Vision:
A Deep Learning Approach to provide walking assistance to the visually impaired

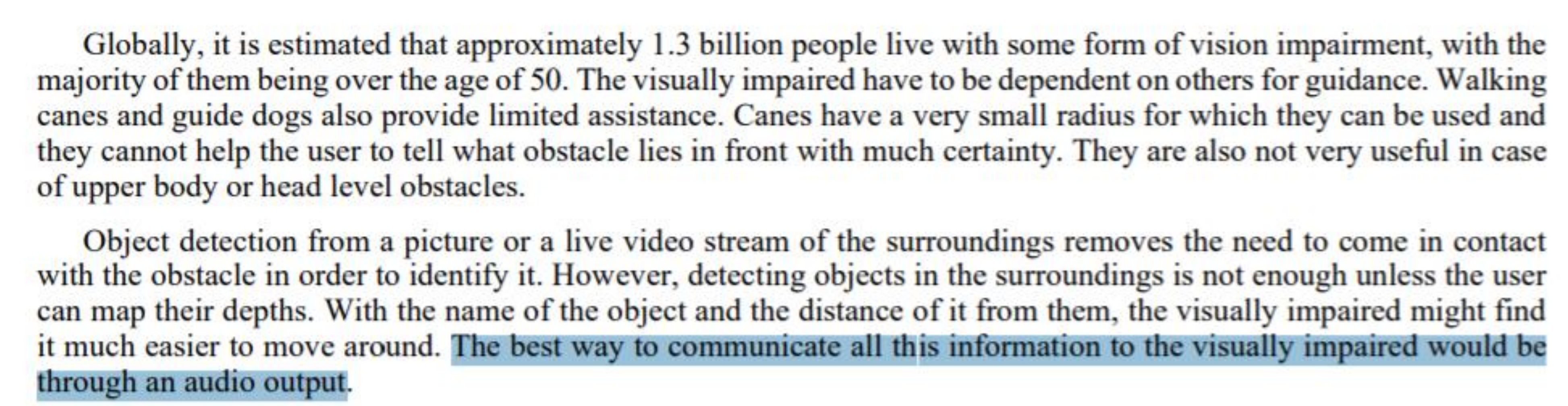
시각장애인에게 보행 보조를 제공하는 딥러닝 연구

강준구(engineerjkk@naver.com)

강준구

논문의 목적 : 시각장애인에게 실제 주변의 물체가 어떤 것인지 알려주고, 그 물체가 몇 미터 앞에 있는지 거리까지 '음성' 으로 알려주는 것입니다.
더불어 단 하나의 카메라만을 사용하므로 비용적으로 절약할 수 있습니다.

논문의 제안 : **Object Detection, Depth Estimation and Text to Speech**



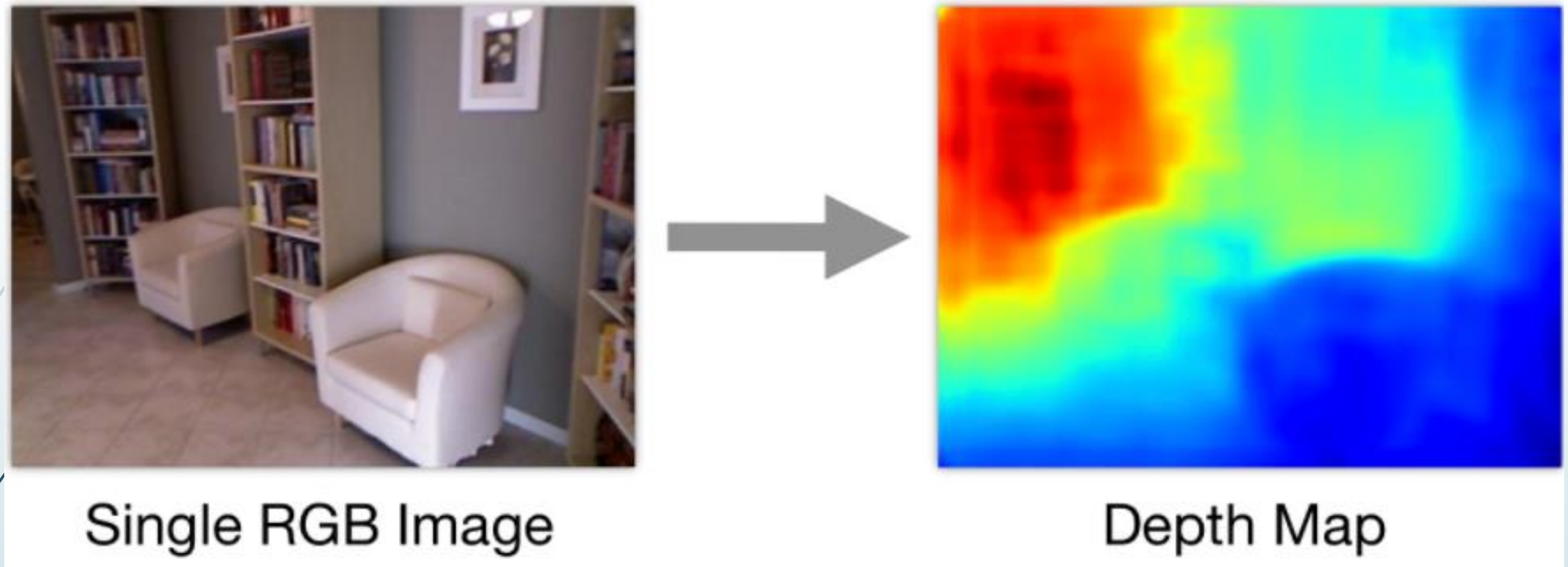
Globally, it is estimated that approximately 1.3 billion people live with some form of vision impairment, with the majority of them being over the age of 50. The visually impaired have to be dependent on others for guidance. Walking canes and guide dogs also provide limited assistance. Canes have a very small radius for which they can be used and they cannot help the user to tell what obstacle lies in front with much certainty. They are also not very useful in case of upper body or head level obstacles.

