Development of Human Machine Interface for Humanoid Robot using vision and speech Interaction with AI

Abdullah Yahya Abdullah Omaisan 1
APcore (Asia Pacific Center of Robotics Enginnering),
Asia Pacific University of Technology & Innovation (APU)
Bukit Jalil, Malaysia
abuod0@gmail.com

Suresh Gobee 2
APcore (Asia Pacific Center of Robotics Enginnering),
Asia Pacific University of Technology & Innovation (APU)
Bukit Jalil, Malaysia
suresh.gobee@apu.edu.my

Ibrahim Sheikh Mohamed3
QSS AI & Robotics
Deparment of Research and Development
Riyadh, Saudi Arabia
ibrahim@qltyss.com

Ts.Vickneswari Durairajah4
Faculty of Engineering,
Masha univiversity
Kuula Lumpur, Malaysia
vickneswari@mahsa.edu.my

***Abstract*—As this humanoid robot designed to revolutionize human-robot interaction through the integration of vision and speech interaction with artificial intelligence (AI). With a focus on enhancing collaboration and safety, this project aims to provide a seamless interface between humans and robots in various industries. The project's purpose is to develop an advanced human-machine interface by implementing object detection, depth estimation, chat-bot, speech recognition, and pose estimation techniques. By leveraging AI technologies, this project offers numerous benefits, including improved efficiency, accessibility, and reduced development time and cost. The potential future applications of this project extend beyond its initial industry focus, with possibilities in healthcare, education, entertainment, and laboratory assistance. Through its comprehensive interaction capabilities and commitment to safety, this project sets the stage for transformative advancements in human-robot interaction and the integration of humanoid robots into everyday life.**

Keywords— Humanoids, Robotics, AI, pose estimation, Depth estimation, Human-Robot Interaction, Chatbot, Embedded AI.

I. INTRODUCTION

Humanoid robots are increasingly being developed to resemble human appearance and behavior, enabling interaction through vision, speech, and physical gestures [1]. These robots are widely applied in fields such as healthcare, education, and customer service, where they act as assistants, companions, or autonomous agents. The advancement of artificial intelligence has significantly enhanced the ability of humanoid robots to understand and respond to human input through multimodal sensing and perception [2], [3], [4]. Humanoid robots are expanding into diverse roles beyond traditional domains. They are being used in elderly care to assist with daily activities, in retail to guide and engage customers, and in museums as interactive tour guides [5]. Additionally, they support mental health therapy, perform household chores in smart homes, and serve as test platforms for studying human-robot social interaction and ergonomics.

This work presents the design and integration of an intelligent system for a humanoid robot, focusing on enhancing its visual understanding, voice interaction, and gesture-based control. The system incorporates an AI-based object detection module that enables the robot to identify specific objects in its environment using a standard 2D camera. In addition, monocular depth estimation is employed to infer scene geometry without the need for expensive depth sensors. A chatbot component, powered by natural language processing and speech recognition, enables the robot to understand and respond to verbal commands. Furthermore, the system includes real-time body and hand pose estimation to interpret human gestures, which are translated into corresponding robot movements.

The developed system is integrated into a humanoid robot platform, Engineering Lab Assistant (ELA2.0), to evaluate its performance in real-world scenarios. This integration allows for comprehensive interaction between the robot and humans without the need for physical control interfaces. The aim is to create a cost-effective and versatile solution that enhances the robot's autonomy and interactivity across multiple use cases. Such a system holds potential for deployment in smart environments, where intuitive and natural interaction between humans and robots is critical.

II. RELATED WORKS

A. Humaniod Robots: Control and Design

Chignoli et al. [6] present a comprehensive system for enabling acrobatic behaviors in humanoid robots, focusing on dynamic motions like flips and spinning jumps. The work introduces a novel humanoid design with specialized actuators capable of impulsive motion, experimentally validated using a custom dynamometer. These actuators are integrated into a kino-dynamic motion planner and simulator for accurate full-body motion planning. A hierarchical control framework combining model predictive control and whole-body impulse control ensures stable landings and responsive feedback. Simulations confirm the system's capability to perform complex motions such as backflips and spinning jumps.

