Building Smart Transportation Hubs with 3D Vision and Video Technologies to Improve Services to People with Disabilities

Jie Gong*, Ph.D., Assistant Professor
Department of Civil & Environmental Engineering
Rutgers, The State University of New Jersey
jg931@rci.rutgers.edu

Cecilia Feeley, Ph.D., Transportation Autism Project Manager
Center for Advanced Infrastructure and Transportation
Rutgers, The State University of New Jersey
cfeeley@rci.rutgers.edu

Hao Tang, Ph.D., Assistant Professor
Department of Computer Science
City University of New York
htang@bmcc.cuny.edu

Greg Olmschenk, Ph.D. student
City University of New York Graduate Center
golmschenk@gradcenter.cuny.edu

Vishnu Nair, Undergraduate Student
City University of New York
vnair000@citymail.cuny.edu

Zixiang Zhou, Ph.D. Student
Department of Civil & Environmental Engineering
Rutgers, The State University of New Jersey
zx_zhou@hotmail.com

Yi Yu, Ph.D. Student
Department of Civil & Environmental Engineering
Rutgers, The State University of New Jersey
yi.yu.civil@rutgers.edu

Ken Yamamoto, Graduate Student
City University of New York
itskenichi@gmail.com

Zhigang Zhu, Ph.D.
Herbert G. Kayser Professor of Computer Science
City University of New York
zhu@cs.ccny.cuny.edu

* Corresponding Author

Word count: (abstract: 228, text: 4850, references: 670) words text + 7 tables/figures x 250 words (each) = 7498 words.

Submission date: 2017/07/30 Revised date: 11/14/2016

ABSTRACT

Large transportation hubs are difficult to navigate, especially for people with disabilities such as those with visual or mobility impairment, Autism Spectrum Disorder (ASD), or simply those with navigation challenges. The primary objective of this research is to design and develop a novel cyber-physical infrastructure that can effectively and efficiently transform existing transportation hubs into smart facilities capable of providing better location-aware services (e.g. finding terminals, improving travel experience, obtaining security alerts). We investigated the integration of a number of novel Internet of Things elements, including video analytics, low-cost Bluetooth beacons, mobile computing, and LiDAR-scanned 3D semantic models, to provide reliable indoor navigation services to people with traveling challenges, yet requiring minimum infrastructure changes since our approach leverages existing cyberinfrastructures such as surveillance cameras, facility models, and mobile phones, and incorporates a minimum number of new and small devices such as beacons to achieve reliable navigation services. We choose two groups of people for our initial study– those with visual impairment and ASD since both groups face difficulties in a crowded and complex 3D environment. Thus two unique features of our solution are the use of 3D digital semantic models and crowd analysis with surveillance cameras for providing the best available paths. We have started a pilot test with people with disabilities at a multi-floor building in New York City to demonstrate the effectiveness of our proposed framework.

Glossary of Terms:
ASD: Autism Spectrum Disorder
BLE: Bluetooth Low Energy
BoW: Bag of Words
BVI: Blind and Visual Impairment
CNN: Convolutional Neural Network
ConvNet: Convolutional Neural Network
DCT: Discrete Cosine Transform
GIST: a low dimensional representation of the scene, which does not require any form of segmentation
IoT: Internet of things
Lidar: Light Detection and Ranging
SfM: Structure from Motion

INTRODUCTION

Transitional spaces such as bus terminals, train stations, airports, and multi-modal transportation hubs have become an increasingly important part of city's infrastructure as we are spending more and more of our lives in these spaces in today's ever connected world. Transportation facility owners are facing growing challenges to accommodate the rising public travel demands while improving quality of service. Future transportation facilities need to be smart, providing efficient, high-quality, and equitable services to the increasingly diverse population. This is especially true for those gigantic transportation hubs because wayfinding in these facilities has always been challenges for people with disabilities, such as individuals with visual impairment and Autism Spectrum Disorder (ASD) and people with difficulties in finding places, particularly persons unfamiliar with metropolitan areas.

In the United States alone, the visually impaired population has reached 6.6 million people and expected to double by 2030 (from 2010 figures) [1]. According to Centers for Disease Control and Prevention (CDC), ASD is the fastest-growing developmental disorder affecting 1 in every 68 people in the US. One common and recurring obstacle that people from both groups face every day is navigation, particularly as related to mobility. Using public transportation services is the best way for them to travel. However, there are also significant hurdles in using them due to their challenges. In 2015, a study conducted at Rutgers University found that according to adult respondents on the spectrum and their family members, 35.1% of these adults with ASD have difficulty in determining directions/route [2]. In this work we will focus the navigation need inside a transportation plaza. For such an environment, finding a traversable path alone cannot solve the problems for these two groups. We will need to consider two challenges these two groups of people in a crowded and complex 3D environment. In a crowded transportation plaza, individuals with ASD will get very nervous when facing too many people, where visually impaired individuals will face great difficulty in avoiding to bump into the crowd. An understanding of the 3D structure of the facility will not only provide the best traversable paths for the users, but also provide semantic information for travelers to avoid certain areas that could be crowded, noisy or dangerous.

