Development of an assistive stereo vision system

Tanmay Shankar
Department of Mechanical Engineering, Indian Institute of Technology Guwahati
tanmay.shankar@gmail.com

Abhijat Biswas
Department of Electrical and Electronics Engineering Indian Institute of Technology Guwahati
abhijatbiswas@gmail.com

Venkat Arun
Department of Computer Science and Engineering, Indian Institute of Technology Guwahati
venkatarun95@gmail.com

ABSTRACT
In this paper, we describe the development and working of an assistive vision system that utilizes stereo computer vision methods to help visually impaired individuals carry out everyday tasks in unknown environments. The proposed device would be capable of a set of behaviors, each of which supports the user in a particular set of tasks. Object detection and hand guidance, face and text detection and recognition are some of the implemented behaviours. A dynamic scheduler and frame selector are also employed to reduce computational cost by processing a frame only when a sufficient degree of change has accumulated. The system interacts with the user via an auditory output channel.

Keywords
Assistive Technology, Visually Impaired, Computer Vision, Stereo Vision, Low Cost.

1. INTRODUCTION
As of August 2014, WHO reports 285 million people with visual difficulties. 80 % of these all visual impairments can be cured or avoided altogether with appropriate treatment and cure. However, surgeries are expensive and skilled doctors are not readily available everywhere, especially in developing countries like India (which accounts for the largest share in the 285 million mentioned). Computer vision, through a system such as the one proposed here, offers a relatively inexpensive alternative to these expensive treatments.

2. RELATED WORK
The bulk of visual systems for assisting visually impaired individuals proposed until now have been focused on navigation for the blind. Literature on vision substituting assistive devices However, there are no applications claiming stereo object detection and hand guidance to facilitate user-object interaction. Some methods [1, 2, 3] for object recognition rely on either external markers on the objects themselves, hand carried sensors or both for detection and locating objects. In these works, the dependence on these specialized markers makes them extremely restrictive due to their functionalities being relevant in only controlled environments. Apart from this, there exist object detectors that use feature extractors (predominantly SIFT or SURF based) and using them to recognize new objects [4, 5].

These have the inherent problem that interest points are extracted from a particular view of the object of interest and lose functionality the moment the object is turned to present a featureless surface. Hence, we propose a shape based segmentation for object recognition in the camera's field of view (FOV). This allows us to detect objects irrespective of the direction in which they are facing in the FOV.

In all of these works, the sensory functionalities of the eyes are substituted by another sensory organ, most commonly touch or hearing. Tactile systems [6] have been proposed and implemented. However, it is not clear as to how effectively such haptic transductions assist in building mental representations of scenes. Other methods [5, 7] suggest sonification to guide users toward objects, which is not a well-developed field. There exist no conclusive results to establish sonification as an intuitive standard for auditory representation of spatial data. Hence, we propose to use speech synthesis and audio output for hand guidance which entails no such ambiguity

3. CONCEPT
This paper proposes an Intelligent Vision System for Blind Enablement (InViSyBlE) to aid visually impaired individuals, which executes a series of independent behaviours. Each of these behaviours aid the user in carrying out a set of tasks that visually impaired individuals typically face difficulty with.

InViSyBlE consists of a light-weight head mounted stereo-camera setup, which is interfaced with a portable computing device. A bracelet embedded with artificial markers is worn by the user for usage in some of the behaviours. The user would be able to toggle between which behaviour is active depending on the situation. The proposed behaviours include – Object Detection; User hand tracking and guidance; Face detection and recognition, and Text detection and recognition.

We use a stereo system, with two constrained cameras to mimic human vision capabilities as closely as possible. This enables the system to obtain depth data from the scene which facilitates the hand guidance behaviour. In order to execute these behaviours in an efficient manner, a dynamic module is implemented for task scheduling, and frame selection.

Emphasis is provided to making the device accessible to visually impaired individuals of all strata of society, keeping in mind blind users in developing countries. The cost of the device peripherals is hence kept as low as possible.

4. BEHAVIOURS
4.1 Object Grasping Behaviour
The first behaviour is aimed at providing users with assistance in grasping day to day objects. By tracking the position of the user's hand, and the 3D pose and orientation of the object in question, a

reliable interface is provided by which the user may guide his or her hand towards the object. This behaviour is implemented in three phases.

4.1.1 Phase 1: Hand tracking and pose estimation.

While literature provides copious documentation on tracking human hand movements and recognizing corresponding gestures, the chosen method to implement the first phase of this behaviour is to provide the user with an Augmented Reality Tag bracelet to wear on their hand.

