ChefMate: A Multimodal System for Real-Time Kitchen Ingredient Management System

Israel Kenneth Asamoah Bamfo
Department of Computer Science
University of Bern
israel.asamoahbamfo@students.unibe.ch

Ghamaathige Sachindu Himash Peiris
Faculty of Science & Medicine
University of Fribourg
ghamaathige.peiris@unifr.ch

Siyu Deng
Department of Computer Science
University of Bern
siyu.deng@students.unibe.ch

Seife Amdemikael
Computer Science Department
University of Neuchâtel
seife.amdemikael@unine.ch

***Abstract*—This report details the development and evaluation of a novel system designed to detect available ingredients using object detection and quantify them through voice inputs. The system captures the ingredients and, with the help of gestures, moves the detected ingredients and spoken quantities into a list. This integrated approach not only simplifies the process of ingredient management but also enhances kitchen efficiency. While this system does not generate recipes, future work could include this functionality alongside other enhancements such as nutritional analysis and inventory tracking.**

The unique combination of these modalities makes our system particularly beneficial for individuals with physical disabilities, busy professionals, and anyone looking to streamline their cooking process. A sample size consisting of 10 participants (Age 23-27) were given an interview to evaluate the effectiveness and user satisfaction. Findings reveals how, by reducing the cognitive load associated with managing ingredients, our system can help users save time, minimize food waste, and enhance their overall cooking experience.

I. Introduction

In contemporary kitchens, efficient ingredient management is a crucial aspect of culinary success. With the advent of smart technology, there is a growing demand for systems that can streamline this process, making it more intuitive and less time-consuming. This project introduces a novel system that leverages object detection, voice recognition, and gesture control to identify, quantify, and organize kitchen ingredients.

The idea behind this project is to create a seamless and interactive experience for users, allowing them to effortlessly manage their ingredients. Traditional methods of inventory management can be cumbersome and time-consuming. Our system aims to address these challenges by providing a hands-free solution that integrates multiple advanced technologies.

The system operates by first using object detection to identify various ingredients in the kitchen. Once an ingredient is detected, the user can verbally state the quantity, which is captured and processed by a voice recognition module. To further streamline the process, users can employ simple gestures to add, remove, or modify items in their ingredient list. This combination of modalities not only enhances the user experience but also ensures greater accuracy and efficiency.

Unlike existing solutions, such as smart refrigerators that rely on RFID technology or mobile apps that use barcode scanning, our system's unique approach lies in its use of real-time object detection and the integration of voice and gesture controls. This makes it particularly useful for individuals who need a more flexible and hands-free method of managing their kitchen inventory.

By simplifying the process of ingredient management, this system has the potential to save users time and reduce food waste, while also making cooking more enjoyable and accessible. As part of the project, we also consider future enhancements, such as recipe generation and nutritional analysis, to further expand the system's capabilities.

This report will detail the system's development, from conceptualization to implementation, and will evaluate its performance across various modalities. Through this, we aim to demonstrate the system's effectiveness and explore its potential applications in modern kitchens.

II. Related Work

In the domain of smart kitchen technology, several projects and systems have been developed to improve ingredient management and enhance the overall cooking experience. These systems often utilize various technologies such as RFID, barcode scanning, and computer vision. Here, we discuss some of the notable related works and highlight how our system differs and advances the state of the art.

Computer Vision-Based Systems Recent advancements in computer vision have led to the development of systems that use image recognition to identify ingredients. Projects such as "GrocerEye" utilize deep learning models to detect and recognize grocery items in real-time using a camera. These systems eliminate the need for manual scanning or RFID tagging, providing a more seamless and automated experience. For example, the GrocerEye project demonstrated the feasibility of using computer vision to recognize packaged goods and produce with high accuracy, leveraging models

like YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) for real-time object detection [1].