Table 1. Difficulty of People with ASD with Different Aspects of Walking

Difficult Aspects of Walking	Responses	Percent of Responses	Percent of Respondents
Difficulty determining directions/route	247	14.2	35.1
Crossing a street	290	16.7	41.3
Judging the distance and/or speed of cars	318	18.3	45.2
Walking in areas without sidewalks (on grass or in streets)	193	11.1	27.5
Dealing with distractions while walking	282	16.2	40.1
Too many people on the sidewalk	64	3.7	9.1
Too many cars or too much traffic	257	14.8	36.6
Other, please specify:	86	5.0	12.2
Total	1737	100.0	NA

While smarter transportation hubs can be built from ground-up by harnessing the latest technology development, retrofitting existing facilities so that to make them smarter may be a much more cost effective choice in many highly developed urban settings. Current emerging mobile computing and Internet of the Things technologies, together with advances in computer vision techniques using in 3D localization and crowd analysis, will provide great opportunities in significantly improving navigation services as well as creating innovative approaches to accommodate passengers and customers. This can be achieved by automatically assisting them to enhance their ability to navigate the complicated plaza with minimum infrastructure changes therefore minimizing cost. In busy transportation plazas such as

NYC Port Authority Bus Terminal, two important features are usually available: the digital 3D model of the facility which doesn't change much, and a huge array of surveillance cameras. In many places, low-cost BLE beacons have also deployed which can help simplify the localization of users. For our BVI and ASD users, the 3D models and the surveillance video will provide the needed 3D and dynamic information. Our solution will focus on the use of the two sources of data and develop (1) semantic modeling and matching technologies, and (2) crowd analysis algorithms together with our available technologies, in order to provide our users personalized traveling guidance inside a transportation plaza. While the support for indoor navigation are evident, many of them are just indoor positioning; few studies have utilized the integration of 3D semantic modeling and crowd analysis with localization functions for provide more effective services to the level of capacities of travelers with normal perception and cognition can reach on their own. Most studies have focused on individual technological solutions such as sign recognition or user location determination, which tend to fail to deliver reliable services in large and complex transportation hubs. On the other hand, travelers with disabilities often have lower tolerance to failure than normal travelers do. For example, blind people often have little tolerance for failure perhaps because of the large cognitive load of travelling while blind and the emotional overlay of blindness. This paper describes and presents preliminary results on a novel cyber-physical infrastructure framework that can effectively and efficiently transform existing transportation hubs into smart facilities that are capable of providing better location- and situation- awareness and personalized navigation services (e.g. choosing the 3D routes, finding terminals, improving travel experience, obtaining security alerts) to the traveling public, especially for the underserved populations including those with visual impairment, ASD, or simply navigation challenges.

RELATED WORK

People who are have normal vision rely almost exclusively on their sight to orient themselves in a new indoor environment, and choose paths that they feel the most comfortable in a crowded and complex 3D transportation plaza, such as using escalators and avoid heavy crowd, or using elevators and go with a crowd. As for people with visual impairment, eyesight is not a useable or reliable perception means, and they need to use alternative sensory tools to collect information to explore the environment. In spite of this need, the majority of the tools available to this population of people are not able to tell them their locations accurately, not even for navigation. For example, a white cane can help them to determine whether an area is walkable or not, but it cannot provide users their location information. Guide dogs may help to lead users to walk along known paths, but users still need other information to reason their locations when they want to change their routes, let alone to say owning a guide dog is expensive. GPS is used for localization in outdoor environments, but GPS signals can rarely be detected indoors or in dense urban areas because GPS signals are weakened and scattered by walls, roofs, and other obstructions [3].

Similarly, ASD individuals welcome technological solutions in order to overcome many of their daily obstacles. Among those obstacles, one common and recurring obstacle is navigation, in particular in indoor settings. Outdoor areas have signs, maps, and GPS-based navigation systems that can help a person navigate to their destination, whereas indoor navigation is often proved to be a much more difficult task. Because of this, lack of adequate navigation capabilities has limited their opportunities to use public transportation services [2]. In many circumstances, ASD individuals may get lost or are unable to find their destinations in a complex building. In situations like these, not all ASD individuals are comfortable enough to seek help from strangers due to several reasons like communication difficulties, language problems, or social issues.

In recent years, researchers and several startups have been working on indoor and/or inner-city positioning systems where GPS signals are not reliable. These include WiFi- or Bluetooth-based navigation approaches, and a few public facilities even have tested such approaches (i.e. at SFO airport by Indo.or [4] and in Washington DC Gallery Place Metro Station by Click-and-Go Wayfinding [5]). Some have proposed localization using the magnetic field [6] while others have suggested using accelerometers and compasses on mobile devices in order to detect the speed and direction of the user [7]. However, these methods are very much prone to error and may not be supported by all devices. Recently, localization using Bluetooth Low Energy (BLE) beacons has emerged as a viable method of positioning considering its wide availability and low cost [8]. A joint CMU-IBM team developed NavCog [9], an iPhone App that navigates blind users to the destinations by giving out turn by turn navigation by using BLE based localization with 1.5 meters accuracy. However, a position system can only provide a very limited service to users with disabilities. The state-of-the-art beacon based methods assume simple indoor layout (i.e. office) that are rarely crowded and changed hence may not work well in public transportation centers. Therefore, the current practical uses of beacons are only for area indicators or non-contact information desks instead of localization. Needed are approaches that would require minimal infrastructure changes and sensor installations. Semantic facility model-based navigation could be a potential solution.

