A NOVEL CROWD-RESILIENT VISUAL LOCALIZATION ALGORITHM VIA ROBUST PCA BACKGROUND EXTRACTION

Zhuorui Yang, Marco F. Duarte, and Aura Ganz

Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003

ABSTRACT

We present a novel egocentric visual localization algorithm for an indoor navigation system, called PERCEPT-V, which is designed to assist the blind and visually impaired users traveling independently in an unfamiliar indoor space. Through the integration of a background extraction module based on Robust Principle Component Analysis (RPCA) into the localization algorithm, we successfully improve the resilience of camera localization to the presence of crowds in the observed scene. Experiments using datasets of videos containing various levels of crowd activity show that the proposed algorithm can increase prominently the reliability of localization performance.

Index Terms— Robust Principal Component Analysis; Background Extraction; Visual Localization; PERCEPT; Assistive Technology

1. INTRODUCTION

According to visual impairment and blindness statistics from the World Health Organization (WHO), there are 285 million people suffering from visual impairment worldwide [1]. Indoor wayfinding in complex public spaces poses a major challenge to blind and visually impaired (BVI) individuals and negatively affects their mobility and the quality of life. To increase the BVI individuals' ability of travelling independently, we developed the PERCEPT indoor navigation system [2,3] using Near-Field Communication (NFC) tags; it was proved to be beneficial to the BVI users. From the experiments conducted with BVI subjects, PERCEPT has shown significant effectiveness on indoor wayfinding by delivering step-by-step audible navigation instructions to users. Although the PERCEPT system provides reliable localization and orientation to users by scanning the tags, the deployment of NFC tags requires changes in the environment, which can be costly.

To make PERCEPT system scalable and cost-effective, we propose to develop an organic computer vision-driven smartphone-based indoor navigation system, which we name PERCEPT-V. For this system, we show that the visual localization algorithm [4] can determine the BVI user's location and orientation in real-time using image or video captured by commercial devices, such as smartphones or wearable cameras. Moreover, the accuracy of the location and orientation estimates is sufficient for BVI users to navigate themselves safely in the space. While existing visual localization algorithms provide sufficiently accurate estimation of location and orientation, we find that there is a new technical challenge to PERCEPT-V: we must increase the reliability of the visual localization algorithm when crowds are present in the observed environment, and our algorithm must be resilient to instability in the framing and view of the images acquired by the BVI users.

In order to address the two aforementioned issues, we propose to integrate a background extraction algorithm into the image processing pipeline to improve the resilience of visual localization to the presence of crowds in the observed scene. In contrast to most existing background subtraction algorithms, which simply identify a mask that identifies and extracts the foreground in the image, we focus on the use of background subtraction algorithms that create a model for all pixels of the background. We refer to such algorithms as background extraction algorithms. Removing the foreground helps prevent spurious matches between features corresponding to crowds (foreground) and the reference navigation space (background).

In this paper, we leverage robust principle component analysis (RPCA) [5], an emerging formulation for background extraction that leverages a low-rank-plus-sparse matrix model, to represent the background in video sequences for the sake of increasing the number of correct keypoint descriptor matches between the target image and the reference images. In this formulation, the images in a video sequence are vectorized and arranged as columns of a matrix, which is then modeled as the sum of two separate components: the first component is a low-rank matrix, while the second one is a sparse matrix. The low-rank matrix corresponds to the background of the video which contains useful visual information of navigation space, as it models a component that is present consistently through different frames and densely present in the image sequence, while the sparse component models activity that is localized in each frame and it is likely to correspond to the moving crowds in the video. While the background of a static camera observing a simple setting can be modeled using a rank-one matrix, variations in illumination and minor camera movement can be accurately modeled by employing background matrices of higher rank; note that even in these cases the rank is usually much lower than the dimensions of the matrix (which almost always will correspond to the number of frames in the video). Furthermore, the low-rank-plus-sparse model allows for the estimation of the occluded pixels in each frame by exploiting the regularity of the background image via the low-rank model applied to the background component matrix. RPCA has potential benefits for low-power systems, such as PERCEPT-V, due to its compatibility with compressive sensing [6], a signal and image acquisition technique that allows for reductions in the dimensionality of the acquired data, and which often can be leveraged into simplified lower-power and lower-storage imaging

This project was supported in part by Grant IIS-1645737 from the National Science Foundation and Grant 80424 from Massachusetts Department of Transportation.

systems. Additionally, the proposed crowd-resilient localization algorithm can increase the applicability of our indoor visual navigation system to more complex and challenging environments, e.g., shopping malls or heavily used transportation hubs.