Julia Berg and Shuang Lu [7] reviewed research on human-robot interfaces in industrial and service robotics, highlighting the growing importance of speech and gesture recognition in industrial settings. They emphasize the development of multimodal, intuitive interfaces to support safe and efficient human-robot collaboration. The study discusses projects like Sherlock, HR-RECYCLER, and KoBo34, which aim to enhance industrial human-robot interaction. The authors conclude that advancing interface technologies and involving users in the design process are key to improving interaction in both service and industrial robotics.

Topology optimization plays a vital role in the development of physical humanoid robots by enabling significant mass reduction without compromising structural integrity. As demonstrated by M. A. Allouzi et al. [8] in their work on Bento-Bot, applying topology optimization to the robot's electronics compartment led to a 41.28% reduction in mass and up to 51.9% lower energy consumption in its DC motors. Lighter materials like PEEK and ABS were selected to replace aluminum, while deformation remained negligible and the safety factor stayed within acceptable limits. This optimization not only improved energy efficiency but also reduced motor stress and operational costs—critical factors for mobile humanoid platforms.

B. Depth Estimation

Recent advances in monocular depth estimation have demonstrated promising results through innovative training strategies and network architectures. Ranftl et al. [9], [10] addressed the challenge of insufficient large-scale ground truth data by introducing MiDaS, a model trained on a combination of diverse datasets. Their work emphasized the importance of dataset mixing and proposed a novel loss function, achieving superior zero-shot performance across multiple benchmarks. The smaller MiDaS variant still outperformed competitors, underscoring the significance of their training approach over architectural size.

Complementarily, Godard et al. [11] explored self-supervised monocular training using stereo and monocular videos. They proposed Monodepth2, which introduced innovations such as minimum re-projection loss, multi-scale sampling, and automatic masking. This approach eliminated the need for dense depth ground truth while outperforming several supervised methods. However, limitations remain when visual assumptions, such as Lambertian surfaces, are violated—highlighting ongoing challenges in robust, generalizable depth prediction from 2D images.

C. Pose Estimation

Human pose estimation (HPE) identifies key body landmarks such as joints, hands, and facial features. Bazarevsky et al. introduced Convolutional Pose Machines (CPM), which use hierarchical deep learning to predict pose configurations from image data [12]. Traditional methods like Pictorial Structures and Flexible Mixture-of-Parts (FMP) represent body parts through rigid or flexible components. Evaluation metrics such as IoU, PCK, and AP help measure performance. However, these models often struggle with occlusions and complex poses. To address real-time needs, Josyula and Ostadabbas proposed BlazePose, a lightweight model that tracks 33 body key points efficiently on mobile devices, enabling applications like fitness tracking and sign language recognition. Similarly, Sung et al. developed a real-time hand gesture recognition (HGR) system using a single RGB camera and MediaPipe, allowing accurate gesture control for interfaces like VR or robotic systems [13]. These advancements highlight the growing role of fast, on-device HPE in human-computer interaction.

D. Chatbot

Generative models like DialoGPT leverage large-scale conversational datasets from Reddit to produce more natural, multi-turn dialogues [14]. Built on GPT-2's transformer architecture, DialoGPT employs maximum mutual information scoring to improve response relevance, but risks generating inappropriate content due to training data bias. Reinforcement learning is being explored to mitigate these issues. Keyner et al. combined dialogue systems with open data access using Rasa framework to assist non-expert users in querying government datasets via intent classification and dialogue management [15]. Tamrakar and Wani reviewed chatbot architectures and applications, noting their growing use in education, marketing, and CAD assistance, highlighting platforms such as Google Dialogflow, IBM Watson, and open-source tools like Rasa and ChatterBot [16].

III. METHODS

A. Sytem Overview

The proposed humanoid robot system is built around a low-power, AI-capable architecture that integrates visual and auditory interaction through deep learning models, controlled via Python-based frameworks. At the hardware level, the NVIDIA Jetson Nano (4GB) is selected as the central computing unit due to its efficient GPU-accelerated inference capabilities and compatibility with AI frameworks such as PyTorch. For perception, a 160° wide-angle camera (IMX219) is used to capture RGB frames for object detection, pose estimation, and depth prediction. Actuation is achieved using a combination of digital and analog servo motors, including a 25kg servo for neck rotation and high-torque CYS-S0650 55kg servos for shoulder movement, all powered via a 6V DC battery.

