Preprint: to appear in ICCHP 2018

Indoor Localization using Computer Vision and Visual-Inertial Odometry

Giovanni Fusco and James M. Coughlan

Smith-Kettlewell Eye Research Institute, San Francisco CA 94115, USA
{giofusco, coughlan}@ski.org

Abstract. Indoor wayfinding is a major challenge for people with visual impairments, who are often unable to see visual cues such as informational signs, landmarks and structural features that people with normal vision rely on for wayfinding. We describe a novel indoor localization approach to facilitate wayfinding that uses a smartphone to combine computer vision and a dead reckoning technique known as visual-inertial odometry (VIO). The approach uses sign recognition to estimate the user's location on the map whenever a known sign is recognized, and VIO to track the user's movements when no sign is visible. The advantages of our approach are (a) that it runs on a standard smartphone and requires no new physical infrastructure, just a digital 2D map of the indoor environment that includes the locations of signs in it; and (b) it allows the user to walk freely without having to actively search for signs with the smartphone (which is challenging for people with severe visual impairments). We report a formative study with four blind users demonstrating the feasibility of the approach and suggesting areas for future improvement.

Keywords: Wayfinding, Indoor Navigation, Localization, Blindness and Visual Impairment.

1 State of the Art and Related Technology

The key to any wayfinding aid is *localization* – a means of estimating and tracking a person's location as they travel in an environment. The most widespread localization approach is GPS, which enables a variety of wayfinding tools such as Google Maps, Seeing Eye GPS[1] and BlindSquare, but it is only accurate outdoors. There are a range of indoor localization approaches, including Bluetooth beacons [1], Wi-Fi triangulation[2] , infrared light beacons [2] and RFIDs [3]. However, all of these approaches incur the cost of installing and maintaining physical infrastructure, or of updating the system as the existing infrastructure changes (e.g., whenever Wi-Fi access points change).

[1] https://itunes.apple.com/us/app/seeing-eye-gps/id668624446?mt=8

[2] https://techcrunch.com/2017/12/14/apple-maps-gets-indoor-mapping-for-more-than-30-airports/

Dead reckoning approaches such as step counting using inertial navigation [4] can estimate relative movements without any physical infrastructure, but this tracking estimate drifts over time unless it is augmented by absolute location estimates.

Computer vision is a promising localization approach, but most past work in this area has either required special hardware [5] or the use of detailed 3D models of the environment [6] that are time-consuming to generate and make the approach vulnerable to superficial environmental changes (e.g., new carpeting, moved tables and chairs). To overcome this limitation in past work we developed an indoor localization system [7] that combines step counting with computer vision-based sign recognition. However, while the step counting algorithm works well for participants who walk with a regular gait, it is unreliable when the gait becomes irregular and halting – which is not unusual when visually impaired people explore unfamiliar surroundings. Accordingly, in our new approach we replaced step counting with visual-inertial odometry (VIO) [8], which functions well even if the user is walking with an irregular gait.

2 Overall Approach

Our localization approach uses a smartphone, held by the user or worn on the chest, to acquire several video frames per second. A 2D digital map of the indoor environment specifies the location, orientation (i.e., flush against the wall or perpendicular to it) and appearance of each sign in the environment. In each video frame, whenever a sign is recognized, the apparent size and perspective of the sign in the image determine the camera's pose relative to the sign and therefore the user's approximate location on the map (Fig. 1a). Our current prototype recognizes only barcode-like signs [9] printed on paper (Fig. 1b), but in the future we will extend the approach to recognize arbitrary flat signs, such as the standard Exit signs that our previous system [7] recognized.

In our experience with blind users exploring an environment, signs are only visible and recognizable in a small fraction of video frames, so it is crucial that we estimate the user's movements in between sign recognitions. We accomplish this using VIO [8], which combines the tracking of visual features in the environment with data from the smartphone's inertial measurement unit (IMU) to estimate changes in the smartphone's (X,Y,Z) location and 3D orientation (roll, pitch and yaw) over time. These movement estimates are projected to the 2D map domain, so that the user's trajectory on the map is estimated continuously over time.

