A Feasibility Study on Smartphone Localization using Image Registration with Segmented 3D Building Models based on Multi-Material Classes

Max Jwo Lem Lee, Li-Ta Hsu
Interdisciplinary Division of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University

BIOGRAPHY (IES)

Mr. Max Jwo Lem Lee is currently a student in the Bachelor of Engineering (Honours) in Aviation Engineering from the Hong Kong Polytechnic University. His research interests include precise positioning in urban environments, indoor positioning, navigation, and unmanned aerial vehicle.

Dr. Li-Ta Hsu received the B.S. and Ph.D. degrees in aeronautics and astronautics from National Cheng Kung University, Taiwan, in 2007 and 2013, respectively. He is currently an assistant professor with Interdisciplinary Division of Aeronautical and Aviation Engineering, Hong Kong Polytechnic University, before he served as post-doctoral researcher in Institute of Industrial Science at University of Tokyo, Japan. In 2012 and 2013, he was a visiting scholar in University College London, U.K and Tokyo University of Marine Science and Technology, Japan, respectively. His research interests include GNSS positioning in challenging environments and localization for autonomous driving vehicle and unmanned aerial vehicle.

ABSTRACT

Accurate smartphone-based outdoor localization system in deep urban canyons are increasingly needed for various IoT applications such as augmented reality, intelligent transportation, etc. This article proposes a multi-material image registration solution for accurate pose estimation in urban canyons where global navigation satellite system (GNSS) tends to fail. In the offline stage, a material segmented city model is used to generate segmented images at each pose (six degrees of freedom of position and rotation). In the online stage, an image is taken with a smartphone camera that provides textual information about the surrounding environment. The approach utilizes computer vision algorithms to rectify and manually segment between the different types of material identified in the smartphone image. The hypothesized poses (candidate) images are then matched with the segmented smartphone image. The candidate image with the maximum likelihood is regarded as the estimated pose of the user. The positioning results achieves 2.0m level accuracy in common high rise along street, 5.5m in foliage dense environment and 15.7m in alleyway. A 45% positioning improvement to current state-of-the-art method. The estimation of yaw achieves 2.3° level accuracy, 8 times the improvement to smartphone IMU.

1. INTRODUCTION

Urban localization is an essential step to the development of numerous IoT applications such as digital management of navigation, augmented reality, commercial related services [1], and an indispensable part of our daily lives due to its widespread application [2]. In the context of outdoor pedestrian localization, the application of global navigation satellite system (GNSS) is the key technology to provide accurate positioning/timing service in open field environments. Unfortunately, its positioning performance in urban areas still has a lot of potential to improve due to signal blockages and reflections caused by tall buildings and dense foliage [3, 4].

Standing at the point of view of how a pedestrian navigate him/herself, we human beings, also locate based on the visual landmarks that consists of different semantic information, and each semantic has a material of its own. Inspired from that, our proposed novel solution is the multi-material image registration utilizing different types of materials that are widely seen and continuously distributed in urban scenes. The proposed method offers several major advantages. Firstly, we can take advantage of building materials as visual aids for precise self-localization. Secondly, with the use of building information modelling (BIM), it does not require pre-surveyed data, hence it is highly scalable and low cost. Thirdly, the semantics of materials are stored as a vector map, making it simple to update and label accurately. Lastly, the proposed method identifies and considers dynamic objects into its scoring system.

The remainders of the paper are organized as follows. Sect. 2 explains the overview of the proposed multi-material image registration approach. Sect. 3 describes the experimentation process and the improvement of the proposed algorithm verified with existing advanced positioning methods. Sect. 4 contains the concluding remarks and future work.