Our contributions can be summarized as follows. We present a novel visual localization algorithm resilient to crowds in the observed scene with RPCA background extraction. Our work considers novel aspects of visual localization that, to the best of our knowledge, have not yet been address in the literature.

The remainder of the paper is organized as follows. Section 2 summarizes a background and discusses related work. The background extraction via RPCA are presented in Section 3. Experimental results are shown in Section 4. Section 5 concludes the paper.

2. BACKGROUND AND RELATED WORK

A. Visual Localization in Wayfinding for BVI individuals

A class of visual systems uses a single camera and attempts to register the image obtained within a spatial model obtained by leveraging a training dataset – e.g., achieved using Structure from Motion (SfM) [7] – for localization purposes. Many of these algorithms leverage robust feature extraction algorithms that are invariant to changes in scale, rotation, scene illumination, etc. The goal is that the same visual features can be detected on the reference images and the currently observed image so that the camera user can be localized with respect to the 3D coordinate system of the environment. This process can be completed using two steps: (i) obtaining information about the 2D-to-3D correspondences between the observed 2D features and registered 3D correspondences via 2D-to-2D matching; and (ii) determining the camera pose with respect to the world coordinate system using the suitable spatial transformation calculated with the 2D-to-3D correspondences. The most popular features for navigation are known as keypoint detectors, where representative examples include speeded-up robust features (SURF) [8] and scale-invariant feature transform (SIFT) [9].

For localization purposes, the keypoint descriptors extracted from the acquired image are compared to those obtained from reference images and registered in 3D coordinate system to search for the best match by measuring the distance. After obtaining the 2D-to-2D matches, the 2D-to-3D correspondences can be found easily. To estimate the pose of the camera when the 2D-to-3D correspondences are available, most methods use random sampling and consensus (RANSAC) [10] to solve the Prospective-n-Point (PnP) problem, which randomly selects the smallest necessary subset of the putative 2D-to-3D correspondences and finds the best geometrical transformation to match the correspondences; the transformation found is then evaluated on all remaining data, selecting the best overall transformation over a fixed number of random draws. The process is repeated until sufficient agreement is observed between different trials or, alternatively, until the number of 2D-to-3D correspondences that agree with the transformation (known as the inlier set) is sufficiently large.

An alternative framework known as simultaneous localization and mapping (SLAM) [11,12] does not require a training dataset; rather, the structure of the environment is established by using the SfM algorithms on the sequence of previously observed images. New images are also matched against previously observed images in the sequence for localization purposes. SLAM is very popular in camera-based navigation of unknown environments. However, such

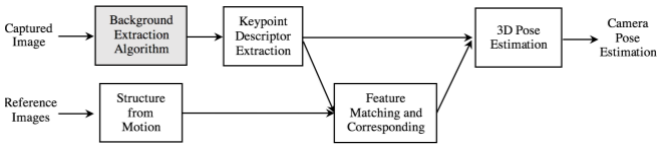

Fig. 1. *Flow chart of data processing pipeline for PERCEPT-V.*

a system is unnecessarily complicated for large public spaces that can be surveyed in advance. Similar methods by [13,14], and [15-17] use SLAM, optical flow, and RBG-D imaging, respectively, for obstacle avoidance.

There is prior research work on navigation systems for BVI users that leverage visual localization approaches both indoors and outdoors [18,19]. While most of these systems cover a wide range of functions, the end devices are inconvenient for daily use because they are heavy, complex, and expensive, which is not a feasible option for a majority of the users.

B. Background Subtraction vs. Background Extraction

There is a rich literature on background subtraction algorithms that are commonly employed in computer vision applications, where the background is not of interest to the application or system [20]. We will use the fuzzy self-organizing background subtraction (FuzzySOBS) algorithm [21] in our examples to compare the performance against the RPCA background extraction. FuzzySOBS poses a statistical model for all pixels of the background image and is one of the best-performing background subtraction algorithms in the literature. In terms of computational efficiency, the running time of FuzzySOBS and RPCA background extraction are $O(mn)$ and $O(m^6 n^6)$, respectively, where m and n are the width and height of the input frame.