Object detection from a picture or a live video stream of the surroundings removes the need to come in contact with the obstacle in order to identify it. However, detecting objects in the surroundings is not enough unless the user can map their depths. With the name of the object and the distance of it from them, the visually impaired might find it much easier to move around. The best way to communicate all this information to the visually impaired would be through an audio output.

- 시각장애인에게 제공할 수 있는 가장 좋은 방법은 오디오를 통해 정보를 제공하는 것입니다. 여기서는 카메라로 단순히 **Object Detection**만해서 정보를 알려주는 것이 아니라 거리도 함께 측정해 알려줍니다.

Depth Estimation

강준구



Depth Map 알고리즘을 활용해 거리를 측정할 수 있으며

Input은 단 한 개의 영상입니다.

차별 점 : 논문 이전의 방식

$$Z = \frac{b * F}{x^l + x^r} \qquad (7)$$

강준구

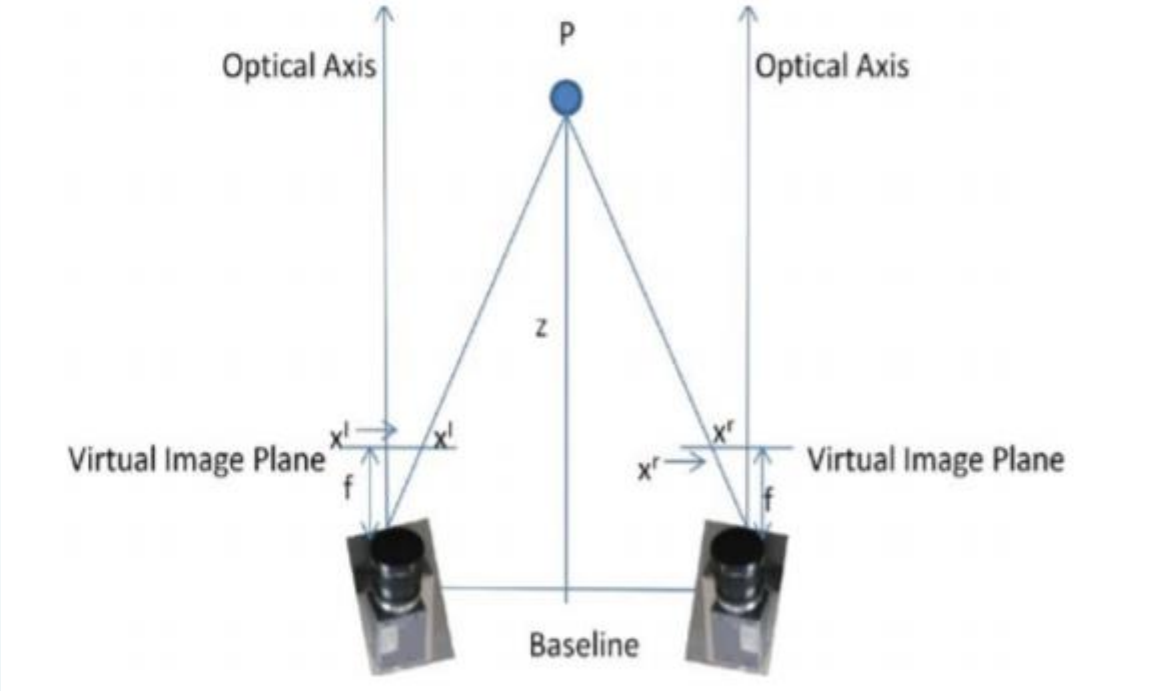


Fig. 4. STEREO VISION

Depth Estimation과는 상반된 **Stereo Vision** 입니다.

두개의 카메라**(Input)**을 사용하여 거리를 측정합니다.

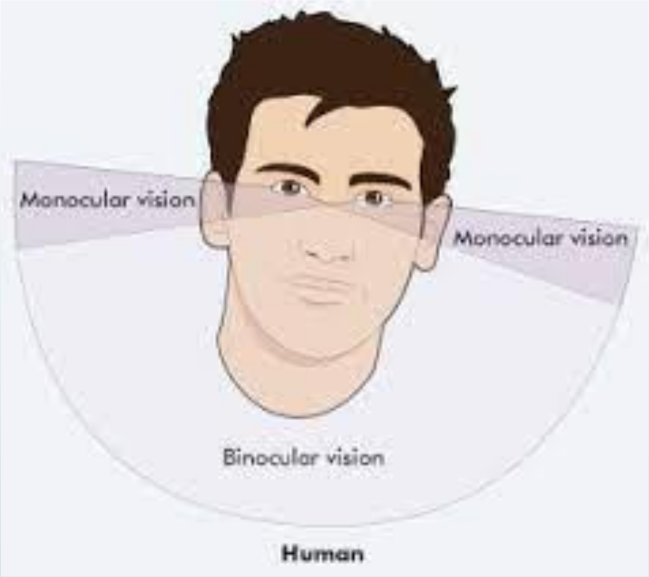
상대적으로 비용이 부담(카메라가 두개이므로)될 수 있는 단점이 있습니다.

