A Benchmark for Visual Positioning from Depth Images

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

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

***Abstract*—In the last years, there has been an increasing interest in research on visual positioning and localization. Several datasets have been published recently. However, little attention has been paid to architectural aspects of scenes that have the advantage of being not influenced by scenery changes.**
Based on depth videos recorded using depth sensor, we propose in this paper a new dataset composed by descriptors of 2D maps associated to computed 3D structures. We show also how to interrogate the dataset using a set of depth frames acquired in visited places. Conducted experiments show the accuracy of the obtained 2D maps and feasibility of the proposed framework.

***Index Terms*—Scene descriptor, 2D map, Visual positioning, Depth videos, Place recognition**

I. Introduction

The RGBD datasets proposed for visual positioning may be classified into two main classes. The first for unsorted list of images intended to evaluate methods that aim to recognize previously visited places [1], [2]. While the second class, includes images of 3D models, which is reserved to assess systems that estimate the 6 degrees of freedom (6DOF) pose of a camera [3][4]. However, all these datasets are mainly aimed at the assessment of visual localization systems on the basis of features extracted from scene images. While the scene architecture has been neglected despite offering a stable overall description of scene which is not influenced by decor changes.

Our objective through this paper is to suggest a framework for the building of an architecture-based scenes descriptors dataset for visual place recognition. To build this dataset, depth videos are acquired by moving Microsoft Kinect sensor in different indoor and outdoor scenes. For each scene, the associated 3D model is computed and the corresponding 2D map is derived and described defining the scene descriptor.

The rest of the paper is organized as follow. In section II, we highlight the different types of visual positioning datasets. The details related to dataset building and scene description are given in III. Section V is devoted to conducted experiments and evaluation. We conclude the paper with some perspectives in VI.

II. Related Works

There is plenty of informations about the environment and numerous types of data that could be extracted (set of frames, 3D reconstructed models, colored points cloud...) in order to be used in localization process.
We can summarize the proposed methods of localization according to the fixed goals on two main classes:

- The first class defines the problem of localization as problem that captures the visual ability of humans and robots to recognize visited places. It can be cast to image retrieval strategy which matches query image with database images and cast the position of query image to the pose of retrieved image [19][7]. So the principal processed information of this class is images.
- The second class includes methods that aim to retrieve 6 DoF position of camera using 3D models that constitute the principal processed data [4][11].

Due to this significant difference between the two classes of visual positioning methods, two different categories of datasets have been published in order to reach the targeted goals using the appropriate information to each class of systems. We can summarizes these datasets on:

- Datasets containing unsorted list of images: often furnish query images acquired under different conditions compared to the database images. They can be used to study the effect of changing conditions on results of visually positioning [6] [9] [10] [16] [27] [15] [13]. This type of dataset is more appropriate to the first class of methods. As examples, we can cite: (1) The Pittsburgh dataset [36] including 254k perspective images from approximately 10.6k panoramas of Google Street View (which leads to a rather large distance between the panoramas locations)[16]. (2) The San Francisco Landmark dataset in [35] was created to encourage the research in landmark recognition with mobile devices. It contains 1.7 million images of buildings in San Francisco. It also includes 803 images taken with a variety of different cameras dedicated for evaluation. The generation process utilizes vehicle-mounted cameras with wide-angle lenses to capture spherical panoramic images. For all visible buildings in each panorama, a set of overlapping perspective images is generated. (3) INRIA Holidays Dataset contains 1491 images.

composed of 500 sets of similar images. Each image set contains 1 query, a total of 500 query images [13].

- Spatially consistent datasets: composed of 3D models and ground truth poses [12] [21] [22] [23]. They do not consider big changes between query and database images due to relying on feature matching for ground truth generation. As examples we can cite: (1) The Cambridge Landmarks dataset [12] is large scale outdoor visual relocalization dataset, it contains 12K images with full 6 DoF camera poses and visual reconstruction of the scene. (2) The Rome and Dubrovnik datasets in [8] contain 3D reconstructed models of some of the most notable landmarks in Rome and Dubrovnik. The Rome dataset has 3D models for 69 different sites generated from images taken by distinct cameras. The Dubrovnik dataset has only one 3D model for one landmark. (3) The first RGB-D dataset for Simultaneous localization and mapping (SLAM) benchmarking was proposed in [17] in which both the RGB-D data produced by a Kinect and the ground truth camera poses estimated by motion capture system was recorded. After that authors have extended the benchmarking of the visual SLAM in[18] [20] by exploring more tests on two different scenes with different trajectories.

