2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia) | 978-1-6654-0857-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCE-Asia53811.2021.9641925

Deep Learning Based Mobile Assistive Device for Visually Impaired People

Chan-Su Lee[1], Jae-Ik Lee[2], Han Eol Seo[1]
[1]*Department of Electronic Engineering, Yeungnam University*
[2]*Haga Cooperation*

chansu@ynu.ac.kr, jilee@multiiq.com , haneol@yu.ac.kr

Abstract

This paper presents a mobile assistive device for the visually impaired. Low vision people had difficulties in object recognition, text reading, and face-to-face communication. We developed a prototype of a portable smart mobile device that supports object detection, text reading, and facial expression recognition using recent deep learning technologies. After applying recent object detection technology and optical character recognition software(OCR), and facial expression recognition, we converted the networks for Android-based mobile devices. We developed a portable mobile device with enhanced battery power and a high-performance AP with a line-connected camera device. The system provides detected object names verbally when looking at the object by face direction and by indicating using a hand finger. The proposed device can assist the communication of the low vision with a normal person in addition to reduce threats or dangers due to unrecognized or mis-detected objects. The proposed system will support social interaction and a better life for visually impaired people.

Keywords: Visual impaired, Assistive device, Deep Learning, OCR, Object recognition, Facial expression recognition

1. Introduction

The number of people with blind or visually impaired is approximately 2.2 billion according to the World Health Organization [1]. The need for assistive devices for navigation and orientation has increased. Many assistive devices such as smart cane, eye substitution, a fusion of artificial vision and GPS, and cognitive guidance systems have been developed [2]. Recently computer vision technology based on deep learning is also applied for the assistance of navigation or location awareness [3-4]. However, there are no assistive devices to help face-to-face communication for the visually impaired.

Understanding facial expressions is very important for efficient face-to-face communication in addition to understanding indicated objects and reading texts presented in front of the person. Therefore, in this paper, we present a prototype system that can recognize facial expressions in addition to object recognition and text reading. For the portability of the device, we developed a customized embedded system based on mobile AP for smartphones.

2. System Overview

The proposed system is composed of three components: video input from an RGB camera, a processor with button input, and audio output using earphones. The audio out is transferred to the user using earphones. The processor and interfaces are customized for a portable and user-friendly button interface. For the deep learning processing on the platform, a high-end AP used for a recent smartphone is used. Android OS is used for easy porting of the developed TensorFlow-based deep learning algorithm to the device. Fig.1 shows the data flow of the overall system.

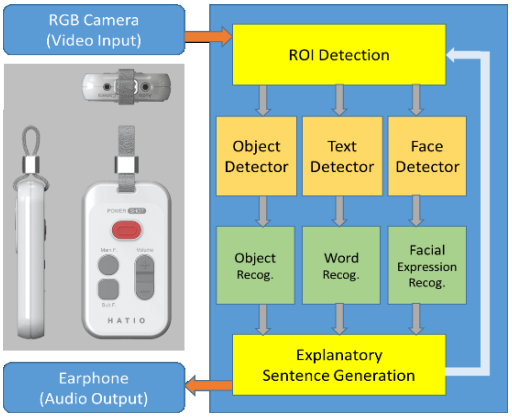

Figure 1: System overview

From camera input, we extract ROI(Region of Interest) for the processing of the image. The ROI is initially the center of the image. However, the ROI is adapted depending on the sub-task to process. In object

detection sub-task, a larger area is used to search for objects. During the text recognition sub-task, the ROI is reduced to the given specific text area. In the facial expression sub-task, higher weights are given to the upper area of the image, where more chances to detect a face. Therefore, the ROI is selected not only by the given image but also by the previous task processed. For efficient management of the ROI, the situation awareness based on detected objects and potential activities of the user can enhance the performance of the system. In addition, the ROI can be selected by finger indication or face direction.

After processing sub-tasks, the device provides an explanatory sentence to the visually impaired user. In the text reading, simply TTS(Text to Speech) can convert the recognized word to the voice. However, object recognition, and facial expression recognition require additional processing to generate text efficiently for the user. Repeated TTS generation of the recognized objects or expressions can be annoying to the user. Therefore, friendly explanatory sentence generation is very important for the acceptability of the device.

3. Implementation and Experimental Results

Mobile Embedding System: We implemented the mobile embedded system based on the Samsung Exynos8895 octa-core CPU for the onboard real-time processing of image data using deep learning. Android 10 is used as the operating system for the device. The embedding system provides Bluetooth connection, and external display connection for debugging propose in addition to camera input and button inputs. Fig. 2 shows the mobile prototype device.