An Augmented reality tag (AR Tag) is a 2D fiducial encoded in low resolution with a unique pattern ID. This facilitates robust pose and orientation estimation of the tag, invariant to scale and rotation changes, provided the initial size and ID of the tag are known. Considering the bracelet is provided to the user, calibration of the behaviour according to this initial data is trivial.

The advantages that the ART-Bracelet provides over typical hand detection are –

- Robust pose and orientation estimation
- Invariance to scale and rotation changes
- Unique identification (for multiple hands in the scene)

For a given frame, the AR tag coordinates are retrieved, and the coordinates are passed through a fixed response filter with an adjustable buffer, to smoothen noise in the data, and minimize discontinuities in coordinate values, as follows:

$$x_{current} = \frac{1}{2} x_{read}(t) + \sum_{i=1}^{n_{buffer}} \frac{1}{2^i} . x_{read}(t-i)$$

4.1.2 Phase 2: Object Detection and Pose estimation

The location of an object in the field of view may be estimated by 2D or 3D techniques. While 2D methods based on feature descriptors provide high performance recognition of a particular object in the scene, especially from a database, they are somewhat inadequate in determining the 3D pose and orientation of these objects. Additionally, 2D methods are typically restricted to recognizing limited views of an object - for example, a juice box whose features were extracted in one view, may not be recognizable by this approach from an alternate spatial configuration. Moreover, depth is of prime importance in this particular application.

Figure 1: Raw camera feed (left) and corresponding disparity map (right).

A 3D approach is hence chosen, based on generating a point cloud from the incoming camera feed, and running appropriate algorithms on it to retrieve objects of the required contour.

Two such algorithms are chosen to perform the object detection itself, i.e. geometric segmentation (GEOSEG) for detection and correspondence grouping (COG) for recognition.

Prior to running these algorithms on the incoming pointclouds, a series of pre-processing steps are executed to improve the performance of the system, in terms of both time taken and robustness of results. The pointclouds are first subject to a pass through filter, or a field of view (FOV) filter, thus eliminating erroneous points that cannot be within the field of view of the assumed camera-sensor model. Pointclouds are then subject to distortion correction, to account for errors in the calibration and alignment of the stereo cameras. Statistical outliers are then removed by the mean neighbour distance technique. These pre-processing steps prove to be valuable in handling corrupted measurement data, and prevent the calculation of irregular curvatures of false surfaces.

The geometrical segmentation behaviour is primarily intended for indoor environments where the user attempts to manipulate an object of a particular form, present on a uniform level surface. Computation is considerably simplified by localizing this level surface, which is done by a planar segmentation algorithm.

The more general GEOSEG algorithm consists of identifying regions in the pointcloud that conform to a particular geometric tolerance. Without loss of generality, a cylinder is considered for GEOSEG. Assuming an adjustable radius value, the segmentation behaviour makes use of the RANSAC algorithm to minimize the distance between the given point cloud feed and a cylinder of the specified dimensions. Multiple instances of cylinders may be detected. The resultant cylinder(s) are then stored in an independent point cloud for post-processing.

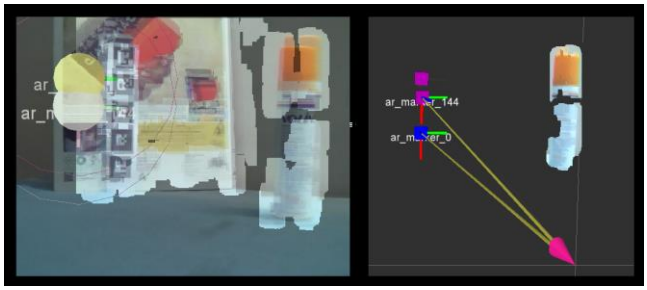

Figure 2: Segmented cylinder and AR Tag coordinates.

Once the object present in the scene is detected, the global coordinates of the object are retrieved by computing the centroid of the points that constitute the object pointcloud. The orientation and the pose are retrieved as 6 DOF transformations with respect to the primary axis of the object. Similar to the AR Tag coordinate output, these orientation and pose estimates are subject to a fixed response filter, with an adjustable buffer, to eradicate noise and smoothen out the output.