Another means to provide indoor localization and navigation services is computer vision based approaches. Previous work [10] explored methods to process images by image matching and estimate the location information. However, image matches are error-prone in the indoor and urban environments with large textureless areas. Some other studies have explored using Structure from Motion (SfM) to create street 3D models in the outdoor environment and recognizing the places utilizing images from Internet [11]. Some researchers use Bag of Words (BoW) [12] or ConvNet features [13] to represent outdoor environments for localization. Among these studies, very few of them focus on indoor scenarios, especially for an assistive localization purpose. In addition, a practical SfM model heavily relies on the richness and distinguishes of environmental features extracted from the images, which is hard to use in environments where few features are available and detected features often tend to be repetitive in space. And the computation for a full search of images in a large transportation facility is also very expensive and thus impractical. Alternatively, crowdsourcing based approaches have been proposed using real-time crowd-annotated maps by a CMU group [14] and using crowd-annotated video by us [15]. Crowdsourcing approaches could be a backup plan for our solution.

The rise of mobile and wearable devices as ubiquitous sensors has greatly accelerated the advancement of both general computer vision research and assistive applications. Farinella et al. [16] uses Android phones to implement an image classification system with DCT-GIST based scene context classifier. Some others apply Google Glass and develop an outdoor university campus tour guide application system by training and recognizing the images captured by Glass camera [17]. Paisios, a blind researcher, creates a smart phone app for the Wi-Fi based blind navigation system [18]. Manduchi proposes a sign-based way-finding system and tests the blind volunteers with smart phones to find and decode the information embedded in the color marks pasted on the indoor walls [19]. Coughlan's team has been worked on various vision solution for outdoor and indoor navigation for the blind [20-22]. However, in spite of the technology promise demonstrated in these studies, few research work exist on designing user-friendly smart phone apps for helping visually impaired people to not only localize themselves but also navigate through an indoor environment with crowd and 3D complexity.

PROPOSED APPROACH

To address the needs of reliable indoor navigation services in major transportation hubs, we propose to integrate video analytics, BLE beacons, mobile computing, and 3D semantic models into a cyber-physical system to provide reliable indoor navigation services to people with traveling challenges. The proposed cyber-physical system is designed to require minimum infrastructure changes as it leverages existing cyber-infrastructures such as surveillance cameras, facility models, and mobile phones, and incorporates a minimum number of beacons to achieve reliable navigation services. Figure 1 shows four essential elements in our proposed framework. In the following, we detail the technical innovations in each of these elements.

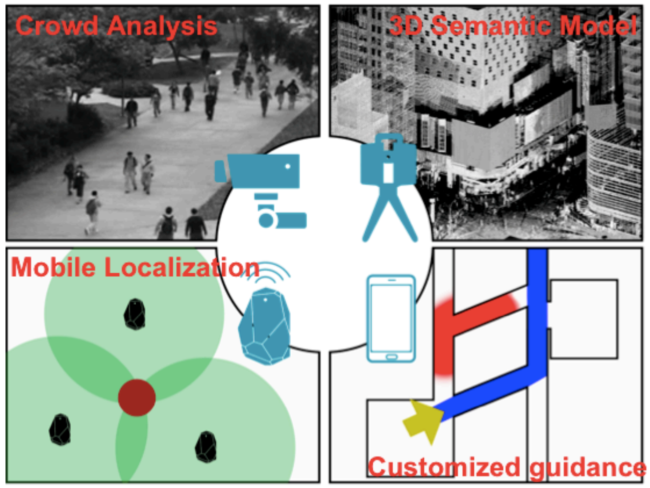


Figure 1. Proposed cyber-physical system to enable personalized indoor navigation assistance to people with special need

3D semantic facility model based localization

We use a 2D-3D registration approach to register smart phone images from multiple users with a pre-built semantic 3D facility model to infer absolute 3D locations of users. We would like to note that we will focus on large transportation plazas where such models have been available for other applications, such as transportation planning, crowd simulation and building renovation. In our proposed framework, we develop a base framework in which facility users and the 3D semantic model of a facility can collaboratively work to realize robust and real-time localizations. To facilitate the process, we create a database of feature distributions for key positions in the facility based on the point cloud data and semantic facility model. We discard or discount those features that are from facility elements that will likely change over the time, and boost those with more permanent installations. The key positions will be determined based on how distinctly features distribute at selected vantage views from candidate key positions. We further assert that passive image capture and registration approach may not be effective whereas providing some directions to the users will greatly accelerate the converging process during localization because of two reasons: (i) Poor coverage—data collected from people without directions will likely have poor coverage of scenes with informative features; (ii) Data quality—without directions, data gathered is uncoordinated, resulting in low quality with more noise, making it difficult to process it, e.g., capturing the scene under different angles/positions, abruptly shaking device during capture, etc. This leads to the need of beacon based localization.

Beacon-based indoor localization

When vision based localization fails, beacon-based indoor localization is the backstop to ensure the availability of adequate navigation services in our proposed framework. We want to emphasize that our approach can utilize any other emerging position solutions such as Wi-Fi-based, mm-wave-based or magnetic based. Moreover, apart from simply deploying a dense network of fixed Bluetooth beacons (or other sensors) with known locations, a unique feature of the proposed work is the utilization of the 3D semantic model for the beacon installation. Both installing and calibrating beacons are tedious and challenging. Therefore, the use of 3D model will make the installation and maintenance of both fixed and mobile beacons more effective. 3D locations of beacons can be planned either interactively or automatically in the 3D digital model for the best coverage, and visualized in a virtual reality display for each installation. When a user comes into the facility, his/her App will be able to detect at least three of the beacons with known locations to obtain a relatively accurate location (from a meter to several meters). Then the 2D-3D registration approach will be used to further refine and track the location of the user.