We have implemented a prototype system on the iPhone 8 smartphone, using OpenCV[3] to perform sign recognition and pose estimation, and Apple's ARKit[4] iOS software to perform VIO. (ARKit is compatible with the iPhone 6s and newer iPhone models; a similar tool, ARCore, is available to perform VIO on newer Android devices.) Currently the system runs as a logging app that captures video in real time, and saves the video and VIO data to the smartphone's memory, which is analyzed on a computer offline. To maximize the chances of capturing images of signs, real-time audio feedback alerts the user whenever the line of sight deviates too far above or below

[3] https://opencv.org/

[4] https://developer.apple.com/arkit/

the horizon. In the future we will implement the system as a standalone wayfinding app that issues directional real-time audio feedback, including turn-by-turn directions and identification of nearby points of interest.

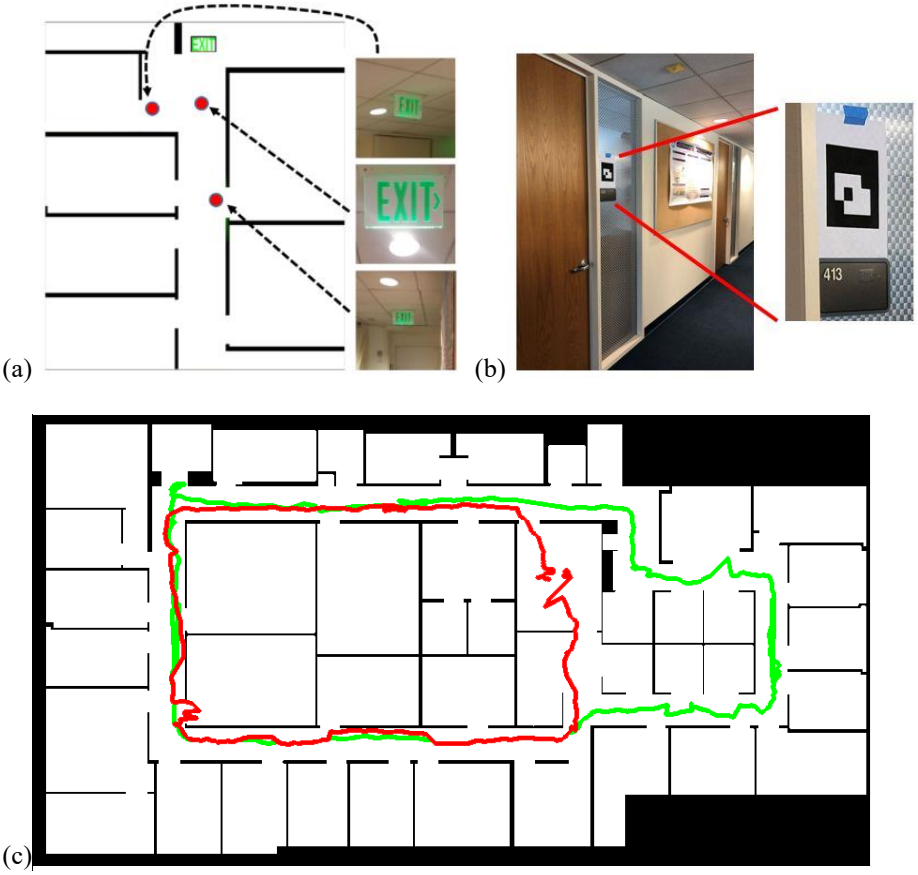

Fig. 1. Overall approach. (a) Signs provide absolute location estimates. Images of an Exit sign (cropped for visibility) taken from three locations are shown on the right, with dashed arrows to corresponding red dots showing approximate location of camera inferred from each image. (b) Prototype system tested in this work uses barcode-like marker patterns, shown here, affixed to the corridor walls instead of Exit signs; we will use Exit signs and other existing signs in place of markers in the future. (c) Map of indoor environment used in our experiments shows sample trajectories estimated along routes R1 (green) and R2 (red).