These are controlled through an I2C-based PWM driver to ensure reliable multi-servo communication. On the software side, Python was chosen for its rich ecosystem of AI libraries, with PyTorch deployed for deep model inference, OpenCV for image processing, and CVZone to interface with Google MediaPipe for skeletal pose and hand gesture estimation. The system includes lightweight AI models: YOLOv5n for real-time object detection, MiDaS and DPT-Small for monocular depth estimation, and a hybrid chatbot model combining NLP-based intent classification and a rule-based response engine, enhanced with Google Text-to-Speech for voice feedback. Table I provides a summary of all key hardware and components integrated into the system.

TABLE I. COMPONENTS LIST

Components List	
Component	**Description**
Jetson Nano (4GB)	Main AI compute module with GPU support
IMX219 Wide-Angle Camera	2D RGB camera for vision tasks
Servo Motors	25kg (neck), 55kg CYS-S0650 (shoulders), analog (arms)
PWM Driver (I2C)	For multi-servo control
DC Motors	For propulsion system
Mecanum wheels	For propulsion system

B. Working Principle

The humanoid robot operates as an integrated multi-modal system for real-time human interaction through vision,

gesture recognition, and voice communication as illustrated in figure 1. The workflow initiates with continuous RGB frame capture via a wide-angle 2D camera. These frames are processed by a YOLOv5n object detection model for scene identification and classification. To simulate natural head-following behavior, the system calculates detected objects' centroids, segments the image vertically, and rotates the robot's neck servo incrementally based on object position.

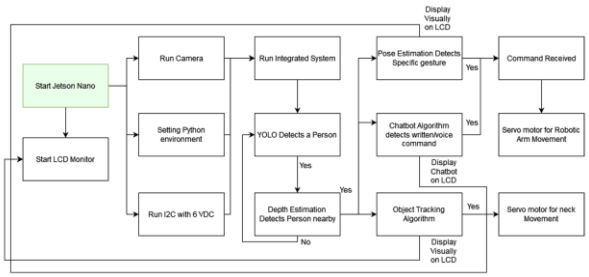


Fig 1. System flow chart

Concurrently, the same image feed undergoes monocular depth estimation using transformer-based models (DPT-Small or MiDaS-Small). These networks generate real-time depth maps from 2D images, enabling spatial understanding without dedicated depth sensors.

For pose estimation, Google MediaPipe (integrated via CVZone) extracts 33 body landmarks and 21 hand key points. Prediction accuracy is quantified using the Mean Per Joint Position Error (MPJPE) metric, which computes the average Euclidean distance between predicted 3D joints and ground truth positions across all joints. Lower MPJPE values indicate higher pose estimation fidelity. This data triggers servo-driven gestures based on recognized body/hand configurations.

In parallel, a hybrid chatbot processes voice/text inputs using an NLTK-trained neural network for intent classification. Recognized intents activate rule-based responses converted to speech via Google Text-to-Speech API.

C. Depth Estimation

Depth estimation models are trained on extracted 3D movie sequences, employing specialized loss functions to address inherent scale and shift ambiguities in monocular prediction. The approach centers on scale- and shift-invariant techniques that align predicted disparities with ground truth using least-squares optimization. This normalization ensures predictions are compared at zero translation and unit scale, while robust loss components prioritize significant errors (e.g., the largest 20% of residuals). The methodology operates primarily in disparity space, chosen for its compatibility with relative depth representations and its natural handling of inverse depth relationships, which mitigates unknown absolute scale challenges.

The final loss integrates two key components: a scale-invariant term and a Normalized Multiscale Gradient (NMG) function. The NMG ensures shift invariance by comparing gradient differences between rescaled predictions and ground truth across four distinct scale levels (K=4). These components are adapted to disparity space and combined using a weighting constant (α=0.5). This composite loss function enables robust depth estimation without requiring known absolute scales or sensor calibration, focusing instead on preserving accurate relative depth relationships and structural consistency across multiple resolutions.

