Scene Description from Depth Images for Visually Positioning

Farah Ibelaiden
Computer Science Departement
USTHB University
Algiers, Algeria
fibelaiden@usthb.dz

Brahim Sayah
Computer Science Departement
USTHB University
Algiers, Algeria
brasays@gmail.com

Slimane Larabi
Computer Science Departement
USTHB University
Algiers, Algeria
slarabi@usthb.dz

***Abstract*—Visually positioning of impaired and blind people is still a challenge for computer vision community. The difficulty still remains in the retrieval process which try to find the best place associated to a query image. We propose in this paper a scene descriptor suitable as a filter in the matching process between query and model frames. From a video sequence of depth images, our method based geometry, represents the scene using 2D map. First, we segment the scene from the depth image into planar regions. Using these regions, the application of registration algorithm allows obtaining 3D model of the scene. Next, we determine the ground plane independently from RGB-D camera's pose and we select walls as the lateral and vertical planes to ground. Their projection on the ground gives the boundaries of the scene which are described geometrically. Finally, the descriptor of the obtained 2D map provides a determining selector of place classes. We applied our method on indoor scenes, the obtained results are presented and discussed.**

***Index Terms*—Visual Positioning, Depth image, Scene descriptor, Ground plane, Wall plane**

I. Introduction

In 2015, the estimated number of visually impaired in the world is 253 million. Of these, 36 million blind and 217 million had moderate or severe visual impairment (MSVI) [18]. This disability has attracted the interest of developers and researchers for the help of this category of persons. As result, a range of systems based on GPS, RGB-D camera were proposed for navigation, positioning and understanding the surrounding.

Visual positioning based on RGB-D camera consists to identify the place by comparing images. Extracted features in both acquired frame and image model, used in the matching process, constitute the kernel of any proposed solution because they are determining factor of the association (image-place) due to image deformation.

Our aim in this work is to propose from depth images a geometrical description of scenes that will serve for visual positioning of visually impaired and blind people. In a primary step, this description is suitable as a filter to select less candidates from the data models for a given acquired frame. In a secondary step, the use of descriptors of the state of the art can achieve the place recognition with more accuracy.

The 2D map of the scene is computed from video sequence of depth images acquired by moving RGB-D camera in the scene. For each selected frame, planar regions are extracted and aligned using the registration algorithm to obtain 3D model of the scene. walls are located and projected on the ground to define 2D map which will be described geometrically.

The rest of the paper is organized as follow. Section II is devoted to related works. In section III we present the plans detection method. We explain in section IV the 2D map construction and in section V we describe how to compute and match the scene descriptors for visually positioning. Experimental results are presented in section VI. We conclude this paper by giving the future works.

II. Related Works

Visually positioning which concerns many problems such as visual based localization of the camera, place recognition, have attracted a significant amount of attention in computer vision and robotics communities. It offers the visual ability of a human or robots to recognize visited places and to retrieve the pose of camera.

Visual positioning is important for applications developed for visually impaired and blind people because it indicates where people are and can improve the efficiency of other softwares devoted to the same people community.

The state-of-the-art of Visual positioning methods may be grouped into three categories:

- Methods that only use RGB image [19] [20].
- Methods that use only depth image [21].
- Methods that combine depth image with other input data: like in [22] [23] [24] where authors combined RGB with depth image, in [25] authors combined IR image with depth image.

There is a similarity between visual positioning and image classification in which extracted features are used for image matching. In visual positioning problem, based on useful features, we search images in the database which indicate the same location. To reach this goal, many image representations have been proposed and used to retrieve the place or the camera pose. Indeed, 2D features are extracted from 2D images and employed to retrieve query image in database images afterward pose of query image is approximated to the pose of retrieved image [15].

Locating features and their matching is a crucial step for visual positioning to overcome challenges like viewpoint and illumination changes for RGB camera, and to overcome the noisy data in case of depth sensor like Microsoft Kinect.