Historically, this kind of datasets were restricted to:
- Large-scale indoor : covering multiple rooms or even whole buildings; as proposed dataset in [24] that contains 6 DOF poses for large scale indoor localization and query photographs captured by mobile phones at different time than the reference 3D map.

- Semantic scene understanding: like in [25] where authors have proposed a dataset including five large-scale indoor areas from three different buildings showing diverse properties in architectural style and appearance.

Through this paper, we seek to propose a new dataset containing scene descriptors calculated using 2D maps (representing architecture of scenes) in order to give access to explore these architectural features in visual positioning process [34].

Note that none of the state of the art datasets can be used to evaluate localization systems based on coarse information of scenes because both first and second class of visual positioning datasets contain extracted features from images which don't describe global information of scenes.

So our contribution is to propose a suitable dataset from depth videos for the localization systems based on global information of scenes.

III. Data aquisition and 3D structure Computation

A. Data acquisition

In order to get high quality of RGBD data we have chosen "kinect V2" to construct our dataset [5]. The videos was acquired by moving a kinect attached to cart in whole scenes intended for dataset, so that each scene will be fully covered in its corresponding video. We note that for scenes of small dimensions and simple structure we have simply surrounded the kinect so that scenes will be completely covered like it is shown through figure 1 (Top); whereas concerning scenes

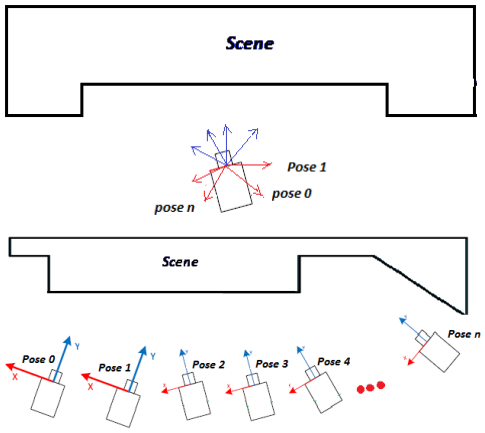

Fig. 1. Cumulation of local transformations in case of (Top) scenes of small dimensions and simple structure (Bottom) scenes of large dimensions and complex structure

of large dimensions and complex structure we have applied rotational and translational movements on the kinect to cover all details of the scenes see figure 1 (Bottom).

We have also acquired query depth videos in order to compute query descriptors that cover parts of scenes for evaluation. We notice that we have considered scenes with different geometric areas (Rectangle, Square, T, L and N shapes) with different dimensions (small $(10m^2)$, medium $(30m^2)$ and large scenes $(55m^2)$). A simplified plan of the working environment is presented in figure 2 for a better explanation.

B. 3D structure computation

The recorded depth video of a given scene is used to compute the 3D structure. Key frames are selected and their plans are identified and aligned using alignment algorithm to construct a complete 3D map of the scene. Figure 3 summarizes the process of 3D structure computation.

The acquired depth video was cut into set of frames to which [34]:
- The pre-processing step has been applied: that includes: (1) Key-frame selection within a fixed distance interval, this reduces execution time and minimizes data storage space. (2) Smoothing selected frames using median filter.
- The construction of polygons by grouping planar regions belonging to the same plane in each key frame has been applied by: (1) Extracting all planar regions considered as rectangular areas in depth image. (2) Clustering planar regions into distinct polygons grouping regions belonging to the same plane.