2. PROPOSED METHOD

An overview of the proposed multi-material image registration method is shown in Fig. 1. The method is divided into two main stages: an offline process, and an online process. In the offline process, the building models are manually segmented into different colors based on the material, which grantees a perfect representation of the materials in the 3D city model. The segmented city model is used to generate an image at each pose. In the online process, the user captures an image with their smartphone, with the initial pose estimated by the smartphone, candidates (hypothesized poses) are spread across a searching grid based on the initial pose. The smartphone image is then rectified and segmented based on the identified materials in the image. The segmented smartphone image is de-rectified and matched with the candidate images using multiple metrics to calculate the similarity scores. The scores of each method are combined to calculate the likelihood of each candidate. The chosen pose is determined by the candidate with the maximum likelihood among all the candidates. The details of the proposed method are described in the following section.

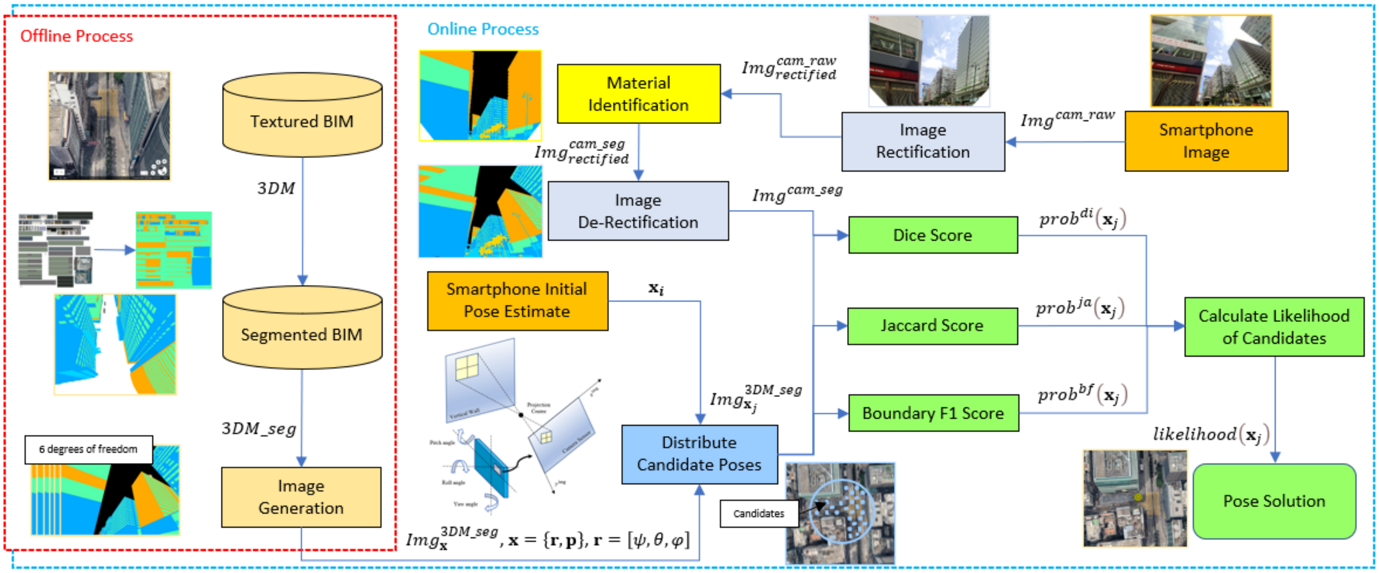

Offline Process
Textured BIM
3DM
Segmented BIM
3DM_seg
Image Generation
6 degrees of freedom
Online Process
Material Identification
$Img_{rectified}^{cam\_raw}$
Image Rectification
Img^{cam_raw}
Smartphone Image
$Img_{rectified}^{cam\_seg}$
Image De-Rectification
Img^{cam_seg}
Dice Score
$prob^{di}(\mathbf{x}_j)$
Jaccard Score
$prob^{ja}(\mathbf{x}_j)$
Boundary F1 Score
$prob^{bf}(\mathbf{x}_j)$
Calculate Likelihood of Candidates
$likelihood(\mathbf{x}_j)$
Pose Solution
Smartphone Initial Pose Estimate
$\mathbf{x}_i$
$Img_{\mathbf{x}_j}^{3DM\_seg}$
Distribute Candidate Poses
Candidates
Vertical Wall
Projection Center
Camera Sensor
Pitch angle
Roll angle
Yaw angle
x^{img}
y^{img}
$Img_{\mathbf{x}}^{3DM\_seg}, \mathbf{x} = \{\mathbf{r}, \mathbf{p}\}, \mathbf{r} = [\psi, \theta, \varphi]$

