Eye of Aurora

Zeqi Li, Weixin Wang, Haohang Yan, Yuankai Huang
Professor Pramod Khargonekar
Department of Electrical Engineering and Computer Science

Background

According to the 2020 report of the World Health Organization, there are about 253 million people with visual impairment in the world [1]. Although many obstacle avoidance tools based on neural network have been developed [2, 3], visually impaired people are vulnerable to depression and anxiety because they ultimately cannot see and integrate into the world around them [1]. Therefore, our project focuses less on navigation, but more on making the user understand what is happening around them.

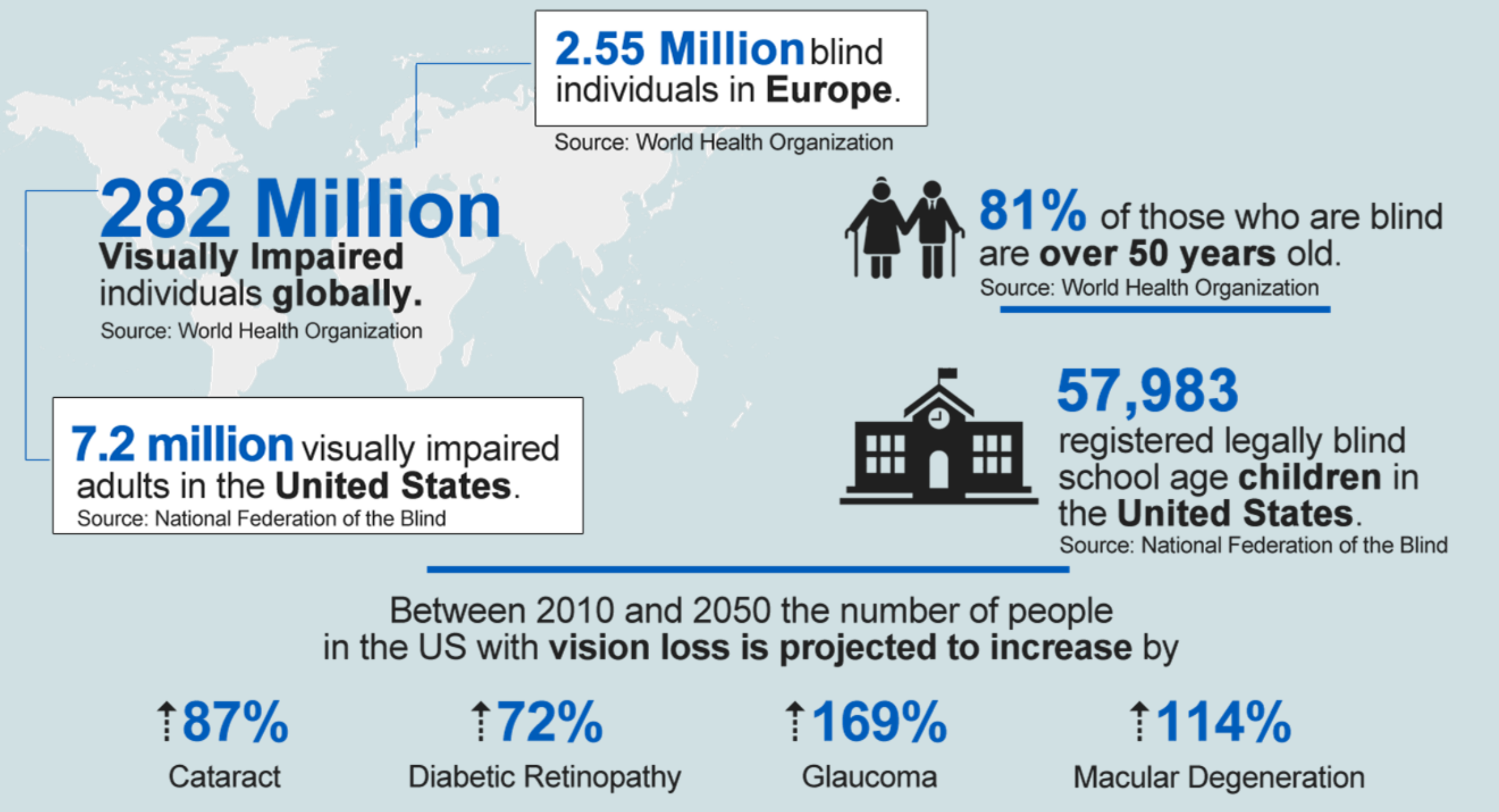

Fig. 1. Visually impaired people in the world
Source: World Health Organization

Project Goal

- The project is to make a pair of camera glasses that can describe the scene in front of the user with sound.
- Convert the camera view into a text description by Image Caption technology.
- Convert the text description into audio output by Text-to-Speech technology.
- Be small, light, and wearable.
- The end goal is to help visually impaired people "see" the world, as shown in Figure 3 and 4.

