Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048312

ARTICLE

Real-Time Object Detection and Face Recognition Application for the Visually Impaired

Karshiev Sanjar[1], Soyoun Bang[1], Sookhee Ryue[2] and Heechul Jung[1,*]

[1]Department of Artificial Intelligence, Kyungpook National University, Daegu, 41466, Republic of Korea

[2]Haga Co., Ltd., Daegu, 38428, Republic of Korea

*Corresponding Author: Heechul Jung. Email: heechul@knu.ac.kr

Received: 04 December 2023 Accepted: 19 March 2024 Published: 20 June 2024

ABSTRACT

The advancement of navigation systems for the visually impaired has significantly enhanced their mobility by mitigating the risk of encountering obstacles and guiding them along safe, navigable routes. Traditional approaches primarily focus on broad applications such as wayfinding, obstacle detection, and fall prevention. However, there is a notable discrepancy in applying these technologies to more specific scenarios, like identifying distinct food crop types or recognizing faces. This study proposes a real-time application designed for visually impaired individuals, aiming to bridge this research-application gap. It introduces a system capable of detecting 20 different food crop types and recognizing faces with impressive accuracies of 83.27% and 95.64%, respectively. These results represent a significant contribution to the field of assistive technologies, providing visually impaired users with detailed and relevant information about their surroundings, thereby enhancing their mobility and ensuring their safety. Additionally, it addresses the vital aspects of social engagements, acknowledging the challenges faced by visually impaired individuals in recognizing acquaintances without auditory or tactile signals, and highlights recent developments in prototype systems aimed at assisting with face recognition tasks. This comprehensive approach not only promises enhanced navigational aids but also aims to enrich the social well-being and safety of visually impaired communities.

KEYWORDS

Artificial intelligence; deep learning; real-time object detection application

1 Introduction

Accessible visual information is critical for improving blind and visually impaired people's independence and safety. There is a real need to develop intelligent, automated solutions to assist individuals with vision impairment or blindness, among the top 10 disorders globally. According to the Korea Blind Union (KBU) [1], as of July 2018, approximately 253,000 South Koreans have a visual disability. A recent study by the Korea Disabled People's Development Institute (KODDI) showed that the number of disabled people is falling. In contrast, the number of registered visually impaired people (VIP) has gradually increased since 2012 [2]. Vision-impaired people usually need others to keep doing their day-to-day activities. Because they cannot be tracked down, they frequently become

victims of unwelcome issues that may lead to emotional distress or uninvited situations. Even a simple daily task like choosing a fruit or vegetable while shopping is undoable for them. Moreover, those who are blind or have low eyesight find identifying others in various social situations difficult. Relying only on voice recognition may be challenging [3,4]. VIP are disadvantaged in many professional and educational circumstances due to their inability to recognize others during group sessions. They even fail to recognize family members they are already familiar with by using the sounds of those people's voices when they are not speaking [5]. Recent advancements in assistive technologies for visually impaired persons have focused primarily on general navigation, obstacle detection, and fall prevention. Various systems leveraging technologies like computer vision, sensor-based solutions, and AI have been developed. These systems, while effective in their respective domains, often lack specificity in tasks such as precise object detection in indoor environments and accurate face recognition in social settings.

A key limitation of current methodologies is their generic approach, which falls short in addressing specific daily challenges faced by VIP, such as identifying different types of fruits and vegetables or recognizing familiar faces in diverse situations. This gap in specificity leads to a lack of performance in real-world applications, where context-specific information is crucial. Additionally, many existing solutions require cumbersome equipment or are not designed for real-time processing, further limiting their practical usability for VIP.

This study aims to assist VIP in everyday life by designing an application that can detect objects, specifically fruits and vegetables, in indoor environments and help recognize family members and close friends in specific conditions. The primary objective of this research is to develop a sophisticated portable device designed assist visually impaired individuals. This device is engineered to accurately detect various types of fruits and vegetables and to recognize faces. The focus is on leveraging advanced object detection and face recognition technologies to enhance the daily life experiences of visually impaired users, providing them with a tool that offers greater independence in navigating their environments. It is anticipated that the results of this research can significantly help to make blind people's lives comfortable, creating a convenient circumstance while they are preparing food at home on their own or shopping at greengrocery stores. All in all, visually impaired people can "see" crops and faces around them. You Only Look Once version 7 (YOLOv7) [6] method, a new state-of-the-art method for real-time object detection, is employed in this application. The "dlib" library (a modern toolkit containing machine learning algorithms and tools) has been utilized for face recognition to gain an accurate and computationally efficient result.

Upon thorough consideration, it becomes evident that research in the field of computer vision should prioritize the development of real-time object detection systems for fruits and vegetables, as well as advancements in face recognition. The contributions of this research are noteworthy and include:

- The development of applications that facilitate rapid, precise, and real-time detection of fruits and vegetables alongside face recognition capabilities.
- An object detection framework capable of identify 20 classes, evenly distributed between ten fruits and ten vegetables, focusing primarily on those commonly used daily.
- The face recognition aspect of the proposed system is designed to learn facial alignments accurately from just a single image of an individual.

The structure of the remainder of this paper is organized as follows: Section 2 delves into related works, focusing on real-time object detection and face recognition technologies for visually impaired individuals. The dataset and methodologies employed in this study are detailed in Section 3, while

Section 4 presents the results. The paper concludes with Section 5, which summarizes the conclusions and outlines potential directions for future research.