강준구

구체적으로는 **Depth Estimation**을 활용하기 위해 **Monocular Vision**을 활용했습니다.

여기서 **Monucular Vision**이란, 한쪽 눈으로만 보이는 부분을 말합니다. **(**한국말로하면 단안시**)**

Binocular vision은 양쪽 눈으로 보이는 교집합 지점입니다.

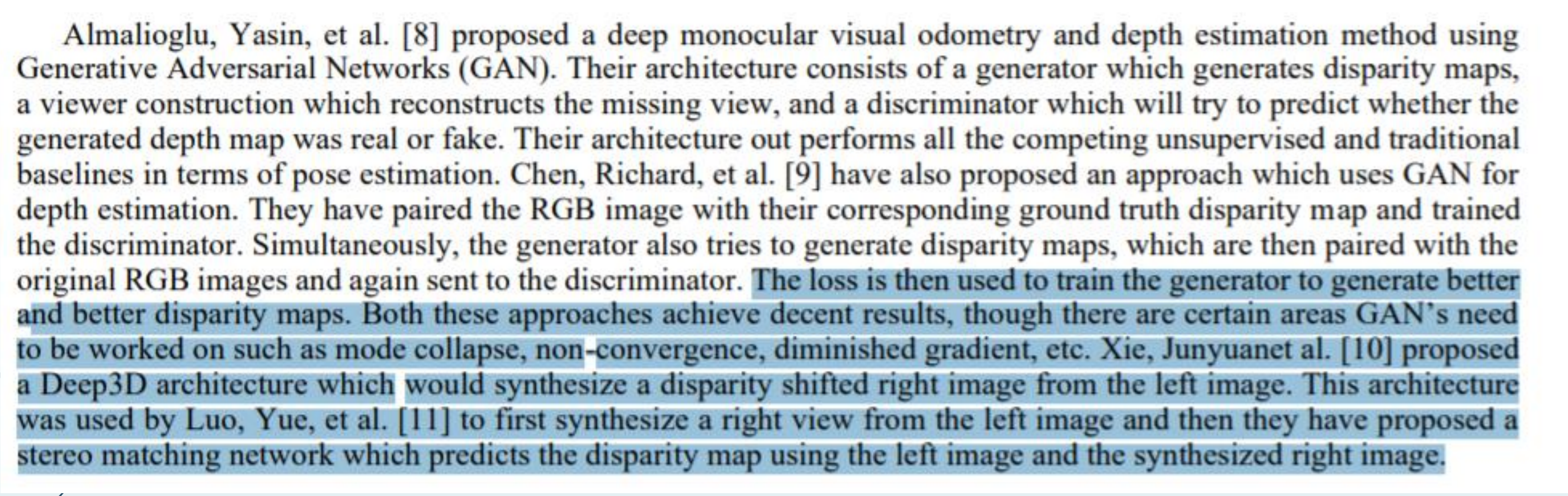


Monocular Vision

그럼 카메라는 하나인데 어떻게 **Monocular Vision**을 만들까요?

GAN

강준구



Almalioglu, Yasin, et al. [8] proposed a deep monocular visual odometry and depth estimation method using Generative Adversarial Networks (GAN). Their architecture consists of a generator which generates disparity maps, a viewer construction which reconstructs the missing view, and a discriminator which will try to predict whether the generated depth map was real or fake. Their architecture out performs all the competing unsupervised and traditional baselines in terms of pose estimation. Chen, Richard, et al. [9] have also proposed an approach which uses GAN for depth estimation. They have paired the RGB image with their corresponding ground truth disparity map and trained the discriminator. Simultaneously, the generator also tries to generate disparity maps, which are then paired with the original RGB images and again sent to the discriminator. The loss is then used to train the generator to generate better and better disparity maps. Both these approaches achieve decent results, though there are certain areas GAN’s need to be worked on such as mode collapse, non-convergence, diminished gradient, etc. Xie, Junyuanet al. [10] proposed a Deep3D architecture which would synthesize a disparity shifted right image from the left image. This architecture was used by Luo, Yue, et al. [11] to first synthesize a right view from the left image and then they have proposed a stereo matching network which predicts the disparity map using the left image and the synthesized right image.

여기서 **GAN** 을 사용하는데요.

