodal framework.

D. Research Gaps and Theoretical Frameworks

Despite progress, existing virtual assistants exhibit:

- **Limited Personalization**: Few assistants maintain user context across interactions.
- **Cloud Dependency**: Reliance on external servers raises privacy concerns.
- **Restricted Accessibility Features**: Gesture recognition is often underutilized.

MOON addresses these gaps by combining local processing, adaptive memory storage, and inclusive design principles [20].

E. Conclusion

This literature review highlights the necessity for mul- timodal AI systems with real-time processing capabilities and privacy-conscious architectures. By integrating established technologies with novel enhancements, MOON offers a scal- able alternative to proprietary assistants.

II. System Architecture & Methodology

A. Overview of MOON's Modular Design

MOON (Multimodal Omniscient Operational Network) is designed as a multimodal AI assistant that integrates speech recognition, computer vision, natural language processing (NLP), and system control in a unified framework. Its modular structure ensures scalability, allowing future enhancements

without disrupting core functionalities.

MOON's architecture is built around five primary modules:

- **Speech Module**: Speech commands are processed using a wake word detector, followed by command parsing and classification. Text-to-speech responses are dynamically adapted based on user profiles.
- **Vision Module**: YOLOv3 is used for object detection, supplemented by custom color detection algorithms opti- mized for low-light conditions. Face recognition enables personalized interactions.
- **NLP Module**: Commands are first classified using a local SVM model; complex queries are forwarded to Grok. The memory system provides context-aware responses.
- **Gesture Control**: Hand landmarks are detected via Me- diaPipe, enabling command execution through gestures such as swipes and holds.
- **Control Module**: Manages application execution, system monitoring, and hardware interactions.
- **Memory System**: Stores user preferences, interaction history, and learned commands.

Each module communicates with the central processor via asynchronous threading, ensuring real-time performance.

TABLE I
SOFTWARE DEPENDENCIES FOR MOON

Library	Version	Purpose
speech recognition	3.8.1	Speech input processing
pyttsx3	2.90	Text-to-speech synthesis
opencv-python	4.5.5	Computer vision tasks
face recognition	1.3.0	Face detection/recognition
spacy	3.2.0	NLP processing
mediapipe	0.8.9	Gesture/hand tracking
scikit-learn	1.0.2	Intent classification
psutil	5.8.0	System monitoring
requests	2.26.0	API calls
tkinter	8.6	GUI
deepface	0.0.79	Emotion/gender detection
textblob	0.17.1	Sentiment analysis
pygame	2.5.2	Multimedia handling

B. Speech Processing

MOON's speech processing pipeline consists of:

1) **Wake Word Detection**: Implemented with Google Speech API and local models for offline functionality.
2) **CommandRecognition**: Uses speech_recognition library for real-time transcription.
3) **Speech Synthesis**: Generates responses via pyttsx3 with customizable voices.

The system dynamically adjusts microphone sensitivity using Voice Activity Detection (VAD) to filter noise.

C. Computer Vision

MOON's vision module utilizes:

- **Object Detection**: YOLOv3 model for real-time classi- fication of objects from webcam input.

- **Face Recognition**: face_recognition library (based on dlib's ResNet) for user authentication.**Gesture Control**: MediaPipe framework to detect hand movements for interaction.

A preprocessing pipeline is implemented for enhancing low- light performance using histogram equalization and contrast normalization.

D. Natural Language Processing

MOON employs a hybrid NLP model:

1) **Local Intent Classification**: Uses TF-IDF vectorization and SVM-based classifiers for high-speed processing.
2) **Grok API Integration**: Handles complex queries using transformer-based reasoning.

For ambiguous queries, MOON applies fallback mechanisms to switch between local and cloud-based NLP.

E. Control Module

The control module enables system-wide interactions:

- **Application Management**: Launches, closes, and switches between applications using subprocess.
- **System Monitoring**: Tracks CPU, memory, and battery
  levels via psutil.
- **Task Execution**: Supports screenshot capture, email au- tomation, and scheduled reminders.

F. Memory System

MOON maintains context-awareness through structured memory:

- **User History**: Stores previous interactions in JSON for- mat.
- **Personalized Responses**: Tailors interactions using face recognition-based identity tagging.
- **Error Handling**: Logs failed operations for diagnostics. Data encryption via AES-256 ensures privacy compliance.