2 Related Works

2.1 Object Detection

Some research works related to object detection and face recognition for people who lost their vision have been done by the research community until now. Kumar et al. [7] developed a fruit and vegetable detection system for blind people. However, this system is not real-time, and the number of classes is not provided in the manuscript. Mahesh et al. [8] trained an object detection system for blind people using YOLOv3 [9] model. This system can detect merely five types of objects. The proposed system by Towhid et al. [10–12] uses traditional machine learning algorithms to detect objects, and the speed of the proposed method is relatively slow for prompt object detection. This paper [13] suggests using the YOLO algorithm to train the Common Objects in Context (COCO) dataset [14], which does not contain everyday usage fruits and vegetables, for the object detection task. Also, the suggested method could be more efficient for real-time applications. Zaidi et al. [15] provided a detailed examination of contemporary object detection algorithms. Their work includes a succinct overview of crucial benchmark datasets and the evaluation metrics employed in detection processes. They also discussed several leading backbone architectures commonly implemented in object detection tasks. Another work described by [16] gives a computer vision idea that converts an object to text by importing a pre-trained dataset model from the Caffe model framework. Then, the sentences are turned into voices. This project [16] attempts to convert the visible world into the audible one, alerting blind people about objects and their spatial locations. In the described system, objects identified within a scene are assigned labels or names and then converted into spoken words through a text-to-speech process. This functionality indicates a sophisticated integration of object recognition and auditory output, allowing for an aural representation of visual data.

2.2 Face Recognition

Even though there are various facial recognition systems available to help people with visual impairments recognize others, almost all of them require a database containing images and names of those who should be monitored. These methods would not be able to aid VIP in recognizing persons with whom they are unfamiliar. Identifying persons is a significant issue for people with vision impairments, prohibiting them from fully participating in many social activities and jeopardizing their sense of privacy and physical protection [17,18]. For instance, when a VIP walks into a conference room, school, or cafeteria, he seems to have no idea who else is there. As a result, VIPs may be hesitant to leave their houses, contributing to anxiety and unhappiness [19]. Face recognition technology has made it possible to improve social events for VIP. Several facial recognition technologies have been developed on smartphone platforms [18,20,21]. For example, Kramer et al. [22] created a smartphone facial recognition application for VIPs. They tested their model in classrooms and conference rooms and discovered that it detected faces with 94% accuracy, even when the faces were facing away from the camera from different angles. One research on face recognition for VIP was held in Meta Inc., [23]. The researchers used the accessibility bot, a research prototype on Facebook Messenger. It uses recent computer vision and tagged photographs from a user's social friends to assist individuals with visual impairments identify their friends. The accessibility bot informs users about their friends' identities, facial expressions, and traits taken by their phone's camera. Krishna et al. [24] used a camera in a pair

of glasses to recognize and identify faces. Nevertheless, it required a user-created collection of face photos and was not thoroughly tested with real users.

In our study, we primarily focus on creating a real-time application that identifies food crops and recognizes the faces of family members. For face recognition, we utilized the "dlib" algorithm, which effectively detects face bounding boxes and categorizes the detected faces as either family or non-family members. We aimed to achieve over 90% accuracy in face recognition, and our results show a commendable performance of 95.64% accuracy, exceeding our initial target. We also explored other methods such as Convolutional Neural Network (CNN) based face recognition [25], which demands extensive data and computational resources; LBPH method [26], which produces long histograms leading to slower recognition speeds; and the EigenFace [27] algorithm, which is sensitive to lighting variations. This study [28] uses the "dlib" library, employing Histogram Oriented Gradients (HOG) for feature extraction and Support Vector Machines (SVM) for differentiating face and non-face regions. Notably, the "dlib" method is solely used for identifying face bounding boxes, while a Deep Neural Network-based classification model categorizes the detected faces. This technique is effective with substantial data and high-performance hardware. However, in our context, "dlib" leverages Euclidean distance for classifying faces into family or non-family categories, offering a computationally efficient solution that measures the distance between the feature vectors of faces, thus allowing for swift and precise recognition even when only a limited number of training images are available.

3 Materials and Methods

3.1 Dataset

AIHub platform generates datasets used in the computer vision domain and is publicly available. A specific dataset [29] was chosen, which contains more than 25,000 fully annotated fruit and vegetable data instances. Additional food crop images were sourced by visiting the local markets to augment this dataset, adding more real-life images to the overall dataset. Consequently, the enriched dataset encompasses approximately 27,000 training and 3000 validation images. Fig. 1 illustrates dataset samples and the number of data samples for each class.

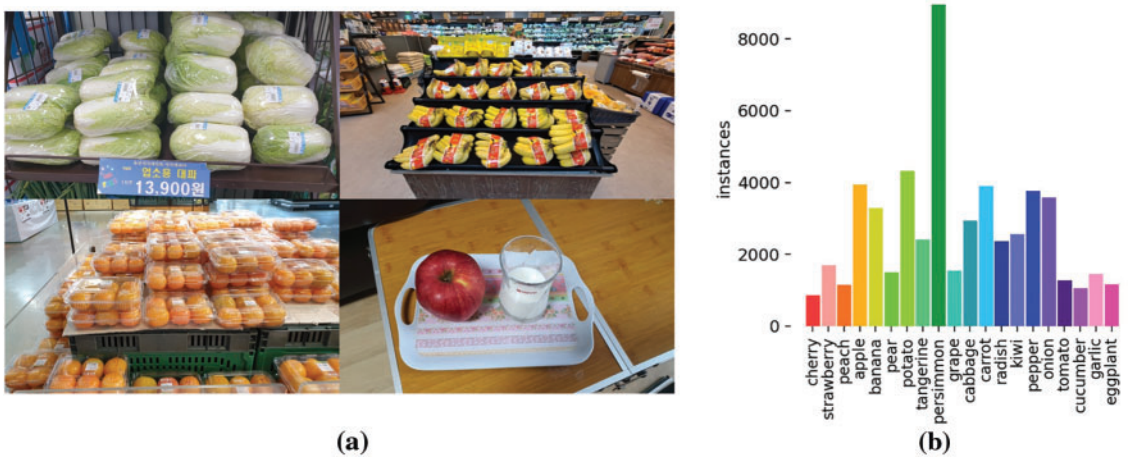

Figure 1: Object Detection Dataset: (a) dataset samples; (b) class names and number of data samples per class