The majority of features concern RGB images, and there are few descriptions extracted from depth images, usually they are used combined with RGB features and rarely alone

Features used in state of the art methods are point features, geometric features or combined features:

• Point features :

- SURF feature: is used in [15] to recover the full 6 DOF camera pose by solving PnP problem of determining the position of a calibrated camera in 6 DOF given the 3D location of features and their corresponding 2D image projections.

- SIFT feature: Torii *et al.* [12] use repeated structures in images by finding spatially localized groups of visual words with similar appearance. SIFT feature is extracted and served to adjust feature weights in the bag-of-visual-words model for efficient indexing. For place recognition Torii *et al.* [11] represent RGB images of outdoor scenes using SIFT across multiple scales. This descriptor is suitable for situations where the scene undergoes a major changes in appearance, due to illumination, change of seasons, or structural modifications over time such as buildings built or destroyed.

- ORB (Oriented FAST and Rotated BRIEF) used as descriptor for indoor scenes in [2] due to the fast time computation. In [4], authors match ORB descriptors extracted from input 2D query images to recover the full 6 DOF camera pose by solving PnP problem.

Point features are often not robust enough for localizing across different weather, lighting or environmental conditions. Additionally, they lack the ability to capture global context, and require robust aggregation of hundreds of points to form a consensus to predict pose [16]. As claimed in [7] these point features are not able to create a representation which is sufficiently robust to challenging real-world scenarios.

These critics conducted to another kind of features that are learned features. A recent image based visual positioning schema was proposed in [2] where target image is matched with database images to get the most similar image for localization computing based on combination of learned features (CNN features) and point features ORB. A. Kendall and R. Cipolla treat the problem of localization as regression of pose estimation with deep learning and show how to automatically learn an optimal weighting to simultaneously regress position and orientation [7].

• Geometric features : Indicate primitive geometric shapes and include semantically meaningful information like:

- 3D Planar surface and line segments: extracted from depth images [3] used for place recognition.

• Combined features : uses an approach to pair various descriptors in order to increase the result of the retrieval step like in [10] where authors have used multifeature fusion of D-CSLBP and HOG features. Disparity information, computed from stereo images, is integrated into complete center-symmetric local binary patterns (CSLBP) to obtain a robust global image description (D-CSLBP).

The majority of used features for visual positioning are extracted from RGB images. We are interested in this work to determine from depth images a new descriptor useful for visual positioning. We have make choice to use depth images because of their simple content and their significantly low computational runtime compared to RGB images.

As contribution, we propose to extract the 2D map of the scene. Planar regions are extracted from depth images, registration process allow determining the 3D model of the scene. Walls are detected and projected on the ground to define the 2D map which is described geometrically.

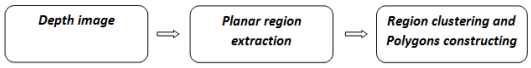

Fig. 1. Planes detection method

III. Planes detection

A. A summary

Fig. 1 summarizes the proposed framework. The first step of our method is the extraction of planar regions. The second step, it consists to construct from depth image the polygons describing areas in the scene which belong to the same plane [14]. For this purpose, we used the method proposed by [13] especially because it offers the possibility to process non convex polygons and polygons with holes.

B. Extracting planar regions

The purpose of this step is to extract all planar regions from depth image. A planar region of the scene is a rectangular area in depth image defined by its left-top pixel, its width and height and contains a set of points belonging to the same 3D plane defined by the equation $ax + by + cz + d = 0$, where the a, b, c parameters defines the normal vector of the plane and the d is the orthogonal distance to this plane from the origin of the reference system. One approach consists to split the depth image into several rectangular regions and verify if each one is planar. Authors in [14] propose a dynamic sized region partitioning method which uses the quad tree algorithm recursively, the idea is to verify if each region is smooth and flat, if it's not then splitting it, starting by the whole image and ending when the region is too small. The depth change indication (DCI) map [6] is used to detect the big changes of depth in the depth image noted I_d. When a pixel have big variation of depth with its neighbors, the region containing it is split. Formally, the DCI is defined by the equation 1.

