Image Sensing System for Navigating Visually Impaired People

Thomas Gonnot and Jafar Saniie

Department of Electrical and Computer Engineering
Illinois Institute of Technology, Chicago, Illinois, USA

***Abstract*— About 285 million people are visually impaired around the world. In the United States, this number is over 1.3 million, which represents 0.4% of the population. However, the cost is about 4 billion dollars a year to build infrastructures and provide proper health care to meet the needs of these people. This paper proposes a system that would help the visually impaired adapt to the environment rather than trying to adapt the environment to them. The development of this system will enable the visually impaired to maintain efficient daily activities in many different environments. This system uses multiple sensors and analyzes the information acquired in real-time. In this paper, we particularly emphasize image processing to provide effective obstacle avoidance, object recognition and 3D reconstruction of the environment.**

I. Introduction

Although the visually impaired represent only a small fraction of the United States population, the need for them to be autonomous is highly important [1] [2]. Beyond health care, the government invests billions of dollars every year to adapt the environment for them [3]. Currently, the visually impaired benefit from a lot of care (for example the sound of the traffic lights or texts translated into Braille); however their mobility becomes highly limited as soon as they move away from a friendly environment.

In the past two decades, a significant level of research has been done to help visually impaired people regain some perception of their environment [4]. Some projects have focused on finding a method to display images using retinal implants [5], others are focused on how to improve the infrastructures for ease of mobility, and a few have explored the possibility of finding a viable substitute to human vision. To address this last point, the principal objective for our investigation is to extract the visual information from the environment, and consequently inform and guide the visually impaired person in real-time. The methods to obtain environmental information vary from using a simple range finder (such as sonar or LIDAR [6]) to the use of a camera system with intelligent image processing. Even though the image processing can potentially extract a significant amount of information, the current developed systems are bulky and cumbersome to wear for an extended period of time.

This paper presents the design of a portable and computationally powerful image processing system, referred to as a Visually Impaired Assisting (VIA) system, to be compact enough to fit on a belt and also smart enough to provide all the information that can be extracted from the environment. This system is based on the fusion of data from sensors and cameras. Such a smart visual system can also be interfaced to a smartphone or GPS for improved navigation. This would allow the user to move around freely, beyond simple object avoidance, and be safely guided toward the destination of interest. Extended mobility requires real-time feedback about changes in the environment and this information can be retrieved with the proposed VIA system.

The VIA system needs to be portable. Hence, it must be designed to be small and easy to wear. For example, the cameras can be integrated in the frame of the eyeglasses. The computation/control unit can be attached to the user's belt, and earphones can be used for navigation feedback. An inertial module on the frame of the eyeglasses monitors the direction and orientation of the camera. Figure 1 shows how a comprehensive visually based navigation system would be worn by a visually impaired person.

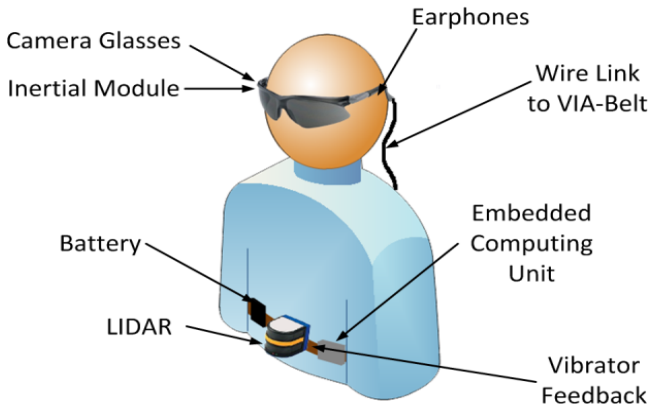

Figure 1. Visually Impaired Assisting (VIA) system

In order to analyze the scene and give feedback to the user, the VIA system is designed to obtain a 3D map of a scene. After this operation, a recognition algorithm segments the scene to extract different objects and reveal their proximities.

978-1-4673-4642-9/13/$31.00 ©2013 IEEE

The objects are recognized using a local database. Alternatively, an online database can be used if a smartphone is connected to the VIA system, providing increased flexibility and accuracy. Once the system has the 3D position and the classification of the objects in the scene, it can determine the path needed to avoid obstacles and provide accurate guidance to reach the destination of interest. This guidance can be either vocal using text-to-speech conversion, or vibrational utilizing the haptic sensitivity of the user. Figure 2 shows the sequence of executions of scene analysis algorithms.