The face recognition part of the study must recognize the family members and close friends of blind people. In this case, pictures of family members are included in the dataset, and the system is required to recognize photos of these family members in different situations, smiling and looking from different angles. The task of face recognition can be implemented in different ways. As the data collection part, we used our laboratory members' face images. Each participant contributed 20 images, resulting in a total collection of 200 face images for the study. Each image has been taken in different situations described above. A single image of an individual will be used to learn face alignments, and others will be used for the testing phase.

3.2 Preliminaries

The most current object detectors combine classification and bounding box regression simultaneously. Classification tries to predict the class of an object in an image region. In contrast, bounding box regression tries to determine the area by predicting the most solid box that includes the thing. Consider the ground truth of bounding box $\boldsymbol{gT}$ related to class label $\boldsymbol{l}$ and detection hypothesis $\boldsymbol{\xi}$ of bounding box $\boldsymbol{b}$. Because $\boldsymbol{b}$ normally contains an item and some background information, determining if the detection is correct can be tricky. The intersection over union (IoU) measure is commonly used to address this.

$$IoU\left(\boldsymbol{b}, \boldsymbol{gT}\right) = \frac{\boldsymbol{b} \cap \boldsymbol{gT}}{\boldsymbol{b} \cup \boldsymbol{gT}} \tag{1}$$

$\boldsymbol{\xi}$ is assigned an example of the class of the object of bounding box "gT" and designated as a *positive* example, if the IoU is greater than a certain threshold μ. The hypothesis $\boldsymbol{\xi}$ class label is a function of μ.

$$\boldsymbol{l}_{\mu} = \begin{cases} l, & IoU\left(\boldsymbol{b}, \boldsymbol{gT}\right) \geq \mu \\ 0, & otherwise \end{cases} \tag{2}$$

If the IoU metric amount does not satisfy the threshold for any object, $\boldsymbol{\xi}$ is assigend as a negative example. In many cases, IoU selects a set of bounding boxes used to train the bounding box regressor.

$$\sigma = \left\{\left(\boldsymbol{gT}_i, \boldsymbol{b}_i\right) | IoU\left(\boldsymbol{b}_i, \boldsymbol{gT}_i\right) \geq \mu\right\} \tag{3}$$

After selecting the set of bounding boxes, non-max suppression [30] algorithm is called to choose the most prominent bounding box for a particular object.

3.3 Methodology

3.3.1 Main Components

The primary purpose of this research is to deliver a portable device to a visually impaired individual that can detect fruits and vegetables and recognize faces. Fig. 2 shows the main mechanisms utilized in this study.

Tensorflow servings, known for its adaptability and robust performance in production environments, is employed as the machine learning serving system for the study's deployment in a production setting. Its architecture and Application Programming Interface (API) remain constant even when deploying new algorithms or experiments. It is compatible with TensorFlow models and can be customized to accommodate other models and data types. The TensorFlow serving system requires the model files with *.pb*, which stands for *protobuf*. The graph definition and model weights are

contained in the protobuf file. The protobuf file can be obtained in two ways. One is to train the model in TensorFlow and get the *.pb* file directly. The second option is to train the model in PyTorch and convert the trained model to the desired model format. In this study, the second option was chosen. The client side is armed with the specific tool comprising a camera, earpiece port and other additional components. The design of the hardware is given in Fig. 3.

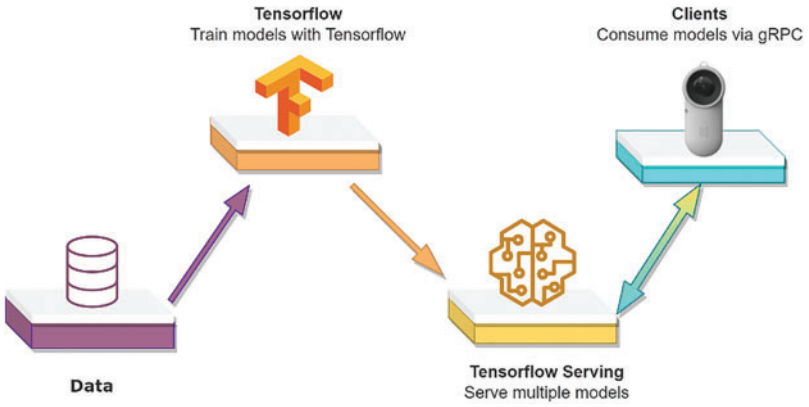

Figure 2: Main components

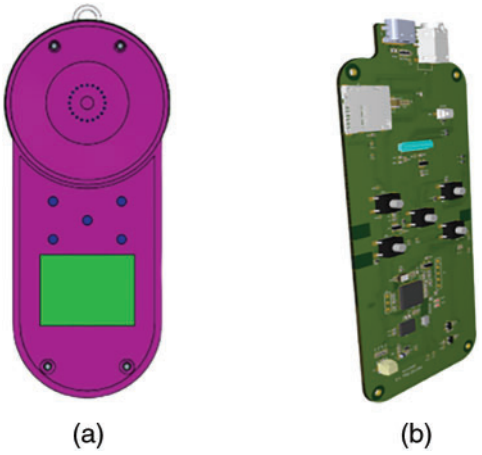

Figure 3: (a) PCB; (b) an external design of the proposed tool

Currently, the development of the printed circuit board (PCB) is in progress. This tool is hung on the user's clothes. It has several buttons to make different types of commands. The client and the server are connected using *Google Remote Procedure Calls (gRPC)*. *gRPC* is a cross-platform, open-source, high-performance remote procedure call framework.

