INTEGRATION OF 2D IMAGES AND RANGE DATA FOR OBJECT SEGMENTATION AND RECOGNITION

Neslihan Bayramoğlu, Oytun Akman*, A. Aydın Alatan and Pieter Jonker*

Department of Electrical and Electronics Engineering, Middle East Technical University,
Balgat 06531, Ankara, Turkey
E-mail: {neslihan, alatan}@eee.metu.edu.tr
www.metu.edu.tr

**Delft Biorobotics Laboratory, Department of BioMechanical Engineering,*
Delft University of Technology, 2628 CD, Delft, The Netherlands
E-mail: {o.akman, p.p.jonker}@tudelft.nl
www.tudelft.nl

In the field of vision based robot actuation, in order to manipulate objects in an environment, background separation and object selection are fundamental tasks that should be carried out in a fast and efficient way. In this paper, we propose a method to segment possible object locations in the scene and recognize them via local-point based representation. Exploiting the resulting 3D structure of the scene via a time-of-flight camera, background regions are eliminated with the assumption that the objects are placed on planar surfaces. Next, object recognition is performed using scale invariant features in the captured high resolution images via standard camera. The preliminary experimental results show that the proposed system gives promising results for background segmentation and object recognition, especially for the service robot environments, which could also be utilized as a pre-processing step in path planning and 3D scene map generation.

Keywords: Scene Analysis, Segmentation; Object Recognition; Range Data; Time-of-flight Camera; SIFT.

1. Introduction

The motivation of the object segmentation and the recognition in vision based robotic applications (e.g. service robots) arises from the fact that there is a significant need for organizing, classifying and searching the content of visual data. In order to manipulate the objects in an environment or to be able to move the manipulator, object-background segmentation and object recognition must be performed. Many research efforts[1–3] are put forward to segment and recognize the objects by exploiting the 3D and 2.5D information captured by the laser scanning devices and the time-of-flight (TOF) cameras. Such data require intense preprocessing step for noise removal and it is extremely difficult to perform robust segmentation, as well as recognition, due to the sparseness of the TOF cameras. In this research effort, we propose a system which integrates the range information with the intensity data to achieve background segmentation and object recognition. Object segmentation is performed on low resolution range images in order to generate candidate regions. After this exploration step, intensity data is used to find scale invariant features on segmented candidate regions to perform object recognition. Also, the performed 3D exploration step is useful for further possible path planning and action planning.

2. System Overview

A standard image sensor measures the intensity of an environment. However, it lacks the available 3D information available in the scene. Moreover, 3D and 2.5D sensors are either very dependent to the scene (lighting, texture) or have drawbacks in terms of noise and spatial resolution. In order to design a robust and accurate system and exploit the intensity and range data simultaneously, 3D and 2D sensors should be integrated. Our experiments were performed on a vision system consisting of a Swissranger SR3000 range imaging camera (by MESA Imaging) and GC1350C color camera (by Prosilica). The SR3000 is a time-of-flight camera with a resolution of 176 × 144 pixels. It returns the depth and intensity images, spanning 47.5×39.6 degrees, with a non-ambiguity range of approximately 7.5m. GC1350C is a gigabit Ethernet color camera with a resolution of 1360 × 1024. It is equipped with a Cosmicar lens with focal length 14 mm (28 degree FOV).Thus the Prosilica camera has a spatial resolution eight times higher than the SR3000 camera. The low resolution 3D range (time-of-flight) camera is preferred to stereo cameras, since it overcomes the feature matching problem during disparity calculations. Also built-in infrared light source of camera makes it robust to different illumination conditions. It provides real time performance and low resolution output which makes it more appropriate as a peripheral sensor.