The extraction process starts by considering the whole depth image which is splitted into several rectangular regions using the quad tree algorithm recursively [37]. Then the smoothness [28] and flatness [29] tests were verified for each region. The split ends when the region is too small.
The smoothness test is assured by calculating the depth change indication (DCI) map [28] used to spot the big variations

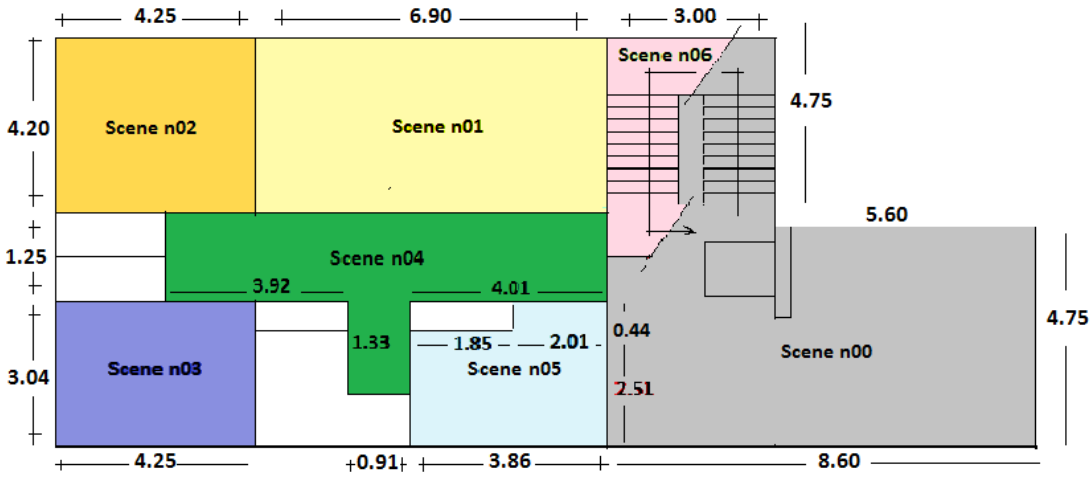

Fig. 2. An example of used indoor working area

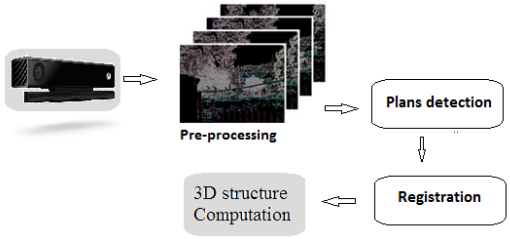

Fig. 3. The 3D structure computation steps

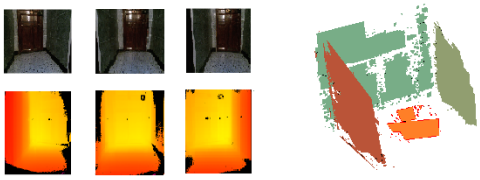

Fig. 4. (Left) Some RGB and depth frames, (Right) Computed 3D structure

of depth in the depth image. So when a pixel hasn't a big diference of depth with its neighbors, the region including it is considered as smooth otherwise it will be splitted.

The resulting polygons of selected key frames were aligned to construct a whole map of scene: Using a registration algorithm [38] which will be used to calculate the geometric transformation that links two successive frames (local transformation). As the coordinate system of the first key frame (defined by the first camera pose) is considered to be the global coordinate system of the scene. The global transformation of a frame is defined as the product of all previous transformations like it is shown in figure 1.

So the calculated transformation is written as matrix including the rotation and the translation. The obtained geometric transformation is used to transform frame polygons and merge all polygons belonging to the same plane; in order to get a complete map representing the scene. Figure 4 illustrates a 3D map constructed from a set of selected depth frames. As the RGB images show, the processed scene is of low light to affirm that the proposed system is precise, even in the case of low light scenes.

IV. 2D MAP COMPUTATION AND DESCRIPTION

A. The 2D map computation

The ground was located using the geometrical method proposed in [31] and the perpendicular polygons with large areas were identified as walls. The calculation of polygon's area has been done in 2D space rather then 3D space (to simplify calculation) by projecting vertices of each polygon on its normal plane and thus we eliminated one coordinate from them, while keeping the same polygon's areas. More explanations are given in [33]. In figure 5 (Top) the 3D map of the scene of figure 4 is shown after walls detection as lateral polygons colored with blue. 2D map representing boundaries of scene area are defined as the projection of walls on the ground (see figure 5). The additional contours around corners, due to noise, are removed by considering intersection points as limits of each segment. We highlight another benefit of selection step which consist on reduction of alignment errors (which increase with the increase in the number of aligned frames) like it is illustrated by figure 5.