$$DCI(u,v) = \begin{cases} 1 & \max\limits_{(m,n)\in F} |I_d(u,v) - I_d(m,n)| \leq f(I_d(u,v)), \\ 0 & otherwise \end{cases} \tag{1}$$

where $F = \{(u-1,v),(u+1,v),(u,v-1),(u,v+1)\}$, $f(I_d(u,v))$ is the smoothness threshold function. Assuming R is an intermediate region in the DCI map, to verify the

planarity of an intermediate region R, it is sufficient to verify it's flatness done by the function $Flat(R)$ defined as [14].

$$Flat(R) = \begin{cases} 1 & \begin{cases} MSE(R) < T_{MSE} \text{ and} \\ Curvature(R) < T_{Cur} \end{cases} \\ 0 & otherwise \end{cases}, \quad (2)$$

$$Smooth(R) = \begin{cases} 1 & if |R| = \sum_{(u,v)\in R} DCI(u,v), \\ 0 & otherwise \end{cases} \quad (3)$$

$|R|$ denotes the size of region R The method proposed in [9], [6] and [14], is used to calculate the MSE (Mean Square Error) and the curvature of a region R in depth image by means of covariance matrix [9] which is a low-dimensional descriptor. $Curvature(R) = \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2}$, where $\lambda_0, \lambda_1, \lambda_2$ are the eigen values of covariance matrix (C), in increasing order. If the MSE and the curvature are under given thresholds T_{MSE}, T_{Curv} (representing respectively MSE and Curvature thresholds that are determined experimentally).

then the region R is considered as flat. According to [8], the best plane fitting a set of 3D points $p_i = (x_i, y_i, z_i)^T, i \in 1...n$, is the one having a normal vector equal to the Eigen vector corresponding to the smallest Eigen value of the covariance matrix C.

Noise elimination, using covariance matrix to estimate the region plane normal and its centroid, provides a natural way to largely filter the noise, due to the statistical averaging during calculation [9].

C. Planar regions clustering and polygons construction

So far we have a set of a planar regions, each is defined by its top-left pixel, width and hight and the plane equation. In this section we use these properties to cluster the above regions into distinct groups each one containing regions belonging to the same plane. Extracting planar regions from the depth image is commonly a part of bigger process like navigation, localization, reconstruction ... etc., most of theses processes cumulate the data from successive frames to construct a description of the whole scene in 3D. For this reason we construct a set of planar polygons for each frame, instead of cloud points, we do this for the three main reasons:

- Reduce the amount of data: a polygon represents with height fidelity all points in its perimeter which is a little set of vertices.
- Polygon is a standard Geometric primitive in 3D: Manipulating polygons (intersection, union, area, tessellation) is more common practice than manipulating planar regions, in fact there is lot of efficient algorithms treating those operations references.
- Reduce the execution time because the manipulation of polygons defined by normals and barycenters is much more efficient than manipulation of all points contained in their area.

We choose clipper [13] for its ability and efficiency to treat non convex polygons and polygons with holes, and the availability of its open-source implementation. To simplify the calculation we do polygons construction on the depth image, then transform the resulting points to 3D reference system using the intrinsic parameters of the camera.

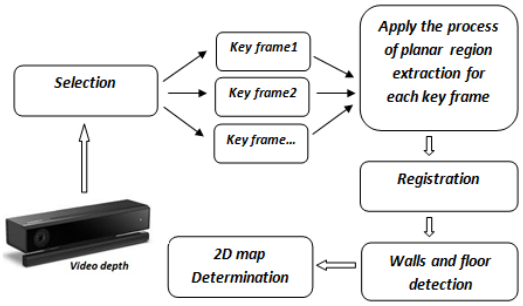

Fig. 2. 2D map construction method

IV. 2D MAP CONSTRUCTION

Fig. 2 represents diagram of 2D map construction method.

A. Key frame selection and Planar region extraction

Is an essential step in which we will select a keyframe every "k" frames for the two main reasons:

- Reduce accumulated error (that increase with the increase of number of frames used in alignment process).
- Reduce execution time which is an important criterion because our system is destined to real time.

In the next, we apply the process of planar region extraction presented previously for each key frame.