Discriminator와 **Generator**간에 발생하는 **Loss**를 통해 안보이는 쪽 이미지

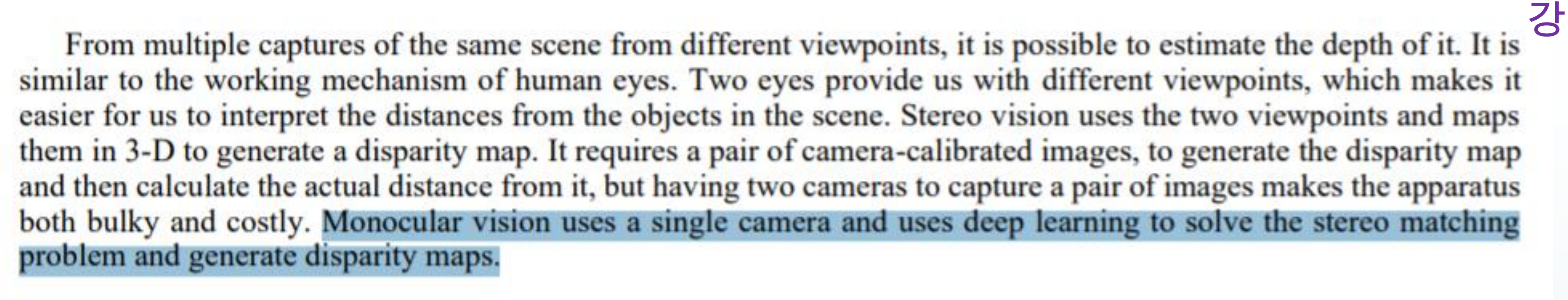
즉, **Disparity Maps**를 생성해 줍니다.

다시 말해 왼쪽 눈에서 바라본 이미지로부터 오른쪽 눈에서 바라본 이미지를 생성해

주는 것이죠.

(Disparity Maps 을 생성하기 위해서는 **KITTI dataset**을 사용했습니다.)

강준구



From multiple captures of the same scene from different viewpoints, it is possible to estimate the depth of it. It is similar to the working mechanism of human eyes. Two eyes provide us with different viewpoints, which makes it easier for us to interpret the distances from the objects in the scene. Stereo vision uses the two viewpoints and maps them in 3-D to generate a disparity map. It requires a pair of camera-calibrated images, to generate the disparity map and then calculate the actual distance from it, but having two cameras to capture a pair of images makes the apparatus both bulky and costly. Monocular vision uses a single camera and uses deep learning to solve the stereo matching problem and generate disparity maps.