Fig. 1. Flowchart of the multi-material image registration based on segmented smartphone image and segmented generated images.

2.1 Textured & Segmented BIM

The city model used in this research is provided by the Surveying and Mapping Office, Lands Department, Hong Kong [5]. Each building model has its own corresponding 2D vector map in JPG format that provides textural information of the building. The building vector maps were manually labelled, in which each pixel in the texture image is assigned a color for the material it represents, which can then be used to simulate a segmented 3D city model as shown in Fig. 1. In this research, we used six classes in total to test the feasibility of the proposed method, each class has their own respective RGB color: Sky (black), Stone (blue), Glass (green), Metal (orange), Foliage (yellow), Others (light blue). The building vector maps were labelled manually with the Image Labeler application, which is part of the Computer Vision Toolbox, MATLAB [6].

2.2 Image generation

Images were generated from the segmented city model to match with the smartphone image. Equation (1) denotes the process.

$$\mathbf{x} = \{lat, lon, alt, \psi, \theta, \varphi\}$$
$$Img_{\mathbf{x}}^{3DM\_seg} = RL\_P(3DM_{seg}, \mathbf{x}) \tag{1}$$

Where $\mathbf{x}$ is the state that defines the pose which holds the three-dimensional position and three-dimensional rotation. RL_P is the function to capture the images from the segmented city model. The format of the images can be described as:

$$Img_{\mathbf{x}}^{3DM_{seg}} = SI(\mathbf{u}_{\mathbf{x}}, \mathbf{v}_{\mathbf{x}})$$
$$SI \in \left\{ \begin{matrix} \text{Sky } (0), \text{Stone } (1), \text{Glass } (2), \\ \text{Metal } (3), \text{Foliage } (4), \text{Others } (5) \end{matrix} \right\} \tag{2}$$

Where $\mathbf{u}_{\mathbf{x}}, \mathbf{v}_{\mathbf{x}}$ are the 2D pixel coordinates of the pixel inside the image generated based on the pose x. SI is the function that assigns each pixel an indexed number to represent a material class. Each image stores its corresponding pose. Fig. 1 shows an example of an image generated from the equirectangular projection based on a defined pose.

Candidate poses are then distributed around the initial estimated pose. The initial rough estimation of the pose is calculated by the smartphone GNSS receiver and IMU when capturing an image with the smartphone. The poses will then be reduced to the specific candidate poses shown in (3).

$$\mathbf{X} = \{\mathbf{x}_0 \cdots \mathbf{x}_s\} \tag{3}$$

Where s is the index of the poses outside of the buildings, that is generated offline and saved in a database. Candidate pose $\mathbf{x}_j$ is extracted from the database **X**, where $\mathbf{x}_j \in$ **X,** and the subscript j is the index of the candidate poses. The corresponding image for each candidate pose is denoted as $Img_{\mathbf{x}_j}^{3DM\_seg}$. The distributed candidate images are used to compare against the smartphone image.