3.3.2 Overall Flowchart

The key goal of this study is to provide a person who is visually impaired with a portable device that can detect fruits and vegetables and recognize faces. First, the images taken by the user's camera are sent to the cloud system. The system, then, identifies whether the face exists in the image or not. If the system detects any face, the "dlib" face recognition task is performed according to the face database provided beforehand. If the system detects any food crop included in the dataset, then the YOLOv7

[6] object detection model is called to detect and classify crops into one of the above-mentioned 20 classes. Before sending audio output to the earpiece placed on the user's ear, the text outputs from object detection and face recognition are converted to speech using *Google Text-to-Speech (gTTS)* model. The overall flow chart of the suggested system is illustrated in Fig. 4.

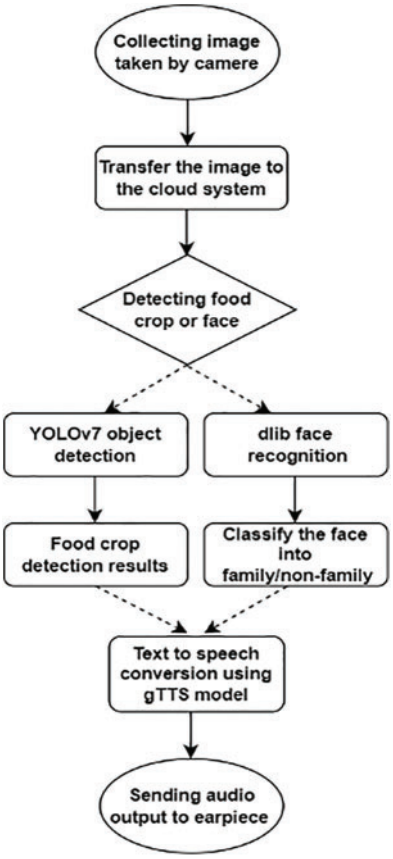

Figure 4: Flowchart of the proposed model

3.3.3 YOLOv7

YOLO is an object identification technique that predicts things inside an image using CNN in a single pass. Several researchers consider the 'YOLO' design a "state-of-the-art" framework. The popularity of YOLO can be attributed to its high inference speed and accuracy. At the time of writing, seven versions of YOLO have been realized by the deep learning community; the first three versions [9,31,32] were developed by the original authors, Redmon and Farhadi, while computer vision researchers contributed the remaining four versions [6,33–35]. We do not pay attention to the first 6 versions of YOLO since our main target is to apply YOLOv7 [6] to food crop detection tasks. To recognize objects in a picture, YOLOv7 employs a multi-scale technique. To clarify, the suggested approaches in this research differ from the present dominant real-time object detectors. In addition to architectural optimization, their suggested approaches will concentrate on training process optimization. The attention will be on several improved modules and optimization strategies that increase the training cost while decreasing the inference cost to improve object detection accuracy. The suggested modules and optimization approaches are referred to as trainable bag-of-freebies.

3.3.4 YOLOv7 Model Training

In this study, we adopted the experimental setup outlined in the original YOLOv7 paper [6]. This decision was made to ensure that our results were directly comparable with the baseline established by the YOLOv7's original research. Our experimental environment consisted of an A6000 GPU, and experiments were conducted using Python 3.9, PyTorch 1.11, and CUDA 11.4. Regarding the specific training details, OneCycle learning rate policy is used with initial and final learning rates set at 0.01 and 0.2, respectively. For optimization, we utilized the Adam optimizer with a weight decay of 0.0005. Additionally, we used a batch size of 64 and trained the model for 300 epochs. These parameters, including batch size and number of epochs, were chosen following the recommendations from the YOLOv7 paper to maintain consistency with their established protocol and to benchmark our results effectively against this standard.

3.3.5 Face Recognition Using Dlib

For face recognition task, the "dlib" library is employed, offering an effective solution that is compatible with embedded systems and does not require the use of complex neural networks. This approach ensures efficient processing suitable for the specific requirements of this application. "Dlib" library is useful and powerful for acquiring face shape predictors by attaining a frontal face detector, 68 face landmarks, and face descriptors. The example image of 68 face landmarks is given in Fig. 5.

