2014 14th International Conference on Control, Automation and Systems (ICCAS 2014)
Oct. 22–25, 2014 in KINTEX, Gyeonggi-do, Korea

Robust Localization Using RGB-D Images

Yoonseon Oh and Songhwai Oh

Department of Electrical and Computer Engineering and ASRI,
Seoul National University
E-mail: {yoonseon.oh, songhwai.oh}@cpslab.snu.ac.kr

Abstract: Visual information extracted from RGB images has been successfully used for mobile robot localization. The main difficulty with localization using RGB images is that visual features from RGB images are not completely invariant against changes in viewpoints and lighting conditions. This problem can be overcome using features from RGB-D images. In this paper, we evaluate two depth features, depth patches and histograms of oriented normal vectors, extracted from RGB-D images for localization of a mobile robot and demonstrate that robust localization is possible under varying lighting conditions.

Keywords: Localization, RGB-D Images, Depth features.

1. INTRODUCTION

Localization for indoor environments is widely studied for personal robotic systems. Localization is a field of study that focuses on finding the position of a robot on a given map. Localization on a 2D map using a laser ranger has shown good performance. For example, Karto is a comprehensive software solution which includes 2D occupancy grid mapping and localization algorithms[1]. In spite of the good performance of 2D mapping using range sensors, localization using inexpensive cameras has been actively studied to reduce the cost. FAB-MAP is a successful mapping and localization algorithm using RGB images and has been successfully applied for topological mapping and localization of a moving vehicle [1].

Visual localization algorithms using cameras are not reliable under varying lighting conditions. The quality of RGB images highly depends on lighting conditions. Thus accurate localization cannot be achieved when it is dark. The use of depth images can be an alternative under different lighting conditions. With the introduction of RGB-D devices, such as Kinect from Microsoft, depth information can now be extracted at a low cost.

Henry et al. [2] extracted RGB features and used depth values of matched features to compute an accurate homography between two successive images. The approach is not robust against different lighting conditions because it assumes that RGB features are extracted reliably. In this paper, we treat depth features like visual features used in FAB-MAP. We develop a depth feature detection algorithm and evaluate the performance of two depth feature models, depth patches and histograms of oriented normal vectors [3]. We demonstrate that the combination of depth features along with RGB features improves the localization performance under different lighting conditions.

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2009348).

[1] http://www.kartorobotics.com

This paper is structured as follows. Section 2 describes observation models and Section 3 introduces a localization algorithm using depth features. Experimental results are discussed in Section 4.

2. DEPTH FEATURES

This paper employs two depth feature models, depth patch (DP) and histograms of oriented normal vectors (HONV) feature models [3]. We designed the DP feature model which compresses raw depth data. The HONV model was suggested for object recognition by Tang et al. [3]. The distribution of depth features is used to represent places in the map. This paper evaluates localization with depth features compared with localization using only RGB information, e.g. SURF features [4].

2.1 Depth Patch (DP)

Each DP feature is a compressed version of a local patch of size $N_w \times N_w$, where N_w is set to 25 in this paper. The depth data of a local patch is compressed into a 13×13 patch by bi-cubic interpolation and rearranged into a 169 dimensional vector to generate a DP feature.

2.2 Histogram of Oriented Normal Vectors (HONV)

The HONV model describes surface properties of a local patch. The normal vector at a point (x, y) on the image D is calculated using the cross product of two tangent vectors in x and y directions. The partial derivative of a tangent vector is the difference in depth values between adjacent pixels.

$$[N_x\ N_y\ 1]^T = \frac{\partial}{\partial x}[x\ y\ D(x,y)]^T \times \frac{\partial}{\partial y}[x\ y\ D(x,y)]^T$$

A normal vector in the Cartesian coordinate system with three dimensions can be represented more compactly by the azimuthal angle ϕ and the polar angle θ in the spherical coordinate system as follows:

$$\phi = \tan^{-1}\left(\frac{-N_y}{-N_x + \alpha}\right),\ \theta = \tan^{-1}\left({N_x}^2 + {N_y}^2\right),$$

where $\alpha \ll 1$. Histograms of ϕ and θ of all pixels in a 25×25 local patch form a HONV feature. A HONV feature is a 128 dimensional vector with 16 bins for ϕ and 8 bins for θ.