Gesture Recognition in Smart Homes Gesture recognition technology has been explored in smart home applications to provide an intuitive and touch-free means of control. For instance, Microsoft Kinect has been used to develop gesture-based interfaces for home automation systems, allowing users to control lighting, media, and other devices with simple hand movements [2]. However, the application of gesture recognition specifically for kitchen ingredient management remains relatively unexplored.

By leveraging the latest advancements in computer vision for real-time object detection, our system can identify a wide range of ingredients with high accuracy. The integration of voice recognition allows users to easily input quantities without physical interaction, and gesture control provides a natural and intuitive way to manage the ingredient list. This multimodal approach not only enhances usability but also makes the system accessible to individuals with physical disabilities or those who prefer a touch-free interface.

III. Hypothesis

The primary research question guiding this project is: "Can a multimodal system combining object detection, voice recognition, and gesture control effectively identify, quantify, and list kitchen ingredients in real-time, thereby enhancing the efficiency and user experience of ingredient management?"

This question aims to explore the feasibility and effectiveness of integrating multiple advanced technologies to create a cohesive and interactive system for managing kitchen inventories. By addressing this question, we aim to assess whether the system can accurately detect various ingredients, capture spoken quantities, and utilize gestures to manage the detected information efficiently.

To answer the research question, we propose the following hypotheses:

- **Hypothesis 1: Multimodal Integration Effectiveness**
  - H1: The integration of object detection, voice recognition, and gesture control will significantly enhance the efficiency and user experience in managing kitchen ingredients compared to single-modality systems.
  - This hypothesis posits that the multimodal approach will create a seamless and intuitive interaction for users, making the process of identifying, quantifying, and managing ingredients more efficient and user-friendly.
- **Hypothesis 2: Multimodal User Satisfaction**
  - H2: Users will report high levels of satisfaction with the multimodal system's usability and overall experience, finding it significantly more effective than traditional methods.
  - This hypothesis suggests that the combination of object detection, voice recognition, and gesture control will lead to higher user satisfaction due to the system's convenience and intuitive interface.

.

To test these hypotheses, the project involved the following objectives:

- **Integrate Object Detection Model:** Implement a deep learning model capable of accurately identifying common kitchen ingredients.
- **Integrate Voice Recognition Module:** Implement a voice recognition system to capture and transcribe spoken quantities.
- **Implement Gesture Recognition System:** Develop a gesture recognition component that can interpret user gestures for managing the ingredient list.
- **System Integration and Testing:** Integrate all components into a cohesive system and conduct rigorous testing to evaluate performance across different modalities.
- **User Studies:** Conduct user studies to gather feedback on the system's usability, efficiency, and overall experience.

IV. Modalities

The ChefMate system leverages three primary modalities to create a seamless and efficient kitchen ingredient management experience: object detection, voice recognition, and gesture control. Each modality plays a critical role in ensuring that the system can accurately detect, quantify, and manage ingredients in real-time, providing users with a hands-free and intuitive interface.

A. *Object Detection*

Object detection is at the core of ChefMate's ability to identify and categorize ingredients. This modality employs advanced computer vision techniques and deep learning models to detect and recognize various kitchen items. ChefMate utilizes state-of-the-art deep learning models, these models are known for their high accuracy and real-time performance, making them ideal for dynamic kitchen environments. The system uses a camera to capture images or video streams of the kitchen workspace. The object detection model processes these images to identify and classify ingredients, such as fruits, vegetables, and packaged goods. The detected items are then displayed on the user interface with labels and bounding boxes, providing visual confirmation to the user.

- **Real-Time Detection:** The models used can process images quickly, allowing for real-time identification of ingredients.
- **High Accuracy:** Advanced deep learning techniques ensure that the system can accurately detect a wide variety of ingredients under different lighting conditions and angles.
- **Scalability:** The system can be trained to recognize new items, making it adaptable to different kitchens and user needs.

B. *Voice Recognition*

Voice recognition enables users to interact with ChefMate through natural language, allowing them to input quantities and manage their ingredient list without physical interaction.

