34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Audible Vision for the Blind and Visually Impaired in Indoor Open Spaces

Xunyi Yu, Aura Ganz, *IEEE Fellow*
Electrical and Computer Engineering Department
University of Massachusetts, Amherst

Abstract—In this paper we introduce Audible Vision, a system that can help blind and visually impaired users navigate in large indoor open spaces. The system uses computer vision to estimate the location and orientation of the user, and enables the user to perceive his/her relative position to a landmark through 3D audio. Testing shows that Audible Vision can work reliably in real-life ever-changing environment crowded with people.

I. Introduction

The World Health Organization (2010) reported that globally the number of people of all ages visually impaired is estimated to be 285 million, of whom 39 million are blind [1]. Difficulties of moving freely and independently, especially in unfamiliar environments without blind accessibility design, is one of the major hurdles that they encounter to lead an independent life and fully integrate into the society. Navigation in large indoor open spaces, e.g. lobby areas of a building, public transit stations, shopping malls, is particularly challenging for the blind and visually impaired population. They must learn to rely on sensory cues from the environment such as tactile, auditory, perception of air currents, light sources, etc. to facilitate orientation while moving through open space from a landmark to another. They also need to remember the layout of the environment, constantly trying to maintain a satisfactory straight line of travel, judge the approximation of distance travelled, locate and keep track of various landmarks they have past, and back-trace to the last known location when they become disoriented. It is a difficult and prolonged process for the blind and visually impaired users to learn to navigate independently in an unfamiliar open space environment, even with the help of experienced Orientation and Mobility instructors.

Large open spaces are also challenging for various navigation aid systems that have been developed to help the blind and visually impaired. The user's location and orientation need to be estimated with both high reliability and accuracy. It is relatively easy for a person with normal vision to verify his location estimate given by a mapping application. But this is not the case for the blind and visually impaired users. Even occasional failures of localization and resulting erroneous navigation instructions can cause serious disorientation, and potentially danger to the user. Also, the blind and visually impaired rely heavily on various landmarks to navigate through an environment. Thus, it is not adequate to just locate the user to a zone level. More accurate location and orientation estimate are needed so that relative position of various nearby landmarks to the user can be calculated with high accuracy. GPS[2], WiFi[3], or active RFID[4] have been used in blind navigation applications. However, they can only reliably locate a user to zone level, and need other means to determine the orientation of the user. A compass is commonly used, but the estimation is subject to interference, and can be off tens of degrees.

An alternative is to use passive RFID tags, and NFC devices that have very short range [5]. When the user touches and scans a tag, reliable and accurate user location and orientation can be determined assuming the user is facing the tag. But the user's location is only available when scanning the tag, and cannot be estimated at a distance from the tag. Visual tags have also been used to locate the blind using smart phones [6]. They can be detected in range, while providing accurate location and orientation of the blind person. Passive RFID and visual tags localization are suitable for school and office buildings, where tags can be easily attached to doors along the hallway, and can be scanned by the user. But in open spaces, it is challenging to find places to attach the tags densely, and for the user to find and scan these tags. Also, the location of all these tags needs to be recorded and maintained in a database, increasing the deployment cost.

Localization using only visual features inherent in the environment without deploying visual tags is another emerging and appealing option. Current Smartphones with increasing processing power can run the computer vision localization algorithm locally [7]. Vision based localization can fail in locations lacking visual features, or repetitive features. They work best along with an alternative localization method that can locate the user to a zone level, and reduce the search space.

As for user interfaces, current navigation aid systems typically give voice navigation instructions to the user once a

Acknowledgement: This project was supported in part by the following grants: DUE- 1003743 from the National Science Foundation and 1 R21 EY018231-01A1 from the National Eye Institute/ National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, National Eye Institute or the National Institutes of Health.

destination is chosen. Voice navigation instructions are very suitable for Manhattan-world style environments where instructions such as "turn left", "walk straight ahead" can be efficient assuming that the user is aligned with a dominant axis of the world. In open spaces however, the voice navigation instructions can be less efficient in conveying the route information to the user. Also, it will be challenging for the user to follow the instructions exactly such as making perfect turns, or moving in straight lines. 3D audio has been used to enable the blind to interact with virtual environments, and has lots of potential to be used in real world navigation applications [8].

In this paper we introduce Audible Vision, a system that uses computer vision to estimate the position and orientation of the user. It delivers the position of landmarks relative to the user through a 3D audio interface. Audible Vision does not provide routing and navigation function on its own, and is designed to complement and enhance existing indoor navigation systems in indoor open spaces by providing the following unique functions: 1) Seamlessly estimate the location and orientation of the user in open spaces with high reliability and accuracy, and 2) Alternative 3D audio based user interface that enables the user to sense the position of landmarks, and directs the user in non-Manhattan-world environments.