4.1.3 Phase 3: Relative transformation

A set of virtual local reference frames are attached to the ART bracelet and the detected object. These reference frames are in connection to the filtered pose and orientation coordinates obtained in phase 1 and 2. The next step is to determine the relative transformation between these reference frames. A homogenous transformation matrix is made use of in computing

the 6 DOF relative transform required for the user to move his hand towards the object.

$$\begin{bmatrix} x_{rel} \\ y_{rel} \\ z_{rel} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & x_{hand} \\ R_{21} & R_{22} & R_{23} & y_{hand} \\ R_{31} & R_{32} & R_{33} & z_{hand} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{obj} \\ y_{obj} \\ z_{obj} \\ 1 \end{bmatrix}$$

The method proposed for this conversion of this data is by interpolation of the coordinates of the object in the point cloud, with respect to the location in the image. This information is subsequently published to an interfacing node with the audio output.

4.2 Text Detection and Recognition

The problem of text recognition from a scene in the wild is non-trivial and is a subject of current interest. To improve the performance of the system, InViSyBlE uses several preprocessing steps before performing actual OCR.

The image is first put through a Canny filter to obtain the edges. A Stroke Width Transform [8] is then performed to isolate candidate characters, i.e. continuous regions of uniform width. Once all such candidate components are obtained, they are chained together to form pairs of such characters, consequently words and finally sentences. At this point, there are only the candidate characters and words in the image. The resulting image is passed to the tesseract OCR engine for individual word and character recognition, which is relayed to the user in a

4.3 Face Detection and Recognition

This behavior allows the user to know if a person appears in his field of view. The Viola-Jones haar-wavelet based face detection framework [9] was used to detect regions of interest in sampled images. A boosted cascade of classifiers was used. For scenes with multiple regions of interest detected, only the largest was passed to the immediate higher level of the software architecture, namely the face recognition module. The face recognition module used the eigenfaces method which is a standard PCA based approach as in [10], for a limited number of people.

5. ARCHITECTURE

5.1 Software

5.1.1 Flow control and frame selection

Since InViSyBlE runs on a mobile platform, power consumption is a concern. Hence our system efficiently chooses which frames to analyse and increases idle time when there is no significant change in the scene. Using this system, the processor usage varies by more than 400% depending on the amount of change in the scene.

The frame selecting algorithm works as follows:

```
prevFrame ← FetchFrame()
avgBlurriness, curBlurriness ← 0, 0
change ← 0
while true
while change < C and curBlurriness <= f*avgChange
        ScheduleDelay()
        avgBlurriness ← (avgBlurriness + curBlurriness)/2
        curFrame ← FetchFrame()
        change ← Change(curFrame, prevFrame)
        curBlurriness ← CalculateBlurriness(curFrame)
SendForProcessing(curFrame)
prevFrame ← curFrame
```

We threshold the maximum tolerable change between two processed frames. That is, we use the Sobel edge detector to obtain the 'edge images' and model change between frames as the average difference between the two 'edge images'. We try to ensure that this is no more than a threshold, C. Using edge images rather than direct pixel wise differences helps guard against noise in poor quality cameras. To determine whether a frame is in sharp focus, we use the average edge strength which we ensure is greater than a fraction of the exponential moving average of previous values. This helps prevent wastage of power on highly motion blurred images. For it to be effective, we ensure that the frame selector is relatively computationally inexpensive.

The 'ScheduleDelay' function makes the process idle for some time,t to save power. The delay is chosen so that the probability of the change exceeding the given threshold is lower than a given threshold, $P_{\text{exceed thresh}}$.

Let Pt(C) denote the probability of a change of C occurring in a time interval t. Assume probability that change C occurs in 1 unit time is a gaussianP1(c) = G, with mean and standard deviation. We will now show by induction that probability distribution at any time t, Pt, is also a Gaussian. With the above expression as the base case, we have the inductive step as follows:

$$P_{t+1}(C) = \int_0^C P_t(x) G_{\mu,\sigma}(C-x)dx \approx \int_{-\infty}^{\infty} P_t(x) G_{\mu,\sigma}(C-x)dx$$

$$\Rightarrow P_{t+1}(C) \approx (P_t * G_{\mu,\sigma})(C)$$

$$P_t(C) = G_{\mu_0,\sigma_0}$$

$$P_{t+1}(C) \approx \left(G_{\mu_0,\sigma_0} * G_{\mu,\sigma}\right) = G_{(\mu_0+\mu),(\sigma_0+\sigma)}$$

$$P_t(C) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(x-t\mu)^2}{2t^2\sigma^2}}$$

μ is estimated by an exponential moving average of measured $C/\Delta t$. σ is similarly measured by an exponential moving average of $(C/\Delta t - \mu)$. We then choose a time such that the probability of the change, $C < P_{exceed\ thresh}$.