ChefMate integrates voice recognition services such as Google Speech-to-Text to capture and transcribe spoken inputs. These services utilize robust natural language processing (NLP) algorithms to convert voice commands into text. After an ingredient is detected, the user can state the quantity (e.g., "two cups of flour" or "three tomatoes"). The voice recognition module captures this input and processes it to extract the relevant information, which is then associated with the detected ingredient in the system's database.

- **Hands-Free Operation:** Users can input quantities and commands without touching any device, which is particularly useful in a busy kitchen environment.
- **Natural Interaction:** Voice recognition provides a natural and intuitive way for users to interact with the system, enhancing the overall user experience.
- **Accessibility:** This modality makes the system accessible to individuals with physical disabilities who may have difficulty using traditional input methods.

C. *Gesture Control*

Gesture control allows users to manage the detected and quantified ingredients through simple hand movements, offering an additional layer of interaction. ChefMate uses sensors to capture and interpret hand and body movements. These devices can accurately track gestures and translate them into commands. Specific gestures are mapped to different actions within the system. The system continuously monitors for these gestures and responds accordingly, providing real-time feedback to the user.

- **Intuitive Interaction:** Gestures provide an easy and intuitive way for users to control the system, reducing the need for complex menus or touch screens.
- **Enhanced Efficiency:** Gesture control can speed up the process of managing ingredients, allowing users to focus more on cooking and less on system interaction.
- **Hygiene and Safety:** In a kitchen setting, minimizing physical contact with devices can help maintain hygiene and reduce the risk of contamination.

D. *Integration of Modalities*

The integration of these three modalities—object detection, voice recognition, and gesture control—forms the backbone of ChefMate's functionality. By combining these technologies, ChefMate offers a comprehensive solution that addresses the limitations of using any single modality in isolation.

- **Complementary Strengths:** Each modality complements the others, providing a robust and versatile system. For instance, while object detection identifies ingredients, voice recognition captures quantities, and gesture control manages the list.
- **User-Centric Design:** The multimodal approach ensures that the system is user-friendly, accommodating different preferences and interaction styles. Users can choose the modality that best suits their current context, making the system flexible and adaptable.
- **Real-Time Interaction:** The seamless integration of modalities allows for real-time interaction and feedback, enhancing the overall user experience and making ingredient management more efficient.

V. CASE / CARE Models

We can categorize these modalities and their functionalities into CASE and CARE models as below.

CASE Model

- **Concurrent:** ChefMate uses object detection and voice input without coreferences in ingredient input.
- **Exclusive:** ChefMate uses gesture recognition after the input of ingredients.

CARE Model

- **Complementarity:** ChefMate uses object detection to detect ingredients and voice input to add the according amount, which is complementary to each other modalities.
- **Assignment:** Gesture recognition is used to accept or decline ingredients, which is not derived by other modalities.

VI. Procedure and Technologies

Since ChefMate can integrate of object detection, voice recognition, and gesture recognition in a multi-modal user interface, we would like to enhance the efficiency and user-friendliness of a cooking assistant application and specific benefits and challenges associated with each modality.

By integrating these three modalities, the ChefMate application aims to create a comprehensive and user-friendly cooking assistant that addresses the limitations of single-mode interfaces. This research seeks to evaluate the effectiveness of this multi-modal approach, identify the specific benefits it brings to the user experience, and understand the technical and practical challenges that need to be overcome to optimize performance.

A. *Technologies*

The ChefMate cooking assistant uses several technologies to create its multi-modal user interface: React, TensorFlow.js, Web Speech API, Handpose model, and Spoonacular API.

React is a popular JavaScript library for building user interfaces, particularly single-page applications where a fast, interactive user experience is crucial. React is used to build the user interface of the ChefMate application. Its component-based architecture allows for the creation of reusable UI components, making the development process more efficient and manageable. React's state management capabilities ensure that the application responds dynamically

to user inputs, maintaining real-time interaction between object detection, voice recognition, and gesture recognition modalities.