B. The 2D map Description

The resulted 2D maps are described based on the geometry of the boundaries using proposed method in [34].

The descriptor D_s is defined as: $D_S = \{\{P_i(\alpha_i, l_i), i = 1..n\}, Location\}$, Where:

- $\{P_i, i = 1..n\}$ is set of located corners in the counter anticlockwise direction.
- S_i is a line segment delimited by the successive corners (P_i, P_{i+1}).
- n represents the number of considered corners.
- Each corner P_i is described with α_i (the angle between

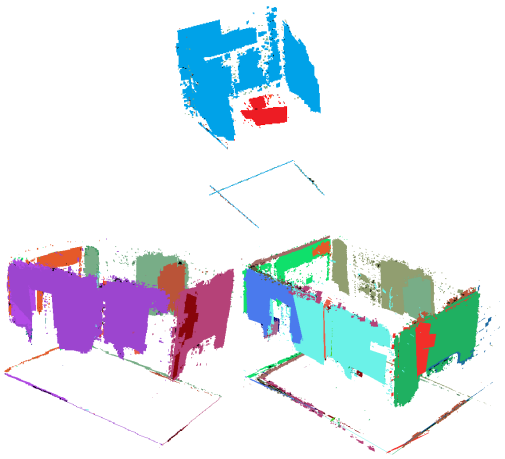

Fig. 5. From Top to bottom: The computed 3D map and associated 2D map, Maps computed with selected frames (left) and using all video frames (right)

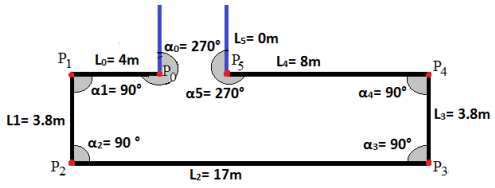

Fig. 6. 2D map calculated for a indoor scene

S_i, S_{i-1}) and l_i (length of segment S_i).
- Location as its name suggests, it denotes the identification of scene.

The corresponding descriptor to the shown 2D map in figure 6 is:
$D_S = \{\{P_0(270,4), P_1(90,3.8), P_2(90,17), P_3(90,3.8), P_4(90,8), P_5(270,0)\}, Scene_{example}\}$. As it is shown the segment S_5 is not demarcated by two corners so the value of its length l_5 is set to zero because it is insignificant.

Algorithm 1 recapitulates the dataset scene descriptor computation.

Data: Depth_video
Result: Dataset_descriptor
while *!end(Depth_video)* **do**
  Planar regions extraction from current frame f_i;
  Clustering each group of planar regions belonging to the same plan into polygon in f_i;
  Align f_i with 3D model constructed with previous frames;
end
2D map computation;
Dataset_descriptor calculation;

