# Deep Learning Based Shopping Assistant For The Visually Impaired

Daniel Pintado, Vanessa Sanchez, Erin Adarve,
Mark Mata, Zekeriya Gogebakan, Bunyamin Cabuk,
Carter Chiu, Justin Zhan, Laxmi Gewali, Paul Oh
University of Nevada Las Vegas
Las Vegas, United States

***Abstract*—Contemporary developments in computer vision and artificial intelligence show promise to greatly improve the lives of those with disabilities. In this paper, we propose one such development: a wearable object recognition device in the form of eyewear. Our device is specialized to recognize items from the produce section of a grocery store, but serves as a proof of concept for any similar object recognition wearable. It is user friendly, featuring buttons that are pressed to capture images with the built-in camera. A convolutional neural network (CNN) is used to train the object recognition system. After the object is recognized, a text-to-speech system is utilized to inform the user which object they are holding in addition to the price of the product. With accuracy rates of 99.35%, our product has proven to successfully identify objects with greater correctness than existing models.**

***Index Terms*—deep learning; object recognition; computer vision; convolutional neural networks;**

## I. Introduction

Visual impairment, characterized by a defective sense of sight for which corrective technology provides no enhancement, affects millions of people worldwide. According to the World Health Organization, 285 million people around the world live with moderate to severe vision impairment. Additionally, because 1 in 3 Americans age 65 and older are visually impaired, a national average of $4 billion is spent annually on benefits for those affected [1].

In today's society, independence is fundamental. However, those who are visually impaired find themselves constantly depending upon others to assist them with simple, everyday tasks. Consequently, the societies of developing countries, which house 90% of those affected, remain vulnerable.

Many have delved into the field of computer science in hopes of coming up with a solution that solves the inconveniences of visual impairment. One such difficulty is the everyday task of grocery shopping. The visually impaired are challenged by the disorienting colors, shapes, and sizes of various products found at the market. Though they work to overcome their disability, shopping with severe vision impairment cannot be done independently. Therefore, research that would contribute to an improved lifestyle for the visually impaired is the development of a convenient yet accurate shopping assistant.

Numerous technologies currently exist to benefit those with a visually impairing disability. The majority of the previously developed algorithms make use of computer vision, an emerging technology. Computer vision concerns the acquisition and processing of data retrieved from visual input. Among the major subdomains of computer vision include object recognition and image restoration, and its applications overlap significantly with machine learning. Machine learning targets the automation of computer algorithms. It adheres to a repetitive cycle of "learning", through which patterns in the data are recognized. Capable of learning without explicit directions, machine learning algorithms specialize in three tasks: classification, regression, and clustering [2].

Deep learning, a subfield of machine learning, has attracted the lion's share of contemporary artificial intelligence research. Building off of machine learning's capability to progressively better its performance, deep learning operates through deep neural networks, or layered algorithms, typically providing unrivaled accuracy [3]. At the intersection of deep learning and computer vision is object recognition. Algorithms in this field detect and identify objects in a sequence of images or videos, categorizing them into classes. Object recognition proves appropriate for the task at hand.

The purpose of this research primarily concerns the heightening of independence among individuals who are visually impaired. Along with increased financial savings, we aim to increase accessibility to those in less fortunate economic situations. Although devices to aid the visually impaired exist, there lacks the existence of a convenient yet accurate system. Thus, to fulfill our purpose, we developed a wearable device programmed to recognize products in a supermarket. A prototype was built using the Raspberry Pi 3 and a deep neural network model.

The remainder of our paper is structured as follows. Section II examines related work, while Section III

delves into the functionality, software, and hardware of our proposed system. Section IV presents and interprets the results. Section V discusses limitations and avenues for future work, and Section VI revisits our contributions and concludes our paper.