Stereo system calibration method of the camera calibration toolbox for MATLAB from Caltech[a] a is used to find the relative positions of both

[a] available at www.vision.caltech.edu/bouguetj/calibdoc

cameras with respect to each other (rotation and translation). Resulting rotation and translation are applied to the interpolated 3D data and 3D coordinates are obtained in the reference frame of the high resolution intensity camera. Low resolution (176 × 144) range images of the TOF camera are upsampled via linear interpolation to obtain coarse 3D representation of the scene. Extra noise removal in the 3D data is not performed since no performance decrease is encountered during image retrieval process due to the noise in the 3D data.

3. Segmentation of Candidate Regions via 3D Range Data

In most of the man-made environments, most of the items have planar, approximately planar or piece-planar geometric structures, such as walls, floors, doors, roads, tables, etc. Moreover, objects targeted for manipulations are usually placed on planar surfaces. When service robots are considered, there are two major planes: background planes (walls and floors) and surfaces that the targeted objects are placed on. Thus defining the planar structures in a scene would be a meaningful effort for the further processing. Hough transformation[4] is a strong and a popular algorithm for detecting 2D shapes such as lines, circles and ellipses in images. Later, Vosselman[5] extended the idea to the 3D space. The extended version of the Hough transform is used to define planar structures in depth map data and LIDAR data.[6,7]

3D Hough transform can be summarized as follows:
1. Plane (parameter) space is discretized, $0 \leq \theta \leq \pi$, $-\pi/2 \leq \phi \leq \pi/2$ and ρ to obtain a 3D parameter matrix, as shown in Fig.1. Each voxel in the 3D matrix represents a plane.
2. Each point in the point cloud votes for several planes (voxels in the 3D parameter matrix) that satisfies the plane equations:

$$xcos\theta_i sin\phi_i + ysin\theta_i sin\phi_i + zcos\phi_i = \rho_i, i = 1, 2, 3, ..., n \qquad (1)$$

3. Voxel with maximum value in the 3D parameter matrix (and its corresponding θ, ϕ, and ρ values) is selected as dominant shape.

Discretization step constitutes the critical part of the algorithm. When the sparse and noisy point cloud data is considered, finer discretization could result in undesirable planes. Moreover the localization of the planes could fail due to the coarse parametrization. Points lying on the same plane coincide in the same voxel of the 3D parameter matrix which gives the parameters of the plane. Thus the voxel having the biggest number of votes

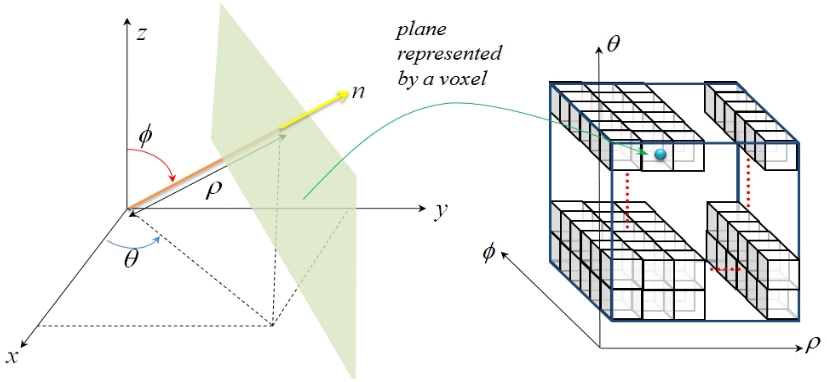

Fig. 1. Plane parametrization

gives the parameters of the plane that contains the majority of the points. Thus we call this plane the "major plane". In typical scenes, major planes usually belong to the background. As mentioned earlier, for service robots, there are two major planes: background planes and surfaces that the objects are placed on. After finding the first major plane, points belonging to this plane are removed to minimize the interference of these points in the following plane detection process. Due to the noise and the finite parameter discretization simple thresholding is used to determine the points belonging to the first major plane in the scene.

For each scene point its Euclidean distance to the considered plane is calculated and the ones closer to the plane than some certain threshold are removed; others remained (Fig.2).