Algorithm 1: Dataset scene descriptor computation algorithm

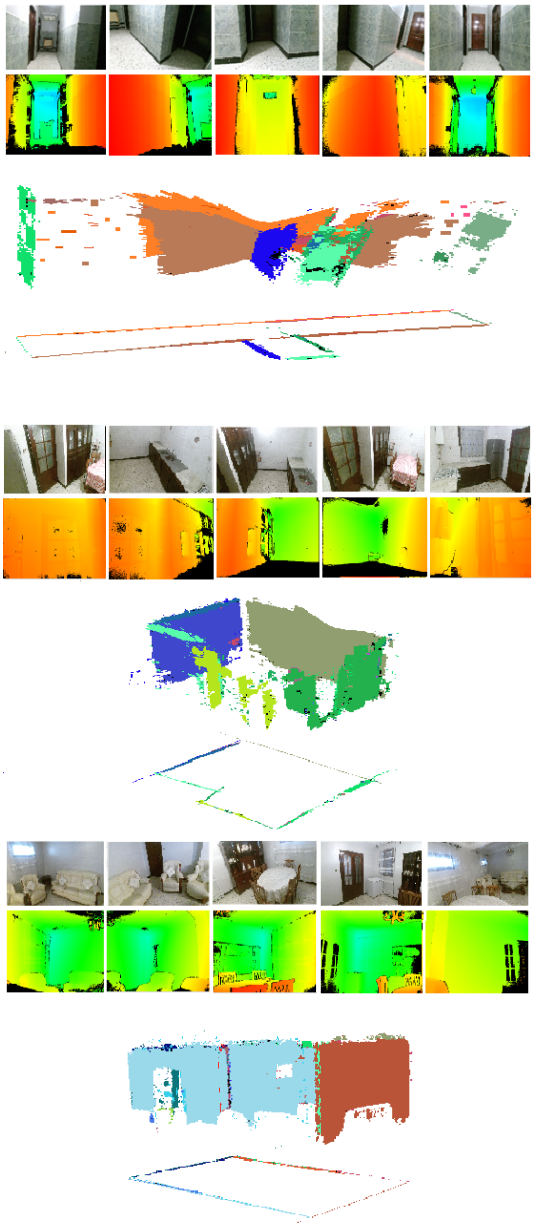

Fig. 7. For each scene, from top to bottom: Some of RGB images, their associated depth images, the computed 3D structure and 2D maps

V. Experiments and Evaluation

A. Building the dataset

1) Data acquisition: As explained in section IV the 2D map computation process passes through several stages, starting by frames selection of a recorded depth video, reconstruction of 3D structure and the associated 2D maps. Figure 7 illustrates the 3D structures and the 2D maps obtained for some indoor scenes.

To study the exactness of 2D map computation, we compared the relative lengths and angles with the known values of the ground truth data. Graph of figure 8 shows the average

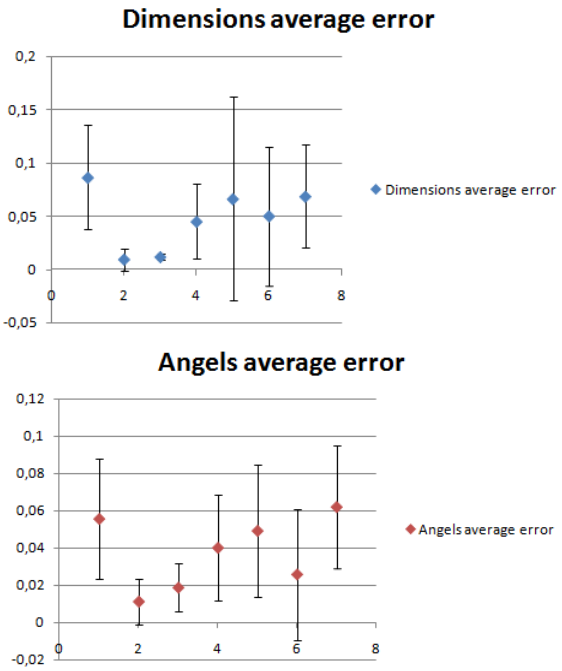

Fig. 8. Average relative error and standard deviation for computed lengths and angles

error and the standard deviation for relative length and angles for a sample of used scenes. The obtained results confirm that the computed 2D maps are accurate and constitute a useful feature to represent the area delimiting the observed scenes.

2) Dataset components: Once the 2D map of each scene is computed, it is described using the geometrical method given in [34]. The descriptor associated to each 2D map of given 3D scene is characterized by the low storage required. For example, considering the example of the 2D map shown by figure 9 (α_i for angles and L_i for dimensions of segments lines) , the descriptor is written as:

$$D_S = \{\{P_0(90.8, 1.55), P_1(268.9, 0.44), P_2(96.2, 2.2), P_3(88.9, 2.87), P_4(89.8, 3.98), P_5(92.2, 2.51)\}, Scene_5\}.$$

Each element of our dataset is composed by:
- Identifier of the scene.
- The associated descriptor.
- A set of frames (RGB) and (Depth) chosen for future improvement of the descriptor.

B. From the Query to the associated recognition of visited scene

A visually impaired equipped with a head mounted depth sensor when is visiting a given place whose descriptor is inserted in the dataset, can interrogate the place recognition system giving only some depth frames of part of the scene. The same method used for 2D map computation of the complete scene is used to compute the partial 2D map associated to subset of acquired depth frames.