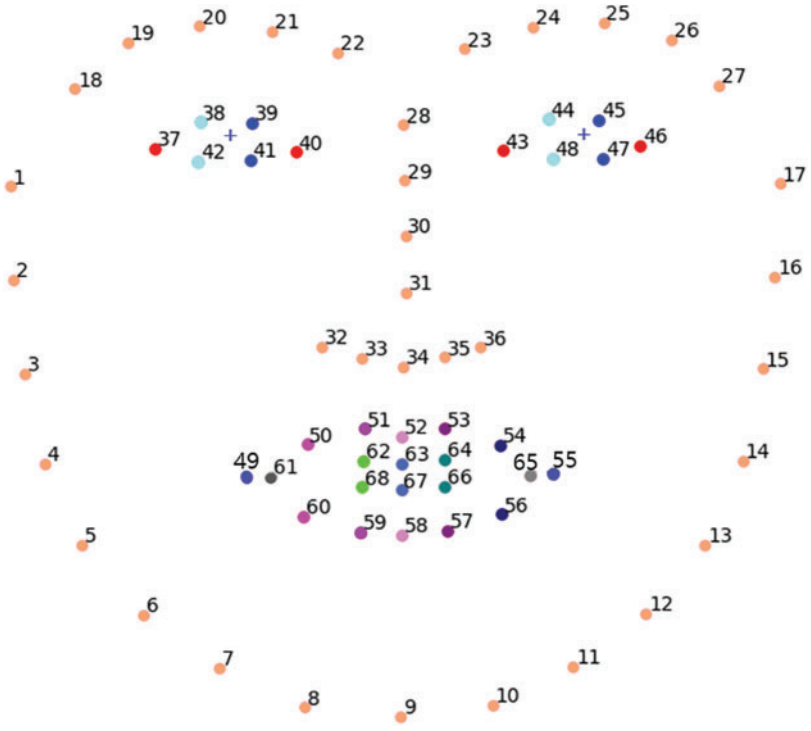

Figure 5: Face landmarks

The figure above shows that "dlib" finds important landmarks like eyes, eyebrows, nose, mouse, chin, lips, forehead, and overall face shape. The information about these parts is essential when comparing two faces. Face recognition is the process of comparing the similarity of the facial features to be identified with the face features in a library in order to determine the identity of the face. Following

the extraction of 68 feature vectors, the *Euclidean distance* between the face feature matrix and the face feature matrix of the training database is calculated. When the distance is less than the stated threshold value, and the feature similarity is high, it is assumed that the two people are the same, or vice versa.

4 Results

In classification tasks, precision is a metric that shows the accuracy of positive results in the model's prediction. Recall measures how effectively the model identifies positive values. However, in object detection, precision denotes the proportion of detected results that are correctly identified. The recall represents the percentage of actual positive results accurately detected and predicted. The detection output of a trained model provides bounding box coordinates and a confidence score that indicates the model's confidence level on each detection. This confidence score allows detections to be sorted in descending order, and precision-recall curves for each class may be constructed based on this ranking. Fig. 6a presents the precision-recall curve for each category.

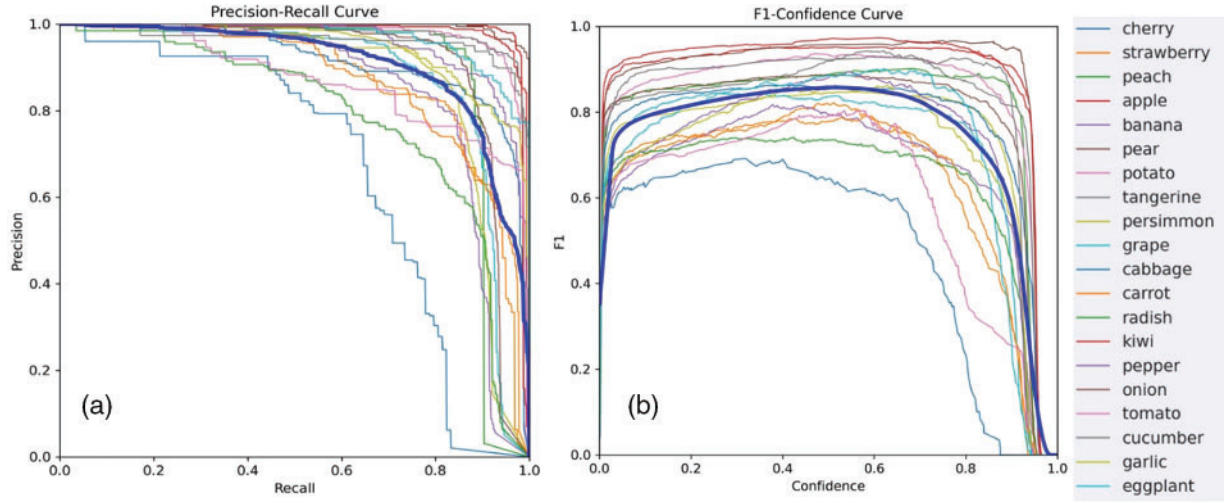

Figure 6: Precision-recall curve (a), F1 score (b), and class types

The average precision (AP) is a helpful statistic for comparing alternative models since it computes the area under the precision-recall curve without considering the confidence score. However, in practice, detections must be performed with a particular level of certainty. The F1 score (F1) is a standard statistic used to estimate the confidence threshold. As demonstrated in Fig. 6b, when the confidence level of a detection grows, accuracy increases, but recall decreases. The aim is to determine that the confidence level maximizes F1 within all classes. A glance at the F1 figure it becomes evident that the maximum value (0.83) of the F1 score is achieved with the confidence threshold of 0.55. Since the goal is to minimize the number of false positives, then the confidence threshold can be increased accordingly. The AP can be computed using the curves in Fig. 6, a number between 0 and 1 that summarizes the precision-recall curve by averaging accuracy across different recall values. It represents the area under the precision-recall curve in essence. When there are several classes, comparing object detectors might be difficult since each class has an AP value. In these circumstances, mean average precision (mAP), the average of AP values for each category, is employed.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

In addition to the above-given metrics, objective, box, and classification losses are used to measure how well the model performs during test and training steps. Fig. 7 illustrates the model performance in object detection.