2.3 Interest Point Detection

An interest point detector selects informative local areas for effective feature extraction. For example, RGB features are shown as yellow dots in Figure 1. Edges are important in both an RGB image and a depth image. Hence, the Sobel edge detector [5] can be used to find interest points from a depth image. However, due to the limitations of the Kinect sensor, we have introduced a post-processing step to generate more consistent interest points. Since objects in a distance are not useful for localization, we eliminate Sobel edges belonging to distant objects. Each pixel belonging to a Sobel edge is an interest point candidate. Since the depth image from Kinect often contains pixels with missing depth data, we eliminate any interest point candidate if the number pixels with missing depth is large for the local patch around the interest point candidate. We then cluster interest point candidates and cluster centers become interest points. The detected interest points are robust against different lighting conditions as shown in Figure 1(b), 1(d), and 1(f).

3. LOCALIZATION

For robot localization, we utilize FAB-MAP [1], which is a visual mapping and localization algorithm using RGB images. It generates a discrete map and finds the robot's current position on the map using RGB observations. The RGB observation is a binary vector which represents existence of visual words. The algorithm shows good performance but it can become ineffective if visual features are not reliably detected. We perform robust localization by replacing RGB data with depth data.

3.1 Observation Vector

From a training set of depth features, we perform k-means clustering algorithm to group the training set into k clusters. Each cluster represent a word. A feature extracted from a test image is assigned to the nearest cluster or word. The existence of words in an image becomes a binary vector and it serves as an observation vector for the localization algorithm.

3.2 FAB-MAP

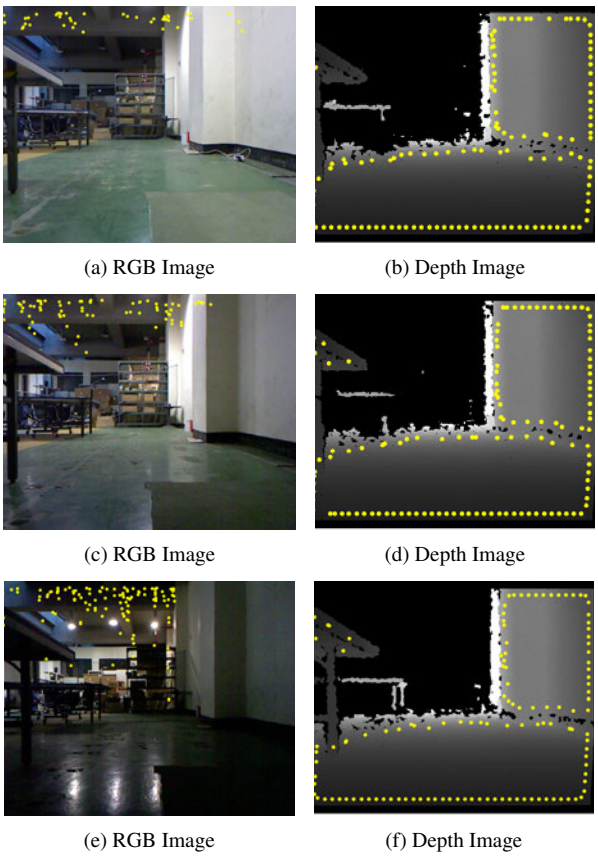
The FAB-MAP algorithm builds a map which is a set of locations. Each location L_i is represented by probabilities of existing words as follows:

$$L_i : \{p(e_1 = 1|L_i), p(e_2 = 1|L_i), \ldots, p(e_{|v|} = 1|L_i)\},$$

where v is a vocabulary set and $|v|$ is the number of words in it. e_j indicates the existence of the word j. An observation at time k is denoted by $Z_k = \{z_1, \ldots, z_{|v|}\}$, where z_j is a binary variable and $z_j = 1$ indicates that the ith word is detected. $Z^k = \{Z_1, \ldots, Z_k\}$ is a set of accumulated observations. Localization is to find the place L_i which maximizes $p(L_i|Z^k)$ given by

$$p(L_i|Z^k) = \frac{p(Z_k|L_i, Z^{k-1})p(L_i|Z^{k-1})}{p(Z_k|Z^{k-1})}.$$

The observation likelihood $p(Z_k|L_i, Z^{k-1})$ is equal to $p(Z_k|L_i)$ due to conditional independence between the current and past observations given the current location. The likelihood is approximated by a Chow-Liu tree [6], which describes a joint probability distribution using a second-order product approximation. Training a Chow-Liu tree and a vocabulary set is done during the mapping phase. More details can be found in [1]. We have used the open source version of the algorithm [2].