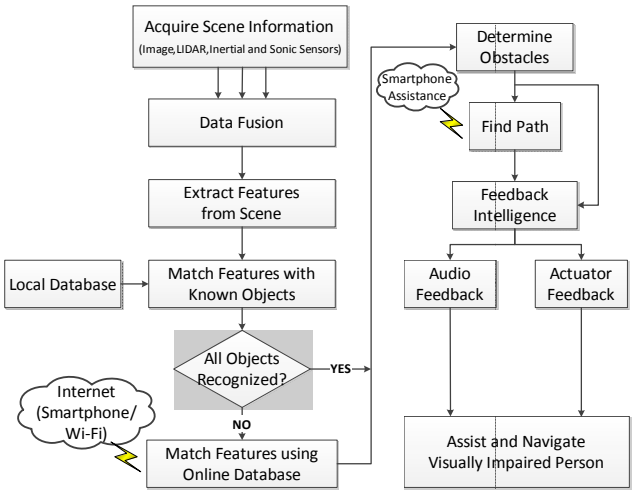

Figure 2. Scene analysis algorithm

In the following sections, this paper presents the image processing used for obstacle avoidance (Section II), object recognition (Section III) and 3D scene mapping (Section IV) for the VIA system.

II. Obstacles Avoidance

It is possible to predict a collision directly from a video stream by analyzing a set of features and their respective movements. In this study, we use Delaunay Triangulation [7] in order to keep track of different features from frame to frame. This creates triangulation groups in which there are no points inside the circumcircle of any triangle. In most cases, this approach is unique and the connections between vertices are maintained regardless of the frame changes. Figure 3 shows an example of Delaunay Triangulation applied to a set of features.

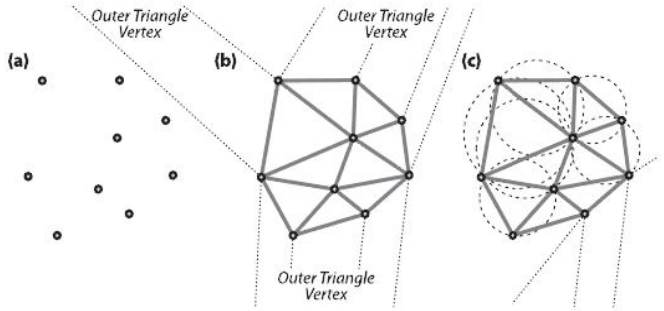

Figure 3. Delaunay Triangulation Calculation Process

The system uses the features extracted from the image and evaluates the optical flow, also known as the motion vectors, of the frame. Each vector is a combination of a rotation and a translation. The knowledge of the features and their motion allows us to separate the translation from the rotation, and to determine a convergence point (called focus of expansion) corresponding to the direction of the object's movement. The faster the features are moving away from the focus of expansion, the closer the object is to the camera position. In order to quantify this, we use the Time-to-Collision (TTC) equation also known as Time-to-Crash [8], $TTC = \Delta_i / \left|\overrightarrow{V^t}\right|$ where Δ_i is the distance of a feature from the focus of expansion, and $\overrightarrow{V^t}$ is the optical flow vector of the feature. If the object's physical size is known by the system, the measurement of the distance from the object can be derived from the physical distance of the features and can be correlated to the motion vectors to predict the collision. The system can then determine if the obstacle is close enough to justify avoiding it. This method however is limited for navigation in a complex environment where several focuses of expansion appear. In this case, the algorithm needs to associate the surrounding features to a particular focus of expansion and compute a time-to-collision factor for each convergence point. This approach may result in error in calculating the time-to-collision factor. Therefore, a better approach is to first segment an image, next identify an object of interest, and then track the motion of the object in order to estimate the time-to-collision factor.

III. Object Recognition

In the VIA system, object recognition is an essential part of the scene analysis within an environment. Detecting an object in the scene enables the user not only to navigate around the object, but also to interact with. For example, a door can be recognized as an obstacle, but the person can open the door to remove the obstacle. The recognition of signs in a street is also a good source of information about the directions to follow, or the dangers to avoid. In this study, we use highly efficient object recognition methods, one being SIFT (Scale Invariant Feature Transform [9] [10]), which utilizes common templates stored in the VIA system database. The other method relies on cascade classifiers and offline learning [11].

The SIFT method consists of two functional parts called extraction and matching. In the extraction step, the image is blurred using Gaussian functions with different blurring intensities. The blurred images are sequentially subtracted from each other and the resulting subtracted images are called Difference of Gaussians (DoGs). These DoG images are used to extract keypoints defined as all minimum and maximum points in the DoG images. Then, the original image is scaled and the process of extracting keypoints for the scaled images is repeated. Figure 4 shows how the DoG images are calculated. The pattern of keypoints is thresholded in order to keep and emphasize the strong image keypoint features.