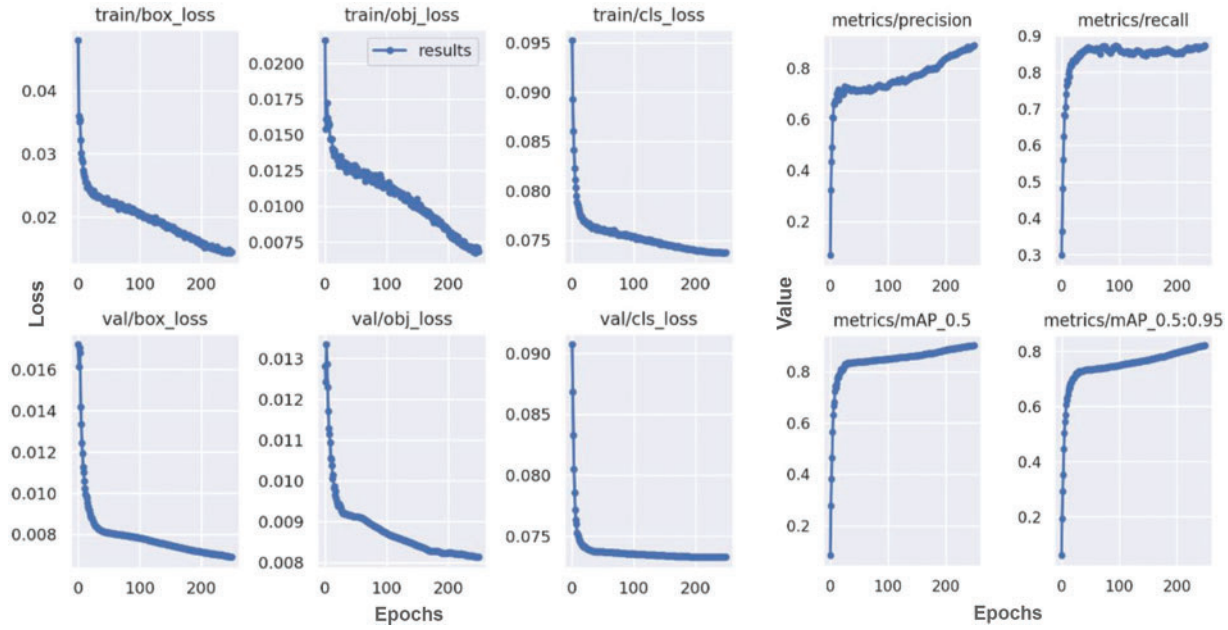

Figure 7: Accuracy results on food crop detection task

The confusion matrix in Fig. 8 represents the prediction probability concerning the actual classes, while the true-positive (TP) results are set on the diagonal. The darker the points are in the matrix, the more the model is robust during test time. But unlike traditional classification task, a confusion matrix in object detection inherited background class, meaning that every anchor box detects either a background or a class. In the confusion matrix given below, the *cherry* class has the lowest TPs representing the prediction of 34% background. The main reason for this result is that the class *cherry* is much smaller in the dataset, and the model confuses this class with background in some cases.

Table 1 below represents a comparative analysis of several prominent object detection models, including Fast R-CNN, Faster R-CNN, YOLO. YOLOX, YOLOv5, YOLOv6 and suggested YOLOv7. The comparison is based on three key metrics: The number of parameters in each model, the frames per second (FPS) and the mean average precision (mAP) percentage.

The comparative analysis highlights the trade-offs between speed, accuracy and model complexity in object detection. YOLOv7, despite its lower speed than YOLOv5 and high complexity, offers the best accuracy among others considered, aligning with the objectives of our study which prioritizes detection precision in real time applications.

