Assistive Technology for the Visually Impaired: Optimizing Frame Rate (Freshness) to Improve the Performance of Real-time Objects Detection Application

Basel Barakat[1][0000-0001-9126-7613], Aiste Steponenaite[2][0000-0002-1988-3419], Gurprit S. Lall[2][0000-0003-3725-4795], Kamran Arshad[3][0000-0002-4447-8335], Ian J. Wassell[4][0000-0001-7927-5565] and Simeon Keates[5]

[1] School of Engineering, University of Greenwich, Kent, ME4 4TB, UK
[2] Medway School of Pharmacy, University of Kent, Chatham, UK
[3] Faculty of Engineering, Ajman University, Ajman, United Arab Emirates
[4] Computer Laboratory, University of Cambridge, Cambridge CB2 1TN, UK
[5] School of Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT
b.barakat@gre.ac.uk

Abstract. It has been 100+ years since the world's first commercial radio station started. This century witnessed several astonishing inventions (e.g. the computer, internet and mobiles) that have shaped the way we work and socialize. With the start of a new decade, it is evident that we are becoming more reliant on these new technologies as the majority of the world population relies on the new technology on a daily basis. As world's population is becoming reliant on new technologies and we are shaping our lives around it, it is of paramount importance to consider those people who struggle in using the new technologies and inventions. In this paper, we are presenting an algorithm and a framework that helps partially sighted people to locate their essential belongings. The framework integrates state-of-the-art technologies from computer vision, speech recognition and communication queueing theory to create a framework that can be implemented on low computing power platforms. The framework verbally communicates with the users to identify the object they are aiming to find and then notify them when it is within the range.

Keywords: Assistive Technologies, Visual Impaired, Artificial Intelligence, Machine Learning, Real-time Objects Detection, Information Freshness.

1 Introduction

The advances in healthcare and vaccines have significantly increased life expectancy in most of the world. For instance, for England and Wales, if a female was born in 1906, her life expectancy was less than 60 years, on the other hand, if she was born in 2016, her life expectancy is 94 years [1]. With such improvements in healthcare, it very important to consider its implications. Vision impairment is currently affecting more than 2.2 billion people worldwide [2], with the majority of the affected people

aged over 50. Hence, it is predicted that in the future more people will be visually impaired. For these people some of their daily activities (such as locating their essentials) might be challenging. Consequently, it is of paramount importance to design technologies assist the visually impaired.

Alongside the advances in healthcare, the last decade has seen an extraordinary innovation in technologies, such as telecommunication networks, robots and artificial intelligence. One of the main effects, is the vast spread of smartphones, tablets, wearable devices and virtual assistants (such as Amazon Alexa and Google Echo). The spread of these technologies has changed the methods people use to communicate and acquire information. On the other hand, the use of these technologies is challenging for the visually impaired. Most of these technologies depend on graphical interfaces to operate and many offer applications that can be used by the visually impaired.

Advances in artificial intelligence have developed several algorithms that help the computers to identify objects in images [3],[4],[5] and [6]. One of the main challenges faced deploying such algorithms is the computing complexity that is reflected in the time required to process the images [7]. Hence, Graphics Processing Units (GPUs) are typically used to minimize the processing time. However, GPUs are expensive and so can be beyond the reach of many visually impaired users. Thus, in this paper we proposed a framework to use the Central Processing Unit (CPU) that exists in a conventional low-cost computer.

In this paper, the proposed algorithm is aimed to empower visually impaired users. When the framework is initiated the smartphone verbally asks the user what the object, he/she is looking for, using the smartphone speakers. Afterwards, the user answers with the object they are interested in finding. The user then moves the smartphone around so that the camera captures the surrounding scene. At the moment, the camera 'sees' the object of interest, the smart phone declare that it has 'found your object'.

This paper is organized as follows; section II presents the main components of the eye and vision mechanism. Section III – presents the effect of visual loss on the economy. Section IV shows the proposed solution, which utilizes a smartphone and a computer. Section V presents the framework optimization and section VI shows the achieved results and the performance of the proposed framework. Section VII concludes this paper.