## II. Related Work

Over the years, many have created their own renditions of devices that grant the visually impaired the ability to complete tasks independently. A neurally-inspired object recognition algorithm called Attention Biased Speeded Up Robust Features (AB-SURF) [4] helps users locate items in a market setting through the use of a head-mounted camera that is positioned onto a pair of eyeglasses. The AB-SURF algorithm achieved an 85% accuracy rate when identifying five or more products within one camera still.

In A Novel Method for Visually Impaired Using Object Recognition [5], the authors propose a system that utilizes two modules, object recognition and color recognition, to provide the user with information on their surroundings. Kumar and Meher conceived the system and have a prototype currently under development. Our idea of a head piece is quite similar to Kumar and Meher's proposed prototype, utilizing Microsoft SAPI for text-to-speech synthesis. The output is played through a speaker on the headpiece. Their color recognition module operates fairly well, with an accuracy rate of 98.33%. Additionally, their program is able to recognize common household products such as bins, mobile phones, and stairs, with an accuracy rate of 93%. However, if the prototype is used in a foreign environment such as the gym, hair salon, or grocery store, the accuracy rate decreases dramatically to an average of 44%.

The Development of Shopping Assistant Using Extraction of Text Images for the Visually Impaired [6] proposes an application that is installed onto a user's Android phone, which would be used in a market setting. After capturing an image of the object on a mobile phone, the text found on the object, such as the label, is detected and extracted by running a color reduction technique. After the text is identified, optical character recognition (OCR) software is used to convert the image of text into a string representation, which is then converted to speech and played through the phone's speaker to the blind user. The designers of this application included an informational feature in which the user is able to receive the description of the product by using a specific search term through voice commands. Another interesting feature was an option for the users to mark the products that they would like to purchase so that it could sum up the prices of the user's marked products and relay it back to the user through speech. The accuracy rate of the proposed system was not disclosed in the paper.

VisualPal [7] is another downloadable, mobile app designed for visually impaired users that recognize objects from captured images through the phone's built-in camera. The proposed system uses a hybrid algorithm comprised of artificial neural networks and Euclidean distance measures. A color detection system converts the image from the RGB color space to HSV (hue, saturation, value) in order to obtain the hue. This value is then compared to a table of hue component values of common colors. After the detection of the object, the user is informed through verbal messages relayed through the phone's built-in speaker. The system is currently being tested on Android smartphones. The accuracy rate of the hybrid algorithm is 97.5%.

A system proposed by Utaminingrum et al. [8] features a multiple phase method that detects text on products. The method includes detection using maximally stable extremal regions (MSER), Canny edge detection, region filtering, and a bounding box. This system produced an 80% accuracy rate.

## III. Methods

The device was developed for people with disabilities ranging from severe visual impairment to complete blindness. This target group was chosen due to the fact that their afflictions decrease their vision to the point where the traditional methods, such as corrective lenses or surgery, do not remedy their vision loss problem.

### *A. Device Prototype*

The developed prototype was assembled using a Raspberry Pi 3 single-board computer, Raspberry Pi camera module, a 3D-printed frame, two push buttons connected to a breadboard, a power bank, and headphones. Figures 2 and 3 illustrate the user and corner views of the prototype respectively. Figure 4 depicts the layout of the breadboard, and Figure 5 depicts the Raspberry Pi system schematic.

Notably, the proposed prototype was created with cost to the end user in mind. A decision was to be made on whether a Raspberry Pi 3 or an Arduino microcontroller would be better suited for the system, but it was concluded that the Raspberry Pi 3 would be a more appropriate choice. The fact that our system would require the running of multiple programs influenced the selection of the Raspberry Pi [9]. In addition, the speed of the central processing unit on the Raspberry Pi was a determining factor, considering its vital role in the efficiency of our program. Because the device required a way of obtaining input from the user, the push button was included as a simple way to capture an image. The specific hardware used to capture images was the

Raspberry Pi Camera Module V2, containing a Sony 8-megapixel sensor capable of capturing 1080p video and still images.