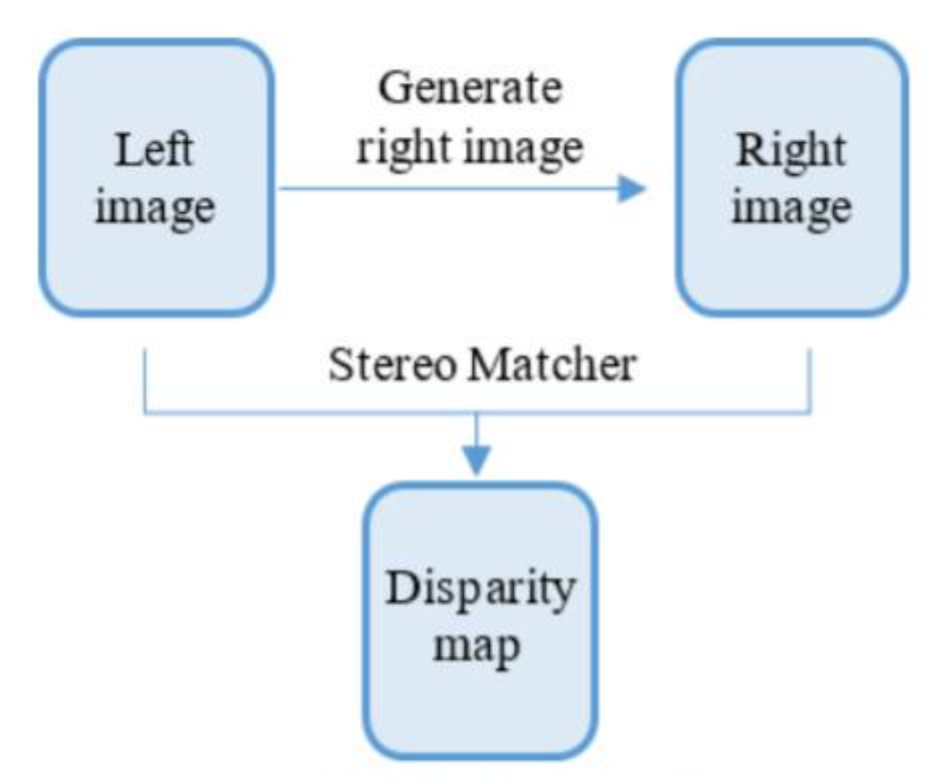
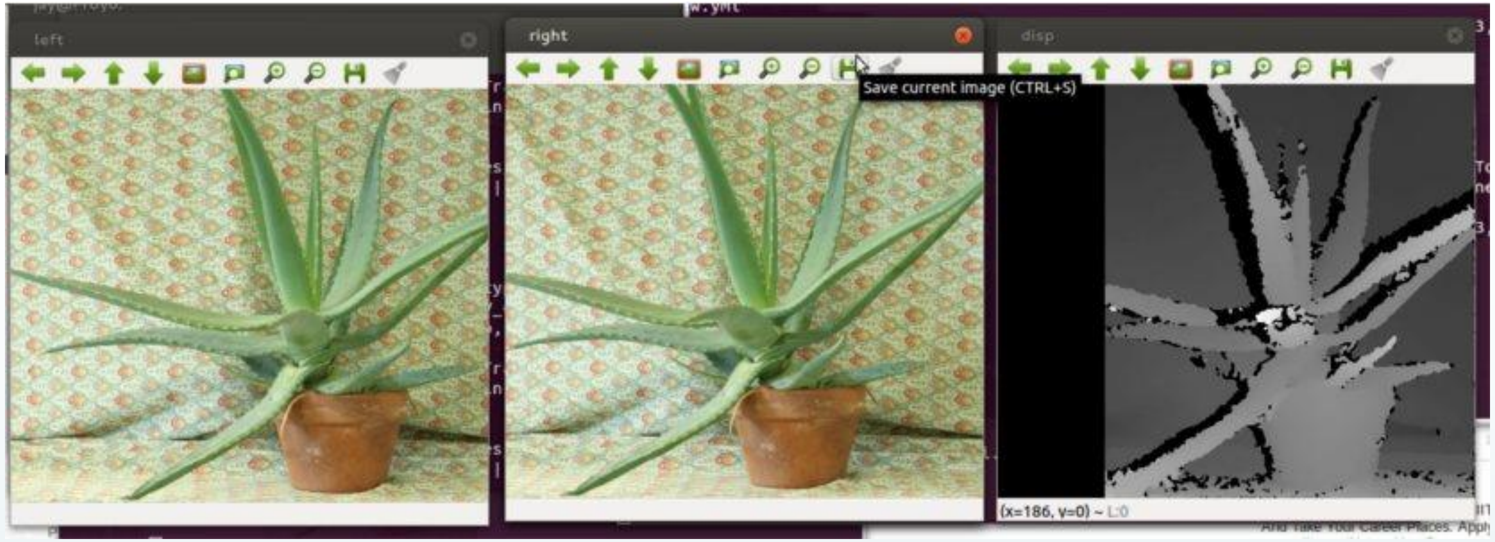
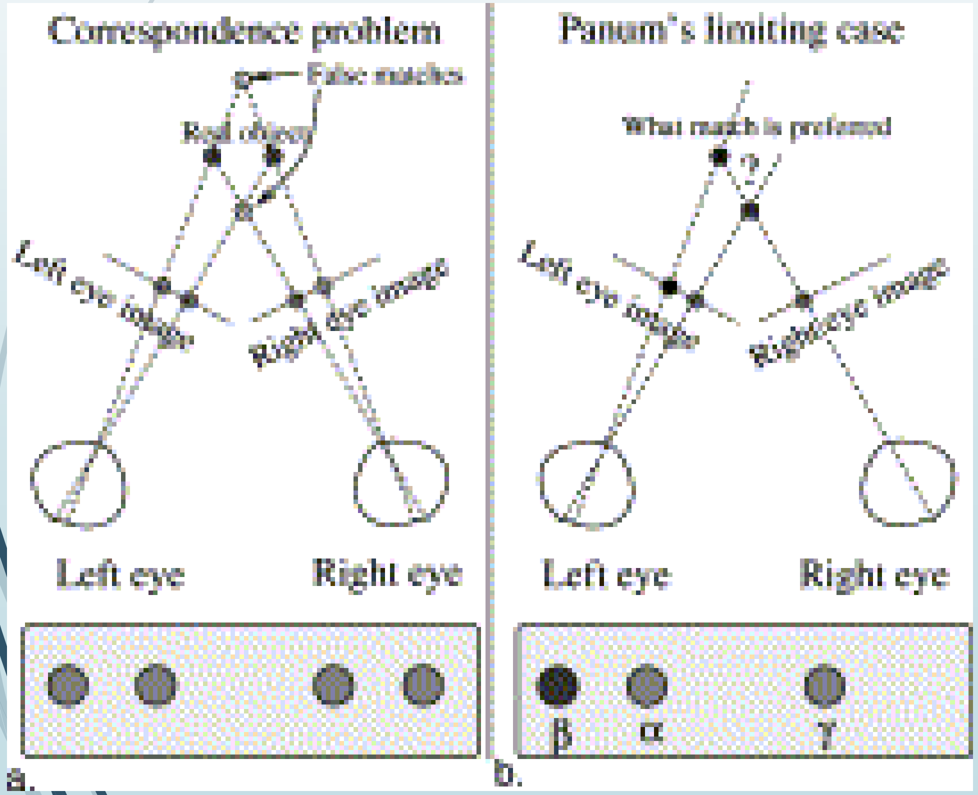


Fig. 5. OVERVIEW OF DEPTH ESTIMATION

Monocular Vision은 하나의 카메라만으로 딥러닝을 활용해 스테레오 매칭 문제를 해결하고 Disparity Map을 생성하는 것입니다.