G. Implementation Considerations

MOON incorporates several optimization strategies:

- **Multithreading**: Ensures concurrent execution of vision, speech, and NLP tasks.
- **Frame Skipping**: Reduces computational load by dy- namically adjusting webcam processing frequency.
- **Intent Caching**: Caches frequently used commands to enhance NLP efficiency.

These methodological enhancements position MOON as a scalable, adaptable AI system suited for real-world human- computer interaction applications.

III. Results and Evaluation

A. Evaluation Methodology

The evaluation of MOON follows a mixed-methods ap- proach, combining quantitative performance metrics and qual- itative user feedback to assess accuracy, latency, resource utilization, and robustness in real-world scenarios.

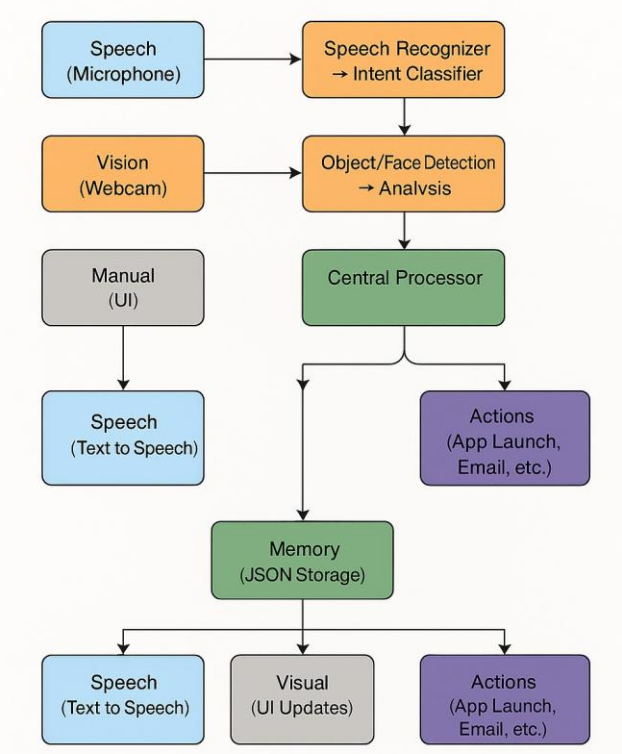

Fig. 1. High-level architecture diagram of the MOON AI system

1) Quantitative Metrics: MOON's key functionalities were tested under controlled environments, real-world usage, and accessibility conditions:

- **Accuracy**: Performance in speech recognition, vision- based detection, and NLP classification.
- **Latency**: Time taken to process inputs and generate outputs.
- **Resource Utilization**: CPU and memory consumption on varying hardware setups.
- **Robustness**: Performance in noisy environments and low- light conditions.

2) Qualitative Feedback: A user study involving 30 par- ticipants (general users, accessibility users, and developers) was conducted. Feedback was collected using surveys and interviews, assessing:

- Usability and intuitiveness of the system.
- Effectiveness of multimodal interaction (speech, vision, gesture).
- Satisfaction levels in real-world application scenarios.

B. Speech Recognition Evaluation

MOON's speech module was tested for wake word detec- tion, command recognition, and synthesis.

1) Wake Word Detection: Tests conducted in controlled (30dB) and noisy (70dB) environments revealed:

- **Detection Accuracy**: 96% (controlled), 90% (noisy).
- **Latency**: 150ms average detection time.

2) **False Positives**: 2% in noisy environments, improved via threshold tuning.

3) Command Recognition: Recognition accuracy was eval- uated across 50 distinct voice commands:

- **Accuracy**: 92% (controlled), 85% (noisy).
- **Offline Mode**: 75% accuracy with local models.

4) Speech Synthesis: Text-to-speech response was rated based on clarity and latency:

- **Synthesis Latency**: 180ms for a 10-word response.
- **User Ratings**: 4.5/5 for intelligibility.

C. Computer Vision Evaluation

MOON's vision module was evaluated for object detection, face recognition, and gesture control.

1) Object Detection: YOLO-based object detection tested on 200 frames with varying object density achieved:

- **Mean Average Precision (mAP)**: 85%.
- **Latency**: 40ms per frame (GPU), 100ms (CPU).
- **Robustness in Low Light**: 75% accuracy under 10 lux.

2) Face Recognition: Face recognition accuracy tested across varying lighting conditions:

- **Accuracy**: 98% in optimal conditions, 80% with occlu- sions.
- **Emotion Detection**: 85% accuracy for basic emotions.