2.3 Smartphone Image Processing

The initial camera rotation information is used to perform rectification on the images as a preparation for further material recognition. The proposed image rectification assumes that the rotation of the camera image is approximately known from the output of the smartphone IMU. From this, horizon and keystone correction can be performed [7]. The first kind of distortion is associated to the roll angle of the camera, whereas the second kind is due to the camera pitch angle. The greater the object elements that are further away from the horizon is, the greater the distortion is. However, the horizon area of the rectified images, which usually contains more distinctive features, provides a more suitable input for classification. Once combined, it can rectify the image such that it is an approximation image taken at horizontal and vertical level shown in Fig. 1. The rectified smartphone image was then labelled manually. In the future, however, we plan to utilize a deep learning neural network to identify the material automatically. After segmentation, the image can be de-rectified with the reverse of the image rectification process.

2.4 Image Registration

In the online stage, the candidate images are compared to the smartphone image. The image registration calculates the score of each candidate image. The target function is to find the candidate image with the largest similarity with respect to the semantic information of materials. A usual approach is to use the region or contours of each material class in the candidate image to compare with the corresponding material class in the smartphone image. The similarity for each segmented material is then weighted according to the number of pixels they occupy in the candidate image to calculate the score of each material. Finally, the score for each material is combined to become the score of the candidate. We considered the score of three metrics, Sørensen–Dice [8], Jaccard [9] for regional metrics, Boundary F1 [10] for contour metric, and calibrated a CDF based on a Gaussian distribution. The scores of each method is used to calculate the corresponding probability value in their respective distributions.

$$prob^*(\mathbf{x}_j) = \frac{1}{\sigma^* \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{score^*(\mathbf{x}_j)} e^{-\frac{1}{2}\left(\frac{x-\mu^*}{\sigma^*}\right)^2} dx \tag{4}$$

TABLE I. Parameters of Gaussian distribution

Method	Standard Deviation	Mean
Dice	0.1813	0.6686
Jaccard	0.1567	0.5399
BF	0.1387	0.4275

Where $*$ is the variable that is dependent on the method, σ is the standard deviation and μ is the mean of the CDF. The combined probability becomes the likelihood of each candidate.

$$likelihood(\mathbf{x}_j) = prob^{di}(\mathbf{x}_j) \cdot prob^{ja}(\mathbf{x}_j) \cdot prob^{bf}(\mathbf{x}_j) \tag{5}$$

A higher priority is given to the candidate image with a higher likelihood. In theory, the candidate image at ground truth should have the maximum likelihood. Thus, the candidate with the maximum likelihood is selected as the chosen candidate indicated in (6).

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}_j}\left(likelihood(\mathbf{x}_j)\right) \tag{6}$$

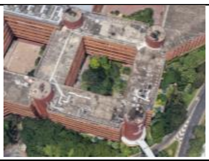
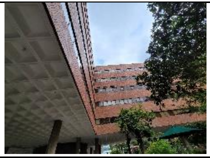
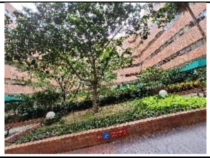
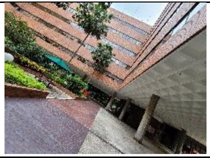
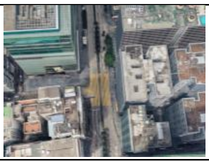
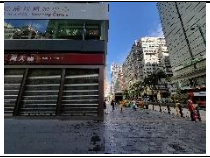
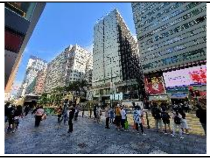
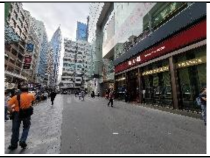
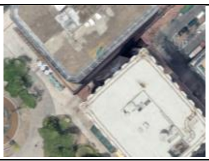
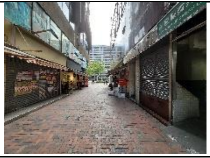
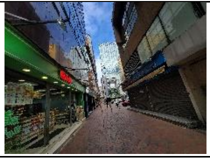
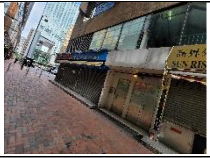
Where $\arg\max_{\mathbf{x}_j}$ is a function that filters the highest total score, and $\hat{\mathbf{x}}$ is the estimated candidate pose with the highest likelihood. The chosen candidate pose stores the latitude, longitude, altitude, yaw, pitch, and roll.