2 The Eye and Human Vision

The eye is an optical instrument allowing us to interpret light and generate images of the world surrounding us. Eyes are adapted to the visible part of the light spectrum; therefore, they can detect wavelengths from 380 to 700 nanometers. Light reflecting off an object located in the vision field is decoded by the eye in a complex series of events. Firstly, light is refracted by the cornea which helps to focus the light. The total refractive power of the eye is 60 diopters, and cornea alone provides 40 diopters [8]. The light then travels through aqueous humor and passes through the pupil. The pupil

is the opening in the center of the colored part of the eye called the iris. Pupil size can be changed through the muscular contraction of the iris which controls the amount of light reaching the retina – the photoreceptive part of the eye. After passing through the pupil, light reaches the crystalline lens. The lens, through changes in shape, is able to focus the light onto a single defined point of the retina. After passing through the lens, light travels through the center of the eyeball filled with jelly like vitreous body until it reaches the retina. A diagram showing the eye main parts is shown in Fig. 1.
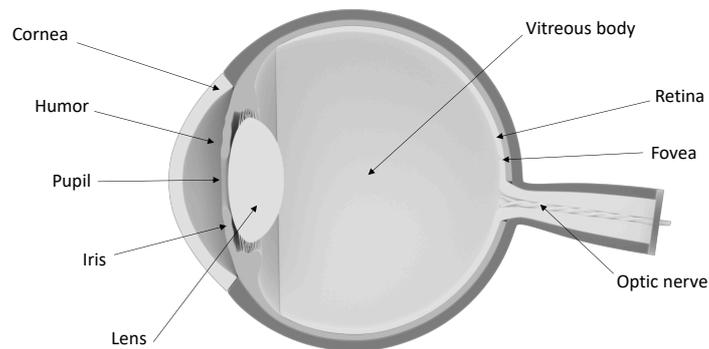
Fig. 1. Diagram of a human eye showing its structures.

The retina is a thin tissue lining the inside of the eyeball. It decodes the image and sends the information to the brain. At the back of it is a layer of photoreceptors known as rods and cones. The human retina has approximately 120 million rods and 6 million cones [8]. Rods are very sensitive photoreceptors; they function in dim light and their image integration time is about 10 milliseconds. Cones, on the other hand, start working at higher light intensities and they can detect rapid light fluctuations, with an integration time approximately 10 times shorter than in rods. For example, when watching a tennis game, it is cones that allow you to see the flying tennis ball. The photoreceptors convert light information to electrical signals traveling through the optic nerve to the visual center at the back of the brain known as the visual cortex [9], [10].

One of the remarkable features of human vision is being able to perceive visual flicker artefacts at rates over 500 Hz when high frequency spatial edges are present [11]. In contrast, traditional movie theatres and televisions present images at a rate of 48-60Hz and a frame rate of 72 Hz was believed to be sufficient to avoid flicker. This proves that retina has incredibly rapid image processing rate that could also potentially help in memory retention of fast changing images. Indeed, work by [12] has shown that after seeing a series of pictures, viewers can detect and retain information about named targets they have never seen before at a rapid serial presentation of 13 milliseconds per image.

3 Visual Impairment in the World and the Associated Economic Impact

Visual recognition of objects is an important part of the behavioral survival instinct for most animals. As humans, we rely heavily on identification of a wide range of visual objects, with the ability to identify them even in unusual orientations, different illumination and in visually complex environments [13]. However, with age or in presence of certain diseases, vision can be affected resulting in difficulties in generating high resolution images with declines in spatial acuity.

A report published by the World Health Association has stated that the estimated number of visually impaired people in the world is approximately 285 million, with the majority of people being 50 years and older [14]. In the UK alone, there are almost 2 million people living with sight loss, accounting for 3% of the population. It is estimated that because of demographic ageing, the prevalence of sight loss and blindness will increase to approximately 4 million by 2050 [15].

There are several conditions that can lead to sight loss and blindness. Age-related macular degeneration (AMD) is the deterioration of the central retina resulting in vision loss, with no currently available effective treatment. Another common condition is that of a cataract, which is clouding of the lens due to protein clumping that can cause blurry vision and faded colors. Luckily, it can be treated by surgical intervention. In diabetic retinopathy, retinal blood vessels are damaged leading to the vision loss. Another common disease affecting vision is glaucoma, raised fluid pressure within the eye, where damage to the optic nerve can cause partial and eventually full blindness. These are just a few the most prevalent visual impairment causes, and most of them are age related.