Video based crowd analysis

A unique component of our proposed approach is integrating crowd analysis into indoor navigation services. Traditional indoor navigation services rarely consider contextual information when providing navigation guidance. However, this could be an important issue for people with disabilities such as visual/mobility impairment or ASD. For example, ASD individuals may prefer to choose paths that have less dense crowds due to psychological factors; people with visual impairment try to avoid large open space due to difficulty to find references for localization; and people in wheel chairs can navigate along paths with less crowds far more conveniently than along those with large crowds. In our proposed framework, we analyze the video feeds in real-time from surveillance cameras in the facility to evaluate the density of crowds in different parts of the facility. The analysis results will be incorporated into path choices.

An adaptive context aware navigation guidance approach

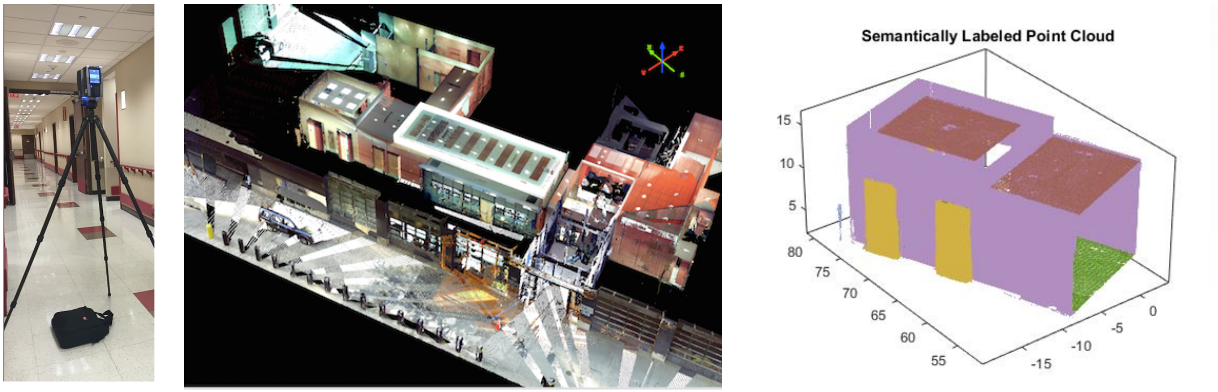
The proposed framework also includes a user-centric, activity-aware and feedback-enabled services with the support of the surveillance camera system to provide human crowd analysis results. In our framework, path planning for a user is made based on the following five factors: 1) Both the user's current location and his/her destination; 2) The user's planned schedule (for example the time to take a bus); 3) The special needs of the user; 4) The semantic 3D models with all the important facility labels; and 5) The crowd analysis results from the surveillance cameras. This is a graph planning problem with multiple cost attributes, and probably the graph and the path need to be updated if the path is not very short. As examples, a visually impaired or wheelchair user should avoid stairs. We will also need to adapt the path based on user's feedback. If an ASD user gets stuck and panics at a certain location, the App will need to re-route the path, probably will also need to put them to wait if certain areas that they have to pass are too crowded and their time still allow them to wait.

PRELIMINARY RESULTS

In order to test the proposed framework, we have started to deploy our technical components at a multi-floor facility in New York City. The facility has high definition surveillance cameras in place and it provides services to people with visual impairment. Estimote beacons [23] are installed in the facility during our study. Although the facility was not a true transportation hub, it provides a great opportunity to test and validate our proposed approach with the easy reach to one of the groups we would like to provide services. The pilot provides foundational knowledge to expand our approach to transit stations and transportation hubs. In the following, we describe the development and testing of the framework at this facility.

3D semantic model and image model registration

As the first step, we utilized a terrestrial laser scanner (Figure 2a) to create a high-fidelity 3D model of the facility; the scanning only took half a day for the building. It is useful to note that more and more facility owners elect to use laser scanners to develop high-fidelity facility models as the baseline data for facility management. It is also important to note that these high-fidelity 3D models can also be built using other spatial data acquisition and processing technologies such as RGBD cameras and structure from motion, although these models are less accurate and only useful for limited applications. In this study, the facility is represented with colorized 3D point cloud (Figure 2b) with dense annotation of building elements (Figure 2c). The creation of dense annotation is realized with a semi-automated segmentation and labeling tool developed as part of this project. Basically, the tool segments the point clouds through a region growing method [24] and the segmented point clouds are manually annotated.



(a) (b) (c)
Figure 2. (a) Facility modeling with terrestrial laser scanning; (b) Colorized point cloud data of the facility; (c) Point clouds with dense annotation of building elements (in this case, elevator doors)

In this component, we also investigated registration of mobile phone image of the user with the 3D semantic model to provide user more accurate location and orientation information to get to his/her desired location. The registration between mobile images and facility point cloud data is solved by determining the projection between corresponding pixels/points. Denote a point as $C = [X, Y,\ Z, 1]^T$, and a pixel as $c = [u, v, 1]^T$. The projection from a 3D point on to a 2D pixel could be expressed as:

$$c = \boldsymbol{A}[\boldsymbol{R}|\boldsymbol{t}]C \qquad (1)$$

Where A includes intrinsic camera parameters, R and T are extrinsic camera parameters, including rotation and translation of the camera, according to the reference coordinate.

$$A = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R = R_x(roll) \cdot R_y(pitch) \cdot R_z(yaw)$$
$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(roll) & \sin(roll) \\ 0 & -\sin(roll) & \cos(roll) \end{bmatrix} \begin{bmatrix} \cos(pitch) & 0 & -\sin(pitch) \\ 0 & 1 & 0 \\ \sin(pitch) & 0 & \cos(pitch) \end{bmatrix} \begin{bmatrix} \cos(yaw) & \sin(yaw) & 0 \\ -\sin(yaw) & \cos(yaw) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$t = [x, y, z]^T$$