TensorFlow.js is an open-source library developed by Google for machine learning in JavaScript. TensorFlow.js implements the COCO-SSD (Single Shot MultiBox Detector) model, which is used for real-time object detection via the camera. This model identifies various ingredients and cooking tools within the camera's field of view.

The Web Speech API is utilized for voice recognition. This API enables the ChefMate system to capture and transcribe spoken inputs, allowing users to input quantities and commands without physical interaction. This hands-free operation enhances the user experience, particularly in busy kitchen environments.

The Handpose model, also provided by TensorFlow.js, is used for gesture recognition. This model detects and interprets user gestures, allowing for intuitive control of the system. Users can use simple gestures to add or remove items in their ingredient list, making the interaction more natural and efficient.

The Spoonacular API is integrated to provide recipe search functionality. This API allows users to find recipes based on the available ingredients, enhancing the utility of the ChefMate system by suggesting potential dishes that can be prepared with the detected ingredients.

By leveraging the strengths of these technologies, ChefMate not only improves the cooking process but also sets a new standard for multi-modal interfaces in the kitchen.

B. Procedure

The development of the user interface for our Multimodal System. The system is designed to facilitate a specific part of the user's cooking, those are includes object detection, voice recognition, and gesture control(As shown in the activity diagram 1)

When the user opens the ChefMate system, firstly needs to allow the use camera and microphone. Then the camera captures real-time visual objects of the cooking elements. An object like a cup or a bowl, and an ingredient like an apple or banana are shown to the camera, and the system detects it using an object detection algorithm. After successful detection, the voice recognition API is activated. Then the user speaks the quantity of the detected object they wish to add, like one or two. So that the system recognizes the spoken quantity and adds it to the drafts.

When the user sees the object and the amount is correct in the draft, the user then shows an "OK" gesture to the camera. The system detects the "OK" gesture and moves the quantity

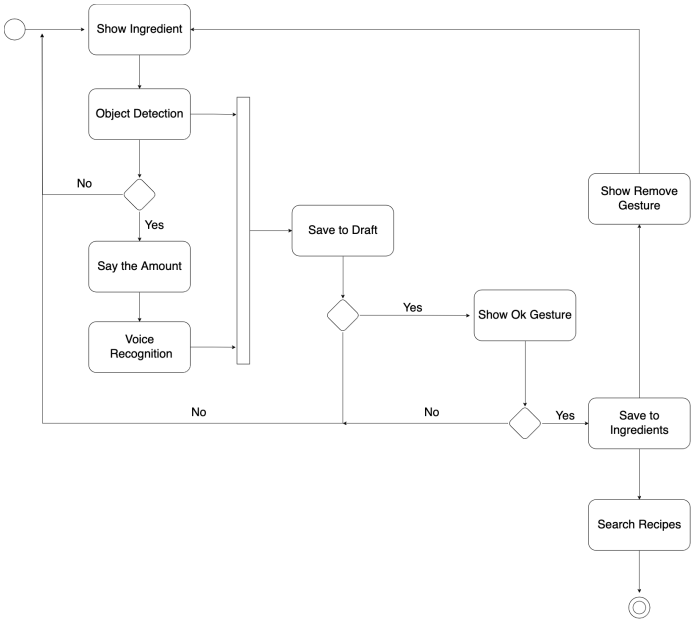

Fig. 1: Activity Diagram

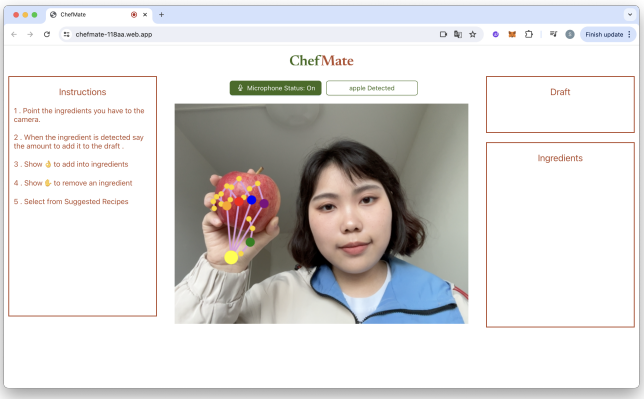

Fig. 2: Show the item and say the quantity

from the drafts to the ingredients list.

