Directing Visual Attention and Object Recognition using 3D Information

Oytun Akman, Boris Lenseigne and Pieter Jonker

Delft Biorobotics Laboratory
Department of BioMechanical Engineering
Delft University of Technology
2628 CD, Delft, The Netherlands
{o.akman, b.a.j.lenseigne, p.p.jonker}@tudelft.nl

Keywords: active vision, range images, computer vision, image stitching, saliency map, object recognition

Abstract

In the field of vision based robot actuation, in order to find and grasp objects in its environment, object recognition is a fundamental task that should be carried out in a fast and efficient way. Although a high resolution imaging environment is convenient for object recognition, it has high computational and memory costs. In this paper, we present a novel method to segment possible object locations using a low resolution range camera so that the further object recognition step, by using a high resolution camera, is only performed in a few candidate regions. A coarse 3D representation of the whole environment is obtained by stitching the range images while the arm is exploring the scene. Exploiting the obtained 3D information of the environment and the objects, a saliency map is built. The high resolution camera mounted on the robot arm is directed to the candidate (salient) regions in the saliency map for more detailed analysis. Finally, object recognition is performed using scale invariant features in the high resolution images.

1 Introduction

In a vision based pick and place robotic framework, a robot arm uses a vision system in order to find and grasp objects in its environment. The system can exploit the motion of a camera to explore the environment for a better recognition and pick/place performance. A similar approach is also employed in the human vision system, in which the gaze is directed to the interesting regions in the environment while the update process continues to stay aware in the entire field of view. The anatomy of the eye (peripheral and foveal distinction in the retina) is one of the major emphases of this approach which concerns the inhomogeneity of the visual system [6, 9]. The goal of this research effort is to enable a robot arm to recognize, localize and pick previously defined objects in a cluttered scene. Features of the targeted objects are extracted from a set of training images and stored in a database with their detailed geometric models. The proposed system is strongly inspired by an active vision concept [9]. Usually, for monitoring/exploring the environment and a better recognition performance, it is convenient to use high resolution cameras. However, using high resolution images for all parts of the scene has high computational and memory costs. Therefore the peripheral-foveal structure of retina perfectly fits to this problem. In this system, a 3D range camera is used as a peripheral vision sensor and a high resolution camera is used as a central vision sensor. Motion of the sensors is achieved by moving the robot arm. This motion serves as peripheral vision, in which the exploration of the environment occurs. During this motion, low resolution 3D images of the environment are captured and stitched to generate a saliency map. The modes of this saliency map represent the most interesting (salient) regions which deserve more detailed investigation in high resolution. After this exploration step, a high resolution camera is directed to the salient regions. Finally, scale invariant features in this view are found to perform object recognition and segmented into objects via graph-based clustering.

2 Related work

Standard approaches for object positioning [1, 4] and pick and place applications [2, 12] use passive vision systems consisting of only one fixed sensor, either a high resolution camera or a stereo vision setup. However, these methods suffer from the high reso-

lution - high computational cost dilemma explained in the previous section. In order to overcome these drawbacks active vision systems, inspired by the human ability to select the relevant aspects of a broad visual input, are intensively studied in literature. From the eighties, many efforts have been made to integrate such systems on mobile robot platforms [16] and a common approach was to use the depth information from stereo cameras to segment the important region [11]. Maki [14] proposed a gaze control model with two modes, a pursuit mode which uses disparity map to estimate the depth of the scene and a saccade mode for directing attention towards the new object using optical flow. Bjorkman [5] suggested hue saliency and 3D size to find interesting objects in a scene. Drawbacks of all these methods are that they can hardly deal with real time constraints, featureless surfaces and weakly illuminated environments. In this paper we propose to use a time of flight sensor that incorporates its own infrared light source and which provides depth information calculated in its hardware. Such a sensor enables us to overcome these limitations and allows us to make a fast (if not real time) system that does not require any specific light or texture hypothesis.

3 System overview

The architecture of the system described in this paper is summarized in Fig.1 and building blocks of the system are explained in detail in the following sections. This block based design makes the system highly modular as every part of the system can be modified in order to cope with new tasks or hardware sensor setups. This modularity makes the system suitable for different pick and place applications.

Our experiments were performed on SCARA (Selective Compliance Assembly Robot Arm) SR8437 robot arm (by Sankyo). The vision system mounted on this arm consists 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 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.