3. EXPERIMENT SETUP AND RESULTS

The experimental images were chosen with the following skyline categorizations: distinctive, symmetrical, insufficient, obscured and concealed. Categorizations were based on the difficulties experienced by current 3DMA GNSS and vision-based positioning methods. The smartphone (Samsung Galaxy Note 20 Ultra 5G smartphone with the ultra-wide-lens 13mm 12-MP f/2.2) was used to capture the images and to record the low-cost GNSS position and IMU rotation. The locations were chosen to test the following environments respectively, dense foliage (Loc. 1), along street (Loc. 2), and alleyway (Loc. 3). The positioning quality of the proposed method was analyzed based on the ideal smartphone image segmentation. The experimental results were then post-processed and compared to the ground truth and different positioning algorithms, including:

1) Proposed Multi-Material Image Registration (Combination of Dice, Jaccard and BF Metrics)
2) Skyline Matching: Matching using sky and building class only [11].
3) 3DMA: Integrated solution by shadow matching, skymask 3DMA and likelihood based ranging GNSS [12].
4) WLS: Weighted Least Squares [13].
5) NMEA: Low-cost GNSS solution by Galaxy S20 Ultra, Broadcom BCM47755.

TABLE II. Locations and images tested with the multi-material image registration method

Loc.	Experimental Images			
1	Overview	1.1 (Obscured)	1.2 (Concealed)	1.3 (Obscured)
				
2	Overview	2.1 (Distinctive)	2.2 (Distinctive)	2.3 (Distinctive)
				
3	Overview	3.1 (Symmetrical)	3.2 (Symmetrical)	3.3 (Symmetrical)
				

The positioning error of the dataset shown in table III range from approximately 5.56 meters for foliage dense environments, 1.97 meters for common high-rise environments, and 15.68 for alleyway environments. Utilizing additional material information from buildings, it outperforms skyline matching twice as much. The inability of skyline matching was due to the presence of foliage obscuring the skyline. Without an exposed skyline, it cannot match correctly and risks increasing the positioning error. 3DMA has shown to correct the positioning to a higher degree, coming behind the proposed method. The positioning error of WLS and NMEA were likely because of the diffraction of GNSS signals passing under the foliage with the combination of high-rise buildings.

In terms of rotational error, the results show that, in an urban environment with features, the material of buildings can be used to estimate the rotation. The yaw, pitch and roll have an accuracy of 2.3, 1.4 and 1.3 degrees, respectively. However, the smartphone IMU pitch and roll estimation is already very accurate compared to the proposed method, and thus the proposed method will only degrade the estimation. Instead, the proposed method succeeds at predicting the yaw accurately within an average of 2.3 degrees. Hence, the proposed method can be considered an accurate approach to estimate the heading of the user in an urban environment.

TABLE III. Positioning and rotational performance comparison of the multi-material image registration and other advanced positioning algorithms.