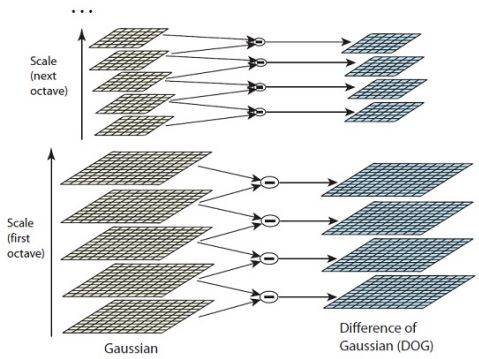

Figure 4. Difference of Gaussian used in the SIFT algorithm

Next, keypoints are analyzed in order to generate feature vectors. The feature vector represents the intensity of the different gradient orientations around each keypoint. In order to recognize an object, the feature vectors of the object must be matched to the desirable vectors stored in the database of the VIA system. This database can be very advanced in terms of recognizing a large number of objects. However, the computation for recognition and storage for a database demands a very high power computing machine and massive storage. Therefore, a central server is required, and this can only be accessed through the means of smartphone. Hence, enabling the VIA system to connect to a smartphone and communicating with a central server, as shown in Figure 2, will provide a much greater flexibility, reliability and accuracy to navigate a visually impaired person.

In the second approach we use a special processing called AdaBoost, a cascade classifier. In this method, the image is segmented in subparts of equal size. Then features are extracted from each subpart, with a special set of features shown in Figure 5.

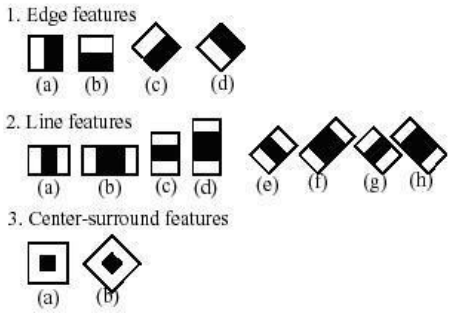

Figure 5. Example of feature used for the classifier, the white part represents the positive weight, and the black represents the negative weight

These features generate results that enter the classifier providing a 'yes' or 'no' answer for object recognition. The cascade classifier generates a 'no' result quickly, improving the overall computation and recognition performance. Once all the subparts are analyzed, the image is downscaled and the process starts over. This continues until the image reaches the size of the classifier input. In order to recognize different objects, the system must include a unique classifier corresponding to each object for recognition. In order to recognize a large number of objects, the training of this classifier becomes computationally heavy. Because of this, the training is mainly done offline, and this method is limited to a small number of objects. However, the object recognition time is faster when compared to the SIFT algorithm.

IV. 3D Scene Mapping

In order to map the recognized objects in the environment, the system can compute the distance of each feature within a scene.

An effective method used to obtain a depth map is the one used by Kinect from Microsoft [12]. An infrared, structured light is projected on the scene, and an infrared camera captures this projection. The system can then compute the distance of each object from the deformation of the structured light projected on them.

The second method consists of using two cameras separated with a constant and known distance. The difference between the two images is used to compute the distance of the objects in the scene. This process, called stereovision, uses the Epipolar theory represented in Figure 6: calculating the 3D position of a feature is done by observing its correspondent point from the two cameras [13].

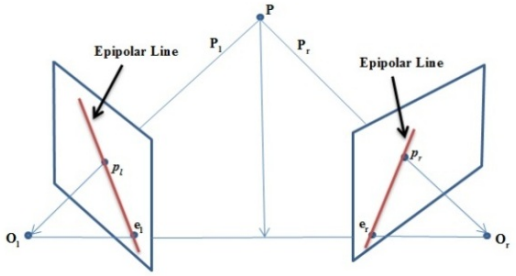

Figure 6. Illustration of the Epipolar theory

Stereovision requires performing a calibration. This calibration will determine the different characteristics of the system, such as the angle formed between the two cameras, and apply the corrections for the inner optical characteristics of the camera. It requires a pattern, typically a checkerboard, which will be presented to the two cameras with various angles. This calibration remains valid as long as the relative position of the cameras does not change.

The stereovision algorithm is composed of three steps. The first step removes the distortions from the camera optics. It also turns the stereo pair into standard form, projecting the images on the plane formed by the two cameras. The next step is to find the position of each feature of the first picture in the second picture. Since the two images have been transformed to a standard form, the search is restricted to the horizontal lines around the feature. The last step is to determine the distance of the feature from the camera by triangulation using the system position and optic constants. A scan of all the pixels in the image generates the depth map profile that can be used to map different objects.

The process of mapping is called SLAM (Simultaneous Localization and Mapping Algorithm). SLAM combines the depth information acquired over time into a map by tracking the motion of the camera and correlating the position of each object in the scene. The result is a 3D map of the scene including all the objects detected.

V. Results

In this study, we present the results from the different algorithms used. For instance, Figure 8 shows the result of the time-to-collision approximation while approaching a closet. The features detected on the closet are shown in Figure 7.