Nonetheless, we believe that standard algorithms for background subtraction will not suffice for our purposes even though they are more computationally efficient than RPCA background extraction. This is due to the fact that these algorithms simply identify the pixels that correspond to activity in the image, but do not provide an estimate of the background for those regions of the field of view. Thus, even if the removal of the foreground also removed the presence of keypoint descriptors associated with them, the effect of occlusions and masking on the extraction of keypoint descriptors is still present. The effect of crowded activity in visual navigation for the blind has only recently begun to be studied [22-24].

3. BACKGROUND EXTRACTION VIA RPCA

As shown in Fig. 1, in the standard approach (without the shaded block), after keypoint descriptors are extracted from the acquired and reference images, a search finds the best match between the descriptors among the reference images that are used to obtain the 3D model of the environment to those from the acquired image. Consequently, the pose estimation module calculates the most likely geometric transformation between the putative 2D-to-3D correspondences, providing an estimate of the location and orientation of the camera. Our proposed architecture adds the one shaded block: a background extraction scheme to remove activity from passerby before keypoint descriptors are obtained. By extracting the background, the proposed localization algorithm

reduces the likelihood of mismatches from features for the foreground (crowds) to the reference images and increases the likelihood of recovering more useful features about background (navigation space) that can match with the registered features and benefit the localization performance simultaneously.

Let the acquired video be comprised of n frames containing m pixels each. We store this video in a matrix $D \in R^{m \times n}$; each column of the matrix D corresponds to a video frame, and each row represents the evolution of a specific pixel over the acquisition time. In RPCA, we consider the following decomposition for the video matrix [25]:

$$D = A + E, \quad (1)$$

where A is a low-rank matrix corresponding to the background and E is a sparse matrix corresponding to the foreground or activity of passerby. This decomposition is motivated by the small number of degrees of freedom for the background and the localized and highly concentrated passerby activity. Note that only A will be subject to the processing used by the visual localization algorithm, while E is discarded.

The exact recovery of the low-rank matrix A of interest from the sum D can be solved by the following convex optimization problem:

$$\arg\min_{A,E} \| A \|_* + \lambda \|E\|_1, \ subject\ to\ D = A + E, \quad (2)$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, $\|\cdot\|_1$ represents the L_1 norm of a matrix, and λ is a positive weighting parameter. We apply the accelerated proximal gradient method [26] to solve the optimization problem in (2). Its convergence rate is $O(k^{-2})$, where k is the number of iterations.

4. EXPERIMENTAL RESULTS

In this section, we reveal the applicability of adopting RPCA background extraction over traditional background subtraction into our localization algorithm. Subsequently, we show the performance improvement for our proposed visual localization algorithm, including the RPCA background extraction. To showcase these improvements, we conduct two experiments. In our first experiment, we test the reliability of feature extraction from background-extracted images using RPCA and FuzzySOBS by measuring the number/percentage of putative matches between the features on processed frames and those on reference images (feature matching block in Fig. 1). In our second experiment, we test the performance improvement (in terms of localization re) achieved by using RPCA background extraction to recover background-generated keypoint descriptors that can be useful for localization purposes. Since there is no benchmark dataset that is publicly available for our case, we collect our own dataset [27] with human crowds in a public space to evaluate the performance of the algorithm.

A. Feature Extraction Reliability

For our first experiment, we use data from video sequences captured at the UMass Amherst Campus Center. The reference dataset consists of images of the center's first floor space that were taken while the center was closed. The test dataset consists of 63 video sequences taken in two groups: 26 sequences were taken during low levels of activity (winter break) and 37 sequences were taken during high levels of activity. The number of frames in each video sequence is 200 and the resolution of each frame is 192 x 108. We performed