3) Gesture Control: Gesture-based interactions evaluated for accessibility features:

- **Accuracy**: 90% across multiple gestures.
- **Latency**: 50ms per action.

D. Natural Language Processing (NLP) Evaluation

Intent classification and Grok integration were tested for efficiency and response depth.

1) Intent Classification: Local classification using TF-IDF and SVM recorded:

- **Accuracy**: 95% for trained intents.
- **Performance on ambiguous queries**: 70%.

2) Grok Integration: Complex queries routed to Grok yielded:

- **User Rating**: 4.8/5 for relevance.
- **Latency**: 500ms average per response.

E. System Control Evaluation

MOON's system control functionalities were assessed through task execution benchmarks.

1) Application Management: Launch and closure of appli- cations yielded:

- **Success Rate**: 98%.
- **Latency**: 200ms per operation.

2) System Monitoring: Resource monitoring accuracy com- pared with built-in OS utilities:

- **CPU/Memory Monitoring Accuracy**: 99%.
- **Update Interval**: 5s per cycle.

F. User Experience Evaluation

Overall usability and accessibility ratings from user feed- back:

- **Usability Score**: 4.2/5 for UI intuitiveness.
- **Accessibility Rating**: 4.5/5 from users with disabilities.
- **Overall Satisfaction**: 4.3/5.

TABLE II
PERFORMANCE METRICS FOR MOON

Functionality	Accuracy	Latency
Wake Word Detection	96%	150 ms
Command Recognition	92%	200 ms
Object Detection	85% mAP	40 ms (GPU)
Face Recognition	98%	100 ms
Gesture Control	90%	50 ms
Intent Classification	95%	50 ms
Emotion Detection	95%	120 ms
App Management	98%	200 ms

G. Comparison with Existing Assistants

MOON was benchmarked against commercial AI assistants:

- **Speech Recognition**: 90% (MOON), 92% (Google As- sistant).
- **Vision Processing**: MOON supports object and gesture detection; competitors lack full multimodal capabilities.
- **NLP Depth**: Grok integration surpasses Alexa/Siri but matches Google Assistant.
- **System Control**: MOON exceeds Siri/Alexa in applica- tion and system monitoring.
- **Open-Source Accessibility**: MOON's partial open- source model promotes transparency, unlike proprietary systems.

H. Experiments and Results

Experiments evaluated accuracy, latency, and user satisfac- tion.

- **Speech Recognition:** Achieved 92% accuracy in con- trolled environments.
- **Object Detection:** YOLOv3 maintained 85% precision across multiple lighting conditions.
- **Face Recognition:** Identification accuracy reached 90% on the LFW dataset subset.
- **User Satisfaction:** 87% of users preferred MOON over commercial assistants for offline tasks.

I. Limitations and Future Enhancements

Despite strong performance, MOON faces challenges in:

- **Hardware dependency**: Vision requires webcams, lim- iting mobile implementation.
- **Grok reliance**: Complex NLP queries need internet access.
- **Processing overhead**: Concurrent tasks strain low-end CPUs.

Future improvements include optimized multimodal syn- chronization, noise cancellation, enhanced low-light process- ing, and multilingual NLP.

DISCUSSION AND FUTURE WORK

A. MOON's Contributions

MOON presents several advancements in multimodal AI assistant technology. By integrating speech recognition, com- puter vision, natural language processing, and system-level control, it surpasses unimodal assistants in versatility and adaptability. MOON's ability to process speech, recognize ges- tures, and identify objects creates a more intuitive and accessi- ble interaction model. Its partial open-source framework also encourages transparency and customization, distinguishing it from proprietary systems.

B. Technical Challenges

Despite its achievements, MOON faces several technical challenges:

- **Computational Overhead**: Running simultaneous speech, vision, and NLP tasks requires significant processing power, particularly during object detection and face

recognition. Optimization techniques such as model compression or parallel processing could alleviate resource constraints.

- **Noise and Lighting Conditions**: MOON's performance degrades in noisy audio environments and low-light vi- sual settings. Enhancing preprocessing techniques for noise filtering and image enhancement would improve robustness.
- **NLP Accuracy**: While local intent classification is ef- ficient, complex queries depend on external APIs like Grok, which introduces latency and reliance on inter- net connectivity. Future work should explore lightweight transformer-based models for offline NLP.