In 2017 the global population of aged 60 years and over was 962 million, which is more than twice compared with the 1980 demographics. It is expected that by 2050 the number of older people will double again reaching nearly 2.1 billion [16]. According to World Bank Population Estimates and Projections data, the proportion of population aged 60 years and older increased by 7.2% in the last 59 years. Another 7.5% increase is predicted to happen by 2050 (Fig. 2). Considering the rapid rate of aged population increase in the world, it is certain that the prevalence of age-related visual impairment conditions will continue to rise which will further increase the economic burden.
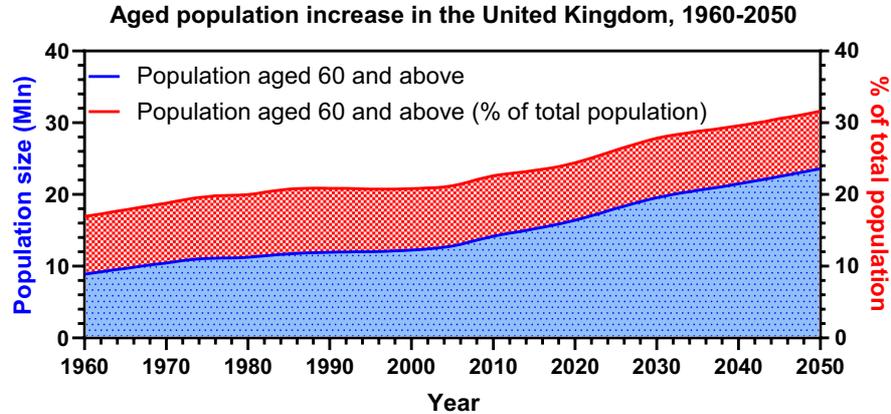
Aged population increase in the United Kingdom, 1960-2050

Fig. 2. Aged population data in the UK. Data collected since 1960 shows linear increase in people population aged 60 and above. Future predictions suggest continuous aged population growth, with another nearly 8 million people will be older than 60 years old than it was in 2019 [17].

In the US, the economic burden of eye disorders and vision loss was estimated to be $139 billion, based on the 2011 expenditure adjusted to 2013 using Consumer Price Index [18]. Whereas in the UK, sight loss and blindness in the adult population cost £28.1 billion to the government. Out of that, £410 million were spent on devices and modifications for visually impaired people [5]. The most common ones include mobility and communication devices, optical aids and home modifications [28]. Often the specialized equipment or software adapted to people's needs is costly and it might be difficult to get, therefore cost-effective, easy to use and reliable tools to assist in day to day life would be of great benefit to the users.

4   Proposed Framework

The proposed system consists of two main elements. The first one, is the smartphone and the second is the personal computer (pc). The smartphone is the first point of contact, that users communicate with, while the pc is the server, where object detection occurs, since it has the necessary computing power. Separating the two nodes minimizes the smartphone energy consumption and gives the developers flexibility when updating the system.

To assist visually impaired users, it is necessary to minimize the dependences on a graphical interface. Hence, in the proposed framework, the users interact with the system verbally. The smartphone asks the user what the object he/she is looking for and the user replays naming the object they are interested in finding. Table 1 contains the list of all the currently recognized objects [19]. The user then moves the

smartphone around so that the camera captures the surrounding environment. At the moment the camera 'sees' the object of interest, the smart phone says 'found your object'.

Table 1. List of Objects recognized by the proposed system [19]

Person	Elephant	Wine Glass	Dining table	Fire Hydrant
Bicycle	Bear	Cup	Toilet	Stop Sign
Car	Zebra	Fork	Tv monitor	Parking Meter
Motorbike	Giraffe	Knife	Laptop	Bench
Aeroplan	Backpack	Spoon	Mouse	Bird
Bus	Umbrella	Bowl	Remote	Cat
Train	Handbag	Banana	Keyboard	Dog
Truck	Tie	Apple	Cell Phone	Horse
Boat	Suitcase	Sandwich	Microwave	Sheep
Traffic Light	Frisbee	Orange	Oven	Cow
Broccoli	Toaster	Hot Dog	Refrigerator	Donut
Carrot	Sink	Pizza	Book	Cake
Chair	Scissors	Sofa	Teddy Bear	Potted plant
Bed	Toothbrush	Hair Drier	Vase	Surfboard
Skis	Clock	Bottle	Tennis Racket	Skateboard
Snowboard	Baseball Glove	Baseball Bat	Kite	Sports Ball

When the application starts the smartphone uses the text to speech functionality that exists on most smart phones to inform the user that the application is running and is ready to start searching for the object of interest by announcing 'What are you looking for?'. The users replay by saying the name of the object (such as a Book). The smartphone then recognizes the user's input using its speech recognition function and then transmitting it to the pc via the network. The computer then initiates the receiving of the video from the smartphone and recognizing the objects in each frame. It is worth mentioning that the streamed video would not be saved anywhere to ensure the privacy of the users.