The paper is organized as follows. Audible Vision components are introduced in the next section. A sample scenario is described in Section III and Section IV evaluates the performance of the vision based localization. Section V concludes the paper.

II. SYSTEM COMPONENTS

In this section, we describe the photo annotation mapping tool, the vision based localization algorithm, and the Android implementation of Audible Vision.

A. Photo Annotation Mapping Tool

We first extract the 3D structure of the environment from image sequences using motion software Bundler[9]. The outputs of the 3D reconstruction are the position and appearance descriptor of feature points in the environment. We use Surf features [10] during the 3D reconstruction as well as in the localization. We assume that the photos are taken with camera approximately up-right, and apply Surf without orientation invariance to increase its discriminative power. The Surf descriptors for the reconstructed 3D points are saved in a kd-tree structure for approximate nearest neighbor search in the localization phase[10]. After the reconstruction phase, the mapping tool will pick a set of photos that cover all feature points of the environment. The sighted user of the mapping tool can put annotations directly on the landmark in the photos, and the tool will automatically establish the correspondence between 3D feature points in the reconstruction and the landmark. The user can optionally mark out areas in photo areas from which feature points are not distinctive, or subject to frequent change, and thus not suitable for localization, e.g. floors and small furniture. Our mapping tool can use up-to-date new photos to incrementally update 3D feature points when new decorations or furniture are added.

We perform reconstruction for individual confined areas, without explicit global registration, which simplifies the mapping process. We perform rotationally aligned local 3D construction with a global coordinate system, so that accelerometers and compass on the phone can be used to assist vision based location and orientation estimation. We do not further divide reconstruction of a confined area into Potentially Visible Sets, which would require knowledge about finer initial location and orientation of the user.

B. Vision based Localization Algorithm

Once we have the 3D reconstruction of an environment, we can locate the blind person using a photo taken by his smartphone. Existing approaches typically assume known camera intrinsic, and solve for full 6 Degrees of Freedom (DoF) camera location and orientation by first analytically solving for the depth of 3 feature correspondences as 3-point pose problem [11]. The approach works well in small rooms, or outdoors where features are spread across the y axis of the images. However, in large indoor open spaces with short ceilings, features crowded on the horizon, making the points and the camera almost co-planar. In this case, the solution of the 3-point pose problem is very sensitive to noisy measurements. Iterative approaches that use more correspondence to solve for the projection matrix of the camera have similar issues, and need large number of RANSAC iterations when the proportion of outliers is high [12].

To circumvent these problems, we take advantages of the accelerometer and compass on the phone, and confine the localization problem to 2D. We assume a projective camera model,

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \mathbf{K}\begin{bmatrix}\mathbf{R} \mid \mathbf{T}\end{bmatrix}\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \qquad (1)$$

where x is the homogeneous coordinate of 2D interest points on the photo, X is the homogeneous coordinate of the corresponding 3D coordinates, $\mathbf{K}$ is the known intrinsic matrix of the camera, $\mathbf{R}$ is the rotation matrix of the camera, and $\mathbf{T}$ is the position of the camera.

The accelerometer on the phone is very accurate when the user is holding the camera still, while the compass can still be off tens of degrees depending on the magnetic sources nearby. Therefore, given the rotation matrix estimate $\mathbf{R}_s$ from sensors on the phone, we can approximate $\mathbf{R}_s'\mathbf{R}$ as a degenerated rotation matrix in 2D, rotating only along z axis, having the following form

$$\mathbf{R}_\Delta = \mathbf{R}_s'\mathbf{R} \approx \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad (2)$$

where θ is the difference between the heading given by the compass and the actual heading of the camera.
We can than transform the coordinates of the feature points in the image into virtual measurements,

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \triangleq \mathbf{R}_s' \mathbf{K}^{-1} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \propto \mathbf{R}_\Delta \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix}. \quad (3)$$

As measurements along z axis are crowded on the horizon, and are very sensitive to residual camera tilt, we only uses measurements along x and y axis. The resulting 2D localization problem can be solved using the following constraints,

$$\frac{x'}{y'} = \frac{\cos\theta X + \sin\theta Y + T_x}{-\sin\theta X + \cos\theta Y + T_y}. \quad (4)$$

Three points correspondence are still needed to solve for x and y coordinates of the camera, and the compass estimation error. Alternatively, we keep θ constant, and solve for location using two points correspondence, which is less complex and takes advantage of initial heading estimation by the compass. We search within the window of the maximum compass error, and use θ and location estimate combination that has most number of inliers as the final estimate.