The 3D Hough transform is executed once again on the remaining points. Then the next major plane and the points belonging to this plane are removed from the point cloud. The resulting point cloud represents the candidate regions of the objects of interest.

For further segmentation, mean shift algorithm[8] is preferred for segmenting these regions into different objects. Mean-shift segmentation algorithm is a straightforward extension of the discontinuity preserving smoothing algorithm. It is a non-parametric multivariate density estimation method. Input data is assumed to be sampled from various density functions. The algorithm finds these function modes and groups data accordingly. Formally, given k input points $x = x_1, x_2, \ldots, x_k$, in d dimensional space R^d, then the density estimation with the window size w is

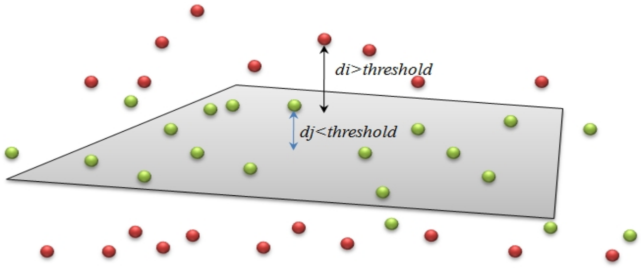

Fig. 2. Point-plane matching

$$f(x) = \frac{1}{kw^n} \sum_{i=1}^{i=k} K(\frac{x - x_i}{w}) \qquad (2)$$

where $K(x)$ is the kernel function (usually Gaussian). The modes of the density function are located at the zeros of the gradient function $\nabla f(x) = 0$. So the window is moved towards the means shift vector until it converges to a stable location (local maximums or minimums). The number of the cluster are determined automatically which is the major advantage of the mean-shift segmentation algorithm.

4. Object Recognition on Segmented 2D Images

The object recognition is performed on high resolution images by using scale invariant features (SIFT).[9] Keypoint matching is achieved by using the method explained in Lowe's seminal work:[9] for each SIFT feature extracted from the incoming high-resolution image, corresponding first and second closest matches in the database are found. Their ratio gives a measure for the quality of match. After assigning object classes to each feature via matching, they are clustered into individual objects by using the position, orientation and scale of every feature. False matches and noise result in errors in model fitting and object loss or false object recognition. Therefore, RANSAC[10] is utilized as a verification step to reject outliers (false matches).

5. Experimental Results and Conclusions

This research aims to recognize and describe the objects appearing in a scene. With the integration of a standard camera and a TOF camera, scene depth and the intensity information is gathered. Background regions in the scene are extracted by assuming that the major planes are backgrounds. Remaining objects are segmented considering their spatial and color information. Then, the recognition of each object is obtained by using its corresponding scale invariant 2D features. A typical result is given in Fig.3. 3D range image and color image are combined together and major planes are eliminated to segment the possible regions occupied by objects. Afterwards, SIFT features are found in the segmented regions and matched with the features in the database. Finally, matched features are segmented into object and a bounding box is drawn which is shown in Fig.3. The proposed algorithm shows good performance on segmenting foreground regions and object recognition. Moreover, it constitutes an initial step for the path planning and 3D scene map generation which are necessary for further object manipulation. As a future work 3D information could be used to extract the shape information besides planar surfaces. 2D shape descriptors other than SIFT features such as SURF,[11] zernike moments, fourier descriptors and 3D shape descriptors such as spherical harmonics can be examined.

Acknowledgments

This work has been carried out as part of the FALCON project under the responsibility of the Embedded Systems Institute. This project is partially supported by the Netherlands Ministry of Economic Affairs under the Embedded Systems Institute (BSIK03021) program. This research is partially supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK).

