ESIGELEC Paper (2025), 1–4

ESIGELEC

ARTICLE TYPE

"Scalpel Please !": Integrating Speech and Vision Recognition for Robotic Assistance in Surgery with Multimodal Approachs

Foued BENIDIR

ISYMED, ESIGELEC, ROUEN, FRANCE
Email: foued.benidir@groupe-esigelec.org.

Abstract

The contemporary healthcare landscape is characterized by persistent personnel shortages and an escalating demand for precise, minimally invasive medical interventions. This research introduces a multimodal robotic pipeline that can address critical challenges in surgical support through advanced technological integration.

Our proposed methodology synthesizes cutting-edge technologies including voice recognition with word segmentation, Intel RealSense camera-based computer vision, and object detection with Yolov8. This comprehensive approach enables accurate interpretation and executions of specifics tool manipulation tasks by establishing a robust communication pathway between verbal commands, visual reconnaissance and robot guidance.

Preliminary experimental results for the object deplacement into a delimited zone demonstrate great performance metrics: object recognition from voice achieved 98.7%, and visual object detection attained 100%. The visual detection component exhibited perfect object from voice recognition identification with 98.1% accuracy, substantively validating the proposed pipeline's potential.

The research concludes that this multimodal robotic pipeline represents a viable technological solution with significant opportunities for future refinement. Potential advancement trajectories include miniaturized tools to detect, performance optimization in challenging acoustic environments, speaker variation adaptation, and natural language processing integration for contextually nuanced object recognition.

Keywords: voice recognition, computer vision, Yolo, Universal Robotic

1. Introduction

The integration of artificial intelligence (AI) and computer vision in medical procedures has evolved significantly since the pioneering work in surgical gesture segmentation (*Iftikhar et al. 2024*). This evolution has been particularly marked since 2021, with the unprecedented acceleration in AI capabilities and their applications in healthcare (*Topol 2019*). Current trends increasingly favor multimodal approaches that combine various AI domains to address complex medical challenges.

Building upon these developments, our research draws inspiration from (*Albiez 2021*), which highlighted significant challenges in surgical environments. Their study demonstrated that performing complex surgical procedures in noisy environments can substantially impact the operating team's performance. Specifically, elevated noise levels can increase stress, deteriorate communication quality, and potentially lead to medical errors. While noise reduction strategies exist, their implementation in super noisy surgical contexts, particularly in orthopedic operations, presents unique challenges. Recent advancements in voice recognition technology, as demonstrated in (*Demir 2024*), offer promising solutions to these challenges.

This article presents an integrated approach combining voice recognition and computer vision technologies to develop an advanced robotic surgical assistant. While this research represents an initial phase of a broader project, it is essential to validate the proposed pipeline at a smaller scale before expanding its scope. Our methodology encompasses several key components:

1. Voice Recognition System with sementic processing: Implementation of a speech recognition algorithm.

3. Computer Vision Integration: Utilization of YOLO-based object detection to precisely locate desired instruments in three-dimensional space, objects previously detected by voice recognition.

4. Coordinate System Transformation: Development of a robust transformation matrix to convert camera-space coordinates to robot-space coordinates.

5. Robotic Control Integration: Implementation of grasp-and-place operations using the RTDE library in order to move the objects into pre-defined zone.

This integrated system aims to demonstrate the feasibility of a voice-controlled robotic assistant capable of understanding verbal commands, identifying required objects, and moving them into a zone. While this research does not attempt to provide an exhaustive validation of the system's clinical implementation, it serves as a proof of concept to establish the technical viability of such an approach. Our primary objective is to demonstrate that the integration of these technologies is not only theoretically possible but practically achievable, thus laying the groundwork for future, more comprehensive studies into points that may requires it. The following sections detail our methodology, present experimental results, and discuss both the current limitations and future possibilities improvements. Through this work, we aim to contribute to the ongoing discussion about multimodal approch, combining AI and robotics in health domain, while maintaining a realistic

perspective on the challenges and opportunities ahead.

2. Methodology

2.1 Voice Recognition

2.1.1 Description

If we decide to be ethical and respect the privacy of the futur users, we need to have a local voice recognition, without any possibility of datas transmisions. So we decided to try with Vosk, selected for its local modularity and extensive exploitation capabilities. Other options were available such as Whisper from OpenAI, used in (*Redmon 2021*) for exemple. But the particularity of Vosk is its simplicity of use, in Python, its display of languages usable (more than 20), and its robustness. Also his small size for his perfomances can be a great compromise for power/efficiency (*Inc. 2025*).

