2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Fusing Lidar Data and Aerial Imagery with Perspective Correction for Precise Localization in Urban Canyons

Jonghwi Kim and Jinwhan Kim

Abstract— **This paper addresses a vehicle localization method that fuses aerial maps and lidar data in urban canyon environments where global positioning system (GPS) signals are inaccurate. The boundaries of buildings are extracted from the aerial map and they are matched to point cloud data provided by the lidar. However, most aerial maps contain perspective projection distortions which can be significant in urban canyons with tall buildings. In this study, a new method to correct such projection distortion is proposed and it is applied to precise localization by fusing the corrected map and lidar data. In order to achieve this, the semantic segmentation of an aerial image is performed using a convolutional neural network, and the mutual information between the lidar measurements and the building boundaries is obtained to measure their similarity. A particle filter framework is employed to localize the vehicle and match the map using the mutual information as the weight of a particle. An experimental dataset is then used to validate the feasibility of the proposed method.**

I. Introduction

In recent years, there has been considerable interest in self-driving cars and unmanned vehicles. One of the most important capabilities of self-driving cars is localization, which is the process of determining the position of the vehicle on a provided map. The global positioning system (GPS) supports localization and navigation; however, the accuracy of the GPS depends on the number and distribution of satellites that are successfully detected by the receiver unit. GPS-based locations can be inaccurate particularly in urban canyons because the signals from satellites are occluded and reflected by the surrounding skyscrapers. Thus, other devices such as cameras and lidars are mounted on the vehicle and utilized to obtain information about the surrounding environment, and vehicle localization is conducted by comparing the obtained sensor measurements with provided maps.

A large amount of research concerning vehicle localization in urban environments has been conducted with various combinations of sensors and maps. Aerial and satellite maps have been utilized for some recent research because they are publicly accessible, well-maintained on the internet, and provide detailed features [1]–[6].

Recently, researchers in [1] presented vision-based localization of a ground vehicle using satellite images. An end-to-end neural network was designed to obtain the position using a satellite image and a ground-level image as inputs. Jende et al. [2] addressed a matching approach to register ortho-projected mobile mapping images and airborne imagery. However, the major drawback of using an onboard camera is the difficulty in measuring the relative distance to an object. Moreover, its considerable dependency on light and weather conditions reduces its robustness.

The authors are with the Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea. stkimjh@kaist.ac.kr; jinwhan@kaist.ac.kr

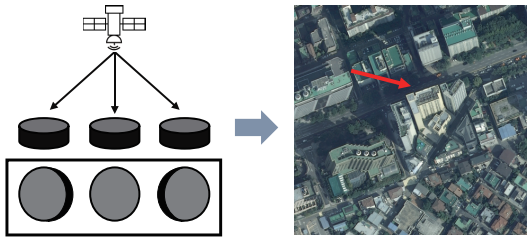


Fig. 1: Illustration of perspective projection distortion in an airborne image. Here, multiple images with different perspectives are stitched into a single aerial map as shown above.

Instead, some studies have adopted a lidar for localization as an alternative to the camera. Kümmerle et al. [3] utilized an aerial map as a prior information in which achieving the global location of the vehicle for compensating the simultaneous localization and mapping (SLAM) errors. The lidar measurements are compared with the edges extracted from the aerial photographs. However, this study did not consider how to manage the perspective projection distortions of the aerial image, which may cause a large error in matching the lidar measurements and the map data in urban environments with tall buildings, as shown in Fig. 1.

In this study, we present a new method for correcting the perspective projection distortion of an aerial map for precise localization in urban canyon areas. The shape of the building is extracted by a convolutional neural network (CNN); the perspective distortion, which renders localization challenging, is corrected by translating the segmented roof to the extracted length and orientation of the wall of the building. The boundaries extracted from the map are then matched with lidar data to maximize the mutual information between the two datasets. Then, a particle filter is applied to conduct localization with map-matching. The key concept of this study is illustrated in Fig. 2.