As point correspondence are subject to error, RANSAC scheme is often used to find hypothesis with most inliers. As we are estimating user localization in 2D as opposed to 6 DoF estimation of camera position, we can afford to directly remove outliers in the solution space, and thus avoid running verification step of RANSAC that need to iterate through all point correspondences for each hypothesis. We first use different point correspondences combinations to generate user location hypothesis. Location estimates from correct correspondence fall in a small area, while estimates from incorrect correspondence will spread across the entire space. Thus, we first divide the 2D space into grids, and find the grid with most votes. With this initial estimate, we use mean-shift algorithm to find the mode of the location estimate distribution [13]. Using the spread of the location estimate hypothesis around the mode, we can get a reliable estimate.

C. Android Implementation

We implemented the Audible Vision system on the Android platform. The implementation includes the following modules: Map Management, Sensor Management, Camera Localization and 3D audio User Interface.

Map Management module: It loads 3D feature point clouds saved in kd-tree and landmark information into the memory, given the user's approximate location. The user's approximate location can be entered by the user when he/she enter the building, or scan a RFID tag located at the building entrance. The map files range from 500KB to 5MB, and can be pre-downloaded into the phone, or pulled from a server dynamically.

Sensor Management module: it estimates the camera rotation using the phone accelerometers and magnetic field sensor. The magnetic field strength is monitored to detect the presence of metal sources that interfere with the compass. This enables the Camera Localization model to adaptively adjust the orientation search window size based on the accuracy of the compass. Also, we use the gyroscope on the phone to track the rotation of the phone after a photo is taken, so that the user can pan the phone around to better perceive the relative position of the landmarks. The rotation of the phone is constantly updated and fed into the 3D audio module so that the user can pan the phone around for better perception of the relative orientations of the different landmarks.

Camera Localization module: it takes photos, extracts Surf features, and uses the initial camera rotation estimate from sensors to estimate the user orientation and location in 2D. The 2D estimate is combined with the initial camera orientation estimate and approximate user height to generate the final 6 DoF pos estimate.

3D Audio module: it updates the initial camera pos estimate obtained by the Camera Localization model with rotation from gyro. Then, the relative positions of the landmarks are calculated. The 3D audio module can cycle through all landmarks within a certain range from the user, and plays back in 3D audio: "landmark X in Y clock direction Z feet away". Repetitive sound patterns, e.g., We use soft-OPENAL port on Android to generate stereo 3D audio from mono sources. The OPENAL uses Head-related transfer function so that the user can perceive full 3D direction and approximate his/her distance of the landmark from 3D stereo audio played back from a headphone. The user can select a particular landmark by turning to it, and receive further route information from separate navigation applications.

III. EVALUATION

To evaluate the performance of the vision based localization and orientation estimation, we conducted experiments in a large open area in the lobby of Umass campus center which spans 150 feet by 50 feet. We took 266 photos on weekends when there are less people in the area for 3D reconstruction, while we test the system on a busy weekday with 91 photos using the back camera of Samsung Nexus Android phone with 35mm equivalent focal length. We used more photos for reconstruction to create enough overlapped areas between the photos. 1280x960 photos are used for reconstruction, while 640x480 photos are used for localization. Some photos of the landmarks in the calibration set and test set are shown in Figure 1. The top row includes photos from the calibration set, and the bottom row displays the test set with landmark positions estimated by the system superimposed on the photo. Landmarks are marked with red circles, with diameter indicating distance, the larger the circle, the closer the distance. Clock position and distance relative to the landmarks are also listed numerically. As we can see, the testing scenario is very challenging as the environment has changed a lot. The area is quite crowded with people, which occlude significant areas of the test photos. There are cases in which the software recognizes that it can not identify the landmarks (i.e., not enough features in the photo), in which case the photos are rejected and the user is asked to take additional photos. In cases the software finds enough features

in a photo, the photo is accepted. The metric of interest is the ratio between the number of accepted photos in which the landmark direction relative to the user was correctly identified and the total number of accepted photos.

Of 91 test photos taken, the system rejected 46 photos due to lack of visual features which can occur when the camera is pointing close up to a featureless area, or when there is too much motion blur. In cases where the photos are rejected, the user is informed and asked to take additional photos. For the 45 photos that were accepted we evaluate the angular accuracy of the relative landmark positions estimated by the system. All landmarks further than 10 feet away from the user within 39 of the 45 photos are estimated within ±15 degrees accuracy. Location estimates are completely off in the remaining 6 photos. To further improve the reliability of the system, multiple photos toward different directions can be taken in the same location. The localization took 3-10 seconds on one photo depending on the number of features extracted. The majority of time is spent in feature extraction and feature matching.

