VizMap: Accessible Visual Information Through Crowdsourced Map Reconstruction

Cole Gleason, Anhong Guo, Gierad Laput, Kris Kitani, Jeffrey P. Bigham
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
{ cgleason, anhongg, gierad.laput, kkitani, jbigham }@cs.cmu.edu

ABSTRACT

When navigating indoors, blind people are often unaware of key visual information, such as posters, signs, and exit doors. Our *VizMap* system uses computer vision and crowdsourcing to collect this information and make it available non-visually. VizMap starts with videos taken by on-site sighted volunteers and uses these to create a 3D spatial model. These video frames are semantically labeled by remote crowd workers with key visual information. These semantic labels are located within and embedded into the reconstructed 3D model, forming a query-able spatial representation of the environment. VizMap can then localize the user with a photo from their smartphone, and enable them to explore the visual elements that are nearby. We explore a range of example applications enabled by our reconstructed spatial representation. With VizMap, we move towards integrating the strengths of the end user, on-site crowd, online crowd, and computer vision to solve a long-standing challenge in indoor blind exploration.

Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User Interfaces - *Input devices and strategies*; K.4.2 [**Computers and Society**]: Social Issues - *Assistive technologies*

Keywords

Blind users; accessibility; crowdsourcing; indoor navigation.

1. INTRODUCTION

Exploring unfamiliar indoor environments can be difficult for blind people. Many use obstacle-avoidance measures, such as navigating with a cane or guide dogs, to make their way through indoor spaces, but these aids do not provide access to critical navigation cues such as signs or the location of doors. Likewise, relevant points of interest (POIs) pertinent to a location (*e.g.*, bathroom signs, trash cans, posters) are equally inaccessible. Sighted users take these visual cues for granted, but most objects lack salient non-visual hints to their presence, making even everyday environments difficult for blind users to access effectively.

In outdoor environments, GPS applications for smartphones have made blind navigation and exploration much easier, as the application can detect the user's location within a few meters and provide information about nearby landmarks. For instance, the popular BlindSquare[1] application describes outdoor points of interest (*e.g.,* businesses and streets) to blind users based on their current location. Indoor localization systems have not yet achieved that level of utility for blind users due to limited maps of points of interest. Indoor environments also suffer from poor GPS signal reception, so researchers have tried many other approaches to find a blind user's position inside a building. Becaon-based approaches use the signals of devices placed throughout the environment, but require instrumenting the building and maintaining the system. Google's Project Tango[2] enables indoor navigation with depth sensors in smartphones, but RGB-D cameras are not yet in widespread use. Impressive projects like Navatar solve this problem by reducing localization drift in dead-reckoning by having the user periodically confirm their location based on nearby landmarks [2].

Computer vision approaches such as Structure from Motion can be used to create models of environments and localize users in them [3], but often have no semantic understanding of the visual information. Prior research has also explored the utility of providing point of interest awareness [4], but annotating all indoor points may be time consuming. Projects like VizWiz proved that crowd workers can provide good semantic labels and human understanding for blind users [1] via a smartphone application. Our system fuses these two approaches to embed the crowd's labels of visual information into a model of the environment. We also explore how this information might be accessed using a smartphone.

2. VIZMAP

We introduce VizMap, a system that captures indoor visual points of interest and makes them available to blind people. VizMap uses OpenMVG's Structure from Motion pipeline[3] to automatically construct a model of an indoor environment, and embeds crowdsourced semantic annotations of visual points of interest. In general, on-site sighted volunteers with smartphones (or wearable cameras) act as distributed physical crawlers to collect video footage of a building's interior. These videos are used to build a 3D point cloud representation of the environment for localization and navigation. Once representative video frames are collected, an online annotation interface allows remote crowd workers to generate semantic labels for important visual cues in key video frames. These labels are then embedded in the underlying 3D point-cloud representation, which

ASSETS '16, October 24–26, 2016, Reno, Nevada, USA.

ACM ISBN 978-1-4503-4124-0/16/10..
DOI: http://dx.doi.org/10.1145/2982142.2982200