II. Semantic Segmentation of an Aerial Image

Because the outline of a building is invariant in time and easily detected by lidar, it can be a robust and distinctive feature for use in matching with lidar measurements. We classified all pixels into a small number of classes using a process known as semantic segmentation to extract the boundaries of the building from an aerial image. There have

978-1-7281-4004-9/19/$31.00 ©2019 IEEE

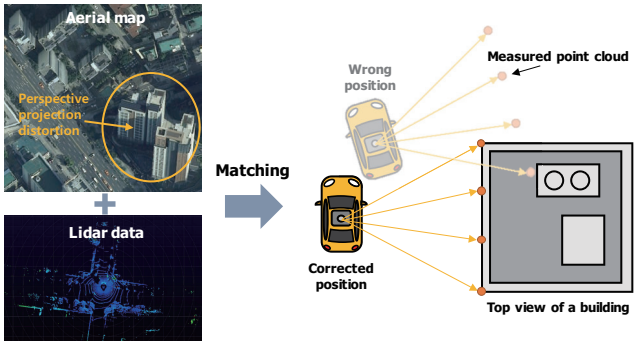


Fig. 2: Illustration of the proposed localization approach.

been multiple studies about semantic segmentation, and we selected the SegNet [7] model as the neural network. The input image is a patch of an aerial image, and it is printed at the same size as the input image through encoder-decoder architecture. The front layers are based on the Visual Geometry Group 16-layer model (VGG-16) [8] which demonstrates good object and feature classification performance, and the back layers restore the spatial information in the segmented output image.

There exist previous studies on semantic segmentation of airborne imagery such as [9] and [10]. However, the majority of the previous studies did not consider perspective projection distortions; thus, they only concentrated on distinguishing buildings from other classes. However, in this study, the building class is subdivided into a roof and a wall to cope with the perspective projection distortion. In addition to the roof and wall classes, the semantic segmentation of the roads and the ground is performed to provide extra information for vehicle localization.

III. Correction of Perspective Projection Distortion

Perspective projection distortion occurs in the majority of publicly available aerial maps and is expressed as occluded road and translated building boundaries. Especially, the maximum distortion in the aerial image used in this study was about 20 m. The degree of distortion can be different in each part of the publicly available aerial maps because ground levels and other factors are calibrated by correcting small patches, which are collaged into a complete image. In addition, the distortions cannot be corrected by applying a single equation to the entire image, because the heights of the buildings are different.

To manage this problem, pixels of buildings are semantically classified into roof and wall classes using SegNet. In order to use valid aerial images, the images are required to undergo a correction process in which the roofs are translated to match the outlines of the walls that intersect the ground, as shown in Fig. 3. More details are introduced as follows.

A. Grouping of Roofs and Walls

The segmentation result is expressed as a group of pixels in a labeled image. We assume that adjacent buildings in the

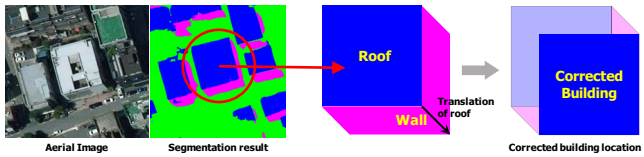
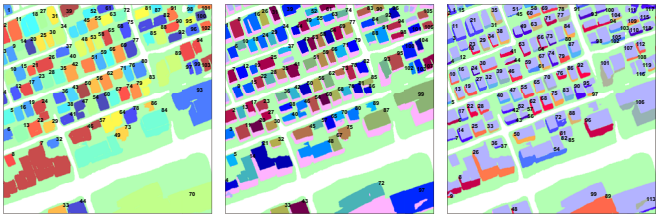


Fig. 3: Illustration of perspective distortion correction.



(a) Clusters of group (b) Clusters of roof (c) Clusters of wall

Fig. 4: Examples of clusters of group, roof, and wall.

aerial images possess the same view direction. Thus, adjacent roofs and walls are classified and clustered as building groups and assigned a number as shown in Fig. 4. After each roof and wall are clustered, they are checked if they belong to the same building. Considering that the buildings were pictured in the same perspective image view, the roof can be matched to multiple walls, but the wall should be assigned to a single roof. Thus, if the wall is adjacent to multiple roofs, it is necessary to accurately select which roof the wall belongs to.

First, the number of roof pixels that are adjacent to the wall is compared with a threshold. If the particular roof pixels adjacent to the wall predominate, this roof is assigned to that wall. If not, based on the assumption that the view to the building is identical within the group, the primary roof-wall direction of the building group is determined by the orientation relationship between the average center position of the already-assigned roof and the wall pixels. Then, the roof that is located in the obtained direction with respect to the wall is assigned to the wall. Fig. 5 depicts the result of this algorithm. The linkage of each roof and wall is represented as a black line, and an identical number on roof and wall represents a classified roof and wall in the same building.