강준구





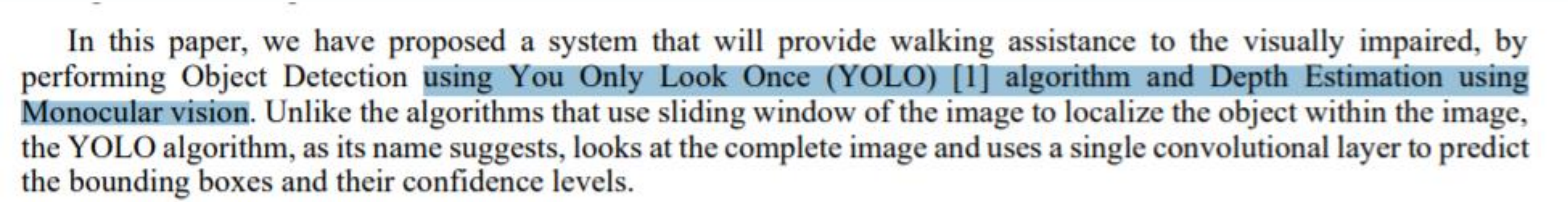
Disparity map 구성도

즉, 정리하면 왼쪽 이미지로 가상의 오른쪽 이미지를 생성하면 Streo효과를 낼 수 있고

그것이 곧 Depth estimation에 활용돼 거리를 알 수 있는 것입니다.

Object Detection

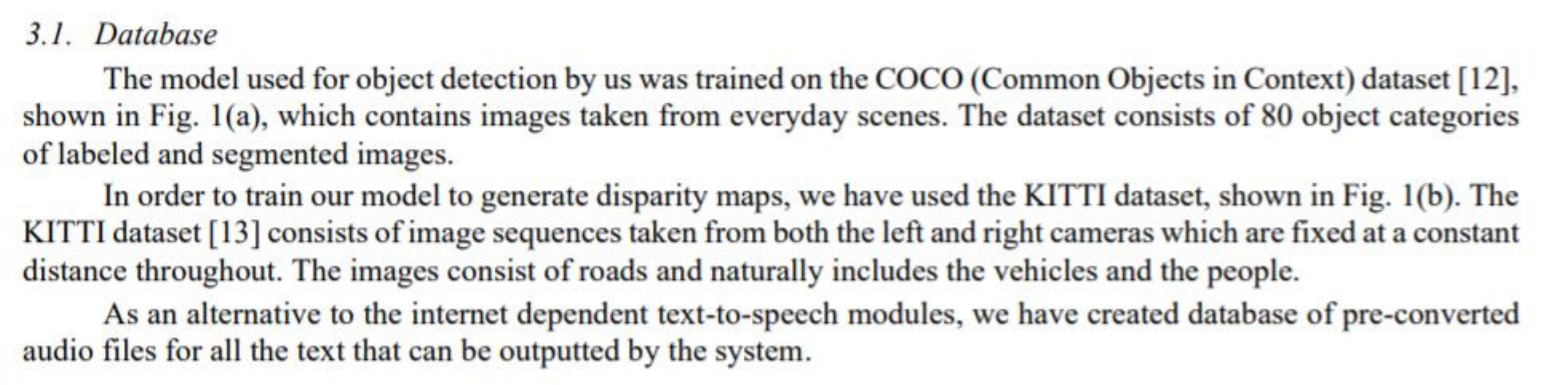
강준구



In this paper, we have proposed a system that will provide walking assistance to the visually impaired, by performing Object Detection using You Only Look Once (YOLO) [1] algorithm and Depth Estimation using Monocular vision. Unlike the algorithms that use sliding window of the image to localize the object within the image, the YOLO algorithm, as its name suggests, looks at the complete image and uses a single convolutional layer to predict the bounding boxes and their confidence levels.

논문에서 다른 객체탐지 방법도 있는데 굳이 **YOLO**를 쓴 이유는 빠르기 때문입니다.

알고 계시듯이 **YOLO**가 이름처럼 **Single Convolutional Layer**를 사용하고 있기 때문이죠.



3.1. Database

The model used for object detection by us was trained on the COCO (Common Objects in Context) dataset [12], shown in Fig. 1(a), which contains images taken from everyday scenes. The dataset consists of 80 object categories of labeled and segmented images.

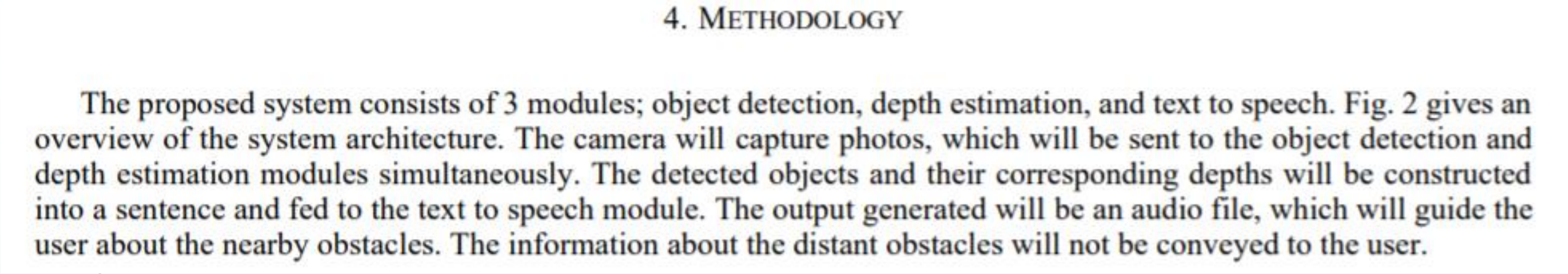
In order to train our model to generate disparity maps, we have used the KITTI dataset, shown in Fig. 1(b). The KITTI dataset [13] consists of image sequences taken from both the left and right cameras which are fixed at a constant distance throughout. The images consist of roads and naturally includes the vehicles and the people.