Loc.	Deviation from Ground Truth Error. Unit: meter.					Deviation from Ground Truth Error. Unit: degree.					
	Multi-Material Image Registration	Skyline Matching	3DMA	WLS	NMEA	Multi-Material Image Registration			Smartphone IMU		
						ψ	θ	φ	ψ	θ	φ
1.1	7.07	22.92	7.96	17.66	36.24	-4	0	-1	-27	-2.0	1.0
1.2	4.34	22.62				3	2	-2	7	0.5	-0.5
1.3	5.28	7.14				3	2	-1	18	-0.5	0.5
1. Avg.	**5.56**	17.56				**3.3**	**1.3**	**1.3**	17.3	1.0	0.6
2.1	0.66	14.80	6.87	23.29	7.94	5	1	-2	11	0.5	-1.0
2.2	1.83	1.58				-3	-1	0	18	2.0	0.0
2.3	3.43	2.89				1	2	-2	19	-2.0	0.5
2. Avg.	**1.97**	6.42				**3**	**1.3**	**1.3**	16	1.5	0.5
3.1	29.89	13.57	18.80	46.58	18.89	2	2	-2	31	1.0	-1.5
3.2	6.61	25.53				0	1	0	28	0.5	-0.2
3.3	10.53	24.80				0	-2	-2	27	-0.5	-0.2
3. Avg.	**15.68**	21.30				**0.6**	**1.7**	**1.3**	28.6	0.6	1.8
All Avg.	**7.74**	15.09	11.21	29.18	21.02	**2.3**	**1.4**	**1.3**	20.6	1.0	1.0

The performance of each metric was also analyzed as shown in the heatmap in Table V. The proposed method using Dice and Jaccard have very large positioning errors in Loc 1, possibly due to the lack of distinctive materials captured in the smartphone image. The tested location is surrounded by buildings of the same shape, size, and material. Therefore, it is a very challenging environment for as the candidate images share a common material distribution. It can be seen in this situation, using the BF achieves a higher positioning accuracy than the Dice and Jaccard, as it calculates the material contour rather than the material region. With the combination of the three metrics, this foliage dense environment proved suitable for the proposed method. Loc 2 demonstrated that the metrics complement each other when combined. As shown in Loc. 2.1, in a scene with diverse materials, the Dice and Jaccard have a higher positioning accuracy and achieve a higher likelihood over BF. Therefore, the combination of the three metrics leans towards the regional based similarities. The poor results of Loc. 3 can be explained by two conditions required for accurate positioning. Firstly, the images ideally should have no segmentation error. This error is not considered in the positioning results, as we are assessing the ideal image segmentation. Secondly, ideally there should be no discrepancies between the smartphone image and the candidate image at ground truth. Loc. 3 suffers from the latter as shown in Table IV.