Denote the projection matrix as $\boldsymbol{P} = \boldsymbol{A}[\boldsymbol{R}|\boldsymbol{t}]$, for each pair of points $C_i = [X_i, Y_i, Z_i, 1]^T$, and $c_i = [u_i, v_i, 1]^T$, equation (1) could be rewritten as:

$$\underbrace{\begin{bmatrix} X_i & Y_i & Z_i & 1 & 0 & 0 & 0 & 0 & u_iX_i & u_iY_i & u_iZ_i & u_i \\ 0 & 0 & 0 & 0 & X_i & Y_i & Z_i & 1 & v_iX_i & v_iY_i & v_iZ_i & v_i \end{bmatrix}}_{G_i} \begin{bmatrix} P_{11} \\ P_{12} \\ P_{13} \\ P_{14} \\ P_{21} \\ P_{22} \\ P_{23} \\ P_{24} \\ P_{31} \\ P_{32} \\ P_{33} \\ P_{34} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (2)$$

The projection matrix $\boldsymbol{p} = [P_{11}, P_{12}, \ldots, P_{43}]^T$ then could be solved by

$$\min_P \|\boldsymbol{Gp}\|^2 \ \ s.t. \|\boldsymbol{p}\| = 1 \qquad (3)$$

After the projection matrix $\boldsymbol{P}$ has been estimated, the intrinsic parameters and extrinsic parameters could be retrieved as follow:

Denote $\boldsymbol{P} = \boldsymbol{A}[\boldsymbol{R}|\boldsymbol{t}] = [\boldsymbol{B}\ \boldsymbol{b}]$, therefore $\boldsymbol{B} = \boldsymbol{AR}, \boldsymbol{b} = \boldsymbol{At}$. Since rotation matrix is orthogonal, we have

$$K = BB^T = \boldsymbol{AR} \cdot (\boldsymbol{AR})^{\boldsymbol{T}} = \boldsymbol{ARR^TA^T} = \boldsymbol{AA^T} = \begin{bmatrix} \alpha^2 + \beta^2 + u_0^2 & u_0 v_0 + \beta\gamma & u_0 \\ u_0 v_0 + \beta\gamma & \beta^2 + v_0^2 & v_0 \\ u_0 & v_0 & 1 \end{bmatrix} = \begin{bmatrix} k_u & k_c & u_0 \\ k_c & k_v & v_0 \\ u_0 & v_0 & 1 \end{bmatrix} \quad (4)$$

Therefore, the intrinsic parameters are computed as:

$$u_0 = K_{13}, v_0 = K_{23}, \beta = \sqrt{k_v - v_0^2}, \gamma = \frac{k_c - u_0 v_0}{\beta}, \alpha = \sqrt{k_u - u_0^2 - \gamma^2}$$

And the rotation matrix and translation vector could be computed as:

$$\boldsymbol{R} = \boldsymbol{A}^{-1}\boldsymbol{B}, \boldsymbol{t} = \boldsymbol{A}^{-1}\boldsymbol{b} \quad (5)$$

Since one characteristic of rotation matrix is $\det(\boldsymbol{R}) = 1$. However, a rotation matrix estimated by equation (5) does not necessarily satisfy $\det(\boldsymbol{R}) = 1$, which will give incorrect rotation angles. To deal with this, a nonlinear optimization procedure is used to estimate the best calibration parameters. Denote a function $\tilde{c}_i = f(C_i, roll, pitch, yaw, x, y, z, \alpha, \beta, \gamma, u_0, v_0)$ that projects a 3D point onto a 2D image plane. The objective function could be defined as

$$\|c_i - \tilde{c}_i\| = \|c_i - f(C_i, roll, pitch, yaw, x, y, z, \alpha, \beta, \gamma, u_0, v_0)\| \quad (6)$$

The best parameters are then estimated by solving the non-linear optimization problem defined as

$$[\widetilde{roll}, \widetilde{pitch}, \widetilde{yaw}, \tilde{x}, \tilde{y}, \tilde{z}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \widetilde{u_0}, \widetilde{v_0}] = \min_P \sum_i \|c_i - f(C_i, roll, pitch, yaw, x, y, z, \alpha, \beta, \gamma, u_0, v_0)\| \quad (7)$$

Figure 3 shows alignment of a user view of the elevation from his mobile phone with the point cloud data.