3 Evaluating Localization Accuracy

Both the sign recognition and VIO algorithms contribute significant noise to the location estimates, so we devised a simple procedure to evaluate the localization accuracy. The procedure establishes an approximate ground truth location at selected reference points in the path taken by the user. This is accomplished by having an experimenter

follow a few meters behind the traveler and take a video of his/her footsteps. This video is reviewed offline, with the experimenter noting each frame when the traveler passes by a reference point (such as a door) and using visual context (e.g., nearby doors and other features) to determine the corresponding ground truth location on the map. The ground truth location for each reference point is then entered by clicking on the corresponding location on the map. We estimate that the ground truth location is accurate to about 1 m, which implies that errors can be estimated to approximately 1 m accuracy.

In our experiments we used about 15-25 selected reference points to evaluate each route, depending on the length of the route and the visibility of landmarks. The data logged by the traveler's app is time-synched with the video so that any reference point in the video is associated with the corresponding logging data. In this way, the ground truth location of each reference point may be directly compared with the corresponding location estimated from the logging data.

Finally, we estimated the participant's average walking speed in each trial simply by dividing the route length by the time it took to traverse the route. We note that in some trials in which the participant used a guide dog, the dog paused at certain points along the path – thus in these cases the average walking speed estimate is an underestimate of the typical walking speed.

4 Formative Study

We conducted a formative study that took place on one floor of the main Smith-Kettlewell office building, with dimensions 39 m x 21 m (Fig. 1c). A total of 41 barcode signs (Fig. 1b) were posted on the walls. In all the experimental trials the experimenter followed the participant from a few meters behind, recorded video of the participant's movements, and issued verbal directions (e.g., "turn left in 5 feet") to guide the participant in real time along the desired route; he also acted as a safety monitor, ready to intervene if necessary to avoid a collision or tripping hazard. Each route (Fig. 1c) was a closed circuit, R1 (77 m long) and R2 (58 m long). Participants were instructed to walk with whatever travel aid they wished to use (white cane or guide dog), and in each trial they either held the smartphone in one hand or else attached it to the strap of a satchel (small shoulder bag) that they wore. Before participants used the system, we explained how it worked, including the audio feedback to help them maintain a horizontal line of sight and the need to avoid moving the camera quickly (which causes motion blur and thereby degrades the image quality) or covering the camera lens with the hands.

The study began with a brief evaluation of the system by one blind participant (P0), which allowed us to test the system, our training procedure and experimental protocol and then refine it before testing with three additional participants (P1-P3). In all, the four participants included two females and two males, with ages ranging from 24-40 years; two of them had no form or light perception, and the other two had very limited form and light perception (e.g., allowing them to perceive large objects and shadows).

Participant P0 used a white cane, but was unable to handle the cane while holding the phone, so we only tested him using the satchel. He completed route R1 in both directions (clockwise and counterclockwise), for a total of two trials.

Next, three blind participants participated in the evaluation phase of the study. Each participant completed eight trials: both routes R1 and R2, traversed in both directions, and in the handheld and satchel conditions. Participant P1 used a white cane, while participants P2 and P3 used a guide dog. The performance evaluation for all participants P1-P3 includes the median localization error, maximum localization error and average walking speed for each participant under each combination of route (R1 and R2) and carrying condition (hand and satchel), aggregated over both walking directions (i.e., two consecutive trials). The results are tabulated in Table 1 and shown graphically in Fig. 2. The system's performance was satisfactory for P1 and P2, with median errors around 1 m, but very poor for P3, with errors from a few meters to tens of meters. We note that reference points for which no localization was available were not included in the performance evaluations; such missing points were rare for P1 and P2 but very common for P3.