B. Registration

To obtain 3D model of the scene, we use the "Iterative Closest Point" (ICP) with the planar areas extracted from each pair of successive images. After the calculation of the geometric transformation that links two images. polygons of the second image will be transformed. In the end, we merge polygons of the whole model that belong to the same plane.

C. walls and floor detection

To be able to extract walls from scene captured with a depth camera we must differentiate walls, floor, objects planes and ceiling. We used in this work the geometrical method proposed in [1] in order to locate the ground. We consider lateral polygons having big areas and normals with small Y component as walls, we use a red-black tree [5] which is an auto-balanced tree to keep polygons meeting the criteria already established in descending order according to their area.

For simplifying the calculation of a polygon's area, we are doing it in 2D rather then 3D, to do so, we are projecting the vertices of each polygon on its normal plane, thus we eliminate one coordinate from them, while keeping the same polygon's areas. More details are given for area computation in [13].

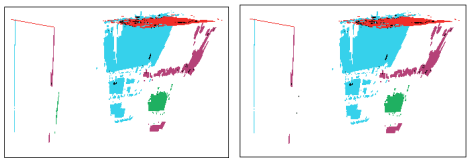

Fig. 3. (Left) An example of encountoured problem while projecting polygons representing walls on the floor plane. (Right)The projection of polygons representing walls on the floor plane considering the polygons having largest surfaces.

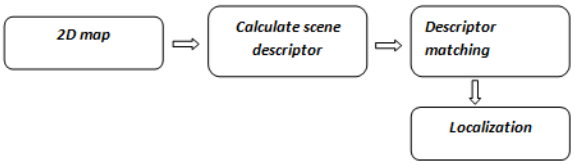

Fig. 4. The Visually positioning process

D. 2D map construction

The projection of polygons representing walls on the floor plane give points.

Fig. 3 illustrates an example of the problems that we have encountered when projecting polygons representing walls (caused by accumulated error and the noisy data of the kinect). That we have solved it by considering the polygons having the largest surfaces as being the walls. For example here in Fig. 3 we notice that:
- Points aligned in Blue represents the projection of polygon (blue) representing the first wall;
- Points aligned in Red represents the projection of the polygon (red) representing the second wall;
- Points aligned in Green and purple represents the projection of the two polygons (green) and (purpule), in this case we consider the purple polygon as being the wall because its surface is greater than the surface of the green polygon.

In the same Fig. 3, we can see the result of the projection after resolving the pre-quote problem.

V. Visually positioning

Fig. 4 summarizes the Visually positioning process. It consists to calculate the descriptor of the scene using projection results of the previous step in order to compare it against descriptors of dataset models. Finally the model having a minimum score in the dataset is considered as appropriate to the query model.

A. The proposed descriptor

The scene is then represented by a 2D map where lines represent the boundaries of walls in the scene. In order to recognize places based on this representation, we propose the extraction of discriminant features suitable for partial shape matching [17]. Let $\{P_i, i = 1..n\}$ be the set of located corners, two corners P_{i-1}, P_i define the segment S_i. Fig. 5 shows an example of the 2D map of scene. Each corner P_i is described

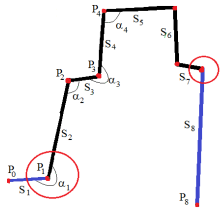

Fig. 5. Example of 2D map computed for a indoor scene. The points P_0, P_8 delimiting the 2D map does not correspond to corners, the lengths of segments S_1, S_8 (in blue color) are then not defined. The starting point will be P_1 or P_7 indicated by a red circle.

using the attributes (α_i, rl_i):
- The inner angle α_i between the connected segments S_i, S_{i+1},
- The lengths l_i of the segment S_i.

For the first and last segments, the corner points may be not detected and then the lengths of associated segments are insignificant. The values of rl_i for S_1 and S_n are set to zero and not considered in the matching process. The descriptor D_q of the 2D map is then a concatenation of all corners descriptors:
$D_q = \{P_i(\alpha_i, l_i), i = 1..n\}$, where n is the number of considered corners. Note that in the same way, the dataset contains the description of each place and its identification. The description associated is built in the same manner as it is done for the descriptor computed from a video sequence of depth images. The unique difference is for the descriptor model all corners points are considered and each one may correspond to starting point of query descriptor.