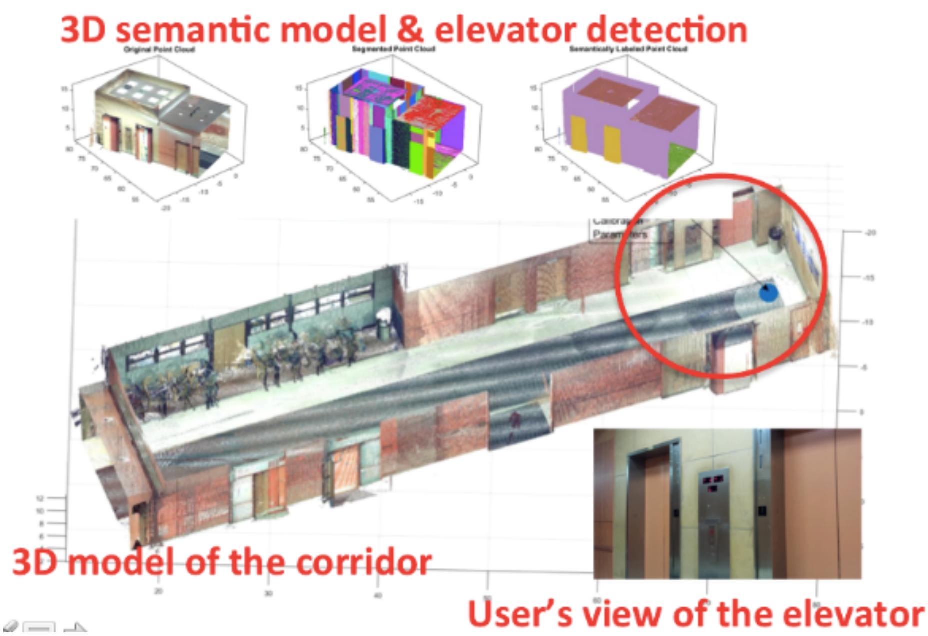


Figure 3. Registration of mobile phone image of the user with the 3D semantic model

Deep learning for crowd analysis

We have been studying deep learning methods to improve the accuracy of crowd density estimation for the low- or mid- density crowd [25], and tackle the high-density crowd using a regression-based method [26, 27]. So far we have obtained very promising results on both crowd counting and crowd density estimation (Figure 4) based on convolutional neural networks (CNNs) [28]. Though other convolutional neural networks have been used for crowd detection [25], our proposed pixel-wise calculation structure of the neural network is novel for the application of crowd density detection. From a high level perspective, the program would take as input a single color frame from the surveillance footage and output a form of "heat map" showing where people are at in the image and how many people there are. The "heat map" visualizes the count of the number of people per pixel of the image. Since a person takes up more than one pixel – and the sum of the total values within the body of a person is 1, the value per pixel is low. Where multiple people are occluding one another, we expect a higher value in that area. That is, even though that specific pixel only shows part of one person, the program should use surrounding pixels to determine that one person is occluding another. From this, any portion of the image can be considered, and within that portion the count of people can be determined. Additionally, these values can be averaged over time to compare the density of people per period of time.

The detection process is performed by a convolutional neural network. A brief explanation of how this works is as follows: an artificial neuron (which exists in the form of code) “looks” at the values in a tiny patch of pixels in the image. Each neuron has a certain pattern of values it “likes” to see in this patch. The closer the patch matches what the neuron likes to see, the higher value the neuron itself outputs. Additional layers of neurons then look at the output of the previous layers, themselves each liking their own pattern from that pervious layer. In this way, early neurons might like to see something like lines while the later layer neurons like to see combinations of lines in certain shapes. Finally, the entire network of neurons is made to like the appearance of people or groups of them. The neurons are trained to like the patterns they do, by training them on manually annotated data with density of people as described above. That is, known input is given to the network, the output is compared with the expected true output, and all the neurons are adjusted to more closely make the network’s output match the expected output.

Our results showed the network had a prediction error of ~10% the count of people per image frame in the camera footage tested. This accuracy is acceptable for general statistics, the crowd avoidance navigation, and crowd simulation verification. This accuracy comes from a small training set of data (due to the large amounts of time required to annotate data). We believe accuracy would improve simply with more ground truth data without any improvements to the network itself. Figure 4 shows example detection results using an early version of the network being applied on publicly available data. The later networks were trained and designed confidential video footage at the facilities we were testing at. The later footage includes more challenging data, particularly in regards to occlusions.

Figure 4. Crowd detection using deep learning

Beacon-based indoor localization

We installed Estimote Beacons in the facility to test the performance of beacon-based indoor localization. With Estimote beacons, we explored two methods of positioning using Bluetooth “beacons”: trilateration and fingerprinting. Our goal was to determine which method would yield a position that was closest to the real position of the device. Trilateration works under two assumptions: (1) We know the ground truth positions of all of the beacons installed, and (2) the distances calculated using the received signal strengths are accurate. This second assumption is problematic because of the interference that may be caused by obstructions and other devices. Fingerprinting, on the other hand, is the process by which a "snapshot" of the area's radio landscape is taken before localization is actually done [29]. Fingerprint-based localization involves comparing the current radio conditions around the device with this snapshot, which consists of multiple "fingerprints." Whereas trilateration required a very high accuracy (for the RSSIs - received signal strength indicators) in order to precisely determine a position, fingerprinting naturally assumes that the RSSIs are error-prone. This is reflected in the algorithm, which defines a margin of error for the measured data RSSI in relation to the fingerprint RSSI. Furthermore, the algorithm also assumes that the client may naturally miss one of the beacons in the fingerprint (potentially from walking around or due to congestion). Thus, we are able to almost guarantee that a position will be computed and that this calculated position is very near to the real position. The resulting fingerprint map using three Estimote Beacons [23] is shown in Figure 5. By comparison between the two approaches, the fingerprinting is a very viable and very robust method of localization and is a preferred approach to provide location-based services inside large, complex transportation hubs.