D. Pose Estimation

The Mean Per Joint Position Error (MPJPE) measures 3D pose estimation accuracy by calculating the average Euclidean distance between predicted and ground truth joint positions. The equation is:

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} ||J_i - J_i^*||_2 \quad (1)$$

where:

- N is the number of joints
- J_i is position of the ground truth
- J_i^* is the humanoid robot joints

E. Object Tracking

1) Multiple Object Tracking Precision (MOTP)

This metric quantifies the localization accuracy of correctly matched detections over time. It calculates the average alignment error between ground truth bounding boxes and their associated tracked hypotheses:

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_i c_t} \quad (2)$$

where:

- c_t is the number matched for a specific time.
- d_t is the distance between the object and the corresponding hypothesis.

2) Multiple Object Tracking Accuracy (MOTA)

It measures the overall tracking performance by accounting for missed detections, false positives, and identity switches. It provides a comprehensive view of how well the tracking algorithm performs over time.

$$MOTA = 1 - \frac{\sum_{i,t}(m_t + fp_t + mme_t}{\sum_t g_t} \quad (3)$$

where:

- m_t is the number of missed targets.
- fp_t is the number of false positive.
- mme_t the number of mismatches or identity switches.
- g_t is the total number of ground truth objects.

3) IDF1

IDF1 Score evaluates the accuracy of identifying and maintaining object identities during tracking. It is defined as the ratio of correctly identified detections (true positives) to the average number of ground truth and predicted detections.

$$IDF_1 = \frac{2IDTP}{2IDTP+IDFP+IDFN} \quad (4)$$

where:

- $IDTP$ is the true positives

- *IDFP* is the false positives
- *IDFN* is the false negatives

IV. Result

A. Robot Design and Assembly

Although the main focus of this project does not lie in the robot's mechanical design, it is essential to include visuals of the hardware to provide a comprehensive understanding of its structure and components. The attached pictures in Figure 2 showcase the physical aspects of the robot, highlighting its design, materials, and overall appearance. While the emphasis remains on the system's working principles and performance, these images aid in visualizing how the physical elements—such as motors, servos, sensors, and embedded hardware—are integrated to support its functionality. This visual context bridges the gap between theoretical system descriptions and their practical implementation, enhancing reader comprehension of how various modules operate within the robot's structure.

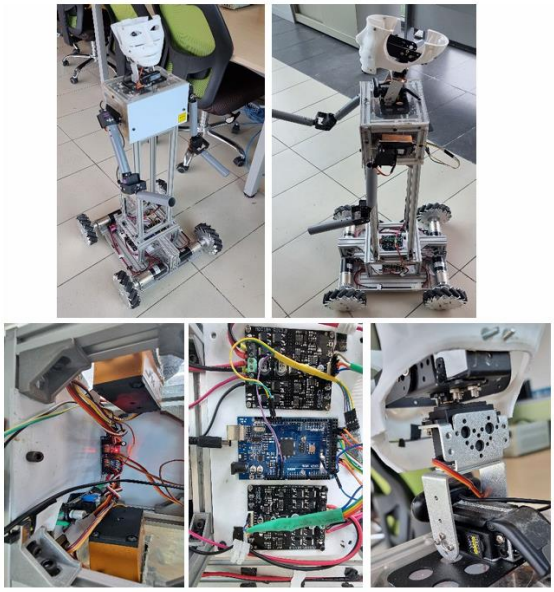
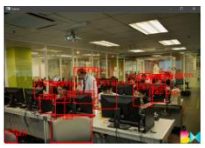


Fig 2. Robot Hardware Assembly

B. Object Detection Evaluation

The object detection component, using the YOLOv5 model, was tested on five images representing different environments to evaluate detection accuracy, inference speed, and non-maximum suppression (NMS) time. Tests were conducted on a CPU-only laptop with 8GB of RAM, which introduced limitations in processing power and led to noticeably extended inference durations. Despite these hardware constraints, the model successfully detected multiple objects—including persons, chairs, laptops, and cars—with confidence scores that demonstrate the model's ability to generalize across varied contexts. Each test image was carefully selected to represent different lighting conditions, object arrangements, and background complexities to assess model robustness. The inference times ranged from approximately 1100 ms to 1500 ms, with NMS times adding minor overhead. These results were essential not only for validating detection accuracy but also for identifying the most efficient YOLOv5 variant suitable for future deployment on embedded GPU platforms like the Jetson Nano.