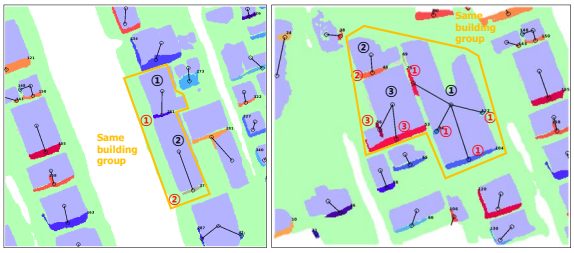
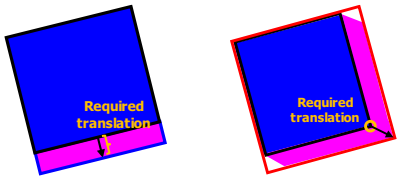


Fig. 5: Assignment of roof and wall.

B. Parametrization of the Roof and Wall as a Rectangle

Because the majority of the buildings in urban environments are designed to be cuboid, the top-view of a building



(a) One-sided wall (b) Two-sided wall

Fig. 6: Classification of wall type as one-sided and two-sided wall and the required translation of each type of wall.

or a roof and wall can be parameterized as a rectangle. A rectangle with the minimum area surrounding each roof and wall cluster is obtained to represent a pixel cluster. The convex hull surrounding the roof and wall is obtained, and the angle between the adjacent vertices of the convex hull, which minimizes the area of the rectangle, is determined to acquire the fitted rectangle [11].

After the roof and wall are parameterized by the rectangle, the wall is classified as a one-sided wall or a two-sided wall depending on the direction and angle of view from above, as shown in Fig. 6. A threshold is applied by checking the proportion of the wall pixel area in the rectangle to classify the wall type. By the type of wall, the length and direction of the required translation is obtained, and it is illustrated as a black arrow in Fig. 6a and Fig. 6b.

C. Estimation of Wall Length and Orientation

Using the parameterized rectangle, the amount and direction of the required translation is determined. For a one-sided wall, the roof needs to be shifted to the intersection between the connected wall outline and the ground. Therefore, a roof adjacent to a one-sided wall is translated by the length of the extracted rectangle from the wall, in the direction of the roof facing the wall. By contrast, a two-sided wall cannot be fully expressed by the rectangle. We emphasize that the farthest vertex of the wall rectangle from the roof is required to correspond to the roof rectangle vertex. The amount and direction of the required translation is defined by the vector between the two vertices, and represented by the black arrow in Fig. 6.

After the roofs are translated by the obtained vector, the map consisting only of corrected roofs is generated for use in vehicle localization. The building boundaries are extracted from the reference map and are utilized as a feature for matching the lidar measurements.

IV. Mutual Information as an Image Matching Metric

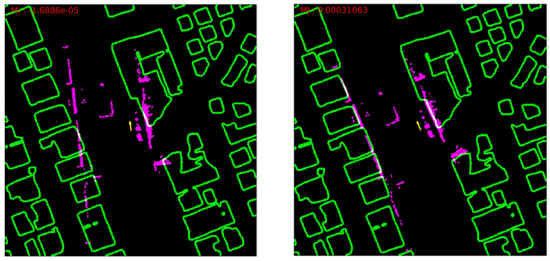
The lidar measurements acquired at the position are required to be matched to the building outlines from the reference map to obtain the position of the vehicle in a provided map. For matching two images, mutual information is used to measure their similarity. The mutual information refers to the measure of the amount of information between two random variables [12]; it is relative to the uncertainty between the two variables. Because of the advantages in removing the effect of outliers and the effectiveness of managing multi-modal sensor data, the mutual information measure has been used for multi-modal image registration. In this study, weighted mutual information (WMI) [13] is used for matching the reference map to the point cloud to manage the problem presented by the majority of the pixels representing background. The process of obtaining WMI between the reference map and the point cloud image is described by

$$WMI(A,B) = \sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)p_{AB}(i,j)\log\frac{p_{AB}(i,j)}{p_A(i)p_B(j)} \tag{1}$$

where A represents the pixel value in the point cloud image, and B represents the gray-scale pixel value of the reference map image. The terms $p_A(i)$ and $p_B(j)$ refer to the normalized histogram of A and B, $p_{AB}(i,j)$ refers to the normalized joint histogram of A and B in the overlapped region, and N is the total number of bins of the histogram. The $w(i,j)$ is the weight matrix for each of the joint variables, and $WMI(A,B)$ denotes the WMI between A and B. The weights are higher in the joint variable where the sensor measurement image overlaps the building line, and they are lower at the overlap of the backgrounds. Thus, the mutual information between two images are maximized when they are well matched and the vehicle is at the correct position, which is expressed by