Headphones that vocalize a price through the text-to-speech program were chosen. To obtain product price information from the website of a grocery store, a web scraping program was employed. In our particular prototype implementation, we used the website of the marketplace Sprouts, but our method can be easily adapted to any market.

To bring all of the aforementioned components together, we 3D-printed a head mount unit that would house the camera, Raspberry Pi 3, push button, and speaker. Due to the fast prototyping that the technology allows, 3D printing was seen as the best method for making an enclosure for the modules.

### *B. Device Operation*

Operation of the device is illustrated by Figure 1. One can understand usage of the device as the performance of three primary functions: (1) capturing image input using the camera, (2) classifying the product within the image, and (3) reporting the price of said product to the user. Pressing the first push button will cause the device to capture an image with the camera. This image is internally rescaled and fed through the predictive model to obtain a product class which describes the object. It then feeds the product class into the text-to-speech software, which produces an audio version of the product name and channels it through the speaker connected to the Raspberry Pi. The Raspberry Pi then reverts to listening for another button press, repeating the process.

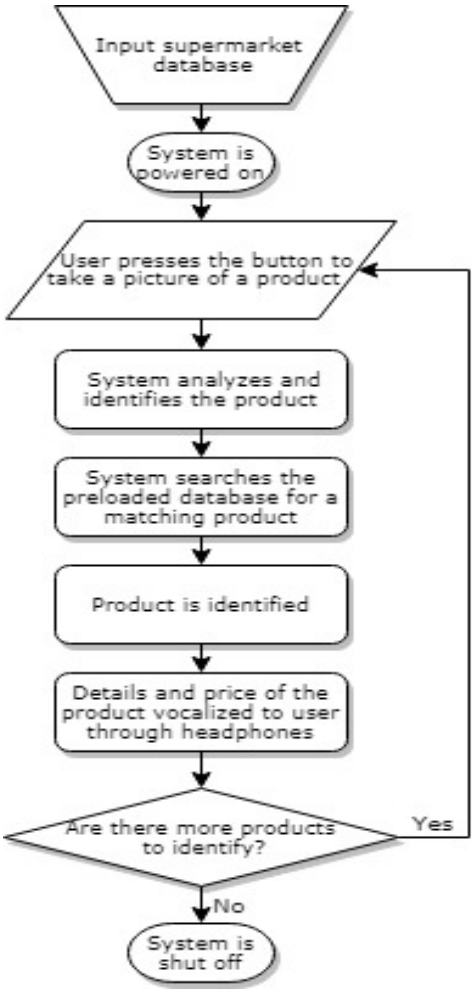


Figure 1. Flowchart of system

### *C. Predictive Model / Convolutional Neural Network*

In deep learning, network architectures can be designed in numerous ways to achieve an accurate predictive model. For our purposes, we utilized a convolutional neural network (CNN) [10]. A convolutional neural network is a category of neural networks specifically tailored for higher efficiency and accuracy in image processing. Like all neural networks, a CNN is composed of neurons that have tunable weights and biases. These interconnected neurons receive input and produce an output as a function of the weights, biases, and an activation function [11]. The output is used to help detect patterns in an image for the final goal of classification.

To aid in the training of a CNN, we utilized two libraries. TensorFlow is an open source software library developed by Google for high performance numerical computation and used frequently for deep learning [12]. Often used in conjunction with TensorFlow is Keras, a high-level deep learning API written in Python which employs TensorFlow as a backend. These two libraries were used to train the CNN with over a thousand training images of produce. As a necessary preprocessing step for the network, all images were resized to 100x100. They were then converted from RGB to grayscale to help identify important features of the image which are unaffected by color. After experimenting with different network hyperparameter configurations, we arrived at the architecture described in Table I which appeared to achieve the highest test accuracy. The network was trained for 10 epochs with a mini-batch size of 100.

When deploying the convolutional neural network onto the device, we used OpenCV, a computer vision library with a focus on real-time applications [13][14]. OpenCV facilitated the preprocessing of images captured by the device camera, including resizing and color transformations, and used the trained model to classify the produce contained within the image [15].