Table IV. Discrepancy between reality and 3D city model

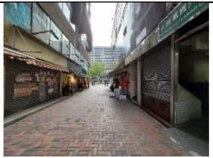
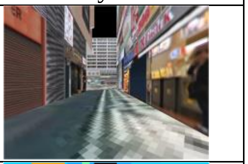
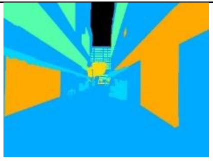
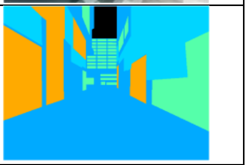
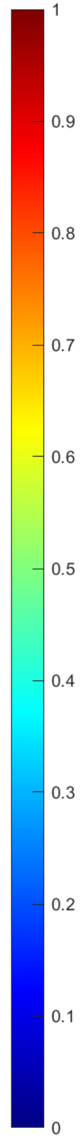
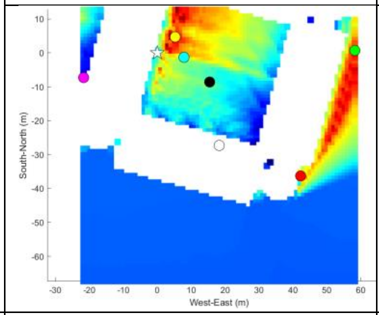
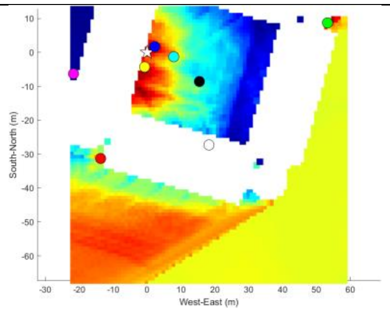
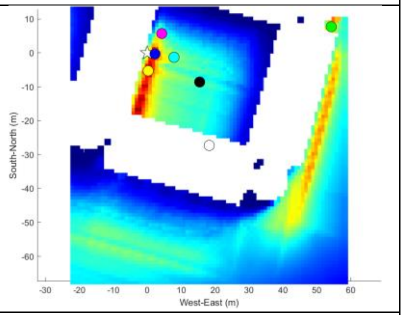
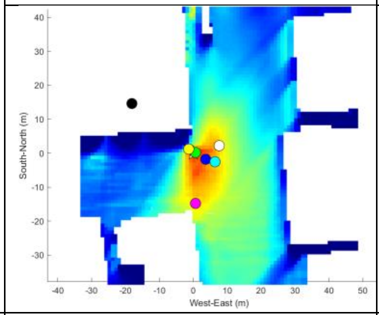
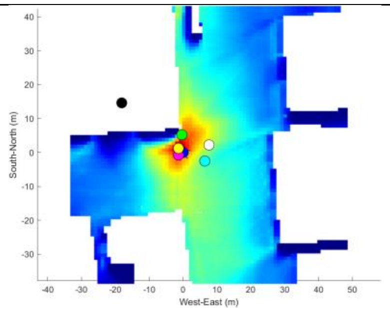
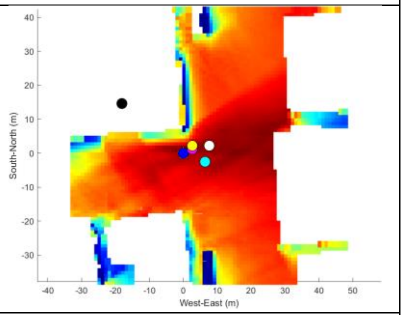
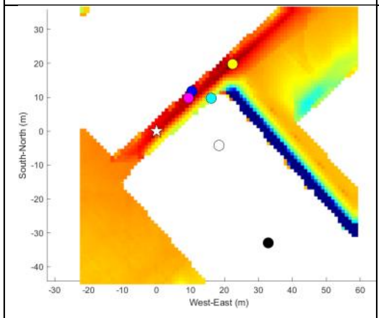
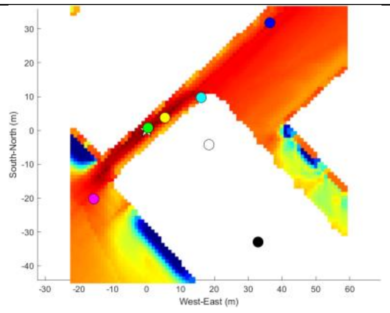
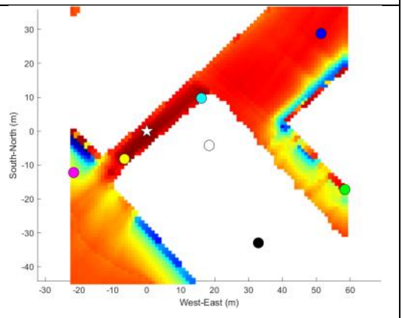
	Reality	3D City Model
Textured		
Labelled		

TABLE V. Heatmap based on likelihood of candidate with the multi-material image registration method

Loc.	Heatmap			
	☆ Ground Truth ● Multi-Material Image Reg. ● Multi-Material Image Reg. (Dice) ● Multi-Material Image Reg. (Jaccard) ● Multi-Material Image Reg. (BF) ● Skyline Matching ● 3DMA ● WLS ○ NMEA			
1	1.1	1.2	1.3	
				
2	2.1	2.2	2.3	
				
3	3.1	3.2	3.3	
				

4. CONCLUSION

This paper proposes a novel multi-material image registration solution for pose (six-DOF) estimation by introducing materials as a new source of information. Provided that the smartphone image segmentation is ideal, our experiments demonstrate that it is possible to outperform existing GNSS and advanced GNSS positioning methods in urban canyons. The experiments show that our approach improves positioning by 45% compared to current state of the art methods and improves the performance of yaw by 8 times compared to smartphone IMU sensors. The pitch and roll estimated by the proposed method, however, achieves a lower performance by half a degree compared to the smartphone IMU sensors. Hence, it is suggested that the proposed method use the already accurate pitch and roll estimated by the smartphone IMU sensors. The elimination of altitude, pitch and yaw estimation will significantly reduce computational load as less candidate images are used for matching.