IV. Conclusion

We introduced the Audible Vision system that can help blind and visually impaired users navigate in large indoor open spaces. The system uses computer vision to estimate the location and orientation of the user, and enables the user to perceive his/her relative position to a landmark through 3D audio. Preliminary testing results show that for landmarks located at least 10 feet away from the user in crowded spaces, the landmarks are estimated within 15 degrees accuracy in 86% of the cases.

REFERENCE

[1] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *Br. J. Ophthalmol.,* 2011.

[2] W. Balachandran et al. ,"A GPS based navigation aid for the blind," in *Applied Electromagnetics and Communications, 2003. ICECom 2003. 17th International Conference on,* 2003, pp. 34-36.

[3] J. Sayah et al., "Localization and guidance in RAMPE/INFOMOVILLE-an interactive system of assistance for blind travelers," in *Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the,* 2009, pp. 243-249.

[4] S. Chumkamon, et al., "A blind navigation system using RFID for indoor environments," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on,* 2008, pp. 765-768.

[5] A. Ganz, et al., "PERCEPT: Indoor navigation for the blind and visually impaired," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE,* 2011, pp. 856-859.

[6] J. Coughlan and R. Manduchi, "A Mobile Phone Wayfinding System for Visually Impaired Users," *Assistive Technology Research Series,* vol. 25, pp. 849, 2009.

[7] C. Arth et al., "Wide area localization on mobile phones," in *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on,* 2009, pp. 73-82.

[8] J. Sánchez and M. Lumbreras, "Virtual environment interaction through 3D audio by blind children," *CyberPsychology & Behavior,* vol. 2, pp. 101-111, 1999.

[9] N. Snavely, S. M. Seitz and R. Szeliski, "Modeling the world from internet photo collections," *International Journal of Computer Vision,* vol. 80, pp. 189-210, 2008.

[10] H. Bay et al. , "Surf: Speeded up robust features," *Computer Vision–ECCV 2006,* pp. 404-417, 2006.

[11] R. M. Haralick, D. Lee, K. Ottenburg and M. Nolle, "Analysis and solutions of the three point perspective pose estimation problem," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on,* 1991, pp. 592-598.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," pp. 726-740, 1987.

[13] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 603-619, 2002.

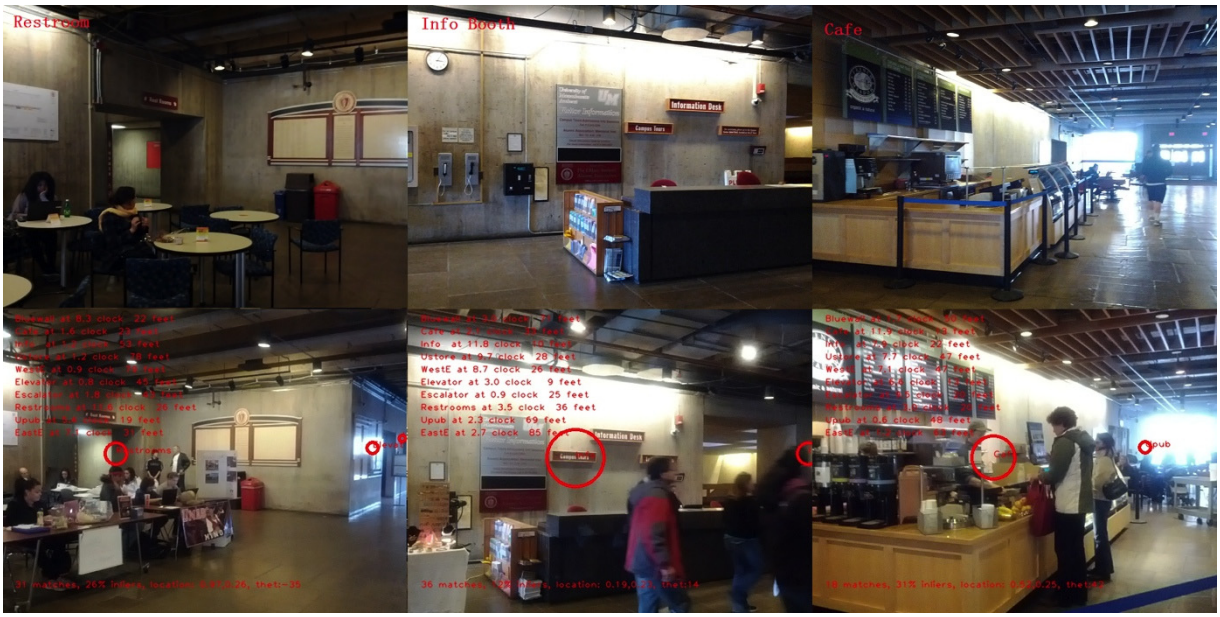

Figure 1. Calibration (top row) and test (bottom row) photos of landmark