After receiving the name of the object, the pc begins detection objects on the frames. The objects detection algorithm used is the 'You Only Look Once' (YOLO V3) object detection algorithm using the 'Tiny' weights [6] implemented in Python using Pytorch [27] that was pretrained on the objects of Common objects in context

(COCO) dataset [19] shown in Table 1. As soon as the object is detected by the algorithm it sends a message back to the smartphone. The smartphone then informs the user that the object has been found. The flowchart of the system is presented in Fig. 3.
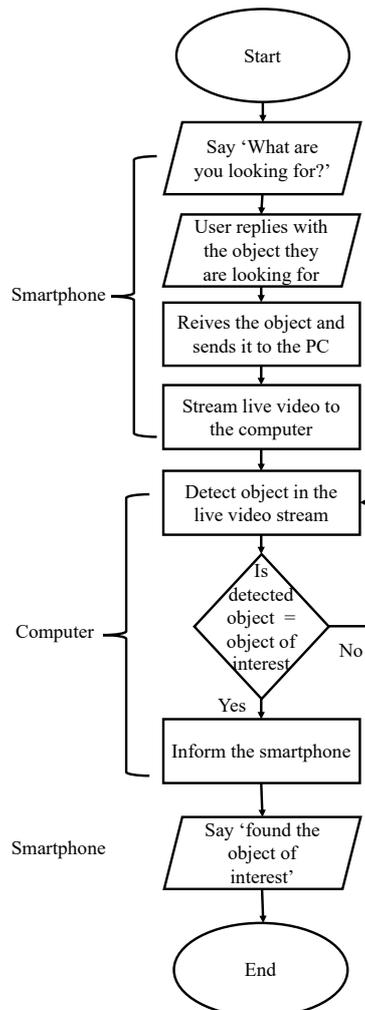
Fig. 3. Flowchart of the framework showing the main tasks performed by the smartphone and the computer.

5   Optimizing the Proposed Framework Response Performance

To optimize the performance of the objects detection, we have modeled the framework as a queue were the video frames per second (fps) represent inter-arrivals rate

(λ) of the queue and the object detection time represents the service time as shown in Fig. 4. Using this model, we derived an optimal expression for the number of fps to maximize the freshness of the frames at the destination (computer) and hence the performance of the framework [20].
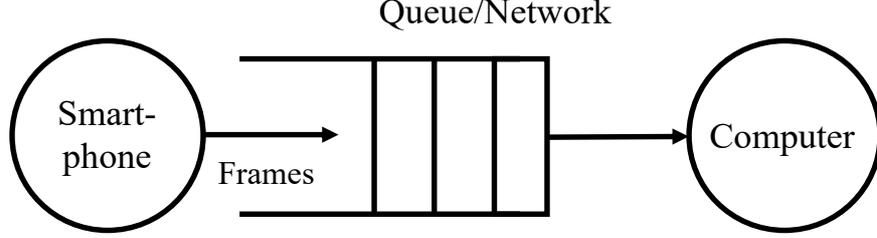

Fig. 4. Queueing system shows the frames generated from the smartphone and received by the computer. The service time of the destination represents the duration required to detect the objects and the inter-arrival time represents the number of frames generated per second.

To evaluate the frames freshness, we use the Information Peak Age (PA) as a metric. PA was introduced in [21],[22] and [23], as a metric representing the information freshness. It is defined as the maximum time elapsed since the preceding piece of information was generated. In other words, PA (P) is:

$$P = E\{X_n + T_n\}, \forall n, \tag{1}$$

where X_n represents the inter-arrival time of frame n or

$$X_n = \frac{1}{\lambda_n}, \tag{2}$$

where λ is the interarrival rate or the number of fps. T_n represents the delay time of a frame or

$$T_n = W_n + S_n, \tag{3}$$

where W_n is the waiting time (queueing time) and S_n is the service time of frame n.