The limitation of the proposed multi-material image registration comes from inaccurate segmentation. As demonstrated in this research, the 3D model was out of date, leading to discrepancies between the smartphone image and candidate image at ground truth. Therefore, it is necessary to update the utilized 3D city model frequently.

ACKNOLWEDGEMENTS

The authors acknowledge the support of Hong Kong PolyU RISUD fund on the project BBWK, “Resilient Urban PNT infrastructure to Support Safety of UAV Remote Sensing in Urban Region”.

REFERENCES

[1] W. Li, Z. Chen, X. Gao, W. Liu, and J. Wang, "Multimodel Framework for Indoor Localization Under Mobile Edge Computing Environment," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4844-4853, 2019, doi: 10.1109/JIOT.2018.2872133.

[2] Y. Zou, H. Liu, and Q. Wan, "Joint Synchronization and Localization in Wireless Sensor Networks Using Semidefinite Programming," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 199-205, 2018, doi: 10.1109/JIOT.2017.2777917.

[3] R. Sun *et al.*, "Improving GPS Code Phase Positioning Accuracy in Urban Environments Using Machine Learning," *IEEE Internet of Things Journal*, pp. 1-1, 2020, doi: 10.1109/JIOT.2020.3037074.

[4] M. J. L. Lee, S. Lee, H. F. Ng, and L. T. Hsu, "Skymask Matching Aided Positioning Using Sky-Pointing Fisheye Camera and 3D City Models in Urban Canyons," *Sensors (Basel)*, vol. 20, no. 17, Aug 21 2020, doi: 10.3390/s20174728.

[5] "Lands Department." The Government of the Hong Kong Special Administrative Region. https://www.landsd.gov.hk/ (accessed 2020).

[6] Matlab. "Computer Vision Toolbox." The MathWorks Inc. https://www.mathworks.com/products/computer-vision.html (accessed.

[7] E. Calore, F. Pedersini, and I. Frosio, "Accelerometer based horizon and keystone perspective correction," in *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, 13-16 May 2012 2012, pp. 205-209, doi: 10.1109/I2MTC.2012.6229434.

[8] T. A. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *Biol. Skar.*, vol. 5, pp. 1-34, 1948.

[9] P. Jaccard, "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1," *New Phytologist*, https://doi.org/10.1111/j.1469-8137.1912.tb05611.x vol. 11, no. 2, pp. 37-50, 1912/02/01 1912, doi: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[10] G. Csurka and D. Larlus, *What is a good evaluation measure for semantic segmentation?* 2013.

[11] A. Armagan, M. Hirzer, and V. Lepetit, "Semantic segmentation for 3D localization in urban environments," in *2017 Joint Urban Remote Sensing Event (JURSE)*, 6-8 March 2017 2017, pp. 1-4.

[12] H.-F. Ng, G. Zhang, and L.-T. Hsu, "A Computation Effective Range-Based 3D Mapping Aided GNSS with NLOS Correction Method," *Journal of Navigation*, pp. 1-21, 2020, doi: 10.1017/S037346332000003X.

[13] E. Realini and M. Reguzzoni, "goGPS: Open Source Software for Enhancing the Accuracy of Low-cost Receivers by Single-frequency Relative Kinematic Positioning," *Measurement Science and Technology*, vol. 24, p. 115010, 10/16 2013, doi: 10.1088/0957-0233/24/11/115010.