As an alternative to the internet dependent text-to-speech modules, we have created database of pre-converted audio files for all the text that can be outputted by the system.

논문에서 사용된 데이터는 **COCO dataset**이며 **80**개의 **class**가 있습니다. **Disparity Maps** 을 생성하기 위해서는 **KITTI dataset**을 사용했습니다. 자동차나 사람들 등을 포함하고 있죠. 또한 미리 오디오 파일로 이뤄진 데이터를 활용해 출력해 줄 수 있습니다.

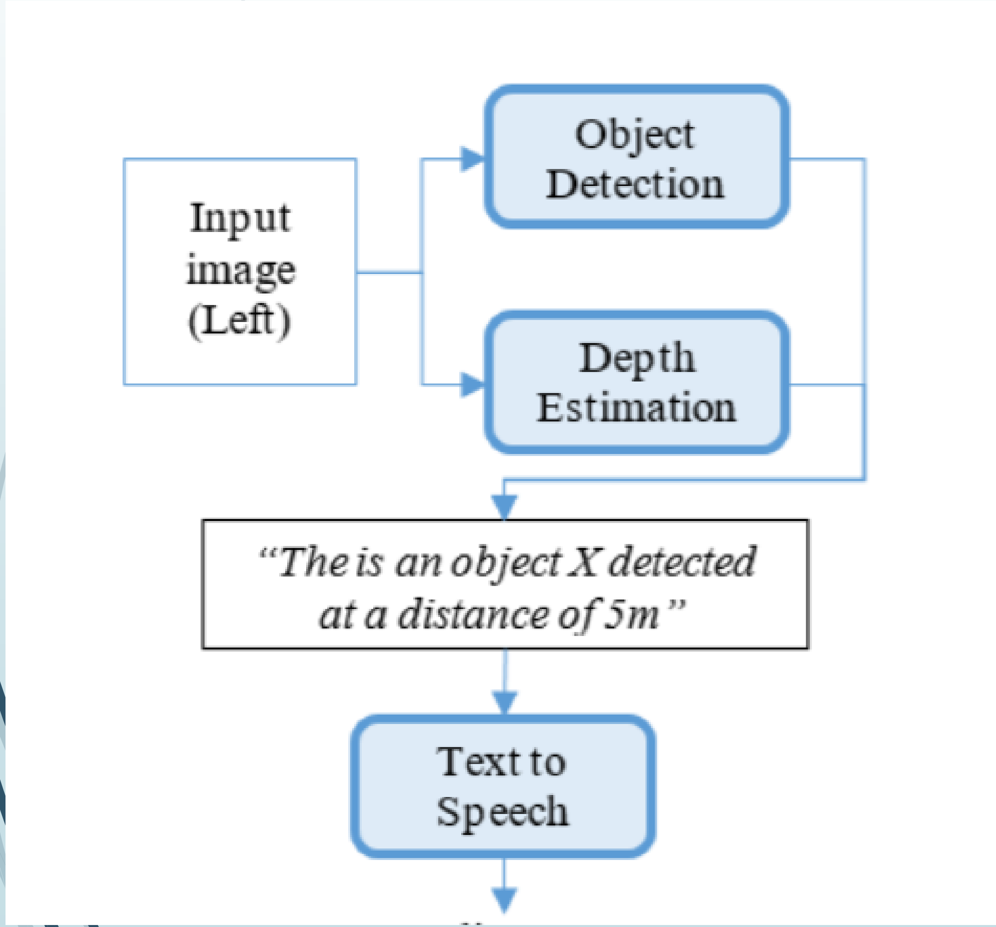
정리

강준구



4. METHODOLOGY

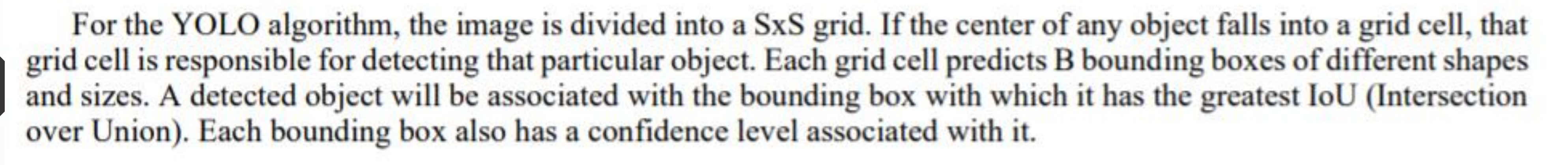
The proposed system consists of 3 modules; object detection, depth estimation, and text to speech. Fig. 2 gives an overview of the system architecture. The camera will capture photos, which will be sent to the object detection and depth estimation modules simultaneously. The detected objects and their corresponding depths will be constructed into a sentence and fed to the text to speech module. The output generated will be an audio file, which will guide the user about the nearby obstacles. The information about the distant obstacles will not be conveyed to the user.



즉, 카메라가 한 프레임을 얻을 때 **Object Detection Module**과 **Depth Estimation Module**에 동시에 전송합니다. 그럼 여기서 생성된 아웃풋을 **Test to speech Module**에 전송하면 오디오 파일이 출력됩니다.