B. 2D map matching

We assume now we have a dataset which contain for each location a shape descriptor model associated to whole place. Retrieve the query descriptor in the set of model descriptors is a the problem of partial matching. It consists to associate a part of a shape to a part of another shape. In this case, the starting point of model descriptor for the comparison is unknown, therefore we need to explore all the starting points.

Let $D_q = \{(\alpha_i, l_i), i = 1..n\}$ be the descriptor of the shape query, where n is the number of corners. Let $D_m = \{(\alpha_j, l_j), j = 1..m\}$ be the descriptor of the shape model, where m is the number of corners. Let $S(D_q, D_m)$ be the similarity measure between the two descriptors D_q and D_m computed as given by equation 4. The comparison starts from the corner P_{j0} for the shape model.

$$S(D_q, D_m) = \sum_{i=1}^{n} \sum_{j=j_0}^{j_0+n} \omega_2 * \|\alpha_i - \alpha_j\| + \omega_1 * \|l_i - l_j\| \quad (4)$$

The weights ω_1, ω_2 are associated to each feature according to the importance of feature in the determination of the similarity measure. The weights were chosen on the basis of the experimentation where ω_2 is smaller than ω_1 because dimensions have more importance than angles. This measure is computed for all possible starting point j_0 and for the two directions (clockwise, anticlockwise). The best measure

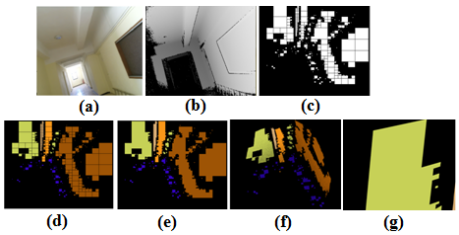

Fig. 6. Planar region extraction.

is selected and saved for the considered 2D map model. This computation is repeated for all candidates descriptors. Finally, the model having a minimum difference measure in the dataset is considered as appropriate model to the query.

VI. Experiments

A. Building the knowledge database

To collect the required data for our dataset that served to test our system of visual positioning, we recorded in our university and in indoor scenes at different locations video sequences of depth images using an RGB-D camera. For each place, we computed the 2D map and the associated descriptor is saved in addition to the identification of the place. We applied our method to compute the descriptor from video sequence of depth images, and the RGB-D camera is moved sufficiently to capture all required data. Note that RGB images were not used at any stage of our processing, we have added them just to be able to visualize the scene. We have tested our system using the two RGB-D cameras (Kinect v1 and V2).

B. Results of planar region extraction

The Fig. 6 represents details of the planar region extraction process where:(a) represent the RGB image. (b) Depth image we notice that the pixels farther from the camera have darker color, while black color represent absence of objects on this point. (c) Is the representation of the planar regions in the 2D space of the image where we can see that each region has a position, width and height. (d) Illustrates the merge of regions belonging to the same plane, a random color has been assigned for each sub-set representing the same region to distinguish it. (e) represents the result of merge in 2D space. (f) represents the result of merge in 3D space. (g) Is a part of polygon built from subset of coplanar region, in this step we build a set of polygons for each captured depth image and we delete all the data relating to the points in order to release memory.

C. Results of registration process

Fig. 7 illustrates the results of registration process. As the camera moves forward, new polygons representing the large plan regions, are built and merged with the previous polygons, so that at the end we will get the 3D model of the scene.

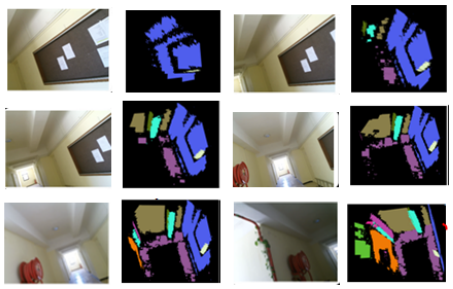

Fig. 7. Registration process.