Table I
PARAMETERS FOR EACH HIDDEN LAYER IN THE CNN

| Layer | Parameters |
|---|---|
| Convolutional | Filters: 64, Window: 11×11, Activation: ReLU |
| Max Pooling | Window: 2×2 |
| Dropout | Dropout: 20% |
| Fully Connected | Nodes: 128, Activation: ReLU |
| Dropout | Dropout: 20% |
| Fully Connected | Nodes: 128, Activation: ReLU |
| Dropout | Dropout: 20% |
| Fully Connected | Nodes: 128, Activation: ReLU |
| Dropout | Dropout: 20% |

Figure 2. 3D-printed prototype, user view

Figure 3. 3D-printed prototype, corner view

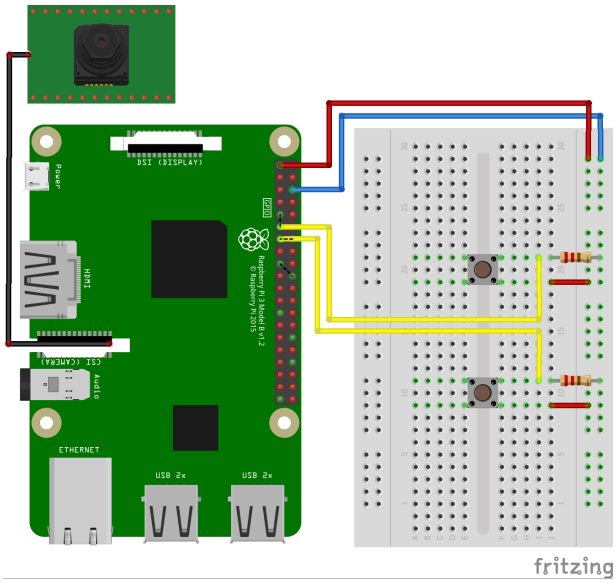


Figure 4. Breadboard diagram

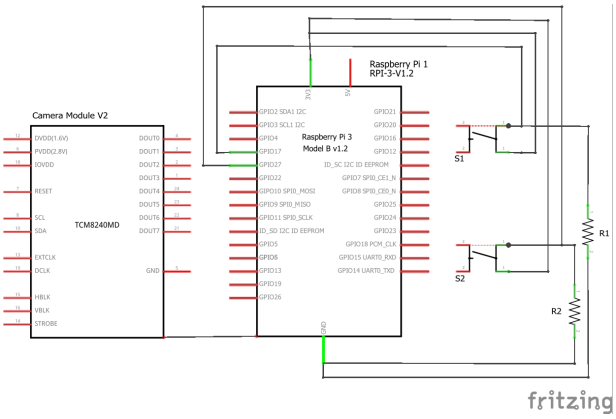


Figure 5. System schematic

## IV. RESULTS

### A. CNN Accuracy

The proposed solution garnered a model testing accuracy of 99.35%, which, as illustrated by the graph in Figure 6, was better than the accuracy of both of the similar prototypes found through previous research. We also depict the accuracy of our CNN model over its various iterations as shown in Figure 7, starting out at around 90 percent and finishing out at 99.35%. The accuracy was improved every iteration due to the adjusting and fine-tuning of hyperparameters, such as the amount of nodes per layer and batch size.