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} WMI(A,B;\mathbf{X}) \tag{2}$$

where $\mathbf{X}$ is the vector consisting of the position and yaw angle of the vehicle in three-degrees-of-freedom (3DOF). Fig. 7 shows that the mutual information value can be used as a similarity metric of the images. The green pixels represent the map data, and the magenta pixels represent the lidar data.



(a) Low WMI (b) High WMI

Fig. 7: Comparison of map matching result according to the mutual information value.

A particle filter framework is utilized to find $\mathbf{X}$ that maximizes the mutual information. The mutual information of each particle is computed and used as the weight to optimize the cost of mutual information. Therefore, as time progresses, only particles with high mutual information will survive and eventually converge to the correct position.

In addition, extra weights are applied for robustness to the yaw angle by assuming that the heading of the road is comparable to the primary angle of the building outline. The

lines from the segmented road edges are compared with those extracted from the lidar measurements.

V. Particle Filter-Based Localization

A. Prediction Step

Each of the particles, $\mathbf{X}^i$, is propagated according to the 3DOF motion model. The state vector is expressed as $\mathbf{X} = [x \quad y \quad \psi]^\mathsf{T}$ where x and y are the position in global coordinates, and ψ is the yaw angle. The kinematic motion model of the vehicle is defined as

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} v\cos(\psi) \\ v\sin(\psi) \\ z_r \end{bmatrix} + \mathbf{w} \tag{3}$$

where v is the linear velocity and z_r is the yaw rate from the wheel encoder. The zero-mean Gaussian process noise that reflects the uncertainty in the system is denoted as $\mathbf{w}$.

B. Correction Step

The weight is obtained based on the mutual information measurement and heading similarity between the road and the lidar measurement. To reflect the characteristics of the mutual information cost, we designed the weight as

$$w^i = \eta e^{\frac{(z^i_{meas} - z_{min})^2}{2\sigma^2}} \tag{4}$$

where z_{meas} is the cost of mutual information and heading similarity, z_{min} is the minimum cost value of particles, and σ denotes the configurable parameter. Here, η is a value for normalization. This representation is designed with references to previous studies [14], [15]. The particles on the buildings are redistributed, and finally the vehicle position is estimated by computing the weighted mean of the particles.

VI. Experimental Setup

A. Dataset

A vehicle navigation dataset with lidar measurements in urban environments is required to show the feasibility and effectiveness of the proposed method. Thus, we selected the Complex Urban LiDAR dataset [16], which includes the data achieved in urban canyons. For the semantic segmentation of an aerial image, we annotated the aerial orthomap of ground sample distance (GSD) 0.51 m from the National Geographic Information Institute (NGII). Six image patches were used for training SegNet, and the size of the patches was 1,132 × 1,409 in the unit of pixels. The image patch of 256 × 256 was randomly sliced and used for an input to the network, and a six-fold cross-validation was conducted for training.

B. Scenario and Experimental Setup

We used the dataset number Urban01 and Urban02 which contain vehicle trajectories in a metropolitan area of Seoul, South Korea. The proposed algorithm was applied to the part of the dataset where the trajectory passes through a region where the perspective projection distortions appear as shown in Fig. 8. The vehicle position was initialized by a GPS, with initial uncertainty. The number of particles was 500.

Fig. 8: Example patch of an aerial orthoimage from NGII [17] and an example experimental scenario trajectory overlaid on the image.

VII. Results

A. Semantic Segmentation

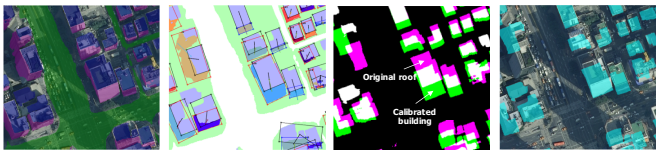
Fig. 9a shows an example of the segmented image overlaid on the aerial image. The building roofs are colored blue, the walls are magenta, and the roads are green. This indicates that the roofs, walls, and roads are segmented well. Table I presents F1 scores over each of the classes.