To optimize the fps rate, it is necessary to accurately model the service time and interarrival time. Hence, the PA value depends on the distribution. For instance, if the inter-arrival time and service time follow an exponential distributions with means $\frac{1}{\lambda}$ and $\frac{1}{\mu}$ respectively, the PA is

$$P^{M/M/1} = \frac{1}{\mu}\left(1 + \frac{1}{\rho} + \frac{\rho}{1-\rho}\right), \tag{4}$$

where the server utilization $\rho = \frac{\lambda}{\mu}$ [23]. For exponential distribution interarrival time and deterministic service time PA is

$$P^{M/D/1} = \frac{1}{\mu}\left(1 + \frac{1}{\rho} + \frac{\rho}{2(1-\rho)}\right). \tag{5}$$

The PA for deterministic service time and inter-arrival time is,

$$P^{D/D/1} = \frac{1}{\mu} + \frac{1}{\lambda}. \tag{6}$$

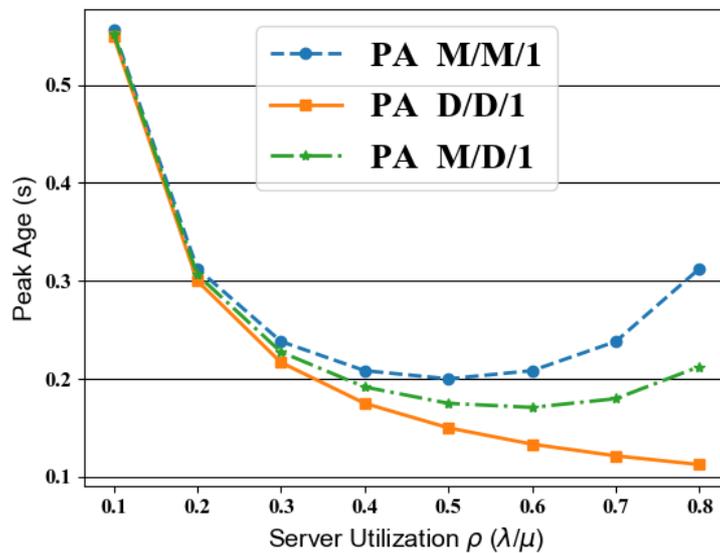Fig. 5, shows the PA for $\mu = 20$; showing the effect of different distributions on the PA value.

Fig. 5. Peak Age versus Server utilization for M/M/1, D/D/1 and M/D/1 queues for service rate (μ) equals to 20 frames per second (fps).

From Fig. 5, it is notable that the distribution of the service time and inter-arrival time plays a critical role in the PA. Thus, we started the optimization procedure by measuring the service time (i.e., the object's detection time) to optimize the number of fps to have the system responsive as shown in the next section.

6   Experiment Results

In this section, the performance of the proposed system is presented. We used a 'low' video resolution of 160x120 to minimize the duration of the communication. The object recognized from the user voice input is communicated to the pc through the network using the Python Socket module [24] and [25]. Initially, we started measuring the time the PC takes to recognize the objects, i.e., the service time. Table 2, presents the measured service time statistics. Although the statistics are insightful, the

service time behavior with time can help us to optimize performance. For instance, if we consider the time series of the first 1000 frames presented Fig. 6; it is notable that the service time is stabilizing and decreasing with time.

Table 2. Service Time Statistics

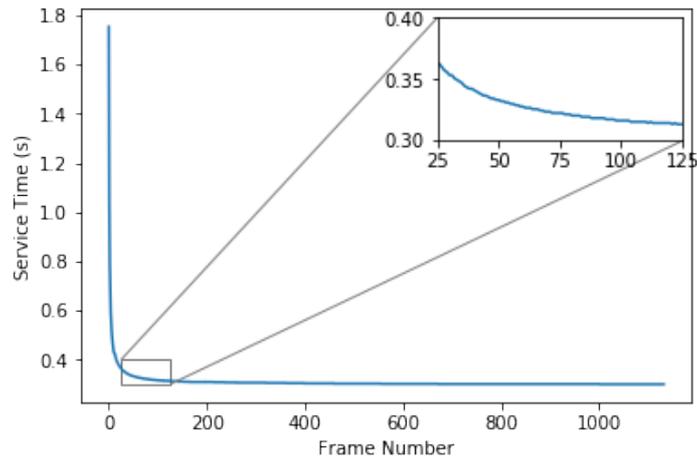| Parameter	Time (s)
Mean	*0.31*
Median	*0.3*
Standard Division	*0.06*
Maximum Value	*1.75*
Minimum Value	*0.3*

Fig. 6. The service Time representing the time for processing and recognizing the objects in each frame.