If an item is incorrect or the quantity is wrong, please use an open palm gesture to indicate that you want it deleted.

On the backend, ChefMate utilizes API interfaces to ensure seamless integration of camera, voice and gesture inputs. The APIs are responsible for processing the voice inputs for ingredient quantities and detecting gestures for various interactions within the website.

C. Modalities

*1) **Voice Recognition**:* This converts spoken words into text and ensures accurate transcription. Users can verbally state the quantities of the detected ingredients. The voice recognition

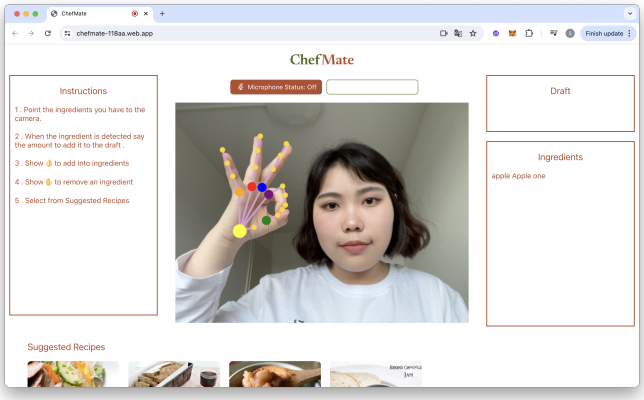

Fig. 3: ”OK” gesture

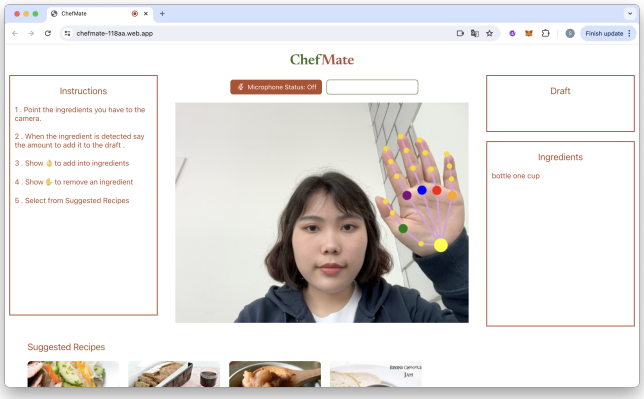

Fig. 4: Remove latest ingredient

module captures these inputs with high accuracy, minimizing the need for manual entry.

*2) **Gesture Recognition**:* This ensures detection and interpretation of user gestures to manage ingredient lists and navigate the app. Users can use simple gestures to add, remove, or modify items in their ingredient list. This touch-free interaction enhances the user experience, especially in a kitchen environment where hands might be occupied or dirty.

*3) **Object Detection**:* With this, the user can turn on the camera and then place the item within range of the camera, which automatically identifies the available materials and quantities, and then generates the available menu.

By integrating these technologies and following this procedure, ChefMate offers a comprehensive solution that enhances kitchen efficiency and user satisfaction. The multimodal approach—combining object detection, voice recognition, and gesture control—creates a seamless and intuitive experience for users, making cooking more accessible and enjoyable.

VII. Feedback & Error Handling

Robust error handling and continuous user feedback are crucial for seamless interaction. When a user inputs an ingredient via object detection, the system first confirms the identification by displaying the detected object on the screen. The user then specifies the amount using voice input, with the system continuously showing the microphone status to ensure clarity and minimize errors in voice capture. Once the ingredient is added, the system provides immediate feedback by announcing the detected ingredient and the specified amount via voice output. This continuous feedback loop ensures the user is always informed about the current state of their input, reducing the likelihood of errors. If the user wishes to remove the latest ingredient from the list, they can do so easily, and the system provides immediate voice feedback acknowledging the removal.