Table II. Object detection result on Jetson board

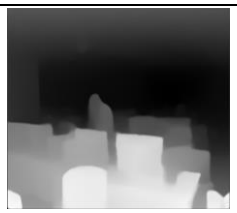
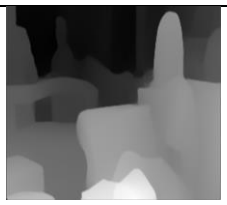
Object Detection				
Detected Object with accuracy		speed	NMS	IMAGE
1	chair 0.87 9x person 0.78 chair 0.74 3x laptop 0.72	1447.5ms inference	12.0ms NMS	
2	Car 0.83	1143.3ms	6.0ms NMS	
3	3x person 0.94 cup 0.82 car 0.61	1131.4ms inference	6.0ms NMS	
4	person 0.81	1478.2ms inference	9.0ms NMS	
5	laptop 0.75 4x person 0.71 chair 0.67 chair 0.6	1122.4ms inference	6.0ms NMS	

C. Depth Estimation Result

Pose estimation was evaluated using three sample images, each featuring a human subject. The MediaPipe-based system effectively identified key-points with varying degrees of accuracy, ranging from 72% to 98%. While full body pose estimation yielded reliable results, hand gesture detection is still under development. This feature will eventually be used for gesture-based command inputs as part of the system's human-robot interaction (HRI) interface.

Table III. Depth estimation Evaluation

Depth Estimation		
Accuracy compared to actual distance	Estimated Distance (CM)	IMAGE

1	600cm	estimated distance: 592	
2	150cm	estimated distance: 36	
3	300cm	estimated distance: 812	
4	100cm	estimated distance: 176	
5	500cm	estimated distance: 694	

D. *Pose Estimation Evaluation*

Pose estimation was evaluated using three sample images, each featuring a human subject. The MediaPipe-based system effectively identified key points with varying degrees of accuracy, ranging from 72% to 98%. While full-body pose estimation yielded reliable results, hand gesture detection is still under development. This feature will eventually be used for gesture-based command inputs as part of the system's human-robot interaction (HRI) interface. The key points were mapped in real time and visually validated on screen to ensure correct alignment with human joint positions. Additional training and fine-tuning are planned to improve detection stability under diverse lighting and background conditions.

TABLE IV. DEPTH ESTIMATION EVALUATION

Pose Estimation		
Detected Key points	Accuracy	IMAGE
1	13	84%	
2	31	98%	
3	10	72%	

E. *Chatbot Evalution*

The chatbot component, trained using a sequential model with 200 epochs on an NLTK dataset, was tested in beta stage. Despite being in early development, the system correctly recognized user intents and responded appropriately to input queries. The model produced accurate outputs with confidence scores exceeding 0.98 and response times ranging from 55 ms to 974 ms. These results indicate strong potential for robust natural language interaction once the model is fully optimized.

TABLE V. CHATBOT EVALUATION

Chatbot Evaluation				
Detected Query		Response	Acuraccy	Speed
1	Hi there [('greeting')] Query: Hi there	Bot: Hello, thanks for visiting APCORE	'0.9994'	974ms
2	How are you? [('Bot-status')] Query: How are you?	I am doing good.	'0.9838'	55ms
3	What is apcore doing? [('APCORE'] Query: What is apcore doing	It is an interesting center for Robotics	'0.9997'	60ms

F. *System Performance Consideration*

All tests were conducted on a CPU-only laptop environment, which lacks GPU acceleration. This significantly impacted inference time and system responsiveness. However, the final deployment target is the NVIDIA Jetson Nano, which features GPU support and

CUDA optimization. Transitioning to Jetson Nano is expected to yield up to a 96.4% performance improvement in AI inference. Despite its limited 4GB RAM, the Jetson platform is suitable for running lightweight AI models. Nonetheless, slower start-up and shutdown times are anticipated due to the cumulative load of running multiple models and managing real-time servo control.

G. Use Case and Deployment

The developed humanoid robot is designed to serve in interactive environments that require human-robot communication through speech, gesture, and vision. Its functionality makes it suitable for applications such as reception assistance, educational demonstrations, and guided interaction in institutional settings. With integrated object detection, pose estimation, and chatbot capabilities, the robot can perform tasks like greeting users, responding to verbal inquiries, and recognizing basic gestures to trigger corresponding actions. These features enable deployment in smart classrooms, tech exhibitions, or support roles in labs and innovation hubs.