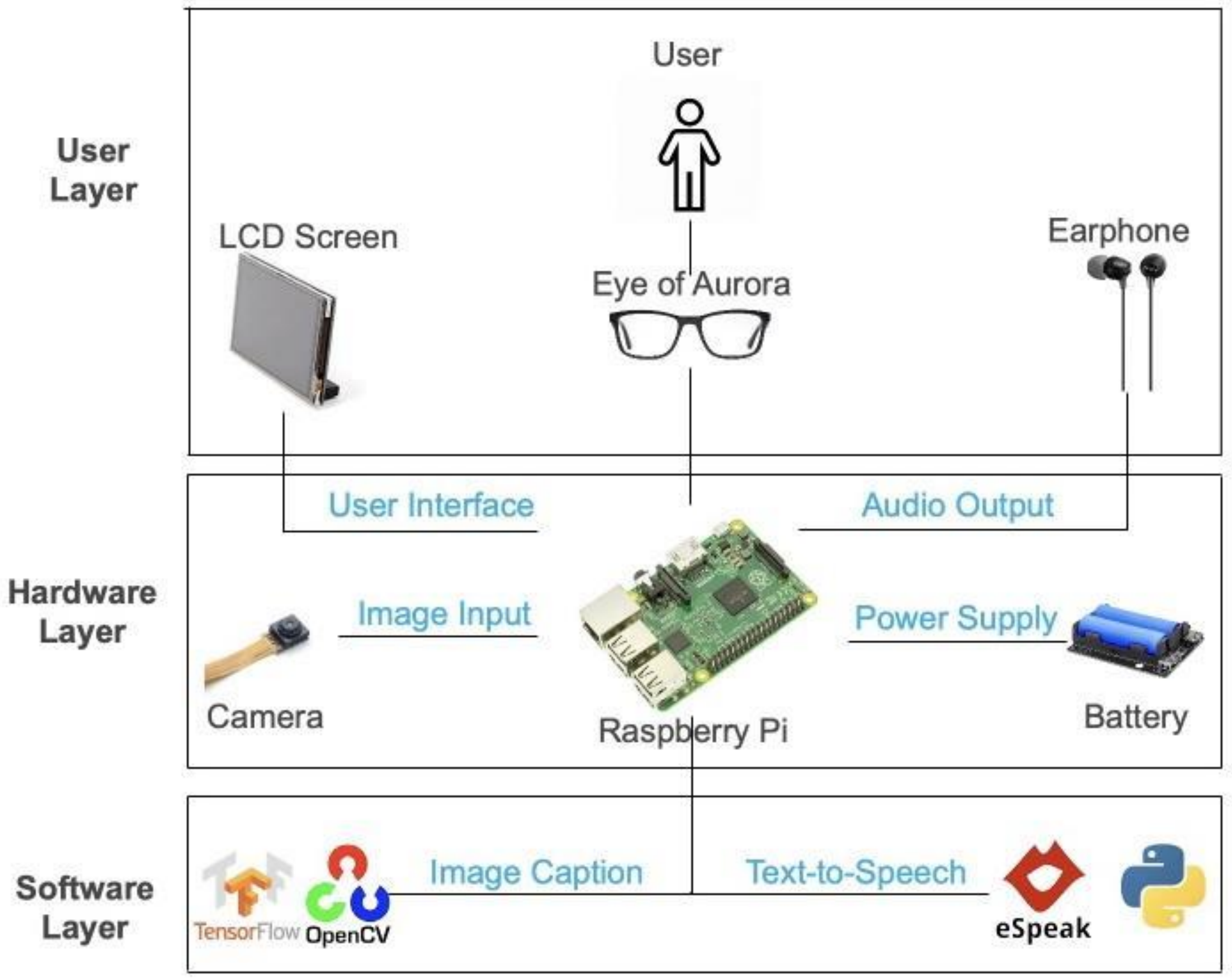

Fig. 2. High-level design of Eye of Aurora

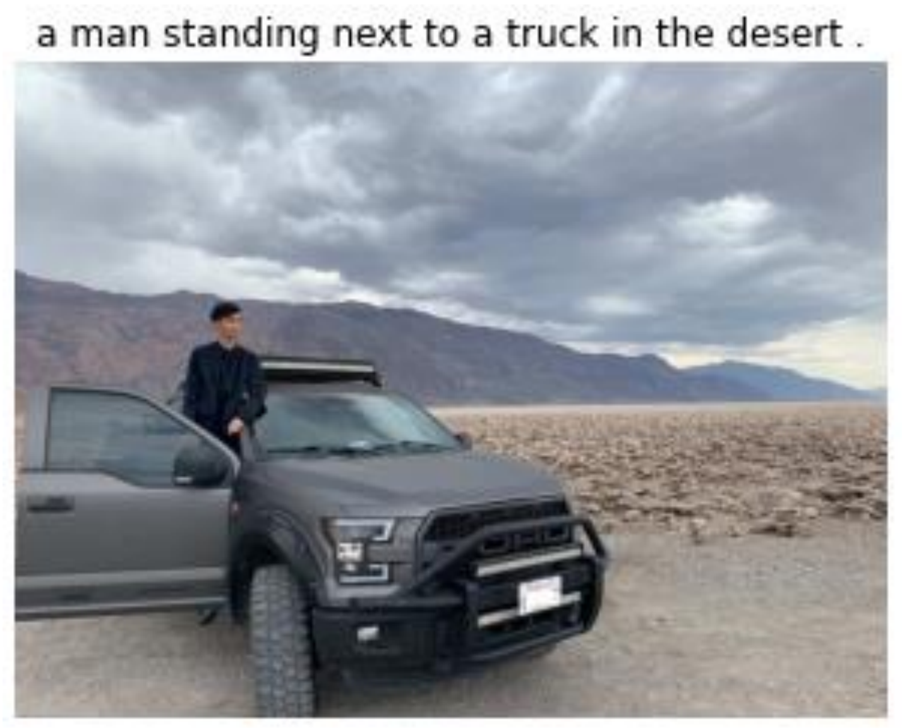

Fig. 3. An image caption example

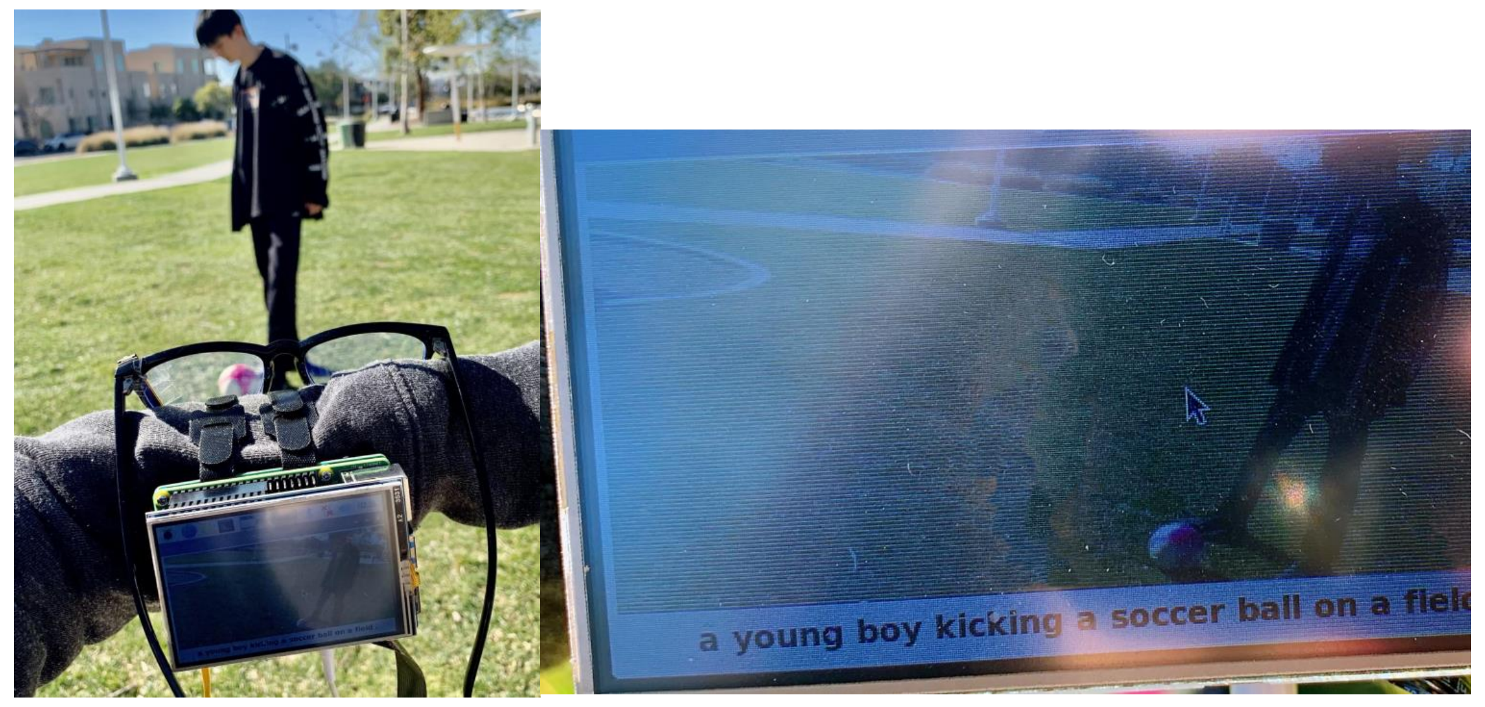

Fig. 4. Eye of Aurora generated "a young boy kicking a soccer ball on a field" to describe what it "saw"

Implementation

We wrote a user interface program that will run automatically when the Raspberry Pi is on. When the user touches the screen, the camera attached to the glasses will take a picture, and then our image caption model implemented with TensorFlow will take the picture as input to produce a natural sentence that can describe the main content of the picture. After that, the text-to-speech program base on eSpeak will speak out that sentence. Finally, the picture and sentence will be displayed on the screen for developing and testing.

Result