TABLE I: F1 scores over each of the classes.

Class	Roof	Wall	Ground	Road
F1 score	80.96	63.85	85.06	77.08

B. Perspective Distortion Correction

The process of perspective distortion correction was performed as shown in Fig. 9b. The black rectangle represents a roof, the blue rectangle represents a one-sided wall, and the red rectangle represents a two-sided wall. In Fig. 9c, the calibrated buildings are indicated in green, in contrast with the original roofs, which are portrayed in magenta. The corrected building map is overlaid on the original aerial map to express the result qualitatively.



(a) (b) (c) (d)

Fig. 9: Experimental results of extracting building outlines: (a) semantic segmentation result, (b) process of perspective distortion correction, (c) comparison between the original roof and corrected building, (d) corrected building overlaid on the aerial map.

The intersection over union (IoU) between the actual digital building map distributed by NGII [17] and the corrected building map was calculated to quantify the degree of distortion correction. Fig. 10 shows the example result for the comparison of the corrected and uncorrected map. The magenta map represents the ground truth building shape map, and the green map represents the corrected or uncorrected map. Table II presents the calculated IoU for five randomly

selected buildings and the mean IoU value within 20 building candidates selected near the vehicle path. According to the results, the corrected map reflects the actual building boundaries better than the uncorrected map.

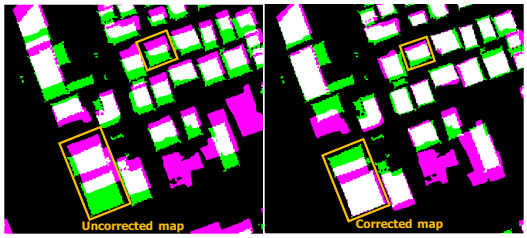


Fig. 10: Example result for comparison of the uncorrected and corrected map.

TABLE II: IoU comparison between the uncorrected map and the corrected map.

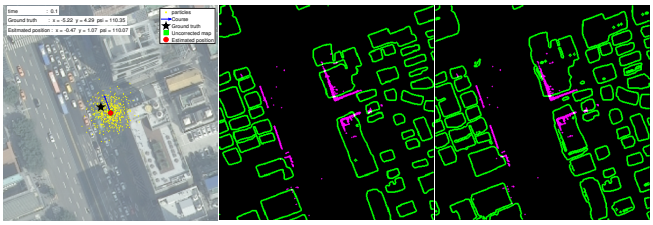
	IoU					Mean IoU
Corrected map	64.6	47.4	71.9	37.3	91.4	57.4
Uncorrected map	37.1	30.5	36.1	38.9	68.8	39.4

C. Map Matching and Localization

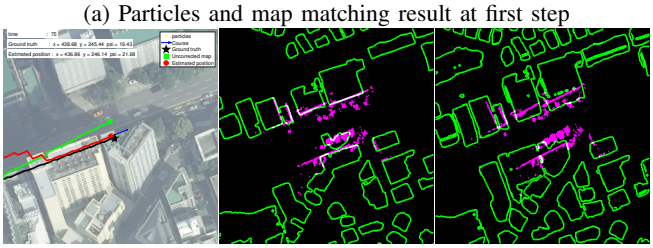
The localization results in the corrected map were compared with the results in the uncorrected map to verify the feasibility of the proposed algorithm. The uncorrected map was generated by extracting the building outlines, which consist of both roofs and walls. The localization results are represented by images that represent the particles and the corresponding lidar measurements on the reference map, as shown in Fig. 11. The left figure displays the particles and corresponding estimated position of the vehicles; the middle and right figures show the matching result between the lidar measurements and the reference map from the corrected map and uncorrected map, respectively. The black star in the left figure indicates the ground truth location of the vehicle. The red circle represents the estimated position of the vehicle by the proposed algorithm, the green square represents the estimated position using the uncorrected map, and the particles are described by yellow points.

At first, the particles are randomly distributed around the initial position with an uncertainty, as shown in Fig. 11a. In Fig. 11b and 11c, the perspective projection distortion is noticeable in that the road is occluded by the building, or the building position is translated. Thus, the localization error certainly appears in the uncorrected map while the error is significantly reduced in the corrected case.