After identifying the behavior of the service time, it is necessary to identify its distribution. To find the distribution that best fits the service time, we used the Fitter package [26]. The Fitter package tests the data against 80 distributions to identify the best representation. Table 3 shows the best five fitting distributions obtained from the fitting. However, it was found that conducting with more tests, the best fitting distribution changes, hence in Fig. 7 we have plotted the histogram of the achieved service times. We see that most of the frames were served in 0.3-0.31 s. Thus, although the service time is not constant, the variation is not significant, and we can assume that the service time is constant and deterministic.

Table 3. The Service Time Data Fitting Results

| Distribution	Sum Square Error	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)
Generalized Pareto	*223.661544*	*5225.025898*	*-1817.183655*
Exponential	*256.870603*	*13496.455939*	*-1667.365434*
Generalized Exponential	*257.189093*	*13607.637665*	*-1644.863641*
Exponentially Modified Normal	*258.963944*	*13501.576277*	*-1651.136959*
Generalized Half-logistic	*270.562111*	*16640.298840*	*-1601.497004*
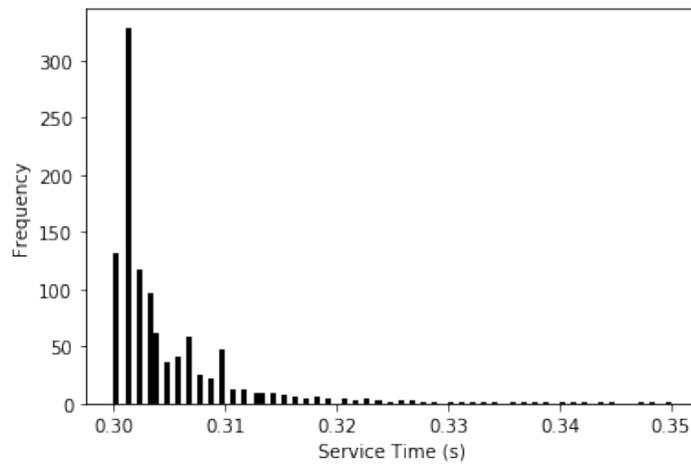
Fig. 7. Histogram of the achieved service time. The data shows that the majority of frames were processed in 0.3 - 0.31 seconds, hence, we can assume that the service time can evidently be considered deterministic.

Since, the service time is evidently deterministic, we can minimize the PA by using the PA M/D/1 queue as shown in Fig. 5. The interarrivals time was assumed to follow exponential distribution hence the frames will be communicated through the network and the delays time can be considered to be following exponential distribution. Thus, we have optimized the fps rate to achieve the minimize the PA. Fig. 8 presents an

experiment where the smartphone was capturing a video of a clock on the computer screen before and after optimizing the frames freshness. We can observe that the difference in the two clocks is approximately 1532 milliseconds. After optimizing the fps, the difference was less than 130 milliseconds as shown in Fig. 9.
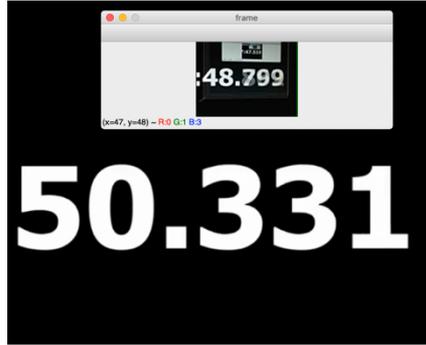

Fig. 8. The difference between the captured time and the actual time before optimizing the frames per second rate.
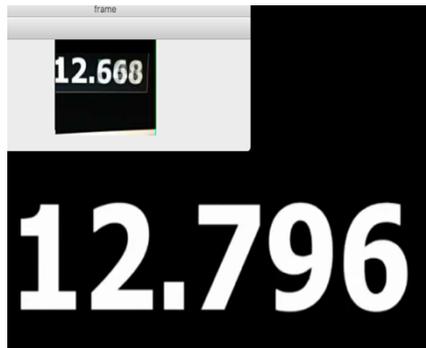

Fig. 9. The difference between the captured time and the actual time after optimizing the frames per second rate.