Table 1. Performance evaluation showing median localization error, maximum localization error and average walking speed for each participant and condition. Errors are in meters, and speeds in meters/second. (Error accuracy is limited to approximately 1 m, the accuracy of the ground truth locations.)

Partic.	R1 satchel	R1 hand	R2 hand	R2 satchel
P1	0.57, 18.34, 0.63	0.49, 1.42, 0.83	0.56, 2.75, 0.86	0.45, 4.01, 0.83
P2	1.22, 3.79, 0.78	0.93, 3.96, 0.91	0.98, 2.86, 0.68	0.87, 3.33, 0.83
P3	24.52, 52.33, 1.05	2.51, 12.11, 1.14	2.63, 4.60, 1.04	1.37, 5.37, 1.16

The poor results for P3 resulted from her very rapid walking speed, which caused motion blur and also severely limited the number of frames in which signs were recognized. Since each sign is visible from only a limited region defined by a certain range of distances and viewing angles, a faster speed implies less time spent traversing this region, and thus fewer chances to glimpse the sign; this resulted in very few sign recognitions in each trial for P3 (in fact, none in one trial). In the future we will explore ways of improving the sign recognition to accommodate faster walking speeds. We also note that we will use signs such as Exit signs in the future, which are visible from a wider range of locations than the signs we used.

All participants expressed interest in an effective indoor wayfinding system, and they commented on possible ways of holding or wearing the smartphone for our approach. P0 suggested many possible ways of wearing the smartphone, including attaching it to a broach or tie clip, or wearing an eyeglass-mounted camera; P1 suggested using a holster to clip the smartphone to the satchel strap, and pointed out the possibility that holding the smartphone while walking might invite theft; P2 felt it should be straightforward to attach the smartphone to the strap of the purse she usually carries with her, but was also pleasantly surprised that it was easy to hold the smartphone while walking; and P3 explained that she would choose to hold the smartphone while walking, since she is already in the habit of doing so (e.g., to text message or use Google

Maps). These comments suggest that different users will seek various ways of holding or wearing the smartphone, depending on individual preferences and circumstances. Thus in the future we will explore multiple options to meet diverse needs, including the option of using an external wearable camera, rather than attempting to find a single approach for all users.

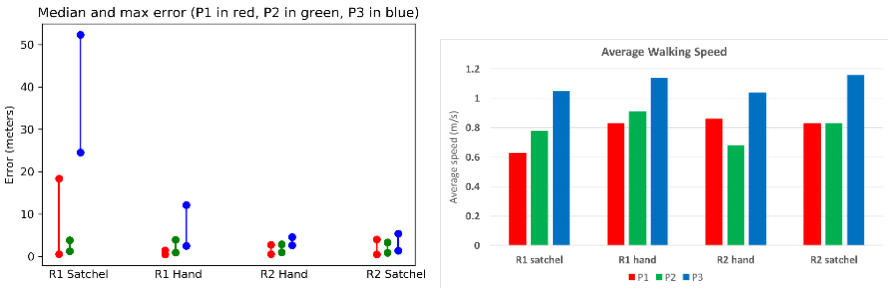

Fig. 2. Graphs of performance evaluation data from Table 1. Data from participants P1, P2 and P3 are shown in red, green and blue, respectively. Left: median and maximum error (meters), indicated by data points connected by vertical lines. Median errors are mostly under 1.5 meters, with the exception of P3, whose rapid walking speed led to severe localization errors (see text). Right: average walking speed (meters/second).

5 Conclusions and Future Work

We have demonstrated the feasibility of a novel indoor localization approach that is usable by visually impaired travelers under a range of normal operating conditions. The approach has the advantage of requiring no added physical infrastructure, and exploits the localization information conveyed by standard informational signs without forcing a visually impaired user to explicitly search for them. We also stress that the approach is applicable to people with normal or impaired vision, who may use a cane or guide dog, and it should also function for people traveling in a wheelchair.