### B. Web Scraper Execution Time

The web scraper graph in Figure 8 illustrates that it takes anywhere from around 240 to 270 seconds to

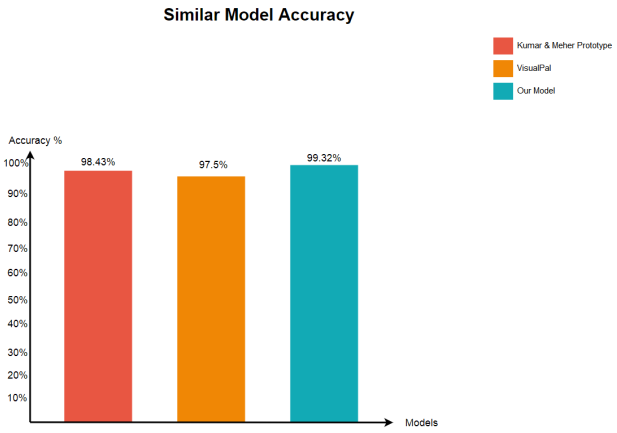


Figure 6. Comparing model accuracy

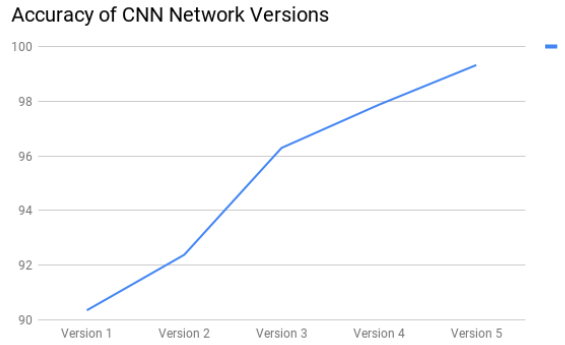


Figure 7. Accuracy of CNN over time

collect the product name and price information from the Sprouts website. Since the process takes a considerable amount of time, it would be better for the user to start the scraping process at home before going out to use the device at the market. Of course, in more sophisticated implementations this time could likely be reduced to a negligible amount.

## V. Discussion

Visual impairment affects millions of people worldwide [16]. Our object recognition device could benefit those millions who struggle with the everyday task of independently completing a purchase at the grocery store. Although more work is needed on this prototype, we achieved our goal of creating a device that aids the visually impaired.

While developing the prototype, we came across several complications which could be overcome in the future. Perhaps our most significant challenge concerned how the user would navigate through and locate certain sections of the store. The need for a guide dog or person to lead the user around the store to the desired products remains. Another challenge was posed by the speed at

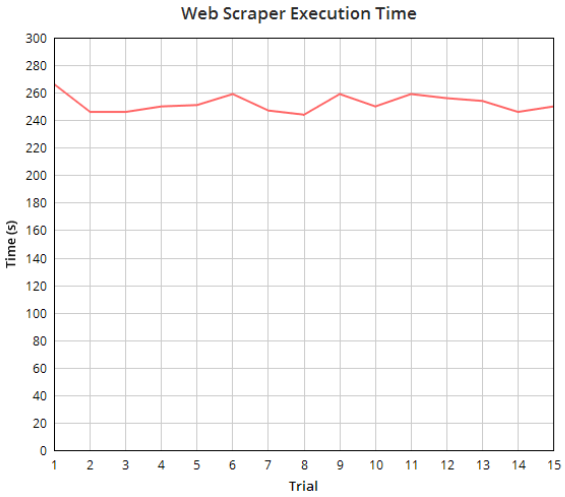


Figure 8. Execution time graph

which the Raspberry Pi computer classifies the object, which is rather slow. To fix this issue, a more powerful single-board computer would have to be used in place of the Raspberry Pi 3. While running the model on a more capable machine would not be an issue, a more powerful unit would cost at least double the price of the Raspberry Pi, adding to the total cost of the device. Additionally, as mentioned before, the execution time of the present implementation of the web scraper proved to be undesirably slow, although this time could be reduced to a negligible amount for instance by accessing a regularly-updating online database.

With further development, the prototype could be trained to identify other products besides food items, such as toilet paper, soap, or many other common objects. This would greatly expand the domain in which our device could function. To achieve this, the current image dataset would need to be significantly expanded to account for all the new categories of products to classify. Also, with increased processing power, the device would be able to identify products in real-time instead of waiting for the user to manually scan for an object. This could be done to aid the user in finding products around the market in addition to the original function of classifying and giving the price for a specific product.