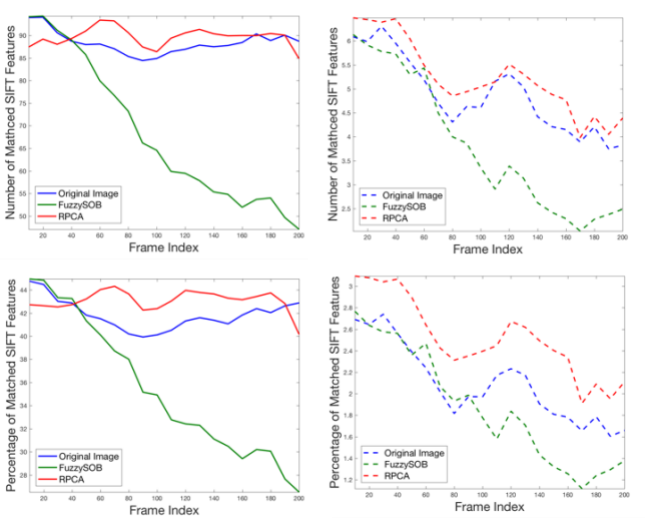

Fig. 2. *Average performance of SIFT descriptor-based image matching for (i) original video sequences, (ii) backgrounds obtained from the FuzzySOBS algorithm, (iii) backgrounds obtained from the RPCA algorithm. Top row: Average number of SIFT feature matches for each frame of the video sequence. Bottom row: Average percentage of SIFT features matched during camera localization. Left Column: Sequences with low and medium-level activity. Right Column: Sequences with high-level activity.*

SIFT feature extraction and subsequent feature matching for three different versions of each video sequence: (*i*) the original video sequence frames, (*ii*) the background frames extracted from the FuzzySOBS algorithm, and (*iii*) the background frames extracted with the RPCA model.

Fig. 2 shows the average number of matches from each of the video frames as a function of the frame index, as well as the average percentage of those descriptors that are successfully matched during the feature matching process. These quantities are averaged over the 26 video sequences containing low activity. The results show that the quality of FuzzySOBS is poor, resulting in a very low percentage of features being matched. Moreover, the quality of the background degrades as further frames are processed. Furthermore, the percentage of keypoint descriptors matched from the RPCA background image is higher than the percentage matched from the original images, which is indicative of the higher proportion of background-generated keypoint descriptors obtained from RPCA.

We also processed video sequences with high levels of activity, which poses a more challenging setting for background subtraction algorithms. The percentage of features that are matched between the captured images and the reference images is very small, as shown in Fig. 2, due to the large number of features obtained from the foreground activity. Since the background extraction algorithms are not completely successful, the percentage of matched SIFT features stays low. Nonetheless, it is still the case that the RPCA background image provides a larger proportion of background-generated keypoint descriptors (in aggregate) than the two alternatives.

B. Improvement of Localization Performance

For our second experiment, we collected an additional set of test data at the UMass Amherst Campus Center. The new test dataset contains 77 video sequences and there are 15,400 frames in total. The number

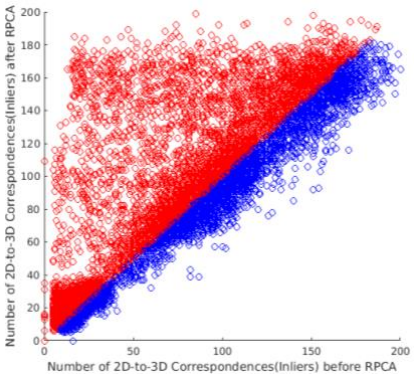

Fig. 3. *Comparison of number of 2D-to-3D correspondences (inlier set) before and after RPCA-based background extraction is applied to video sequence frames. Red and blue marks represent increases and decreases in the number of correspondences due to RPCA, respectively.*

of frames in each video sequence is 200 and the resolution of each frame is 480 x 270. The goal for this experiment is to determine the impact of RPCA background extraction on the performance of visual localization algorithm as measured by the number of 2D-to-3D correspondences in the inlier set returned by the pose estimation (last block in Fig. 1) using RANSAC. An increase in the number of inlier correspondences is indicative of improved localization performance, given that there are more background-generated keypoint descriptors: these descriptors are recovered by RPCA and represent a single perspective hypothesis. We checked this impact by plotting the number of correspondences from the background image obtained by RPCA to the number of correspondences from the original image without processing by RPCA.

As shown in Fig. 3, we observe significant increments on the number of correspondences for many images after applying RPCA background extraction, as evident by the large number of points on the upper triangle of the figure. For the rest of the frames, almost all of the marks in the figure are close to the diagonal, implying that any negative effects on a frame from applying RPCA are minor. Based on our observation, the slight reduction in number is due to the presence of the blurriness caused by either the capturing or the background extraction. Among 15,400 frames, there are 9,339 frames (60%) whose number of inlier correspondences after RPCA is larger or equal than that before RPCA.