2.2 Computer Vision

2.2.1 Description

Now that we have our voice recognition path, we need to be able to find those objects in the room. We desire that our objects can be detected in the space quickly and without any misunderstanding. Through computer vision, those things can be done really fast. Yolov8 is one of the latest releases from Ultralytics open-source pretrained computer vision recognition libraries and CNN. Indeed, YOLO (You Only Look Once) is a real-time object detection system that approaches the problem in a unified manner. It divides an image into a grid and, for each cell, predicts bounding boxes and associated class probabilities. This method enables rapid and efficient detection of objects within the image (*Redmon 2021*).

2.2.2 Method

Since we know what we will use for recognizing our objects, we have to find out the method to segmentate and identify them into the space. Two options exist to have depth and be able, through Yolo, to find coordinates. First one explored is to calibrate two cameras in order to create Stereovision. Following (*Chen et al. 2020*), this calibration process consists to use 2 points of view, from two exact same model of camera, in order to create intersects parameters and get the real-life delay between two cameras, emphasizing the importance of synchronizing the cameras to ensure accurate stereo imaging.

It also discusses the combination of baseline distance and checkerboard tile size for calibration, noting that these factors influence the accuracy of depth measurements. The study highlights challenges such as lens distortion and the need for precise synchronization to achieve accurate 3D reconstruction. But this option, based on the pipeline that we want to prove, is hard to still hold in the time. The second option, considered lately, is to use an infrared camera. Those cameras have the brilliant capability of having the depth of the object in each frame (*Corporation 2021*). In order to keep it simple and not reinvent the wheel, one camera is great and has many great reviews.

This is the Intel RealSense d435i, with advanced object detection capabilities, and pre defined calibration. The Yolov8n algorithm enables precise spatial identification and tracking. With those technologies, and only by using the camera referential, we can have a realy precise space positionning, with an accurate bounding box.

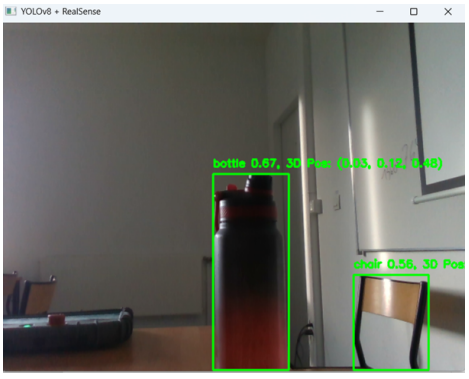

Figure 1. Exemple of bottle detected by Yolov8n, and his estimated position (48cm in deepness)

2.3 Robot

2.3.1 Description

The robot we have chosen for this project is the Universal Robot UR5. It was selected primarily due to its ease of use, intuitive programming interface, and availability. The UR5's user-friendly design and accessible software tools make it an ideal choice for rapid development and prototyping, especially in environments where flexibility and speed are critical. Its programming is straightforward, with drag-and-drop features and an extensive library of pre-built functionalities, enabling seamless integration with various systems. Additionally, the UR5's availability allowed for efficient implementation without delays, ensuring the project could proceed on schedule. These factors combined make the UR5 a versatile and reliable option for tasks requiring precision and adaptability.

2.3.2 Method

Since we received data from the camera, we needed to define a final position. To do this, we placed the robot in a pre-defined position, recorded its coordinates (x, y, z, RX, RY, and RZ for angular rotation), and integrated these commands into the internal script to check if they could be reached without any obstacles. Afterward, we disabled the local control system to operate solely through the Python script, allowing direct connection to the camera data. The method is based on controlling the UR5 robot through Python scripts, as outlined in the UR5 user manual (*Robots 2025*). Now that we are ready, we can start calibration, explained in the next section.

3. Calibration Between Camera and Robot Frames

3.1 Description and method used

For our validation test, we decided to place a camera in such a way that we could see the robot, the object to be moved, and the target area. This setup requires extrinsic calibration of the camera, also known as Eye-To-Hand calibration. Indeed, the

camera has its own frame of reference in which the object has coordinates. However, this frame is not understandable by the robot. Therefore, we need to perform a frame transformation to map the object's coordinates from the camera's reference frame to the robot's reference frame.