Figure 7. Three frames and features at different distances from a closet while moving forward to it

As the camera approaches the closet, the features move toward the outside of the picture at an increasing speed. The resulting time-to-collision is presented in Figure 8. In this case, we slowed the approach to the closet. As we can see, even close to the closet, the features' movements are quite visible.

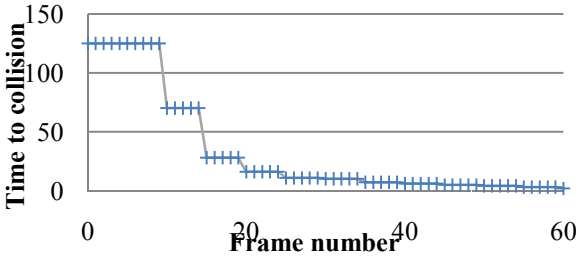

Figure 8. Resulting Time-to-Collision while approaching the closet

In order to test the object recognition algorithm, pictures of the object to be recognized are presented to the algorithm alongside a video stream. The algorithm then analyses each frame of the video stream and returns the position and orientation of the object if found in the frame. Figure 9 shows the result of the recognition of a door using the SIFT algorithm and the recognition a helicopter using a cascade classifier.

Figure 9. Left: example of detection of a door. Right: example of detection of a helicopter

The last algorithm presented is the 3D-reconstruction. The system uses two cameras and generates a depth map of the scene. As shown in Figure 10, the generated map is represented in grey levels, white (here the hand) being the closest.

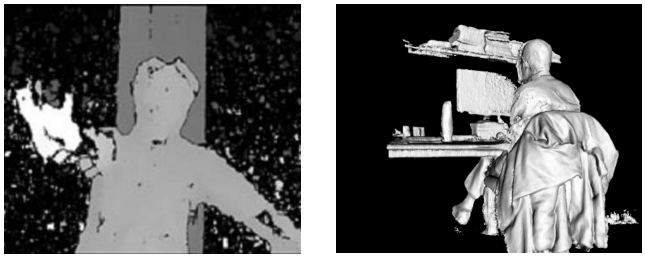

Figure 10. Left side: Depth map obtained by combining 2 cameras Right side: 3D reconstitution of a scene using a Kinect sensor

VI. Conclusions

This paper presented an image processing concept for a Visually Impaired Assisting device. We also introduced and tried several algorithms to avoid obstacles, recognize objects or reconstruct a scene in 3D. The results show that these algorithms are efficient enough for the system to perform its task: helping the visually impaired by enhancing their mobility.

References

[1] National Federation of the Blind, [Online]. Available: http://www.nfb.org/blindness-statistics. [Accessed 2012].

[2] "U.S. & World Population Clocks," United State Census Bureau, [Online]. Available: http://www.census.gov/main/www/popclock.html. [Accessed 2012].

[3] C. YP, B. LJ and J. JC, "Federal budgetary costs of blindness.," 1992, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1614382. [Accessed 2013].

[4] J. A. Brabyn, "New Developments in Mobility and Orientation Aids for the Blind," *IEEE Transactions on Biomedical Engineering,* Vols. BME-29, no. 4, pp. 285-289, 1982.

[5] W. Mokwa, "Retinal Implants to Restore Vision in Blind People," *16th International Conference on Solid-State Sensors, Actuators and Microsystems,* pp. 2825-2830, 2011.

[6] M. Bansal, B. Matei, B. Southall, J. Eledath and H. Sawhney, "A LIDAR Streaming Architecture for Mobile Robotics with Application to 3D Structure Characterization," *2011 IEEE International Conference on Robotics and Automation,* pp. 1803-1810, 2011.

[7] M. d. Berg, O. Cheong and M. v. Kreveld, in *Computational geometry: Algorithms and Applications*, Springer-Verlag, 2008, pp. 191-218.

[8] N. Ancona and T. Poggio, "Optical flow from 1D correlation: Application to a simple time-to-crash detector," *Fourth International Conference on Computer Vision,* pp. 209-214, 1993.

[9] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision,* pp. 91-110, 2004.

[10] T. Lindeberg, "Scale Invariant Feature Transform," 2012. [Online]. Available: http://www.scholarpedia.org/article/SIFT.

[11] P. Viola and M. Jones, "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade," *International Journal of Computer Vision,* pp. 1311-1318, 2001.

[12] J. MacCormick, "How does the Kinect work?," [Online]. Available: http://users.dickinson.edu/~jmac/selected-talks/kinect.pdf. [Accessed 2013].

[13] S. Mattoccia, "Stereo Vision: Algorithms and Applications," 2012. [Online]. Available: http://www.vision.deis.unibo.it/smatt/Seminars/StereoVision.pdf.