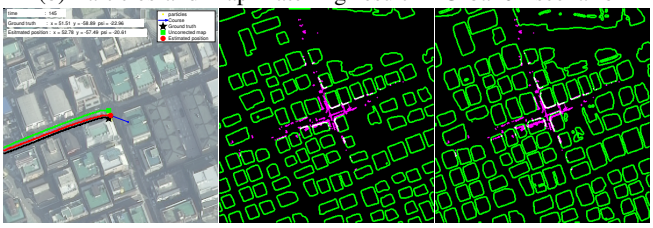
Fig. 12 shows the localization result as the trajectories of the vehicle. The black line denotes the ground truth location of the vehicle. The blue line displays the path generated by dead reckoning, the red line represents the path generated by the proposed algorithm, and the green line represents the path generated by the uncorrected map. The result demonstrates that the red line follows the black line better than the green and blue line. This indicates that both the correction and localization algorithm worked well.



(a) Particles and map matching result at first step



(b) Particles and map matching result in Urban01 scenario



(c) Particles and map matching result in Urban02 scenario

Fig. 11: Examples of distribution of particles, map matching result from the corrected map, and result from the uncorrected map.

In Fig. 13, the localization errors over time are compared. Sudden errors occasionally exist due to the error in correcting building outlines. The convergence of the estimated location of the vehicle is represented while the dead-reckoning error continuously diverges.

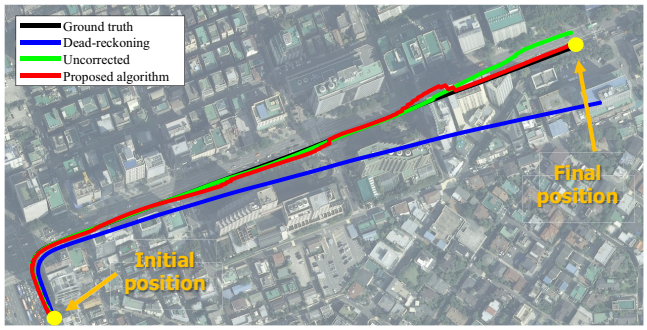
In Table III, the comparison of mean absolute error of position between dead-reckoning, the localization in the uncorrected map, and the proposed method is shown. The mean value of the position difference between the ground truth and the proposed algorithm is lower than that of dead-reckoning and the localization in the uncorrected map.

TABLE III: Mean position error comparison between dead-reckoning (DR), the localization in the uncorrected map, and the proposed method. The localization error of the proposed method compared to the uncorrected map is represented in percentage.

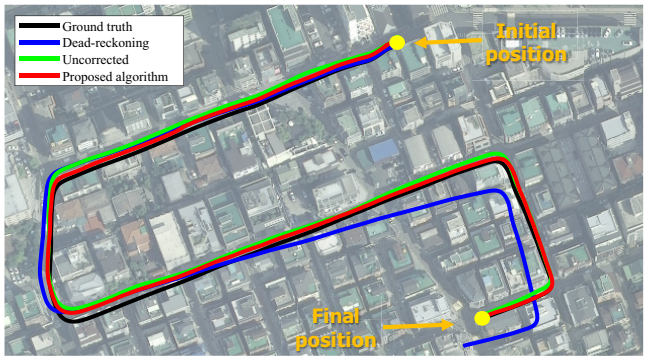
	Urban01		Urban02	
	Dist. [m]	ψ [°]	**Dist. [m]**	ψ [°]
Proposed	2.68(79.8%)	1.63(81.9%)	2.51(58.5%)	1.04(72.7%)
Uncorrected	3.36(100%)	1.99(100%)	4.29(100%)	1.43(100%)
DR	22.23	3.23	8.70	3.17

VIII. Conclusion

In this study, we proposed a vehicle localization method for urban environments using an aerial map and lidar. To



(a) Urban01



(b) Urban02

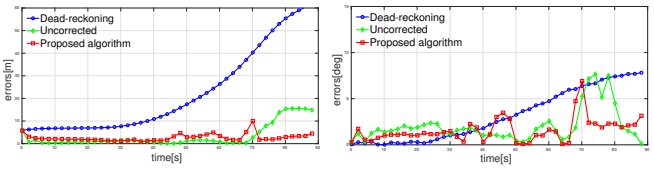
Fig. 12: Comparison of vehicle trajectories by GPS, dead-reckoning, the method without the corrected map, and the proposed method.