5.1.2 Process flow diagram

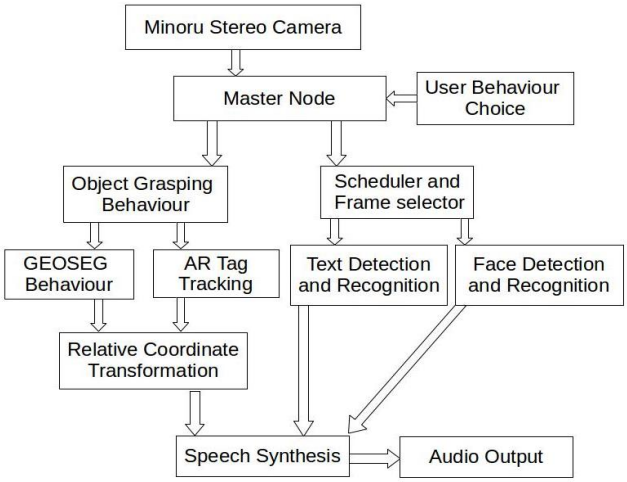

Figure 3: Process flow diagram showing the flow of control for a typical system run.

5.2 Hardware choice: low-cost peripherals

Since InViSyBlE is designed to be low cost, an inexpensive set of stereo cameras is chosen. Aural output is relayed through a pair of ordinary earphones. The device is supported by a portable computing device.

Figure 4: The device peripherals. A headband mounted stereo camera that can be worn on the forehead.

6. FUTURE WORK

Future work can focus on trying to push computation or database intensive workload to remote servers. This should be done intelligently to minimize power usage due to the network and response latency. Our scheduling algorithm can be used here to select appropriate frames to send. Further, in case of a pure camera pan, the frame selector can send only the novel parts of the image to the further layers for object/face/text detection and recognition while a motion estimate can be used as a basis for tracking in all behaviours to save on computation.

The AR tag hand guidance system can be enhanced by detecting best regions to grasp on an object using DNNs [11]. Further, implementing everything on a GPU will help increase speed while simultaneously saving power. Proper research on the HCI elements of the system will help improve the usability of the system. Though systems for navigation for the blind exist, it is crucial for them to be very accurate for them to be used in generic scenarios without danger. The ultimate aim is to have a highly robust device with an intuitive interface that can help the blind to perform all activities without restrictions.

7. ACKNOWLEDGMENTS

We thank Amit Sethi for his valuable inputs and having fruitful discussion toward developing this system.

8. REFERENCES

[1] Rabia Jafri, Syed Abid Ali and Hamid R. Arabnia: Computer vision-based object recognition for the visually impaired using visual tags. *Proceedings of the 2013 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV '13)*

[2] Murad, M., Rehman, A., Shah, A.A., Ullah, S., Fahad, M., Yahya, K.M.: RFAIDE—an RFID based navigation and object recognition assistant for visually impaired people. *7th International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, pp. 1–4 (2011)*

[3] Hub, A., Diepstraten, J., Ertl, T.: Design and development of an indoor navigation and object identification system for the blind. *Proc. Int. SIGACCESS Conf. Computers and Accessibility (2004)*

[4] Winlock, T., Christiansen, E., Belongie, S.: Toward real-time grocery detection for the visually impaired. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 49–56 (2010).*

[5] Boris Schauerte, Manel Martinez, Angela Constantinescu, and Rainer Stiefelhagen. 2012. An assistive vision system for the blind that helps find lost things. *Proceedings of the 13th international conference on Computers Helping People with Special Needs - Volume Part II.*

[6] Akhter, S., Mirsalahuddin, J., Marquina, F.B., Islam, S., Sareen, S. A smartphone-based haptic vision substitution system for the blind. *2011 IEEE 37th Annual Northeast Bioengineering Conference (NEBEC), Fairfax, VA, USA, pp. 1–2 (2011)]*.

[7] J. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh, "VizWiz::LocateIt - enabling blind people to locate objects in their environment," *3rd Workshop on Computer Vision Applications for the Visually Impaired (CVAVI 10), San Francisco, California, 2010.*

[8] B. Epshtein, E. Ofek and Y. Wexler, Detecting Text in Natural Scenes with Stroke Width Transform, *Computer Vision and Pattern Recognition (CVPR) 2010*

[9] P. Viola and M. J. Jones. 2004. Robust Real-Time Face Detection. *Int. J. Comput. Vision 57, 2 (May 2004), 137-154.*

Matthew Turk and Alex Pentland, Eigenfaces for Recognition, *Journal of Cognitive Neuroscience 1991 3:1, 71-86*

[10] Lenz, Ian, Honglak Lee, and Ashutosh Saxena. "Deep learning for detecting robotic grasps." *arXiv preprint arXiv:1301.3592 (2013).*