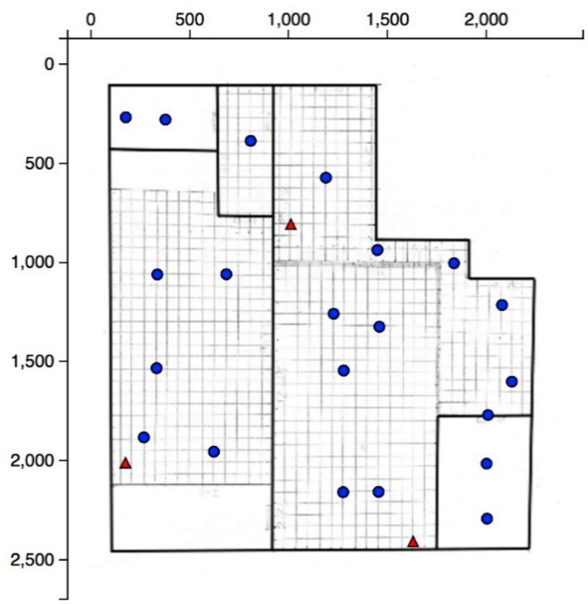


Figure 5. Server-generated map of fingerprints (blue circles) and beacons (red triangles) in the test area. Grid lines on hand-drawn floor plan represent tiles on floor. Axes represent pixel coordinates. 76 fingerprints were taken at 21 locations (average: ~3.6 fingerprints per location). In this visualization, the unit of both axes is in pixels.

Path planning and navigation assistance

The path planning element is encapsulated in a mobile application which leverages user location information (computed from the registration of image captured by user and the 3D facility model, and beacon-based localization), semantic facility model or simply a floor plan of the facility, and crowd analysis results to make decisions on paths that consider users' personal need. The mobile application is capable of providing multi-mode sensory feedback such as vibration and voice to users to achieve assistive navigation. Figure 6 shows an example test scenario where a user used the mobile app to navigate our studied facility using both the beacon-based localization and floor-plan-based path planning algorithms we developed on Android smart phones. Our study has shown that the app is capable of providing personalized travel guidance utilizing semantic 3D model, crowd analysis results, and strategically placed beacons. Even though a more formal evaluation is our next step, the initial feedback from our visually impaired users and service providers at Lighthouse Guild is very positive. The estimated Key Performance Indicators (KPIs) of our formal evaluation include: (1) The cost saving in using the 3D-model in assisting both sensor installation and maintenance, which will be measured against a tedious manual approach; (2) Average time to find a terminal reduced, by measuring navigation time, compared to baseline; (3) Increase of number of users with special needs, by measuring number of people downloading and using the apps; and (4) User experience satisfaction by using questionnaires to measure user experience in terms of navigation, waiting times, and safety concerns.

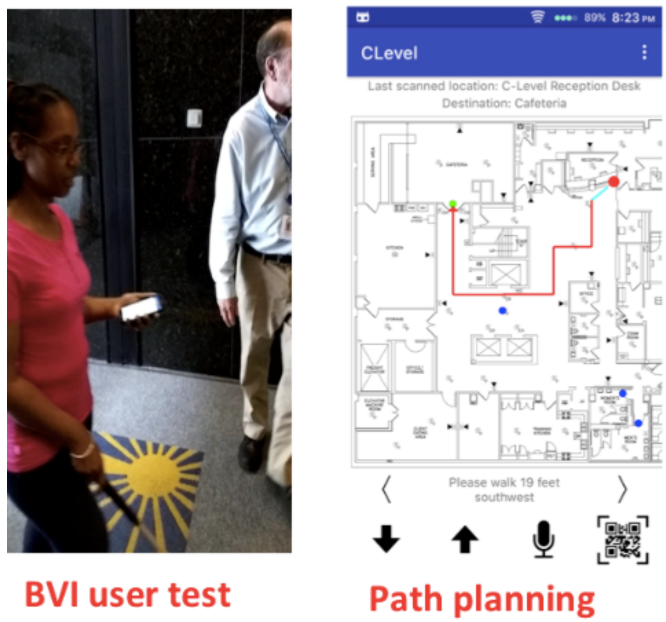


Figure 6. Beacon-based localization and floor-plan-based path planning.

CONCLUSION

This project investigated a novel cyber-physical infrastructure framework that can effectively and efficiently transform existing transportation hubs into smart facilities that are capable of providing better location-aware services (e.g. finding terminals, improving travel experience, obtaining security alerts) to the traveling public, especially for the underserved populations including those with visual impairment, ASD, or simply those with navigation challenges. We have started to conduct our pilot test at a multi-floor building in New York City to evaluate the feasibility of our proposed framework. This initial test has demonstrated that it is feasible to integrate our proposed Internet of Things elements (including video analytics, BLE beacons, mobile phone apps, and LiDAR-scanned 3D digital models) into a coherent framework to provide navigation services to people with disabilities. The results of detailed evaluation on user performance and system performance (i.e. robustness, user friendliness, and battery drain) will be reported in future publications.

Future improvements we have already identified would include using the 3D model to automatically determine information about the surveillance camera scene (such as camera pose and environment structure). This will not only improve the accuracy of the network, but more importantly provide a way in which the network can be generalized to all cameras in a facility without specific training the network to each individual camera. This could also make the network viable for completely different facilities and useable in any location, which will be our follow-on work. Since this framework is generic in the sense that it integrates wireless sensing, optical sensing, and mobile devices, new advancement in each technology domain can be easily incorporated into the framework. For example, mmwave is a new promising wireless localization technology as an alternative to BLE beacons, and it deserves further investigation. Another area of interest is the best method for selecting the user's current coordinates during fingerprinting. Existing services automatically assume that the location that the user selects on the floor plan is correct. However, there is no way for the user to actually know if they are correct or are off by inches or feet. Thus, a better method for self-localization during fingerprinting is also certainly a future area of research. Lastly, but not the least, it is in our team's agenda to test this framework in several public transit hubs in New York City and New Jersey in addition to the test in the pilot test building.