## VI. Conclusion

Through the use of a convolutional neural network and a Raspberry Pi 3, we were able to build and configure a wearable object recognition device that assists the severely visually impaired in a market setting, with accuracy exceeding the marks set by previous research. Our product appeals to millions worldwide, and in the future, we hope to expand upon the prototype presented in this paper. With the answer to a problem that affects so

many, our prototype provides a solution to an everyday task made impossible by a disability.

## ACKNOWLEDGMENT

This research was supported in part by the Department of Defense under the Army Educational Outreach Program (AEOP) and the National Science Foundation under the Research Experiences for Teachers (RET) program. The authors would also like to thank the UNLV Writing Center, the UNLV Department of Communications Studies, and Carter Chiu for their mentorship and support for this research.

## REFERENCES

[1] C. M. Mangione, S. Berry, K. Spritzer, N. K. Janz, R. Klein, C. Owsley, and P. P. Lee, "Identifying the content area for the 51-item national eye institute visual function questionnaire: results from focus groups with visually impaired persons," *Archives of Ophthalmology*, vol. 116, no. 2, pp. 227–233, 1998.

[2] G. Bebis, D. Egbert, and M. Shah, "Review of computer vision education," *IEEE Transactions on Education*, vol. 46, no. 1, pp. 2–21, 2003.

[3] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[4] K. Thakoor, S. Marat, P. Nasiatka, B. McIntosh, F. Sahin, A. Tanguay, J. Weiland, and L. Itti, "Attention biased speeded up robust features (ab-surf): A neurally-inspired object recognition algorithm for a wearable aid for the visually-impaired," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2013, pp. 1–6.

[5] R. Kumar and S. Meher, "A novel method for visually impaired using object recognition," in *2015 International Conference on Communications and Signal Processing (ICCSP)*, April 2015, pp. 772–776.

[6] A. Farhath, R. Amruthavarshini, S. Harshitha, A. Saranya, and R. Velumadhavarao, "Development of Shopping Assistant Using Extraction of Text Images for Visually Impaired," in *2014 Sixth International Conference on Advanced Computing (ICoAC)*, Dec 2014, pp. 66–71.

[7] S. Bagwan and S. Sankpal, "VisualPal: A mobile app for object recognition for the visually impaired," in *2015 International Conference on Computer, Communication and Control (IC4)*, Sept 2015, pp. 1–6.

[8] F. Utaminingrum *et al.*, "Text detection and recognition using multiple phase method on various product label for visual impaired people," in *Sustainable Information Engineering and Technology (SIET), 2017 International Conference on*. IEEE, 2017, pp. 398–404.

[9] S. Resnikoff, D. Pascolini, D. Etya'Ale, I. Kocur, R. Pararajasegaram, G. P. Pokharel, and S. P. Mariotti, "Global data on visual impairment in the year 2002," *Bulletin of the world health organization*, vol. 82, pp. 844–851, 2004.

[10] X. Du, Y. Cai, S. Wang, and L. Zhang, "Overview of deep learning," in *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, Nov 2016, pp. 159–164.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[12] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[13] A. Fernandes, L. Moreira, and J. Mata, "Machine vision applications and development aspects," in *2011 9th IEEE International Conference on Control and Automation (ICCA)*, Dec 2011, pp. 1274–1278.

[14] R. Simpson, "Computer vision: an overview," *IEEE Expert*, vol. 6, no. 4, pp. 11–15, Aug 1991.

[15] Y. Zhao, H. Shi, X. Chen, X. Li, and C. Wang, "An overview of object detection and tracking," in *2015 IEEE International Conference on Information and Automation*, Aug 2015, pp. 280–286.

[16] S. Resnikoff, D. Pascolini, D. Etya'Ale, I. Kocur, R. Pararajasegaram, G. P. Pokharel, and S. P. Mariotti, "Global data on visual impairment in the year 2002," *Bulletin of the world health organization*, vol. 82, pp. 844–851, 2004.