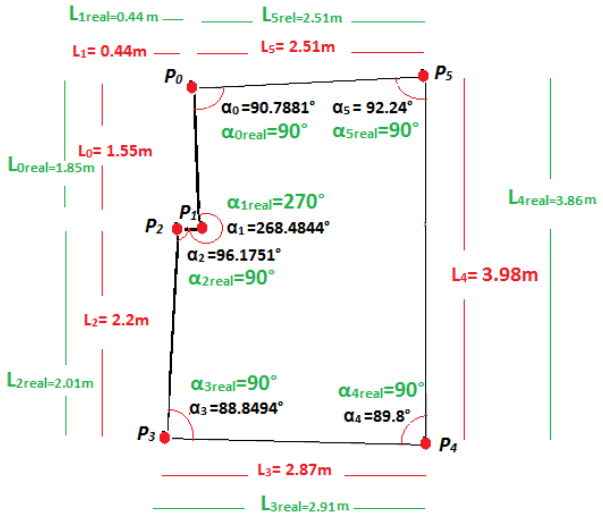

Fig. 9. Attributes of the corners of the computed 2D map model

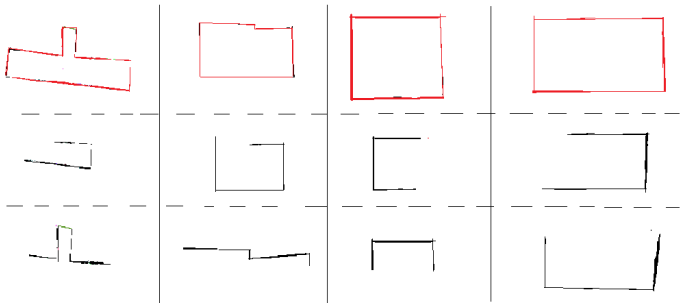

Fig. 10. 2D map model without processing of the obtained contour segments (first row), queries 2D maps (second and third rows)

The next step is then to find the best score by matching the partial descriptor of visited place with complete descriptors of all scenes models. Figue10, illustrates four 2D maps models and two 2D map queries associated to depth frames taken at differents position in each scene.

VI. Conclusion and futur works

We presented in this paper a framework for construction of visual place recognition dataset using depth sensor on the basis of the proposed 2D map computation method. We can then progressively build a dataset of a given region by visiting its different places and acquiring depth frames covering all their areas. Their 3D structures and 2D maps are then computed and used to calculate scenes descriptors. We also gave how to interrogate this dataset by collecting some depth frames of visited place.

With the release of this dataset, we expect the appearance of more localization systems based on the architectural aspects of scenes, that have the advantage of being invariant to the frequent changes of scenery brought in daily life caused by the modification of objects present in scenes. Because the 2D map used for scene description is independent of scenery changes, but needs to be improved in order to avoid ambiguity in the

matching of descriptors (query-models). We plan to add to the 2D map descriptor all informations related to doors, windows, stairs and perhaps some indices on the walls inferred from RGB images.

REFERENCES

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

[2] 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.

[3] Chen, Kuan-Wen and Wang, Chun-Hsin and Wei, Xiao and Liang, Qiao and Chen, Chu-Song and Yang, Ming-Hsuan and Hung, Yi-Ping. Vision-based positioning for internet-of-vehicles. IEEE Transactions on Intelligent Transportation Systems.364–376, 2017.

[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] Fankhauser, Péter and Bloesch, Michael and Rodriguez, Diego and Kaestner, Ralf and Hutter, Marco and Siegwart, Roland.: Kinect v2 for mobile robot navigation: Evaluation and modeling.:International Conference on Advanced Robotics (ICAR), 2015.

[6] Philbin, James and Chum, Ondrej and Isard, Michael and Sivic, Josef and Zisserman, Andrew Object retrieval with large vocabularies and fast spatial matching. conference on computer vision and pattern recognition, 2007.

[7] 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).

[8] Li, Yunpeng and Snavely, Noah and Huttenlocher, Dan and Fua, Pascal.: Worldwide pose estimation using 3d point clouds.: In european conference on computer vision. pp.15-29 (2012).

[9] Philbin, James and Chum, Ondrej and Isard, Michael and Sivic, Josef and Zisserman, Andrew. Lost in quantization: Improving particular object retrieval in large scale image databases. IEEE conference on computer vision and pattern recognition, pp. 1-8 (2008).