Future work will proceed in several directions. The first will entail enlarging the range of conditions under which our approach functions, such as a rapid walking pace, which may require improved sign recognition and/or a faster frame rate. We will extend the sign recognition algorithms to accommodate arbitrary planar signs, and will focus on maximizing their ability to recognize signs from a wide range of viewing angles and distances.

In addition, we will incorporate more information in the localization algorithm, including (a) sign visibility, enforcing the fact that a sign cannot be seen through an intervening wall or other opaque barrier, and (b) traversability constraints, reflecting the impossibility of walking through walls. A natural way to incorporate this additional information is to maintain and evolve a distribution of multiple location hypotheses over time using a framework such as particle filtering [10], which is standard in robotics applications. The particle filtering framework will be useful for explicitly representing uncertainty about the user's location, which may result from detections of signs at long

distances (camera pose estimation is noisy when the sign appears small in the image), ambiguity about which sign has been detected (e.g., multiple Exit signs in a building may have identical appearance), or when no sign has been recognized for some length of time. Another source of absolute localization information that we will incorporate, when available, is the signals from Wi-Fi and from Bluetooth beacons.

This paper has focused on our development of a localization algorithm, but a considerable amount of work will also be required to transform this into a full wayfinding system that runs as a standalone app, with a fully accessible user interface that provides turn-by-turn directions and/or information about nearby points of interest. We will develop a streamlined approach for creating digital maps (which in some cases will have to be scanned in from paper maps) and annotating them with sign information, including photos of non-standard signs to train the sign recognition algorithms. While we envision that the mapping process will be accomplished primarily by sighted assistants (perhaps using crowdsourcing), we will explore the possibility of making some annotation features accessible to visually impaired users, for example, enabling them to add their own points of interest to an existing map. Finally, ongoing testing of the system with both blind and low vision users will guide the entire development process to ensure that the wayfinding system is maximally effective and easy to use.

6 Acknowledgments

The authors were supported by NIDILRR grant 90RE5024-01-00.

References

1. D. Ahmetovic, C. Gleason, K. Kitani, H. Takagi & C. Asakawa. (2016). NavCog: turn-by-turn smartphone navigation assistant for people with visual impairments or blindness. Web for All Conference. ACM.
2. L.A. Brabyn & J.A. Brabyn, An Evaluation of 'Talking Signs' for the Blind. Human Factors, vol. 25, no. 1, pp. 49–53, Feb. 1983.
3. A. Ganz, S. R. Gandhi, C. Wilson & G. Mullett. (2010). INSIGHT: RFID and Bluetooth enabled automated space for the blind and visually impaired. In 2010 Annual Int'l Conf. of the IEEE Engineering in Medicine and Biology.
4. G. Flores & R. Manduchi. (2018). Easy Return: An App for Indoor Backtracking Assistance. CHI 2018.
5. F. Hu, Z. Zhu, and J. Zhang, (2014). Mobile Panoramic Vision for Assisting the Blind via Indexing and Localization. Second Workshop on Assistive Computer Vision and Robotics, in conjunction with ECCV 2014.
6. C. Gleason, A. Guo, G. Laput, K. Kitani & J.P. Bigham. (2016). VizMap: Accessible visual information through crowdsourced map reconstruction. ASSETS 2016.
7. A. Rituerto, G. Fusco & J.M. Coughlan. (2016). Towards a sign-based indoor navigation system for people with visual impairments. ASSETS 2016.
8. J. Kelly & G.S. Sukhatme. (2011). Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. The International Journal of Robotics Research, 30(1), 56-79.

9. S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, & M.J. Marín-Jiménez. 2014. "Automatic generation and detection of highly reliable fiducial markers under occlusion". Pattern Recogn. 47, 6 (June 2014), 2280-2292. DOI=10.1016/j.patcog.2014.01.005
10. S. Thrun, W. Burgard & D. Fox. (2005). Probabilistic Robotics. MIT Press, 2005.