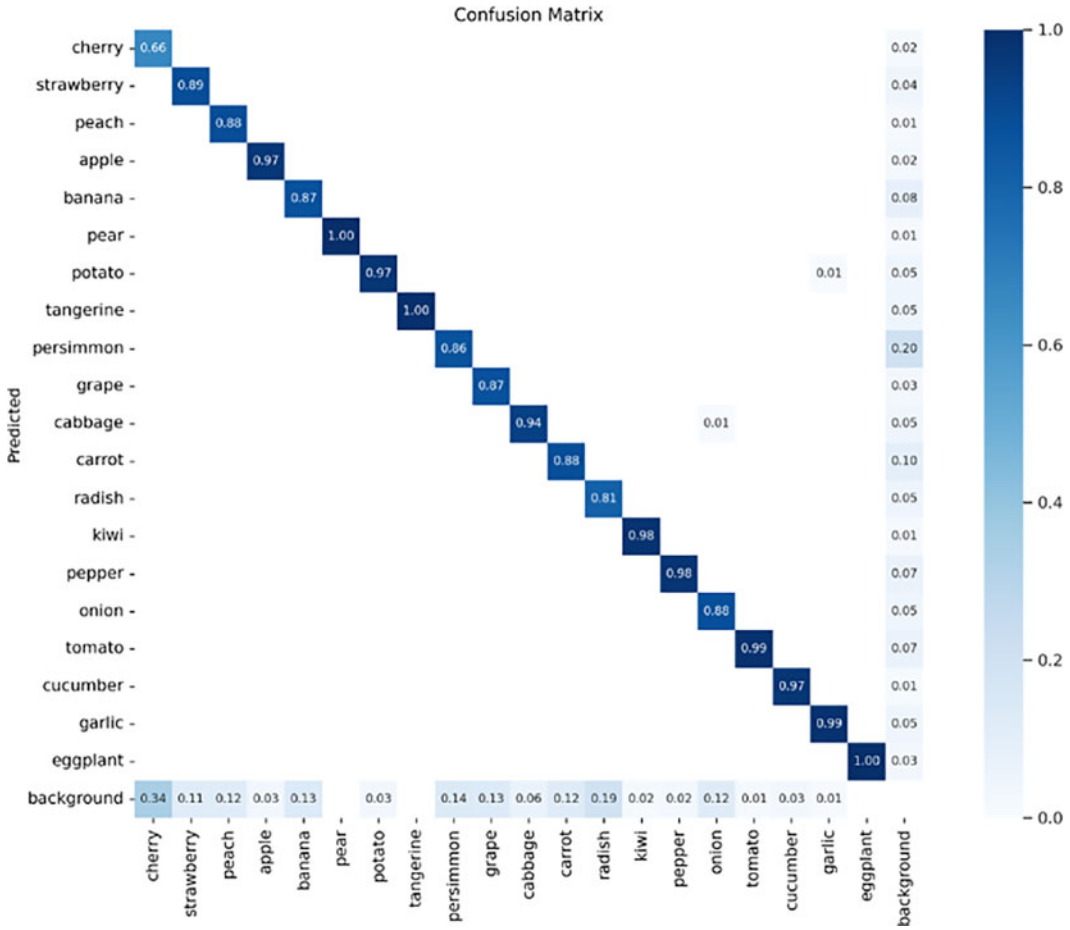

Figure 8: Confusion matrix

Table 1: Comparative analysis of object detection models

Model	No. of parameters	FPS	mAP (%)
Fast R-CNN [36]	138 M	0.5	70.0
Faster R-CNN [37]	25.6 M	7	73.2
YOLO [32]	61 M	33.4	45.92
YOLOX [38]	25.3 M	41	68.47
YOLOv5 [34]	21.2 M	70	64.1
YOLOv6 [35]	27.7 M	57	77.34
YOLOv7 [6]	30.4 M	64	83.27

The detection results for images captured in local markets visited during the research are presented in Fig. 9. The outcomes showcase the practical application of the developed system in real-world environments.

Figure 9: Fruit and vegetable detection results from the local markets

The system will identify a person as a family or non-family member in the face recognition task. Fig. 10 gives some examples of faces recognized as family member.

Figure 10: Face recognition results recognized as family members (with corresponding names)

There are some conditions that the system cannot recognize a person, such as when a person wears a mask or sunglasses or faces and investigates different points of view. Fig. 11 illustrates some of the examples of recognizing as non-family.

Figure 11: Face recognition results recognized as non-family members

5 Conclusion and Future Work

In this study, we suggested a system that simplifies the mobility of VIP, and it can be beneficial in food crop detection and face recognition for VIPs' daily lives. Our model has gained 83.27% mAP in object detection and over 95% accuracy in face recognition, which are reasonably decent results. This research is anticipated to help to improve blind people's lives by making them able to "see" their surroundings. In future developments of this system, the integration of voice commands and audio feedback will be considered to enhance the user experience. This addition aims to make the application more intuitive and accessible, allowing users to interact with the system using natural language commands and receive auditory information about their surroundings. The future scope in object detection will be counting the number of objects in the frame and calculating the distance between a user and a thing to give more precise information to the user. There is also potential to expand the range of detectable objects to include various household items, further aiding in everyday tasks. The face recognition system can be improved by life-long learning when a new person introduces oneself. In such a scenario, a new person will tell one's name, and the system will save his photo and name from learning the face's alignments for future use.

Acknowledgement: Not applicable.

Funding Statement: This work was partially supported by the Korea Industrial Technology Association (KOITA) Grant Funded by the Korean government (MSIT) (No. KOITA-2023-3-003). Also, this research was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) Support Program (IITP-2024-2020-0-01808) Supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: Heechul Jung, Sookhee Ryue; data collection: Soyoun Bang, Sookhee Ryue; analysis and interpretation of results: Heechul Jung, Soyoun Bang, Karshiev Sanjar; draft manuscript preparation: Karshiev Sanjar; supervisor of this research work: Heechul Jung. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that supports the findings of this study are openly available in 'The Open AI Dataset Project (AI-Hub, S. Korea)' at www.aihub.or.kr.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] "Republic of Korea–World Blind Union–Asia Pacific," 2018. Accessed: Dec. 20, 2022. [Online]. Available: https://wbuap.org/archives/category/countries/republic-of-korea

[2] "Korean disabled people's development institute (KODDI)," 2019. Accessed: Dec. 20, 2022. [Online]. Available: https://www.koddi.or.kr/eng/greeting.jhtml

[3] L. Mesquita, J. Sánchez, and R. M. C. Andrade, "Cognitive impact evaluation of multimodal interfaces for blind people: Towards a systematic review," in *Int. Conf. Universal Access Human-Comput. Interact. (UAHCI)*, Las Vegas, NV, USA, 2018, pp. 365–384.

[4] R. S. Mulky, S. Koganti, S. Shahi, and K. Liu, "Autonomous scooter navigation for people with mobility challenges," in *IEEE Int. Conf. Cognitive Comput. (ICCC)*, San Francisco, CA, USA, 2018, pp. 87–90.

[5] I. Tobibul, A. Mohiuddin, and S. B. Akash, "Real-time family member recognition using Raspberry Pi for visually impaired people," in *IEEE Region 10 Symp. (TENSYMP)*, Dhaka, Bangladesh, 2020, pp. 78–81.

[6] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 7464–7475.