[10] Tolias, Giorgos and Avrithis, Yannis. Speeded-up, relaxed spatial matching. International Conference on Computer Vision, pp.1653-1660 (2011).

[11] Kendall, A., Cipolla, R.: metric 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).

[12] Kendall, Alex and Grimes, Matthew and Cipolla, Roberto. Posenet: A convolutional network for real-time 6-dof camera relocalization.: international conference on computer vision, pp.2938-2946, (2015).

[13] Jegou, Herve and Douze, Matthijs and Schmid, Cordelia.:Hamming embedding and weak geometric consistency for large scale image search. In: 2008 European conference on computer vision, pp. 304-317.

[14] Chen, David M and Baatz, Georges and Köser, Kevin and Tsai, Sam S and Vedantham, Ramakrishna and Pylvänäinen, Timo and Roimela, Kimmo and Chen, Xin and Bach, Jeff and Pollefeys, Marc and others: City-scale landmark identification on mobile devices in CVPR 2011, pp.737-744.

[15] 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)

[16] 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).

[17] Sturm, Jürgen and Magnenat, Stéphane and Engelhard, Nikolas and Pomerleau, François and Colas, Francis and Cremers, Daniel and Siegwart, Roland and Burgard, Wolfram. Towards a benchmark for RGB-D SLAM evaluation (2011).

[18] Sturm, Jürgen and Engelhard, Nikolas and Endres, Felix and Burgard, Wolfram and Cremers, Daniel:A benchmark for the evaluation of RGB-D SLAM systems.2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.573-580(2012).

[19] 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.

[20] Handa, Ankur and Whelan, Thomas and McDonald, John and Davison, Andrew J.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. : 2014 IEEE international conference on Robotics and automation (ICRA).: pp.1524-1531(2014).

[21] Li, Yunpeng and Snavely, Noah and Huttenlocher, Daniel P. Location recognition using prioritized feature matching.: European conference on computer vision.: pp.791-804 (2010).

[22] Sattler, Torsten and Leibe, Bastian and Kobbelt, Leif.: Improving image-based localization by active correspondence search.: European conference on computer vision. pp.752-765 (2012).

[23] Snavely, Noah and Seitz, Steven M and Szeliski, Richard. Photo tourism: exploring photo collections in 3D. ACM Siggraph 2006 Papers, pp. 835-846, 2006.

[24] Taira, Hajime and Okutomi, Masatoshi and Sattler, Torsten and Cimpoi, Mircea and Pollefeys, Marc and Sivic, Josef and Pajdla, Tomas and Torii, Akihiko.:InLoc: Indoor visual localization with dense matching and view synthesis.: Proceedings of the IEEE, Conference on Computer Vision and Pattern Recognition. pp. 7199-7209, 2018.

[25] Armeni, Iro and Sener, Ozan and Zamir, Amir R and Jiang, Helen and Brilakis, Ioannis and Fischer, Martin and Savarese, Silvio. 3d semantic parsing of large-scale indoor spaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1534-1543, 2016.

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

[27] Zamir, Amir Roshan and Shah, Mubarak.: Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs.: In journal of pattern analysis and machine intelligence. pp. 1546-1558 (2014).

[28] 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).

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

[30] 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).

[31] 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.

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

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

[34] Farah Ibelaiden, Brahim Sayah, Slimane Larabi. Scene Description from Depth Images for Visually Positioning. International Conference on Communications, Control Systems and Signal Processing. March 2020.

[35] Chen, David M and Baatz, Georges and Köser, Kevin and Tsai, Sam S and Vedantham, Ramakrishna and Pylvänäinen, Timo and Roimela, Kimmo and Chen, Xin and Bach, Jeff and Pollefeys, Marc and others. City-scale landmark identification on mobile devices. Proceedings of IEEE. CVPR. pp.737-744, 2011.

[36] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[37] Hunter, Gregory M and Steiglitz, Kenneth. Operations on images using quad trees. IEEE Transactions on Pattern Analysis and Machine Intelligence. pp.145-153, 1979.

[38] Diebel, James and Reutersward, Kjell and Thrun, Sebastian and Davis, James and Gupta, Rakesh. Simultaneous localization and mapping with active stereo vision. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566). pp.3436-3443, 2004.