References

1. H. Chen and B. Bhanu, 3d free-form object recognition in range images using local surface patches *Pattern Recognition Letters* **28**, 2007.
2. E. Hameiri and I. Shimshoni, Using principal curvatures and darboux frame to recover 3d geometric primitives from range images *3D Data Processing Visualization and Transmission, International Symposium on* 2002.
3. G. Hetzel, B. Leibe, P. Levi and B. Schiele, 3d object recognition from range images using local feature histograms *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, (CVPR)* **2**, 2001.
4. P. Hough, Method and means for recognizing complex patterns *U.S. Patent 3.069.654* 1962.
5. G. Vosselman, Building reconstruction using planar faces in very high density height data *ISPRS Conference on Automatic Extraction of GIS Objects from Digital Imagery, International Archives of Photogrammetry and Remote Sensing* **32**, 1999.

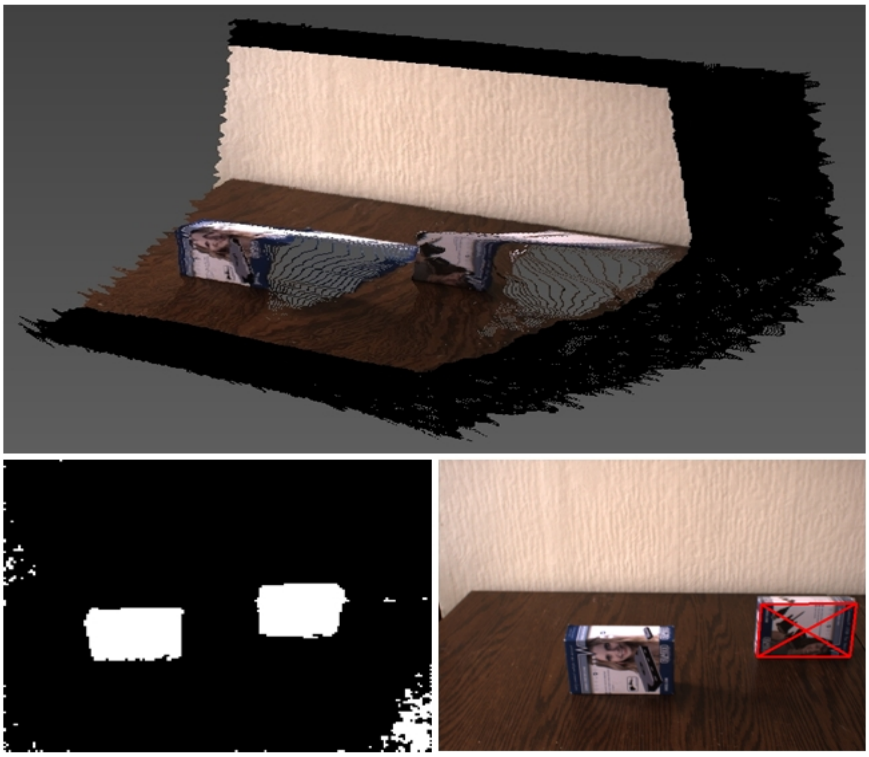

Fig. 3. Experimental results left to right and top to bottom, combined depth and color image, segmented regions and recognition result

6. K. Okada, S. Kagami, M. Inaba and H. Inoue, Plane segment finder: Algorithm, implementation and applications *Robotics and Automation (ICRA). IEEE International Conference on* **2**, 2001.
7. A. Sarti and S. Tubaro, Detection and characterisation of planar fractures using a 3d hough transform *Signal Processing* **82**, 2002.
8. D. Comaniciu and P. Meer, Mean shift: a robust approach toward feature space analysis *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **24**, 2002.
9. D. G. Lowe, Distinctive image features from scale-invariant keypoints *International Journal of Computer Vision* 2004.
10. M. A. Fischler and R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography *Communications of the ACM* **24**, 1981.
11. H. Bay, A. Ess, T. Tuytelaars and L. V. Gool, Speeded-up robust features (surf) *Computer Vision and Image Understanding (CVIU)* 2008.