After optimizing the system response time, we have tested it on several objects to ensure the usability of the proposed system. Fig. 10 presents some the captured objects with different backgrounds and lighting. The proposed framework source code and instructions can be found in [29].
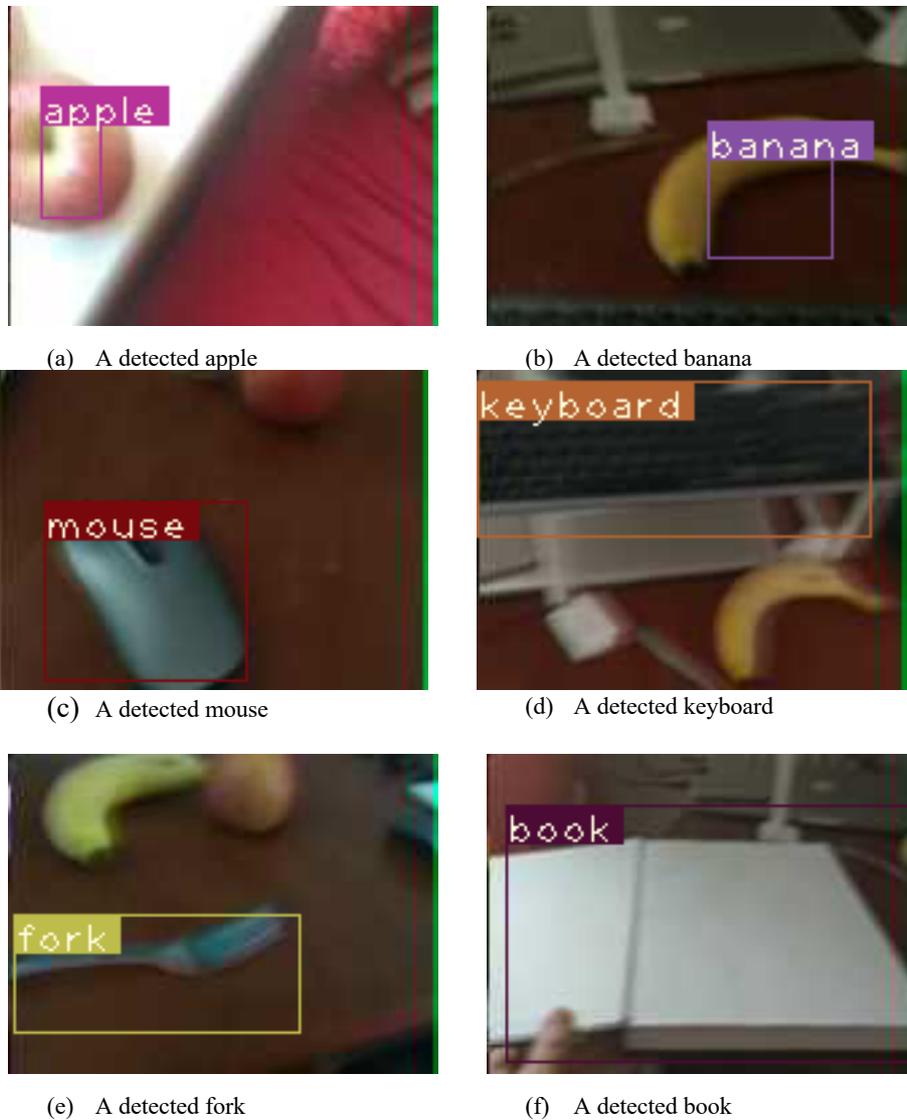
(a)　A detected apple

(b)　A detected banana

(c)　A detected mouse

(d)　A detected keyboard

(e)　A detected fork

(f)　A detected book

Fig. 10. Examples of detected objects using the proposed system

7　Conclusions

In this paper, we have presented a system that can help visually impaired users in navigating their surroundings. The system runs on a smartphone and a computer. The smartphone is used for communicating with the users and the computer is used for the

object's detection for each frame. The system is able to identify 82 objects that users use in their everyday life. We optimized the number of frames captured from the camera to optimize the performance of the system. In the future work, the system would be able to adapt its frames rate according to the network and processing power of the pc.