[1] http://blindsquare.com/

[2] https://get.google.com/tango/

[3] https://github.com/openMVG/

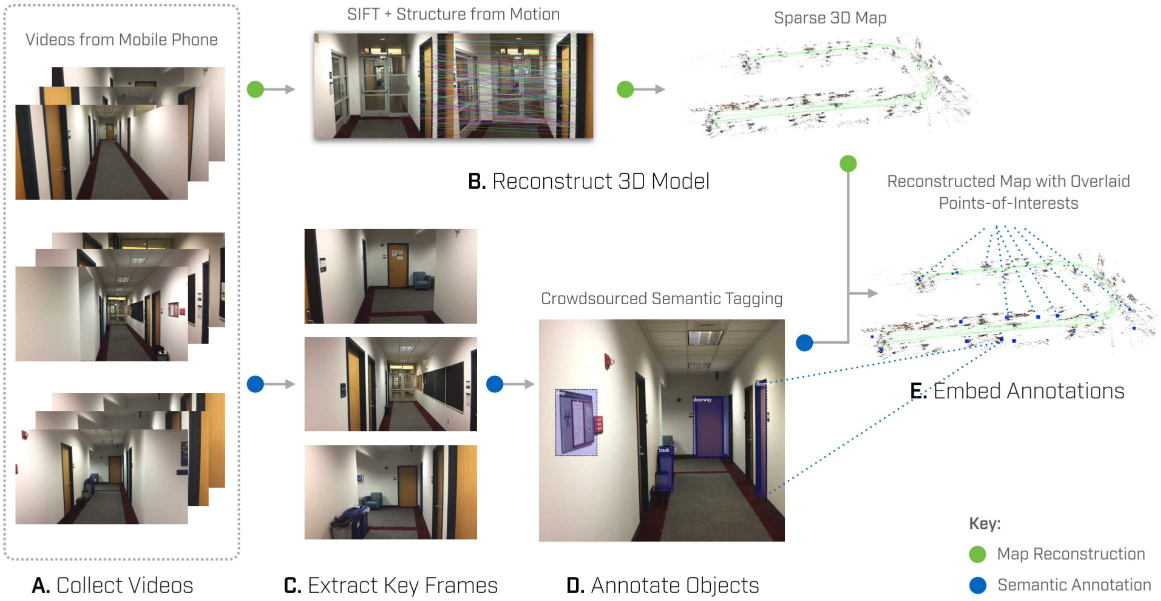

Figure 1: VizMap system infrastructure. VizMap collects videos from sighted volunteers (A) and constructs a sparse 3D model of the environment (B). At the same time, clear key frames are extracted (C) for the crowd to annotate points of interest (D). Finally, the crowd labels are embedded into the generated points cloud (E), shown here as blue squares.

a blind user with a smartphone can interact with by simply taking a photo. Using SIFT features from the captured photo, VizMap determines the user's indoor location and heading to sub-meter accuracy. With that information, it is simple to retrieve all labels in the user's vicinity (*e.g.*, 3 meters) and display them relative to their current direction.

VizMap leverages the strengths of the on-site crowd to provide access to the environment; the online crowd to offer always available sight and general intelligence; computer vision to deliver automation, speed, and scalability; and the end user to build a mental model of the environment based on the embedded points of interest. Our approach also takes advantage of the ubiquitously available smartphones instead of instrumenting the environment.

3. EXAMPLE APPLICATIONS

VizMap produces a point-cloud representation of an indoor environment with embedded semantic labels. To explore interaction techniques with this model, we designed and built three prototype applications using the VizMap infrastructure on an iPhone, taking advantage of the built-in accessibility and screen reader features (*e.g.*, VoiceOver). Whenever the blind user takes a photo in the mapped environment, the image is sent to a server which performs real-time localization and finds the user's orientation.

- **Nearby POIs:** The server finds all POIs within a predefined radius of the user. The orientations for the retrieved POIs are then computed relative to the user and read aloud in a clockwise fashion via VoiceOver (*e.g.*, "3 o'clock: door").
- **Fine-grained interrogation:** VizMap determines all POIs directly within the user's field of view and allows the user to interact with these POIs by tapping buttons on screen, using an interface similar to RegionSpeak [5].
- **Dynamic messages:** In a third prototype, a sighted or blind user can take a photo and embed a message into the environment in order to mark dynamic information like ongoing construction or social events.

A blind user evaluated these three applications (the second in a Wizard of Oz fashion) in a campus corridor, but she had difficulty taking a good photo, as many of the bare walls lacked enough feature points to localize. Future work will allow the user to continuously pan the camera until localization succeeds, removing the need to aim. Dead-reckoning and other approaches will provide smooth localizations between these successful frames. Alternatively, the annotated 3D map could use different indoor localization methods, such as beacons, in supported buildings.

4. REFERENCES

[1] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 2010.

[2] N. Fallah, I. Apostolopoulos, K. Bekris, and E. Folmer. The user as a sensor: navigating users with visual impairments in indoor spaces using tactile landmarks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012.

[3] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[4] R. Yang, S. Park, S. R. Mishra, Z. Hong, C. Newsom, H. Joo, E. Hofer, and M. W. Newman. Supporting spatial awareness and independent wayfinding for pedestrians with visual impairments. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2011.

[5] Y. Zhong, W. S. Lasecki, E. Brady, and J. P. Bigham. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.