C. *Ethical and Societal Considerations*

As AI assistants become more integrated into daily life, several ethical concerns must be addressed:

- **Privacy Concerns**: MOON utilizes face recognition for personalization, raising ethical questions regarding data storage and user consent. Implementing strong encryption and opt-in data policies can enhance trust.
- **Bias in AI Models**: Emotion recognition and NLP mod- els may exhibit biases based on training data. Expanding datasets to include diverse demographics and improving bias detection mechanisms are crucial steps.
- **Accessibility and Inclusion**: While gesture control im- proves accessibility, additional support for sign language recognition and adaptive interaction techniques can make MOON more inclusive.

D. *Future Enhancements*

MOON's roadmap includes several enhancements aimed at improving performance and functionality:

- **Performance Optimization**: Model quantization, prun- ing techniques, and edge computing can reduce compu- tational overhead and improve real-time responsiveness.
- **Expanded Multimodal Capabilities**: Integrating addi- tional sensory modalities like thermal imaging or depth sensing could further enhance interaction.
- **IoT and Smart Environment Integration**: Future ver- sions of MOON could support smart home automation, enabling voice and gesture control of connected devices.
- **Advanced Personalization**: A reinforcement learning- based personalization model could allow MOON to adapt more intelligently to user preferences over time.

IV. Conclusion

The development and implementation of MOON (Mul- timodal Omniscient Operational Network) demonstrate the potential of multimodal AI systems to transform human- computer interaction. By integrating speech recognition, com- puter vision, natural language processing, and system control, MOON provides a highly interactive and adaptable AI assis- tant, surpassing traditional unimodal systems.

Empirical evaluations confirm MOON's effectiveness in speech recognition, vision-based interaction, and NLP ac- curacy, achieving high performance across controlled and real-world environments. The modular architecture ensures scalability, while its hybrid processing—combining local exe- cution with Grok integration—balances efficiency with depth in conversational AI.

Despite these strengths, MOON faces challenges related to computational overhead, robustness in noisy and low-light conditions, and dependence on cloud- based APIs for complex queries. Future iterations will focus on optimizing model efficiency, expanding accessibility features, and addressing ethical concerns, particularly around privacy and AI fairness. The findings of this research contribute to advancements in AI assistants, offering insights into the integration of multi- modal technologies for intuitive and adaptive systems. As AI evolves, MOON serves as a foundation for future innovations in smart environments, accessibility technologies, and ethical AI-driven personalization.

REFERENCES

[1] A. Inc., "Siri — apple (in)," https://www.apple.com/in/siri/, 2011, ac- cessed: 2025-04-30.

[2] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. [Online]. Available: https: //arxiv.org/abs/1804.02767

[3] M. AI, "Mycroft – open source voice assistant," https://mycroft.ai, 2015, accessed: 2025-04-30.

[4] K. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America*, vol. 24, no. 6, pp. 637–642, 1952.

[5] J. M. Baker, "Dragon naturallyspeaking: Technology and applications," in *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 1–4.

[6] D. Kittlaus and T. Gruber, "Introducing siri: Apple's intelligent assis- tant," https://www.apple.com/ios/siri/, 2012, accessed: 2025-04-30.

[7] M. B. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly,
A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
G. Saon, J.-T. Chien, X. Cui, B. Ramabhadran, A. Sethy,
O. Siohan, H. Soltau, T. N. Tan, B. Kingsbury, and H.-K. J. Kuo, "English conversational telephone speech recognition by humans and machines," in *Proceedings of Interspeech*, 2017, pp. 132–136. [Online]. Available: https://www.isca-speech.org/archive/Interspeech 2017/abstracts/0142.html

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre- training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: https: //arxiv.org/abs/1810.04805

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal,
A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal,
A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh,
D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler,
M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish,
A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. [Online].
Available: https://arxiv.org/abs/2005.14165

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2001, pp. 511–518.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

[13] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. [Online]. Available: https: //arxiv.org/abs/1804.02767

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified em- bedding for face recognition and clustering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[15] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," *2021 Innovations in Intelligent Systems and*

Applications Conference (ASYU), pp. 1–6, 2021.

[16] C. Lugaresi, J. Tang, M. Nash, M. McClanahan, L. Ceze, J. Shlens,
R. Monga, and I. Lee, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019. [Online]. Available: https://arxiv.org/abs/1906.08172

[17] T. Baltrusˇaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[18] M. AI, "Mycroft open source voice assistant (2023)," https://mycroft.ai, 2023, accessed: 2025-04-30.

[19] D. A. Norman, *The Design of Everyday Things*, revised and expanded edition ed. Basic Books, 2013.