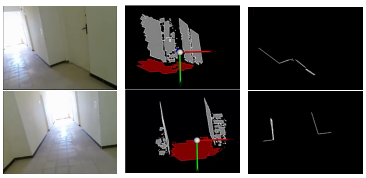

Fig. 8. From right to left: Processed frame, the planes located in 3D: floor (red color), walls (white color), projected wall on the floor plane

D. Results of walls detection

Fig. 8 illustrates the obtained results using the proposed framework. Indoor scene has served as tested data. Our method runs in real time and allows detecting floor and walls. The projection of walls on floor gives the lines describing the scene boundaries. Fig. 9 represents an other example of walls detection:
(a) is the RGB image representing the 3 detected planes that's shown in (b) with their projection in (c). Fig. 10 shows the projected walls and the ceiling plane located as the highest parallel plane to the ground. Note that even the ground in this frame is not seen, the data of previous frames is used.

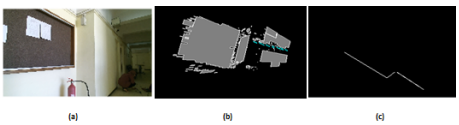

Fig. 9. From right to left: Frame processed, the plane located in 3D, their projection.

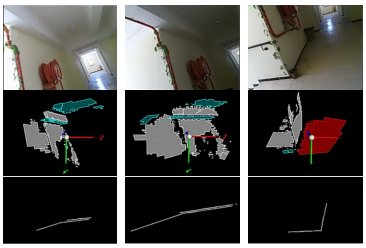

Fig. 10. Processed successive frames, the planes located in 3D: ground for the third frame (red color), ceiling (blue color in the first two frames), walls (white color), projected wall on the floor plane. The inner angle for this case is equal to 270○.

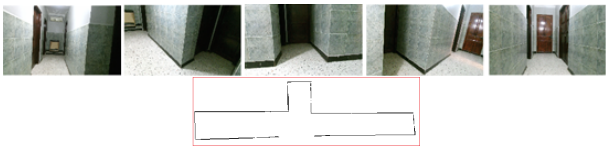

Fig. 11. The 2D map computed of the first indoor place from depth images.

Fig. 12. RGB images representing the second location of indoor scene and its corresponding 2D map.

E. Results of 2D map construction

Fig. 11 represents the 2D map computed from set of depth frames taken in the first indoor location. Here we notice that the 2D map contains an aperture because the video sequence of the depth frames did not capture that part of the scene.

Fig. 12 the 2D map constructed from sequence of depth images taken in the second location of our indoor scene with some RGB images describing the location.Fig. 13 illustrates the computed 2D map of the two places of indoor scene.

VII. Conclusion

We proposed a new scene geometry based descriptor of the scene computed from video sequence of depth images acquired by moving RGB-D camera in the scene. For each selected key frame, planar regions are extracted and aligned using the registration algorithm to obtain 3D model of the scene. The located walls are projected on the ground plane and define the 2D map of the scene. A geometrical description of the 2D map is proposed to be used as first filter in order to select a restrictive data representing the candidate places. When the places have a discriminant geometry, our descriptor facilitates the comparison of query and model images using extracted features from images. As future works we plan to build datasets with high number of places and to study the possibility to include the information of windows and door in the 2D map in order to have a more discriminant descriptor.

References

[1] Chayma Zatout, Slimane Larabi, Ilyes Mendili, Soedji Ablam Edoh Barnabe. Ego-Semantic Labeling of Scene from Depth Image for Visually Impaired and Blind People.ICCV 2019-EPIC. Seoul, November 2, 2019.

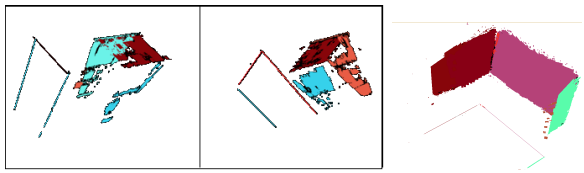

Fig. 13. From right to left: 3D reconstruction and the 2D map computed for the first location of indoor scene from different positions, 3D reconstruction and 2D map computed for the second location.