Furthermore, since our interest focuses on recovering as many background-generated keypoint descriptors as possible for the localization algorithm when the background is almost covered by passerby (e.g., frames containing less than 10 correspondences, reflecting such scenarios), we analyzed the result particularly for these worst cases to check if RPCA background extraction can benefit the visual localization algorithm for this type of circumstances. Figure 4a shows a histogram for the increment of the number of correspondences due to the use of RPCA. The figure shows that the benefits of applying RPCA are much larger than the losses, since the range of increments is from -20 to 140. Among the 737 worst-case frames, 716 frames (97%) have equal or number of correspondences after applying RPCA background extraction.

Figure 4b shows the cumulative error distribution function for localization with and without RPCA background subtraction in the processing pipeline. The figure presents the substantial

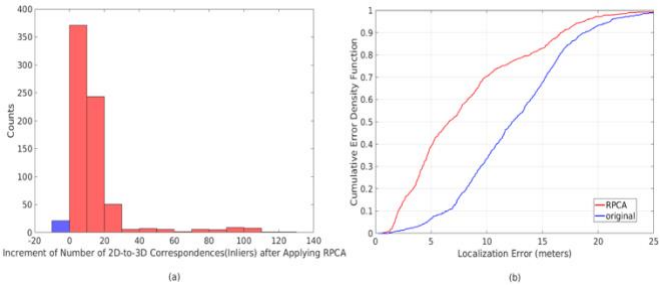

Fig. 4. *Left: Histograms for increment in number of 2D-to-3D correspondences (inlier set) after RPCA. The blue column denotes a decrease in the number of correspondences. Right: CDF of localization error in meters.*

improvement of localization performance after RPCA is applied due to the increment of number of 2D-to-3D correspondences, which occurs thanks to the higher number of background-generated keypoint descriptors recovered by the RPCA background extraction. Besides,
the average localization error for these worst cases reduces from 12.50m to 8.14m after RPCA is applied, reflecting that 35% of the localization error is reduced in average. These results imply that the increase in the number of correspondences after applying RPCA results in an improvement of localization performance.

5. CONCLUSIONS

In this paper, we propose PERCEPT-V, an indoor navigation system for the BVI users based on a novel visual localization algorithm resilient to crowds in the observed scene. We addressed the new challenges in localization faced by the system using RPCA background extraction to increase the localization reliability. Unlike popular background subtraction algorithms, RPCA background extraction enables us to model the background more accurately by leveraging a low-rank-plus-sparse decomposition of a matrix representation of the tested video sequence. With more useful information available in the extracted background, the proposed visual localization algorithm can increase its reliability in the presence of crowds.

Our experimental results indicate two positive findings. First, we show that RPCA background extraction outperforms FuzzySOBS in obtaining an accurate background model for PERCEPT-V. Second, we demonstrate an improvement of localization reliability when using RPCA background extraction that is due to the recovery of more background-generated keypoint descriptors that can be matched to those from the reference images. We anticipate future work in the direction of finding the best-performed RPCA algorithm implementation by comparing among all candidates [28]. We also remain open to the introduction of additional modules that can augment the background extraction to further improve the performance of PERCEPT-V. We will also consider how to discern between background and foreground regions within the set of keypoint descriptors obtained from each frame. The expectation is to distinguish between these classes of keypoint descriptors (and possibly more refined classes) by leveraging both signal processing and machine learning schemes for keypoint descriptor classification.

Acknowledgement: We thank Binru Cao for the help with building the test datasets.

7. REFERENCES

[1] World Health Organization. *Visual Impairment and Blindness* [Online]. Available: http://www.who.int/mediacentre/factsheets/fs282/en/

[2] A. Ganz *et al*., "PERCEPT: Indoor navigation for the blind and visually impaired," in *IEEE Int. Conf. Engineering in Medicine and Biology Society (EMBC)*, Boston, MA, 2011, pp. 856-859.

[3] A. Ganz, J. M. Schafer, Y. Tao, C. Wilson and M. Robertson, "PERCEPT-II: Smartphone based indoor navigation system for the blind," in *IEEE Int. Conf. Engineering in Medicine and Biology Society (EMBC)*, Chicago, IL, 2014, pp. 3662-3665.

[4] Z. Yang and A. Ganz, "Egocentric Landmark-Based Indoor Guidance System for the Visually Impaired," *Int. J. E-Health and Medical Communications*, vol. 8, no. 3, pp. 55–69, June 2017.