The calibration process involves determining a transformation matrix that can map the coordinates in the camera's frame to those in the robot's frame. This transformation typically includes a **rotation matrix** and a **translation vector**. Following (*Corke 2011*), the process can be broken down into the following steps:

1. **Capturing Corresponding Points**: First, we identify several points in the environment that are visible to the camera. These points should also be easy to locate in the robot's coordinate system. It is important to select points spread across the workspace to ensure better calibration accuracy.
2. **Record Camera and Robot Coordinates**: For each corresponding point, we record the position of the point in both the camera's coordinate system and the robot's coordinate system. This will allow us to compare and map the two different frames of reference.
3. **Calculate Transformation Matrix**: Using the data we recorded, we calculate the **transformation matrix**, which tells us how to convert the coordinates from the camera's frame to the robot's frame. This is done by solving a system of equations that minimizes the error between the known positions in both frames, often using techniques like **least squares optimization**.
4. **Apply Transformation**: Once we have the transformation matrix, we can use it to convert the coordinates of any point from the camera's reference frame to the robot's reference frame. The transformation is done by multiplying the camera's coordinates with the transformation matrix. Mathematically, this is expressed as:

$$\begin{bmatrix} x_{robot} \\ y_{robot} \\ z_{robot} \\ 1 \end{bmatrix} = \mathbf{T}_{camera\_to\_robot} \cdot \begin{bmatrix} x_{camera} \\ y_{camera} \\ z_{camera} \\ 1 \end{bmatrix}$$

Here, $\mathbf{T}_{camera\_to\_robot}$ is the transformation matrix, and the right-hand side represents the point's coordinates in the camera's frame, which will be converted into the robot's frame.

Once the transformation matrix is obtained, we can use it to map any detected object coordinates from the camera's reference frame into the robot's frame, enabling precise control of the robot for tasks such as object manipulation.

4. Robot Manipulation

Once calibration is completed, we can manipulate the robot. We will use the RTDE library, which is specifically designed for the UR model. By connecting to the robot via Ethernet, retrieving its IP address, and linking it to our computer, we can execute the Python remote control code. The robot's movement is controlled by sending commands that use the transformation matrix to correctly interpret the object's position in the robot's frame.

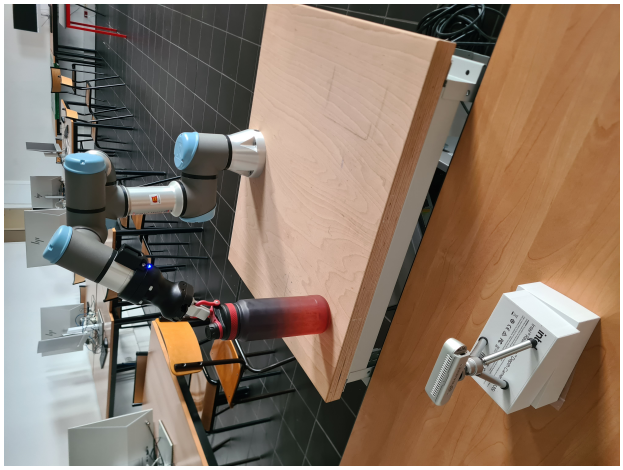

Figure 2. Test setup showing the robot, object, and target area. This setup allows us to capture the object's position in the camera frame and transform it into the robot's frame.

We followed the procedure outlined in (*Que et al. 2024*) for robot manipulation, ensuring smooth integration between the camera data and the robot's actions.

5. Experimental Results

5.1 Test of Vosk

To ensure the reliability of Vosk, we designed a benchmark, given the limited number of papers available on the subject. To simplify the test and keep in mind that this pipeline will ultimately be used in an operating room, we focused on recognizing a limited set of objects. For the test, we selected 5 objects: a bottle, pen, phone, book, and ball. The microphone used was built into an MSI GF Thin 65 laptop, powered by an Intel Core i7 9th Gen processor and an Nvidia RTX 2060ti graphics card. Although the conditions were not optimal, they were intentionally chosen to simulate a noisy environment. In these tests, we performed 10 iterations per object at 5-second intervals, achieving an overall accuracy of 98.7

5.2 Vision Recognition

For the vision tests, we used the YOLOv8n model, which contains 20 million parameters, making it sufficient for recognizing a limited set of objects. In optimal conditions for object recognition, as described in *Redmon 2021*, the model achieved 100% accuracy in object detection. To further evaluate spatial accuracy, we conducted 25 trials to measure object distances, with 21 successful measurements, resulting in an 84

5.3 Voice Recognition and Computer Vision

With both the voice recognition and vision recognition systems performing well individually, we combined them for a total of 50 use cases. In 49 out of 50 cases, the object detection

was accurate after voice recognition, resulting in an overall accuracy rate of 98%. The testing conditions were identical to those used in the previous tests.