Throughout this process, the system employs a draft mode that allows users to review and correct their inputs before finalizing them. This draft mode is displayed on the screen, showing detected ingredients and their amounts, enabling users to make adjustments as necessary. Additionally, a comprehensive instructions list on the screen guides users through each step of the interaction, ensuring clarity and ease of use.

VIII. Evaluation

We have evaluated Chefmate through paired T-Test Analysis and User Interviews based on the two hypothesis derived with. To evaluate the Hypothesis 1, We designed a T-test to compare the modalities and for Hypothesis 2, we gathered qualitative data regarding efficiency,ease of use and learning and user interface to evaluate the user satisfaction.

A. *Paired T-Test*

In this study, we aimed to evaluate the performance differences between various modalities used for ingredient detection and input. The modalities tested included Voice Recognition, Gesture Recognition, Object Detection, and Text Input. We collected time data from 10 users performing actions using these modalities and conducted paired t-tests to determine if there were significant differences in the times taken to complete tasks using different modality pairs

Hypothesis:

- **Null Hypothesis (H0) :** There is no significant difference in the mean times taken to perform tasks between the two modalities.
- **Alternative Hypothesis (H1) :** There is a significant difference in the mean times taken to perform tasks between the two modalities.

Data Collection:

The time data (in seconds) for each user performing tasks with different modalities is as follows in Table I:

User	VR	GR	OD	TI
1	1.0	2.0	1.1	1.2
2	1.2	2.5	1.3	1.4
3	1.4	1.8	1.4	1.3
4	1.1	2.2	1.2	1.1
5	1.3	2.4	1.5	1.2
6	1.5	1.9	1.0	1.3
7	1.2	2.1	1.3	1.4
8	1.4	1.7	1.2	1.0
9	1.0	2.3	1.4	1.1
10	1.3	2.0	1.5	1.2

TABLE I: Collected time data (in seconds) for each user

Methodology:

- Voice Recognition : The time for voice recognition was measured by calculating the duration it took for each user to say the phrase ”two pieces” or the word “Banana”. This included the time from the start of the utterance to the end of the phrase/word.
- Gesture Recognition : The time for gesture recognition was evaluated by measuring the duration it took for each user to perform a gesture that corresponds to an action. The time ranged between 1 and 3 seconds.
- Object Detection : The time for object detection was measured by calculating the duration it took for the system to recognize an object after the user presented it.
- Text Input : The time for text input was measured by calculating the duration it took for each user to type the required text input. This included the time from the start of typing to the completion of the text input.

Results:

The paired t-tests were performed for the following modality pairs:

1) Voice Recognition (VR) vs. Gesture Recognition (GR)
2) Object Detection (OD) vs. Text Input (TI)

The results are summarized in the table II:

Modality Pair	T-Statistic	P-Value
Voice vs Gesture	-7.113586	0.000056
Object vs Text	1.048690	0.321665

TABLE II: Paired T-Test Results

Interpretation and Conclusion:

1) Voice Recognition vs. Gesture Recognition
   - T-Statistic: -7.113586
   - P-Value: 0.000056

   Conclusion: The p-value (0.000056) is much smaller than the common significance level of 0.05. Therefore, we reject the null hypothesis. This suggests that there is a statistically significant difference in the time taken between Voice Recognition and Gesture Recognition, with Gesture Recognition taking significantly longer. Figure 5
2) Object Detection vs. Text Input
   - T-Statistic: 1.048690
   - P-Value: 0.321665

   Conclusion: The p-value (0.321665) is greater than 0.05. Therefore, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in the time taken between Object Detection and Text Input. Figure 6