ACKNOWLEDGMENT

This research was performed under two appointments (one for Zhu/Gong and one for Tang) to the U.S. Department of Homeland Security (DHS) Science & Technology (S&T) Directorate Office of University Programs Summer Research Team Program for Minority Serving Institutions, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by ORAU under DOE contract number DE-SC0014664. This work is supported in part by the National Science Foundation through Award EFRI - 1137172, VentureWell (formerly NCIIA) through Award 10087-12, NIST/US Ignite Global City Teams Challenges(GCTC) [30]. The work is also supported by Rutgers Center for Advanced Infrastructure and Transportation – a national University Transportation Research Center supported by USDOT. We would like to thank the following individuals for their partnerships in the GCTC effort: Barbara Campbell of the NYS Commission for the Blind, William H. Seiple of Lighthouse Guild, Ed Hoff of NJ Transit and Michael Lysicatos of Passaic County Department of Planning & Economic Development. All opinions expressed in this paper are the authors' and do not necessarily reflect the policies and views of DHS, DOE, NSF, USDOT, VentureWell or ORAU/ORISE

REFERENCES

1. Varma, R., et al., *Visual Impairment and Blindness in Adults in the United States: Demographic and Geographic Variations From 2015 to 2050.* JAMA ophthalmology, 2016.
2. Feeley, C., et al., *Assessment of Transportation and Mobility Adults on the Autism Spectrum in NJ*. 2015, NJ Department of Health.
3. Agarwal, N., et al., *Algorithms for GPS operation indoors and downtown.* GPS solutions, 2002. **6**(3): p. 149-160.
4. Indoor.rs, *Indoors*. 2016.
5. *Click-and-Go*. Available from: http://www.clickandgomaps.com/.
6. Li, B., et al. *How feasible is the use of magnetic field alone for indoor positioning?* in *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*. 2012. IEEE.
7. Collin, J., O. Mezentsev, and G. Lachapelle. *Indoor positioning system using accelerometry and high accuracy heading sensors*. in *Proc. of ION GPS/GNSS 2003 Conference*. 2003.
8. Gruman, G., *What You Need to Know about Using Bluetooth Beacons.* Smart User (blog), InfoWorld, July, 2014. **22**.
9. Ahmetovic, D., et al. *NavCog: a navigational cognitive assistant for the blind*. in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'16). ACM*. 2016.
10. Hu, F., Z. Zhu, and J. Zhang. *Mobile Panoramic Vision for Assisting the Blind via Indexing and Localization*. in *European Conference on Computer Vision*. 2014. Springer.
11. Sattler, T., et al. *Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2015.
12. Cao, J., T. Chen, and J. Fan, *Landmark recognition with compact BoW histogram and ensemble ELM.* Multimedia Tools and Applications, 2016. **75**(5): p. 2839-2857.
13. Sünderhauf, N., et al. *On the performance of ConvNet features for place recognition*. in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. 2015. IEEE.

14. Min, B.-C., et al. *Incorporating information from trusted sources to enhance urban navigation for blind travelers*. in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 2015. IEEE.
15. Olmschenk, G., et al. *Mobile crowd assisted navigation for the visually impaired*. in *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015 IEEE 12th Intl Conf on*. 2015. IEEE.
16. Farinella, G.M., et al., *Representing scenes for real-time context classification on mobile devices.* Pattern Recognition, 2015. **48**(4): p. 1086-1100.
17. Altwaijry, H., M. Moghimi, and S. Belongie. *Recognizing locations with google glass: A case study*. in *IEEE Winter Conference on Applications of Computer Vision*. 2014. IEEE.
18. Paisios, N., *Mobile accessibility tools for the visually impaired*. 2012, Citeseer.
19. Manduchi, R. *Mobile vision as assistive technology for the blind: An experimental study*. in *International Conference on Computers for Handicapped Persons*. 2012. Springer.
20. Manduchi, R. and J.M. Coughlan. *The last meter: blind visual guidance to a target*. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014. ACM.
21. Murali, V.N. and J.M. Coughlan. *Smartphone-based crosswalk detection and localization for visually impaired pedestrians*. in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. 2013. IEEE.
22. Rituerto, A., G. Fusco, and J.M. Coughlan. *Towards a Sign-Based Indoor Navigation System for People with Visual Impairments*. in *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. 2016. ACM.
23. Estimote. *Esimote*. 2016; Available from: http://estimote.com/.
24. Rusu, R.B., *Semantic 3D object maps for everyday manipulation in human living environments.* KI-Künstliche Intelligenz, 2010. **24**(4): p. 345-348.
25. Zhang, C., et al. *Cross-scene crowd counting via deep convolutional neural networks*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
26. Lempitsky, V. and A. Zisserman. *Learning to count objects in images*. in *Advances in Neural Information Processing Systems*. 2010.
27. Chen, K., et al. *Feature Mining for Localised Crowd Counting*. in *BMVC*. 2012.
28. Jia, Y., et al. *Caffe: Convolutional architecture for fast feature embedding*. in *Proceedings of the 22nd ACM international conference on Multimedia*. 2014. ACM.
29. Subhan, F., et al. *Indoor positioning in bluetooth networks using fingerprinting and lateration approach*. in *2011 International Conference on Information Science and Applications*. 2011. IEEE.
30. zhu, z., et al., *Towards Smart Transportation Hub: Services to Persons with Special Needs Requiring Minimal New Infrastructure*, in *GCTC Expo, June 13-14, 2016* 2016, 2016 (A YouTube video https://youtu.be/_hxdXqTVG7I).