[2] Chen, Y., Chen, R., Liu, M., Xiao, A., Wu, D., Zhao, S.: Indoor visual positioning aided by cnn-based image retrieval: Training-free, 3d modeling-free. Sensors 18(8), (2018).

[3] Cupec, R., Nyarko, E.K., Filko, D., Kitanov, A., Petrović, I.: Place recognition based on matching of planar surfaces and line segments. The International Journal of Robotics Research 34(4-5), pp. 674-704 (2015).

[4] Feng, G., Ma, L., Tan, X.: Visual map construction using rgb-d sensors for image-based localization in indoor environments. Journal of Sensors, 2017.

[5] Hinze, R.: Constructing red-black trees, pp. 89-99, (Sept 1999).

[6] Holzer, S., Rusu, R.B., Dixon, M., Gedikli, S., Navab, N.: Adaptive neighborhood selection for real-time surface normal estimation from organized point cloud data using integral images. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2684-2689, (Oct 2012).

[7] Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5974-5983, (2017).

[8] Poppinga, J., Vaskevicius, N., Birk, A., Pathak, K.: Fast plane detection and polygonalization in noisy 3d range images. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3378-3383, (Sep 2008).

[9] Porikli, F., Tuzel, O.: Fast construction of covariance matrices for arbitrary size image windows. In: 2006 International Conference on Image Processing. pp. 1581-1584, (Oct 2006).

[10] Qiao, Y., Zhang, Z.: Visual localization by place recognition based on multifeature (d-λlbp. Journal of Sensors, 2017.

[11] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1808-1817, (2015)

[12] Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 883-890, (2013).

[13] Vatti, B.: A generic solution to polygon clipping. Commun. ACM 35(7), pp. 56-63, Jul 1992.

[14] Xing, Z., Shi, Z.: Extracting multiple planar surfaces effectively and efficiently based on 3d depth sensors. IEEE Access, 2018.

[15] Yongshik, M., Soonhyun, N., Daedong, P., Chen, L., Anshumali, S., Seongsoo, H., Krishna, P.: Capsule: A camera-based positioning system using learning. 2016 29th IEEE International System-on-Chip Conference (SOCC) pp. 235-240, 2016.

[16] Zeisl, B., Sattler, T., Pollefeys, M.: Camera pose voting for largescale image-based localization. In: In International Conference on Computer Vision (ICCV) (2015)

[17] Saliha Bouagar, Slimane Larabi. Efficient descriptor for full and partial shape matching , Multimedia Tools Appl. 75(6), pp. 2989-3011 (2016).

[18] Global Blindness and Visual Impairment Data 2015. The international Agency for the Prevention of Blindness. 2015.

[19] Knopp, Jan and Sivic, Josef and Pajdla, Tomas. Avoiding confusing features in place recognition. European Conference on Computer Vision, pp. 748-761, 2010.

[20] Chuang, Chi-Hung and Chen, Ying-Nong and Fan, Kuo-Chin. Image Feature Point Matching for Indoor Positioning. International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). pp. 139-140, 2017.

[21] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence. 37(9), pp. 1904-1916, 2015.

[22] Pujar, Karthik and Chickerur, Satyadhyan and Patil, Mahesh S. Combining RGB and Depth Images for Indoor Scene Classification Using Deep Learning. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). pp. 1-8, 2017.

[23] Song, Xinhang and Jiang, Shuqiang and Herranz, Luis and Chen, Chengpeng. Learning Effective RGB-D Representations for Scene Recognition. IEEE Transactions on Image Processing. 28(2), pp. 980-993, 2019.

[24] Chuhang Zou and Zhizhong Li and Derek Hoiem. Complete 3D Scene Parsing from Single RGBD Image. CoRR 2017.

[25] Zheng, Yali and Luo, Peipei and Chen, Shinan and Hao, Jiasheng and Cheng, Hong. Visual search based indoor localization in low light via rgb-d camera. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering. 11(3), pp. 349-352, 2017.