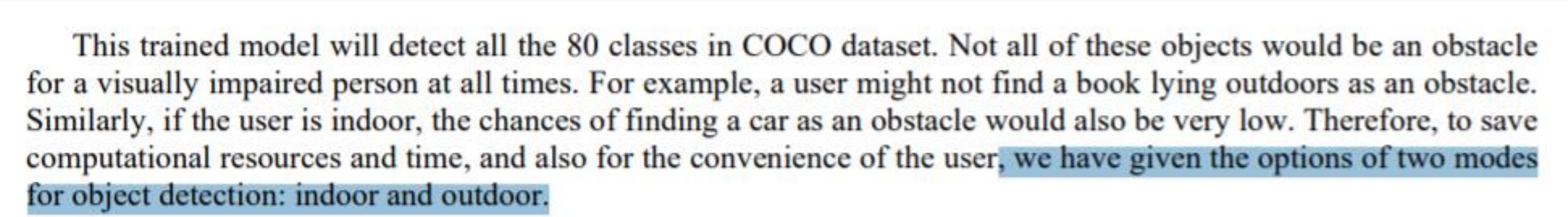
그럼 사용자에게 근처에 장애물이 있음을 알려줄 수 있죠. 또한 멀리 있는 장애물은 굳이 알려주지 않습니다.

강준구



For the YOLO algorithm, the image is divided into a SxS grid. If the center of any object falls into a grid cell, that grid cell is responsible for detecting that particular object. Each grid cell predicts B bounding boxes of different shapes and sizes. A detected object will be associated with the bounding box with which it has the greatest IoU (Intersection over Union). Each bounding box also has a confidence level associated with it.

간단하게 **YOLO** 알고리즘에 한번 더 되짚어 보면, 객체에서 센터를 기준으로 **Grid Cell**이 만들어 집니다. 각 **Grid Cell**은 **8**개의 **Bounding Box**가 제각기 이루어지며 그중 가장 **IoU**(정확성)이 높은 박스가 사용됩니다. 각 **Bounding Box**는 **Confidence** 값이 동반되죠.



This trained model will detect all the 80 classes in COCO dataset. Not all of these objects would be an obstacle for a visually impaired person at all times. For example, a user might not find a book lying outdoors as an obstacle. Similarly, if the user is indoor, the chances of finding a car as an obstacle would also be very low. Therefore, to save computational resources and time, and also for the convenience of the user, we have given the options of two modes for object detection: indoor and outdoor.

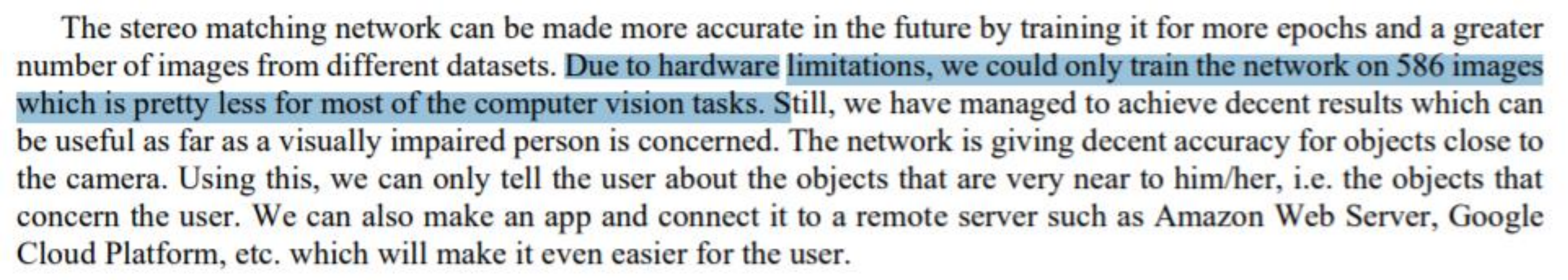
논문에서 **80**개의 클래스가 있는 **COCO** 데이터셋을 활용하였지만 사실 언제나 이 모든 데이터 셋이 필요한것도 아닙니다.

컴퓨터 자원을 생각해서라도 구분해서 클래스를 사용할 필요가 있습니다. 예를 들어 사용자가 밖에서 눕혀 있는 책을 볼일도 없고, 실내에서 자동차를 볼일도 없을 것입니다.

따라서 **Object Detection** 모듈에서 실내와 실외를 구분해서 사용할 필요가 있다고 합니다.

강준구

Future Work



The stereo matching network can be made more accurate in the future by training it for more epochs and a greater number of images from different datasets. Due to hardware limitations, we could only train the network on 586 images which is pretty less for most of the computer vision tasks. Still, we have managed to achieve decent results which can be useful as far as a visually impaired person is concerned. The network is giving decent accuracy for objects close to the camera. Using this, we can only tell the user about the objects that are very near to him/her, i.e. the objects that concern the user. We can also make an app and connect it to a remote server such as Amazon Web Server, Google Cloud Platform, etc. which will make it even easier for the user.

“하드웨어적으로 한계가 있어 많은 데이터들을 학습시키지 못한 것이 아쉽고,
App이나 아마존 서버들을 활용하면 사용자들이 더욱 더 사용하기 쉬울 것 같다.”

라며 끝내고 있습니다.

감사합니다