Table 1: Evaluation of our image caption model

	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Our model	0.505158	0.278898	0.199664	0.103846
Img2txt (Google)	0.561677	0.319980	0.223410	0.114370

We also trained our own image caption model. We used 1000 images of flickr8k as our testing datasets and computed BLEU [4] scores to evaluate the performance. We compared our model and the Img2txt model trained by Google. As shown in Table 1, our model's BLEU [4] scores are comparable to Google's model img2txt.

Improvement

- Reduce the time gap between taking pictures and processing Image Caption by improving the efficiency of the model.
- Extend the dataset and vocabulary to allow the model to identify more objects and behaviors.
- Add voice control as a new function.

Reference

[1] Demmin DL, Silverstein SM. Visual Impairment and Mental Health: Unmet Needs and Treatment Options. Clin Ophthalmol. 2020;14:4229-4251. Published 2020 Dec 3. doi:10.2147/OPTH.S258783
[2] Hengle, A. Kulkarni, N. Bavadekar, N. Kulkarni and R. Udyawar, "Smart Cap: A Deep Learning and IoT Based Assistant for the Visually Impaired," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1109-1116, doi: 10.1109/ICSSIT48917.2020.9214140.
[3] Liu OC.A., Li SK., Yan LQ., Ng SC., Kwok CP. (2020) A Visually Impaired Assistant Using Neural Network and Image Recognition with Physical Navigation. In: Han M., Qin S., Zhang N. (eds) Advances in Neural Networks – ISNN 2020. ISNN 2020. Lecture Notes in Computer Science, vol 12557. Springer, Cham. https://doi.org/10.1007/978-3-030-64221-1_24
[4] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: A method for automatic evaluation of machine translation. In ACL, 2002.

THE HENRY SAMUELI SCHOOL OF ENGINEERING
UNIVERSITY of CALIFORNIA · IRVINE