Figure 2: Developed prototype device

Object detection and sentence generation: We employed YOLO-SPP [5] for the implementation of real-time object detection. YOLO-SPP model is implemented in the TensorFlow framework, which is supported in our mobile device. Real-time processing is possible using the YOLO-light model. The original YOLO3 model is too heavy to run in the device. It is implemented using TensorFlow Lite, which is developed for a mobile device with a similar performance to the conventional desktop system with limited resources. The module recognizes objects with 96.46% accuracy for 500 test images. Figure 3 shows some examples of object detection such as chair, laptop and cup from the proposed model.

Figure 3: Examples of object detection (Chair, Laptop, and Cup)

Text Recognition and TTS generation: We employed Tesseract for text recognition, which is an open-source OCR engine that has gained popularity among the OCR community. Internally, Tesseract does various image processing operations. However, the character recognition performance depends a lot on image pre-processing. We detected potential areas for text recognition such as paper, memo, plate, and signboard. We first applied adaptive binarization after Gaussian filtering to remove noise. Then, we estimate the bounding box of the text region by detecting the edge and corner area. After applying cropping operation to the rectangle area, we applied tesseract OCR algorithms in the OpenCV program on the Android platform. Our system recognizes character by 98.00% for Korean character text recognition evaluation.

Face detection and facial expression recognition:

For the facial expression recognition, we first detect faces using the Dlib library. The Android porting process was required to use the Dlib face detection algorithm in the mobile device. The detected face was resized to feed to the deep learning facial expression recognition.

Table 1: Confusion matrix for FER 2013 facial expression database

	Ang.	Dis.	Fear	Hap.	Sad	Sur.	Neu.
Ang.	**0.66**	0.02	0.10	0.03	0.11	0.02	0.06
Dis.	0.11	**0.77**	0.03	0.01	0.13	0.04	0.01
Fear	0.08	0.00	**0.64**	0.02	0.13	0.08	0.05
Hap.	0.01	0.00	0.01	**0.90**	0.02	0.02	0.03
Sad	0.08	0.00	0.14	0.02	**0.64**	0.01	0.11
Sur.	0.02	0.00	0.08	0.04	0.02	**0.82**	0.03
Neu.	0.07	0.00	0.06	0.06	0.15	0.01	**0.65**

For facial expression recognition, we use MobileNet SSD [6] for the mobile implementation of the network. For the facial expression recognition, we used FER2013 database [7] and CK+ database [8] for the training and testing of the neural network. We used a pre-trained model at this moment. The evaluation of facial expression recognition using FER 2013 shows 72.6% recognition accuracy. Table 1 shows confusion matrix for the FER 2013 facial expression recognition.

.

4. Conclusions and Future works

This paper presents a portable assistive device for visually impaired people. Deep learning technology such as YOLO, Tesseract, and MobileNet are used for object detection, text reading, and facial expression recognition. The system can support social interaction such as communication, and dialogue for visually impaired people.

At this moment, the detection routine for the object, text, and face are separated and different modules are used for each sub-task. In the future, we combine them as a single detection system, which can combine all the detection routines in a single network and directly interpret the situation and process required recognition in the following. In addition, the explanatory sentence generation routine is also enhanced by employing deep-learning-based scene descriptions and language generation techniques.

Acknowledgement:

This research was partially supported by Basic Science Research Program through the National Research Foundation of Korea((NRF) funded by the Ministry of Education (2021R1A6A1A03040177)

References

[1] World Health Organization. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs282/en/

[2] W. Elgannai and K. Elleithy, ”Sensor-Based Assistive Devices for Visually-Impaired People: Current Status, Challenges, and Future Directions”, Sensors, vol. 17, 2017, pp.565.

[3] M. Poggi and S. Mattoccia, ”A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning,” IEEE Symposium on Computers and Communication (ISCC), Messina, 2016, pp. 208-213

[4] Y. Lin, K. Wang, W. Yi, and S. Lian, ”Deep Learning based Wearable Assistive System for Visually Impaired People”, In ICCV/ACVR, 2019.

[5] Z. Huang, J. Wang, ”DC-SPP-YOLO: Dense Connection and Spatial Pyrimid Pooling Based on YOLO for Object Detection”, https://arxiv.org/abs/1903.08589.

[6] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, ”MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, https://arxiv.org/abs/1704, 2017.

[7] Goodfellow I.J. et al. ”Challenges in Representation Learning: A Report on Three Machine Learning Contests”, In Proceedings of ICONIP, Lecture Notes in Computer Science, vol 8228, 2013.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, ”The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94-101, 2010.