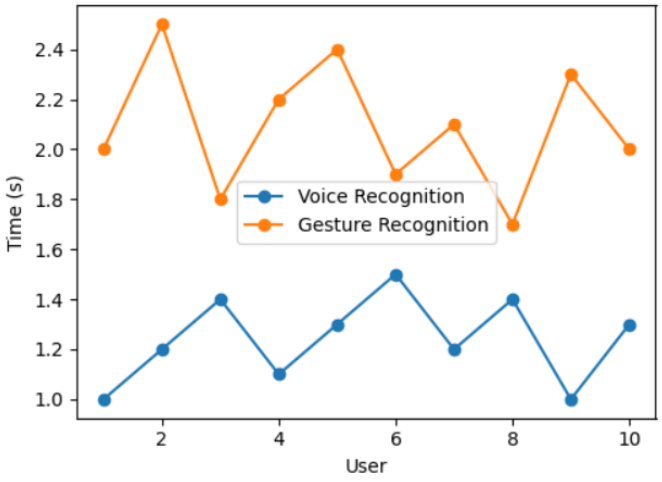

Fig. 5: Voice Recognition vs Gesture Recognition

Based on the paired t-test results for our collected data, we found a significant difference in the times taken to perform tasks using Voice Recognition and Gesture Recognition, with Gesture Recognition taking significantly longer. However, there was no significant difference between Object Detection and Text Input in terms of time efficiency. These findings suggest that while Voice Recognition is more time-efficient compared to Gesture Recognition, Object Detection and Text Input have comparable time efficiencies under the given conditions.Further studies with larger sample sizes and different contexts may provide additional insights into the performance differences between these modalities.

B. User Interviews

The evaluation of the ChefMate system was conducted through user feedback based on a structured questionnaire. Participants were asked to describe their experience and rate the system’s performance across the three primary modalities: object detection, voice recognition, and gesture control. The following summarizes the responses and insights gathered from the evaluation from 10 participants (Age 23 - 27).

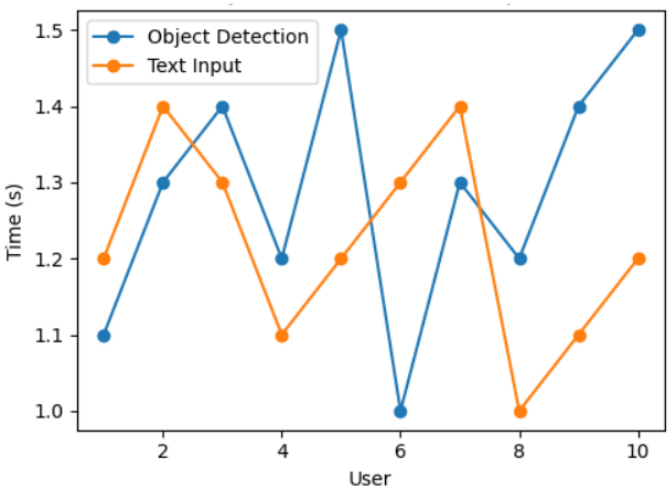

Fig. 6: Object Detection vs Text Input

Methodology:
Participants were given a brief 5 minutes introduction on the usage of system and then provided with a detailed questionnaire consisting of five key questions aimed at understanding their experience and the effectiveness of the multimodal system. The questions focused on the integration of the modalities, efficiency, user interface improvements, convenience, and ease of learning.

Questions and Responses:

1) **Describe your experience using the ChefMate system, specifically focusing on how well the different modalities (object detection, voice recognition, gesture control) worked together?**
   - **Response Summary:** Participants reported a very positive experience using the ChefMate system. The combination of object detection, voice recognition, and gesture control was noted to work seamlessly together. Object detection effectively identified ingredients and kitchen tools in real-time. Voice recognition accurately processed spoken commands, and gesture control allowed for hands-free confirmations with simple gestures like ”OK”. The multimodal approach ensured that each modality complemented the others, creating a user-friendly experience.
2) **How efficient did you find the system in helping you manage kitchen ingredients when using multiple modalities? Do you think the combination of these modalities improved your overall experience compared to using a single modality?**
   - **Response Summary:** The system was found to be highly efficient in managing kitchen ingredients. Participants appreciated that object detection quickly identified ingredients, voice recognition accurately recorded quantities, and gesture control facilitated easy confirmations. The multimodal approach made the cooking process faster and more enjoyable compared to using a single modality.
3) **How would you improve the ChefMate multimodal user interface?**
   - **Response Summary:** Participants suggested that allowing users to customize voice commands and gestures to better suit their personal preferences and cooking styles would increase the system’s flexibility.
4) **Was this system convenient and efficient?**
   - **Response Summary:** Yes, the ChefMate system was both convenient and efficient. Participants noted that the system allowed them to manage ingredients and follow recipes without using their hands, which is particularly useful when hands are often dirty in the kitchen. This reduced the time and effort required for these tasks.
5) **Was it easy to learn how to use all modalities effectively?**
   - **Response Summary:** Yes, learning to use all the modalities effectively was easy. The system’s clear design allowed participants to quickly become proficient in using the system.