References

1. Office for National Statistics, What is my life expectancy? And how might it change?. Retrieved February 27, 2020, from https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlife expectancies/articles/whatismylifeexpectancyandhowmightitchange/2017-12-01

2. World Health Organization, Blindness and vision impairment. (2019). Retrieved from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

3. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

4. Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

5. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S. and Lew, M.S., 2016. Deep learning for visual understanding: A review. Neurocomputing, 187, pp.27-48.

6. Redmon, J. and Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv pre-print arXiv:1804.02767.

7. R. Huang, J. Pedoeem and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 2503-2510.

8. S. H. Schwartz, Visual Perception: A Clinical Orientation (4th Edition). 2009.

9. P. Artal, "Image Formation in the Living Human Eye," Annu. Rev. Vis. Sci., vol. 1, no. 1, pp. 1–17, 2015, doi: 10.1146/annurev-vision-082114-035905.

10. H. Kolb, "How the retina works," Am. Sci., vol. 91, no. 1, pp. 28–35, 2003.

11. J. Davis, Y. H. Hsieh, and H. C. Lee, "Humans perceive flicker artifacts at 500 Hz," Sci. Rep., vol. 5, p. 7861, 2015, doi: 10.1038/srep07861.

12. M. C. Potter, B. Wyble, C. E. Hagmann, and E. S. McCourt, "Detecting meaning in RSVP at 13 ms per picture," Attention, Perception, Psychophys., vol. 76, no. 2, pp. 270–279, 2014, doi: 10.3758/s13414-013-0605-z.

13. N. K. Logothetis, "Visual Object Recognition," Annu. Rev. Neurosci., vol. 19, no. 1, pp. 577–621, 1996, doi: 10.1146/annurev.neuro.19.1.577.

14. World Health Organization, "Global data on visual impairments 2010," 2010.

15. L. Pezzullo, J. Streatfeild, P. Simkiss, and D. Shickle, "The economic impact of sight loss and blindness in the UK adult population," BMC Health Serv. Res., vol. 18, no. 1, pp. 1–13, 2018, doi: 10.1186/s12913-018-2836-0.

16. U. N. Department of Economic and Social Affairs, Population Division, "World Population Ageing 2017 - Highlights (ST/ESA/SER.A/397)," 2017.

17. World Bank, "Population Estimates and Projections data", (2019). Retrieved from https://datacatalog.worldbank.org/dataset/population-estimates-and-projections

18. J. Wittenborn and D. Rein, "Cost of Vision Problems The Economic Burden of Vision Loss and Eye Disorders in the United States," 2013.

19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014, September. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

20. Barakat, B., Keates, S., Wassell, I. and Arshad, K., 2019, July. Adaptive status arrivals policy (ASAP) delivering fresh information (minimise peak age) in real world scenarios. In International Conference on Human-Computer Interaction (pp. 419-430). Springer, Cham.

21. Huang, L. and Modiano, E., 2015, June. Optimizing age-of-information in a multi-class queueing system. In 2015 IEEE International Symposium on Information Theory (ISIT) (pp. 1681-1685). IEEE

22. Costa, M., Codreanu, M. and Ephremides, A., 2016. On the age of information in status update systems with packet management. IEEE Transactions on Information Theory, 62(4), pp.1897-1910.

23. Barakat, B., Yassine, H., Keates, S., Wassell, I. and Arshad, K., 2019, May. How to Measure the Average and Peak Age of Information in Real Networks?. In European Wireless 2019; 25th European Wireless Conference (pp. 1-5).

24. Sarker, M.F., 2014. Python Network Programming Cookbook. Packt Publishing Ltd.

25. Ferrill, P. and Grammens, J., 2011. Pro Android Python with SL4A. Apress.

26. {fitter}, fit data to many distributions. Retrieved from https://github.com/cokelaer/fitter

27. Ketkar, N., 2017. Introduction to pytorch. In Deep learning with python (pp. 195-208). Apress, Berkeley, CA.

28. A. Lafuma et al., "Evaluation of non-medical costs associated with visual impairment in four European countries: France, Italy, Germany and the UK," Pharmacoeconomics, vol. 24, no. 2, pp. 193–205, 2006, doi: 10.2165/00019053-200624020-00007.

29. BaselBarakat/Assistive-Technology-for-the-Visually-Impaired-Optimizing-the-Frames-Freshness-of-Real-time-Objects. Retrieved March 1, 2020, from https://github.com/BaselBarakat/Assistive-Technology-for-the-Visually-Impaired-Optimizing-the-Frames-Freshness-of-Real-time-Objects