utilize an aerial map as a reference map for localization, building outlines were detected by SegNet, which was trained to semantically classify all pixels into four classes — building roof, building wall, ground, and road. After the segmentation was performed, a series of procedures was proposed by correcting the perspective projection distortion that inevitably exists in aerial maps. The positions of the buildings were calibrated in order for the roofs to be translated to match the outlines of the walls that intersected the ground. The lidar measurements and the corrected aerial map were then matched using their mutual information as a similarity measure. The particle filter framework was employed using the mutual information value as the weight to find the position. This procedure was implemented and verified using an experimental dataset. The localization results of the proposed method were demonstrated and compared with the ground truth data, dead-reckoning data, and the trajectory obtained from the uncorrected map.

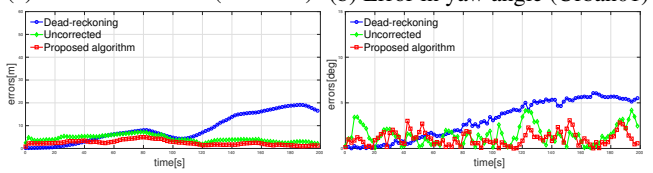
REFERENCES

[1] D.-K. Kim and M. R. Walter, "Satellite image-based localization via learned embeddings," in *IEEE International Conference on Robotics and Automation*. IEEE, 2017, pp. 2073–2080.

[2] P. Jende, F. Nex, M. Gerke, and G. Vosselman, "A fully automatic approach to register mobile mapping and airborne imagery to support the correction of platform trajectories in GNSS-denied urban areas," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 141, pp. 86–99, 2018.



(a) Error in distance (Urban01) (b) Error in yaw angle (Urban01)



(c) Error in distance (Urban02) (d) Error in yaw angle (Urban02)

Fig. 13: Localization error comparison between the dead-reckoning, the localization in the uncorrected map, and the localization in the corrected map.

[3] R. Kümmerle, B. Steder, C. Dornhege, A. Kleiner, G. Grisetti, and W. Burgard, "Large scale graph-based SLAM using aerial images as prior information," *Autonomous Robots*, vol. 30, no. 1, pp. 25–39, 2011.

[4] H. Roh, J. Jeong, and A. Kim, "Aerial image based heading correction for large scale SLAM in an urban canyon," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2232–2239, 2017.

[5] L. de Paula Veronese, E. de Aguiar, R. C. Nascimento, J. Guivant, F. A. A. Cheein, A. F. De Souza, and T. Oliveira-Santos, "Re-emission and satellite aerial maps applied to vehicle localization on urban environments," in *IEEE International Conference on Intelligent Robots and Systems*, 2015, pp. 4285–4290.

[6] M. Javanmardi, E. Javanmardi, Y. Gu, and S. Kamijo, "Towards high-definition 3D urban mapping: Road feature-based registration of mobile mapping systems and aerial imagery," *Remote Sensing*, vol. 9, no. 10, p. 975, 2017.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 140, pp. 20–32, 2018.

[10] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2016, pp. 1–9.

[11] J. D'Errico, "A suite of minimal bounding objects," 2019. [Online]. Available: https://kr.mathworks.com/matlabcentral/fileexchange/34767

[12] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012, pp. 19–20.

[13] S. Guiasu, *Information theory with applications*. McGraw-Hill New York, 1977, vol. 202.

[14] E. R. Arce-Santana, D. U. Campos-Delgado, and A. Alba, "Image registration guided by particle filter," in *International Symposium on Visual Computing*. Springer, 2009, pp. 554–563.

[15] R. Käslin, P. Fankhauser, E. Stumm, Z. Taylor, E. Mueggler, J. Delmerico, D. Scaramuzza, R. Siegwart, and M. Hutter, "Collaborative localization of aerial and ground robots through elevation maps," in *IEEE International Symposium on Safety, Security, and Rescue Robotics*. IEEE, 2016, pp. 284–290.

[16] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban LiDAR data set," in *IEEE International Conference on Robotics and Automation*. IEEE, 2018, pp. 6344–6351.

[17] "National geographic information institute," 2019. [Online]. Available: www.ngii.go.kr