[7] A. S. Kumar, P. M. Venkatesh, S. S. Siva, S. Reddy, S. Nagul and S. Sayeed-Uz-Zama, "Fruits and vegetables detection for blind people," *Int. J. Innov. Res. Technol.*, vol. 9, no. 1, pp. 475–476, 2022.

[8] T. Y. Mahesh, S. S. Parvathy, S. Thomas, S. R. Thomas, and T. Sebastian, "CICERONE- A real time object detection for visually impaired people," in *IOP Conf. Series: Mat. Sci. Eng.*, Kanjirapalli, India, 2021, pp. 45–53.

[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv.1804.02767, Apr. 2018.

[10] V. K. Sharma and R. N. Mir, "A comprehensive and systematic look up into deep learning-based object detection techniques: A review," *Comput. Sci. Rev.*, vol. 38, no. 1, pp. 1–29, Nov. 2020.

[11] Y. Xiao *et al.*, "A review of object detection based on deep learning," *Multimed. Tools Appl.*, vol. 79, no. 33–34, pp. 23729–23791, 2020. doi: 10.1007/s11042-020-08976-6.

[12] N. Laptev, V. Laptev, O. Gerget, and D. Kolpashchikov, "Integrating traditional machine learning and neural networks for image processing," in *31st Int. Conf. Comput. Graph. Vis.*, Tomsk, Russia, 2021, pp. 896–904.

[13] M. Rajeshvaree, R. Karmarkar, and V. N. Honmane, "Object detection system for the blind with voice guidance," *Int. J. Res. Appl. Sci. Technol.*, vol. 6, no. 2, pp. 67–70, 2021.

[14] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, 2014, pp. 740–755.

[15] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar and B. Lee, "A survey of modern deep learning based object detection models," *Digit Signal Process.*, vol. 126, no. 1, pp. 1–19, 2022. doi: 10.1016/j.dsp.2022.103514.

[16] Q. Lin and S. Qu, *Let Blind People See: Real-Time Visual Recognition with Results Converted to 3D Audio*. Stanford, CA, USA: Stanford University, Mar. 2016.

[17] T. Ahmed, R. Hoyle, K. Connelly, D. Crandall, and A. Kapadia, "Privacy concerns and behaviors of people with visual impairments," in *Proc. 33rd Annual ACM Conf. Human Factors Comput. Syst.*, Seoul, Republic of Korea, 2015, pp. 3523–3532.

[18] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[19] L. E. Dreer, T. R. Elliott, D. C. Fletcher, and M. Swanson, "Social problem-solving abilities and psychological adjustment of persons in low vision rehabilitation," *Rehabil Psychol.*, vol. 50, no. 3, pp. 232–238, 2005. doi: 10.1037/0090-5550.50.3.232.
[20] S. Chaudhry and R. Chandra, "Design of a mobile face recognition system for visually impaired persons," arXiv:1502.00756, Feb. 2015.
[21] Z. X. Wang, J. Y. Yan, C. Pang, D. Chu, and H. Aghajan, "Who is here: Location aware face recognition," in *Proc. Third Int. Workshop Sens Appl. Mobile Phones*, Toronto, Canada, 2012, pp. 68–73.
[22] K. M. Kramer, D. S. Hedin, and D. J. Rolkosky, "Smartphone based face recognition tool for the blind," in *Annual Int. Conf. IEEE Eng. Med. Biol. Soc.*, Buenos Aires, Argentina, 2010, pp. 4538–4541.
[23] Y. Zhao, S. Wu, L. Reynolds, and S. Azenkot, "A face recognition application for people with visual impairments: Understanding use beyond the lab," in *Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2018, pp. 1–14.
[24] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *Workshop Comput. Vis. Appl. Visually Imp.*, Marseille, France, 2008, pp. 68–81.
[25] D. Wang, H. Yu, D. Wang, and G. Li, "Face recognition system based on CNN," in *Int. Conf. Comput. Inform. Big Data Appl.*, Guiyang, China, 2020, pp. 470–473.
[26] U. N. Akshata and N. Guinde, "LBPH algorithm for frontal and side profile face recognition on GPU," in *Proc. Third Int. Conf. Smart Syst. Invent. Technol.*, Tirunelveli, India, 2020, pp. 776–779.
[27] R. Rosnelly, S. S. Mutiara, S. Ade Clinton, A. Mulkan, K. Sandy, Husen, "Face recognition using eigenface algorithm on laptop camera," in *8th Int. Conf. Cyber IT Service Manag.*, Pangkal, Indonesia, 2020, pp. 1–4.
[28] J. R. Lee, K. W. Ng, and Y. J. Yoong, "Face and facial expressions recognition system for blind people using ResNet50 architecture and CNN," *J. Inform. Web Eng.*, vol. 2, no. 2, pp. 284–298, 2023. doi: 10.33093/jiwe.2023.2.2.20.
[29] P. Junseok, "Image data for small object detection," 2021. Accessed: Oct. 21, 2022. [Online]. Available: https://bit.ly/aihub_small_object
[30] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6469–6477.
[31] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517–6525.
[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788.
[33] A. Bochkovskiy, C. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934, Apr. 2020.
[34] J. Glenn, "YOLOv5," 2022. Accessed: Oct. 22, 2022. [Online]. Available: https://github.com/ultralytics/yolov5.
[35] C. Li *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," arXiv:2004.10934, Sep. 2022.
[36] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1440–1449.
[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. 28th Int. Conf. Neural Inform. Process. Syst.*, Montreal, Canada, 2015, pp. 91–99.
[38] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," arXiv:2107.08430, 2021.