(a) RGB Image (b) Depth Image (c) RGB Image (d) Depth Image (e) RGB Image (f) Depth Image

Fig. 1 (Left) Detected SURF features. (Right) Detected depth features. Dots are detected interest points. A depth detector provides consistent interest points under different lighting conditions.

4. EXPERIMENTS

4.1 Datasets

Datasets used in experiments consist of RGB-D images and the ground truth data for localization. They

[2]Open Source FAB-MAP: `http://www.robots.ox.ac.uk/~mobile/wikisite/pmwiki/pmwiki.php?n=Software.FABMAP`

have variations in camera viewpoints and lighting conditions. Table 1 explains datasets used in experiments. $Freiburg2\_pioneer\_slam$ series are open RGB-D SLAM datasets from the Computer Vision group at Technische Universität München [7]. Some photos from the dataset are shown in Figure 2(a)-2(f). *Laboratory1* and *Laboratory2* are collected from our laboratory and seminar rooms. *Laboratory1* is collected by a robot moving around the laboratory arbitrarily. *Laboratory2* contains 452 images at 96 places with various lighting conditions to evaluate robustness of depth features (Figure 2(h)-2(l)). *TestSet1*, and *TestSet2* are datasets different from the training sets. Datasets contain image sequences with small overlaps since the FAB-MAP algorithm does not perform well if there is a large overlap between successive images. A vocabulary set is trained using *TrainingSet1* and Chow-Liu trees are generated using *TrainingSet2*. The performance of localization is evaluated using *TestSet1*, *TestSet2*, and *TestSet3*.

4.2 Parameters

The performance of FAB-MAP depends on its sensor model and the number of words. The sensor is modeled with the true positive rate, $p(z = 1|e = 1)$, and the false positive rate, $p(z = 1|e = 0)$. The true positive rate is the probability that there exists a vocabulary word when the sensor detects it. The false positive rate is the probability that there is no vocabulary word when the sensor detects it. $p(z = 1|e = 1)$ was selected from $\{0.34, 0.36, 0.38, 0.4, 0.43, 0.45, 0.5\}$, and $p(z = 1|e = 0)$ was selected from $\{0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12\}$. The algorithm is more sensitive to the false positive rate than the true positive rate. Those parameters are chosen for the best performance from the training dataset. From our training set, we found that the true positive rate of 0.45 and the false positive rate of 0.12 gave the best performance. The number of words K was selected from $\{512, 1024, 2048, 4096\}$ and the best parameter for each case is denoted in Table 2.

Table 1 Datasets

Dataset	Source	Note
TrainingSet1	$Freiburg2\_pioneer\_slam$	vocabulary
	Laboratory1	
TrainingSet2	$Freiburg2\_pioneer\_slam$	Chow-Liu tree
TestSet1	$Freiburg2\_pioneer\_slam2$	321 images
TestSet2	$Freiburg2\_pioneer\_slam3$	381 images
TestSet3	*Laboratory2*	452 images

Table 2 Localization Accuracy (%). The size of vocabulary is shown in parenthesis.

Dataset (K)	RGB	HONV	DP	RGB+ HONV	RGB+ DP
TestSet1	70.09 (2048)	46.11 (1024)	71.03 (2048)	73.83 (2048+ 1024)	**84.74** (4096+ 2048)
TestSet2	61.42 (1024)	40.68 (2048)	57.48 (2048)	64.83 (4096+ 512)	**71.39** (4096+ 512)
TestSet3	48.45 (2048)	72.12 (1024)	**88.05** (512)	73.67 (2048+ 1024)	84.29 (4096+ 1024)