5.4 Robotic Test

The robot tests presented additional challenges, particularly regarding precise calibration to ensure accurate transformation between coordinate frames. During the calibration process, 5 reference points were selected, but even with these, the error remained too significant. This indicates that further improvement is needed in the calibration procedure.

In 30 full trials, which incorporated voice transcription, object recognition, and robot movement, 15 successful transformations of coordinates into the correct frame were achieved. However, the robot failed to execute 4 movements due to residual noise or calibration errors, which resulted in the robot being blocked. For the remaining 11 trials, the robot achieved an average positional precision of 5 cm, corresponding to an overall success rate of 25%.

The calculation of the transformation matrix and filtering using the least squares method did not yield entirely satisfactory results. This underscores the need for enhanced filtering and noise reduction, particularly when transforming coordinates between the camera and robot frames. Future work will focus on improving these aspects to increase system reliability and precision.

6. Conclusion

In this research, we proposed a multimodal robotic pipeline that integrates voice recognition and computer vision for precise object manipulation in a surgical environment. Through various tests, we demonstrated the feasibility of using Vosk for reliable local voice recognition, achieving an impressive 98.7% accuracy. Vision recognition with YOLOv8n also showed great promise, achieving perfect object detection and strong spatial accuracy. The combination of both modalities yielded an overall accuracy rate of 98% in object detection, proving that the multimodal system works effectively for the intended purpose.

However, challenges emerged during the robotic manipulation phase, particularly regarding the calibration process between the camera and robot frames. Despite using 5 reference points for calibration, the error remained too significant for accurate transformation, highlighting the need for further improvements in filtering and noise reduction techniques. Additionally, the system's performance in terms of positional precision and successful robot movements is still under development, as reflected by the 25% success rate in the robotic tests.

Future work will focus on enhancing the calibration process, improving filtering methods, and refining the overall system for better reliability and precision in dynamic, noisy environments. This study sets the foundation for further exploration into multimodal robotic systems that could revolutionize the assistance offered in surgical settings, ultimately contributing to the ongoing advancement of AI and robotics in healthcare.

References

Albiez, Benjamin. 2021. Aj april 2021. *Issuu,* https://issuu.com/walkermanagement/docs/aj_04-2021_web/48.

Chen, Xin, Xiangfei Wu, Shiqi Gao, Xiaomei Xie, and Ya Huang. 2020. Synchronization and calibration of a stereo vision system. In *Global oceans 2020: singapore – u.s. gulf coast,* 1–6. https://doi.org/10.1109/IEEECONF38699.2020.9389422.

Corke, Peter. 2011. *Robotics, vision and control: fundamental algorithms in matlab.* Springer.

Corporation, Intel. 2021. Intel realsense depth camera d400 series calibration guide. Accessed January 2025, https://www.intelrealsense.com.

Demir, Kubilay Can. 2024. Voice recognition in noisy environments. *arXiv,* https://arxiv.org/abs/2406.14576.

Iftikhar, M., M. Saqib, M. Zareen, and H. Mumtaz. 2024. Artificial intelligence: revolutionizing robotic surgery: review. *Annals of Medicine and Surgery* 86, no. 9 (August): 5401–5409. https://doi.org/10.1097/MS9.0000000000002426.

Inc., Alpha Cephei. 2025. *Vosk speech recognition toolkit.* Accessed: 2025-01-28. https://alphacephei.com/vosk/.

Que, Haohua, Wenbin Pan, Jie Xu, Hao Luo, Pei Wang, and Li Zhang. 2024. "pass the butter": a study on desktop-classic multitasking robotic arm based on advanced yolov7 and bert. Submitted on 27 May 2024, *arXiv preprint arXiv:2405.17250* (May). https://arxiv.org/abs/2405.17250.

Redmon, Joseph. 2021. You only look once (yolo) - real-time object detection. *arXiv,* https://arxiv.org/abs/1506.02640.

Robots, Universal. 2025. *Ur5 user manual.* Accessed: 2025-01-28. https://www.universal-robots.com/products/ur5-robot/.

Topol, Eric J. 2019. *Deep medicine: how artificial intelligence can make healthcare human again.* Basic Books.