To visually represent the evaluation results, a bar chart was created based on participant responses. The chart includes ratings on a scale of 1 to 5 for the following criteria: overall experience, efficiency, convenience, and ease of learning.

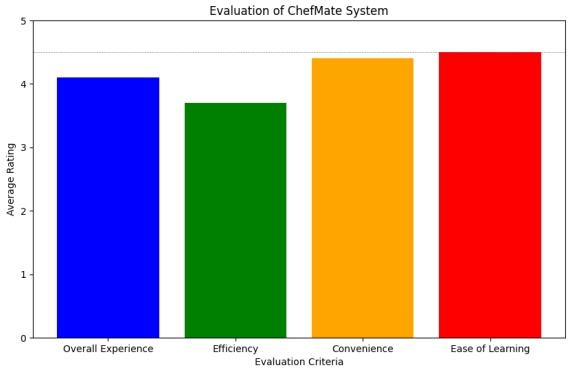

Fig. 7: Evaluation of ChefMate System

IX. Discussion

Strengths and Contributions: The evaluation results indicate that the ChefMate system effectively integrates object detection, voice recognition, and gesture control to create a robust and user-friendly kitchen management tool. Each modality complements the others, providing a comprehensive solution that addresses the limitations of single-modality systems. The high ratings in overall experience, efficiency, convenience, and

ease of learning underscore the system's reliability and user-friendliness.

Areas for Improvement: While the feedback was overwhelmingly positive, participants suggested enhancements such as customizable voice commands and gestures to better suit individual preferences. These improvements could further increase the system's flexibility and user satisfaction.

Future Work: Future enhancements could include nutritional analysis, which could provide users with valuable information about the nutritional content of their ingredients, aiding in healthier meal planning. Advanced inventory management and tracking capabilities could help users monitor ingredient usage and expiration dates, reducing food waste.

Additionally, integrating ChefMate with smart kitchen appliances and cloud-based data synchronization could create a more interconnected and automated kitchen environment. This would enable seamless communication between devices and provide users with a more comprehensive cooking assistant.

X. Conclusion

The development and evaluation of the ChefMate multimodal system have demonstrated its potential to revolutionize kitchen ingredient management. By effectively integrating object detection, voice recognition, and gesture control, ChefMate offers a hands-free, intuitive, and efficient solution that enhances the cooking experience. The system's high accuracy and positive user feedback underscore its reliability and user-friendliness.

While the current system provides a robust foundation, ongoing refinement and expansion will further enhance its capabilities and appeal. By addressing the identified areas for improvement and exploring additional functionalities, ChefMate can continue to evolve into a comprehensive and indispensable tool for modern kitchens. This project not only meets current needs but also paves the way for future innovations in smart kitchen technology, ultimately contributing to a more efficient and enjoyable cooking experience for users worldwide.

XI. GitHub Repository Link

https://github.com/SachinduHimash/ChefMate
Hosted Site : https://chefmate-118aa.web.app/

References

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[2] Jamie Shotton, Andrew Fitzgibbon, Matthew Cook, Toby Sharp, Mark Finocchio, Richard Moore, ..., and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, 2011.