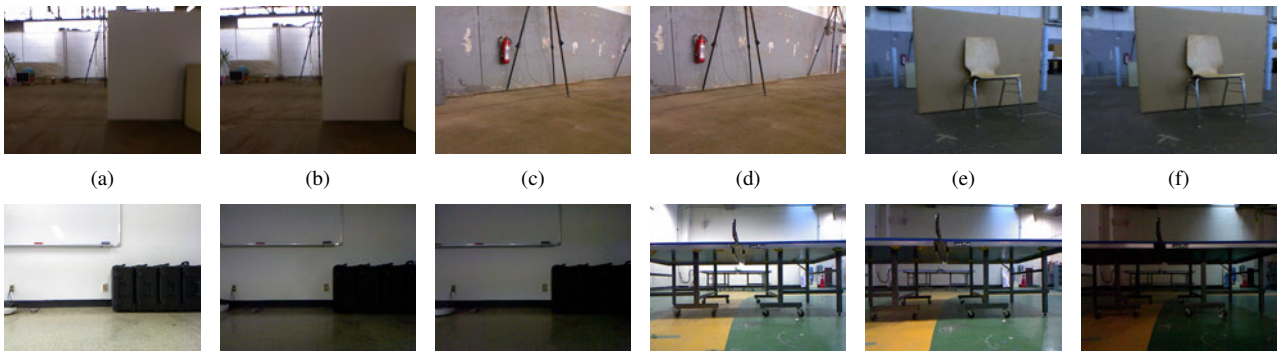
4.3 Results

Localization is evaluated using the *localization accuracy*, which is the ratio between the number of correctly localized places and the number of all visited places. The estimated location $\hat{L}_i$, which maximizes $p(L_i|Z^k)$, is correctly localized if it matches the ground truth location.

The results are shown in Table 2. We find that localization with depth features shows an excellent performance. A detector extracts depth features at reliable interest points in spite of various lighting conditions and viewpoints. The high localization accuracy of depth features shows that those features represent properties of locations well in comparison with using only RGB features.

The interest point detector for depth features provides invariant informative points against lighting conditions while a detector for RGB features does not. Yellow points in Figure 1(a), (c), and (e) are interest points extracted by a SURF detector from RGB images. Three images have different interest points though they are taken at the same place and this is due to different lighting conditions. Since different interest points generate different visual features, they can not provide useful information for reliable localization. On the other hand, the proposed detector extracts stable interest points as shown in Figure 1(b), 1(d), and 1(f). In addition, the detector selects only nearby points, while the RGB detector selects distant points on the ceiling. The detector using depth information selects more important features for reliable localization.

Table 2 shows the localization accuracy of combinations of features and the number of words K. When the lighting condition is not stable and the viewpoint changes, a combination of RGB features and depth features improves the localization performance (*TestSet1* and *TestSet2*). Localization with only depth features works poorly, because depth features are not robust to changes in camera viewpoints. However, the combination of HONV features and RGB features localizes correctly 3.74% better than RGB features on *TestSet1*. The



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

Fig. 2 Photos from datasets used in experiments. *TestSet1*: (a) and (b). *TestSet2*: (c) and (d). *TrainingSet2*: (e) and (f). *TestSet3*: (g) - (l).

combination of DP features and RGB features localizes correctly 14.65% better.

The localization accuracy of RGB features is less than 50% on *TestSet3*. Depth features shows the localization accuracy of 72.12% and 88.05% for HONV and DP, respectively. When depth features are combined with RGB features, the localization accuracy is increased up to 84.29% for DP. Depth features are not powerful when they are used alone but they can provide additional information when combine with other visual features as shown for the *TestSet3* dataset.

5. CONCLUSION

This paper suggested a robust localization algorithm using depth features collected from RGB-D images. FAB-MAP, an RGB visual slam algorithm, is modified to use RGB-D images to improve localization robustness. We have evaluated two depth feature models, DP and HONV. In experiments, we have demonstrated that the localization performance can be improved by combining depth features and visual features under varying lighting conditions.

REFERENCES

[1] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[2] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments," *International Journal of Robotics Research*, vol. 31, no. 5, pp. 647–663, 2012.

[3] S. Tang, X. Wang, X. Lv, T. X. Han, J. Keller, Z. He, M. Skubic, and S. Lao, "Histogram of oriented normal vectors for object recognition with a depth sensor," in *Proc. of the Asian Conference on Computer Vision (ACCV)*, 2012.

[4] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proc. of the European Conference on Computer Vision*, May 2006.

[5] R. C. Gonzalez and R. Woods, *Digital Image Processing*. Addison-Wesley Publishing Company, 1993.

[6] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[7] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.