[5] C. Guyon, T. Bouwmans, and E. Hadi Zahzah, "Robust Principal Component Analysis for Background Subtraction: Systematic Evaluation and Comparative Analysis," in *Principal Component Analysis*, Intech, 2012, pp. 223–238.

[6] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to Compressed Sensing," in *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[7] Ccwu.me. *VisualSFM : A Visual Structure from Motion System -Documentation*. [online]. Available: http://ccwu.me/vsfm/doc.html

[8] H. Bay et al, "Speeded-up robust features (SURF)," Comput. Vision Image Understanding, vol. 110, (3), pp. 346-359, 2008.

[9] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE Int. Conf. Computer Vision*, Kerkyra, Greece, Sep. 1999, pp. 1150–1157.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[11] V. Pradeep, G. Medioni, and J. Weiland, "Robot vision for the visually impaired," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, San Francisco, CA, June 2010, pp. 15–22.

[12] P. F. Alcantarilla, J. J. Yebes, J. AlmazÅLan, and L. M. Bergasa, "On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, Saint Paul, MN, May 2012, pp. 1290–1297.

[13] R. Manduchi and J. Coughlan, "(Computer) vision without sight," *Commun. ACM*, vol. 55, no. 1, pp. 96–104, Jan. 2012.

[14] A. RodrÅLıguez, J. J. Yebes, P. F. Alcantarilla, L. M. Bergasa, J. AlmazÅLan, and A. Cela, "Assisting the visually impaired: Obstacle detection and warning system by acoustic feedback," *Sensors*, vol. 12, no. 12, pp. 17476–17496, Dec. 2012.

[15] N. Bourbakis, S. K. Makrogiannis, and D. Dakopoulos, "A system-prototype representing 3D space via alternative-sensing for visually impaired navigation," *IEEE Sensors*, vol. 13, no. 7, pp. 2535–2547, July 2013.

[16] S. Pundlik, M. Tomasi, and G. Luo, "Collision detection for visually impaired from a bodymounted camera," in *IEEE Int. Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Portland, OR, June 2013, pp. 41–47.

[17] A. Tamjidi, C. Ye, and S. Hong, "6-DOF pose estimation of a portable navigation aid for the visually impaired," in *IEEE Int. Symp. Robotic and Sensors Environments (ROSE)*, Washington, DC, Oct. 2013, pp. 178–183.

[18] Y. H. Lee, and G. Medioni, "RGB-D camera based wearable navigation system for the visually impaired," *Computer Vision and Image Understanding*, vol. 149(C), pp. 3–20, 2016.

[19] S. M. Jonas *et al.*, "Imago: image-guided navigation for visually impaired people," *J. Ambient Intelligence and Smart Environments*, vol. 7, no. 5, pp. 679-692, 2015.

[20] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vision Image Understanding,* vol. 122, pp. 4-21, 2014.

[21] L. Maddalena and A. Petrosino, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing & Applications,* vol. 19, *(2),* pp. 179-186, 2010.

[22] T. S. Leung and G. Medioni, "Visual navigation aid for the blind in dynamic environments," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Columbus, OH, June 2014, pp. 579–586.

[23] A. Caldini, M. Fanfani, and C. Colombo, "Smartphone-based obstacle detection for the visually impaired," in *Int. Conf. Image Analysis and Proc. (ICIAP)*, Genoa, Italy, Sep. 2015, pp. 480–488.

[24] T. Schwarze and M. Lauer, "Robust ground plane tracking in cluttered environments from egocentric stereo vision," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, Seattle, WA, May 2015, pp. 2442–2447.

[25] E. J. Candès, X. Li, Y. Ma and J. Wright, "Robust principal component analysis?". *J. ACM,* vol. *58, no.* 1, pp. 1-37, 2009.

[26] M. Chen, A. Ganesh, Z. Lin, Y. Ma, J. Wright and L. Wu, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *J. Marine Biological Association Uk*, vol. 56, no.3, pp. 707-722, 2009.

[27] Datasets [Online]. https://github.com/zhuoruiyang/Crowd-Resilient-Visual-Localization-Algorithm.git

[28] The Perception and Decision Laboratory. *Low-Rank Matrix Recovery and Completion Via Convex Optimization* [Online]. Available: http://perception.csl.illinois.edu/matrix-rank/sample_code.html