The robot was successfully deployed and operated at APCORE (Asia Pacific Centre of Robotics Engineering) in Asia Pacific University (APU). During this on-site operation, the robot demonstrated stable performance in object recognition, tracking, and conversational response, validating both its hardware integration and software pipeline in a real-world academic environment. This successful demonstration confirms its readiness for broader institutional use and highlights its potential as an interactive AI-driven platform for research and education.

V. Conclusion

This project presented the development and initial testing of a humanoid robot capable of interacting with humans through vision, gesture, and speech using AI-based models. Key modules such as object detection (YOLOv5), depth estimation (DPT-Small), pose recognition (MediaPipe), and a hybrid chatbot were successfully implemented and evaluated. Initial results demonstrated acceptable performance, with future optimization planned for deployment on the Jetson Nano platform. The robot was successfully operated at APU's APCORE, confirming the system's potential for use in educational, reception, and interactive environments. Further enhancements will focus on improving accuracy, reducing latency, and expanding gesture-based control.

Acknowledgment

I would like to thank all those who have contributed to the successful completion of this project. I sincerely appreciate the support provided by Asia Pacific University of Innovation and Technology (APU) and APCORE for this project. I am particularly grateful to Ts Suresh Gobee and Ts Vickneswari Durairajah provided continuous guidance, advice, and support throughout the entire project process. Their professional knowledge and insights are crucial to the success of the project.

References

[1] L. Cao, "AI Robots and Humanoid AI: Review, Perspectives and Directions," 2024. doi: https://doi.org/10.48550/arXiv.2405.15775.

[2] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, and M. A. Laribi, "Recent advancements in multimodal human–robot interaction," *Front Neurorobot*, vol. 17, May 2023, doi: 10.3389/fnbot.2023.1084000.

[3] A. Pajaziti, X. Bajrami, and G. Pula, "Communication and Interaction between Humanoid Robots and Humans," in *Collaborative Robots [Working Title]*, IntechOpen, 2021. doi: 10.5772/intechopen.97334.

[4] A. J. Romero-C. de Vaca, R. A. Melendez-Armenta, and H. Ponce, "Using Social Robotics to Identify Educational Behavior: A Survey," *Electronics (Basel)*, vol. 13, no. 19, p. 3956, Oct. 2024, doi: 10.3390/electronics13193956.

[5] E. Coronado, T. Shinya, and G. Venture, "Hold My Hand: Development of a Force Controller and System Architecture for Joint Walking with a Companion Robot," *Sensors*, vol. 23, no. 12, p. 5692, Jun. 2023, doi: 10.3390/s23125692.

[6] M. Chignoli, D. Kim, E. Stanger-Jones, and S. Kim, "The MIT Humanoid Robot: Design, Motion Planning, and Control For Acrobatic Behaviors," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.09025

[7] J. Berg and S. Lu, "Review of Interfaces for Industrial Human-Robot Interaction," *Current Robotics Reports*, vol. 1, no. 2, pp. 27–34, Jun. 2020, doi: 10.1007/s43154-020-00005-6.

[8] M. A. Allouzi *et al.*, "Design Changes, Saving Materials and Energy in Food Serving Robot Via Topology Optimization," in *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/CASE56687.2023.10260438.

[9] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision Transformers for Dense Prediction," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.13413

[10] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer," Jul. 2019.

[11] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," Jun. 2018.

[12] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," Jun. 2020.

[13] R. Josyula and S. Ostadabbas, "A Review on Human Pose Estimation," Oct. 2021.

[14] F. Bonnefoy *et al.*, "From modulational instability to focusing dam breaks in water waves," Oct. 2019, doi: 10.1103/PhysRevFluids.5.034802.

[15] S. Keyner, V. Savenkov, and S. Vakulenko, "Open Data Chatbot," Sep. 2019.

[16] R. Tamrakar, S. Vallabhbhai, and N. Wani, "Design and Development of CHATBOT: A Review," 2021. [Online].Available: https://www.researchgate.net/publication/351228837