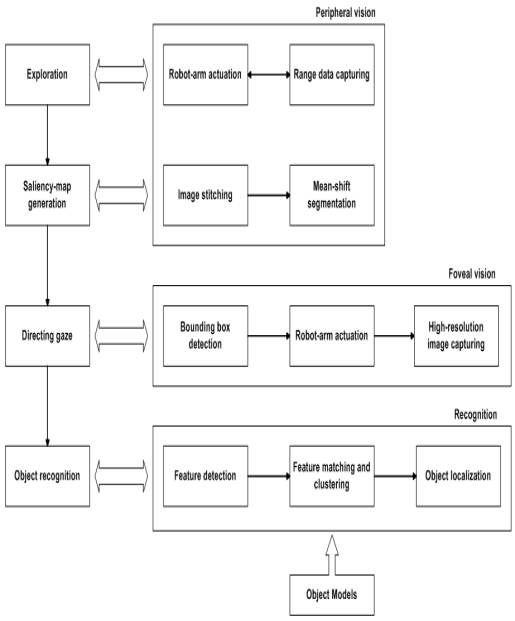

Figure 1: Overall system architecture

4 Active search

The natural way of directing attention by visual means is first searching the environment by fast eye moves (saccades) and head moves, and then fixating the gaze to the interesting regions in the environment. In the proposed method, active search is achieved by using a low-resolution range camera. First, the environment is explored by moving the camera via arm movements over all the possible regions where pick/place actions can take place. While moving the camera, both low resolution range images and intensity images output by the range sensor are captured.The intensity images are stitched to each other to have a broader view of the environment. The generated 3D mosaic view provides a coarse 3D representation. This representation is converted into a saliency map by mean-shift segmentation [7]. Segmented regions are labeled as interesting regions for further high resolution inspection.

4.1 Constructing peripheral view

An image stitching algorithm is utilized to construct a coarse mosaic view (peripheral view) of the environment. A 3D mosaic view of the environment is generated by first stitching 2D intensity images and then combining each pixel in the intensity image with its 3D information coming from the corresponding range image. Because the intensity images encapsulate invariant (eg. scale,rotation) interest points yielding a high matching performance between im-

ages [15], stitching the *intensity* images is preferred instead of stitching the range images directly. As an offline preprocessing step, the range camera and the high resolution camera are calibrated by using Zhang's camera model and calibration algorithm [17] while the robot arm is fixed in a predefined position. Afterwards, during the motion of the arm new images are captured via the range sensor. For each incoming image, SURF features [3] are found and these features are matched with the features in the previous image. For image stitching, SURF-64 which has a descriptor vector of length 64, is used. The matching is performed by considering the Euclidean distance between 64 dimensional feature descriptors. Based on the set of matched points between the current image and the new image, the projective transformation (homography) is calculated by using the RANSAC [10] algorithm. Images are stitched and projected on an imaginary ground-plane view. An imaginary ground-plane view is generated by using the projection matrix P of the range camera. P satisfies the following equation

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ p_1 & p_2 & p_3 & p_4 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

where (X, Y, Z) are the 3D coordinates of a point in the scene and (u, v) are the 2D coordinates of the same point in the image plane. The planar scene assumption, in which it is assumed that all points in the scene have a Z coordinate equal to zero, cancels the contribution of p_3 and the mapping between scene and imaginary view becomes a 3×3 homography matrix H. H maps the ground-plane points $(X, Y, 0)$ to the image plane points (u, v). Therefore, the inverse of the H matrix, H^{-1}, can be used to warp stitched images onto the imaginary ground-plane view. Then, the created ground-plane view image is combined with range information to generate the final 3D mosaic view.

4.2 Generating the saliency map and directing gaze

Interesting/important regions in the generated 3D mosaic view should be labeled/segmented in order to direct the gaze to those regions. One significant measure of 'importance' is to have similar dimensions with the requested objects. Therefore, representative and distinctive features of these objects were extracted from a set of training images and stored in a database with their detailed geometric models. This training step is performed off-line.

Depth values in the 3D mosaic view should be homogeneous inside the boundaries of an object in both spatial and range domains, and discontinuous at the boundaries. To segment these homogeneous regions the mean shift procedure-based image segmentation [7], which is a straightforward extension of the discontinuity preserving smoothing algorithm, is used. After running the mean shift filtering procedure for the mosaic image, all the information about the 3-dimensional convergence points is stored. Pixels converging to the same convergence point (mode) are clustered into same region and their values are replaced with the value of the mode pixel. Then, clusters which are closer than resolution h_s in the spatial domain and resolution h_r in the range domain are grouped together and assigned a label. Also spatial regions containing less than M pixels are eliminated. In our setup, we used spatial resolution $h_s = 5$ *pixels* and range resolution $h_r = 1$ *cm* and minimum region pixel count $M = 50$. Because the dimensions of the segmented regions (width, height and depth) in the 3D mosaic view represents the dimensions of the objects or homogeneous surfaces, they can be used as a measure between the objects in the scene and objects in the database. Therefore, a circumscribed rectangle of the minimal area for each region is found and utilized to extract the dimensions of the region. Saliency of a region with similar dimensions of objects should be higher than other regions. Hence, a value representing the saliency is given to each region according to its dimensional similarity with objects in the database. After generating the saliency map, maximas in the map represent the possible object locations.

A high resolution analysis is more convenient since it captures more information about the salient features on targeted objects. Therefore, high resolution camera should be directed to the possible object locations. The regions in the constructed saliency map represent the most salient regions where the object can be situated. After segmenting the important regions, their metric coordinates can be found by using H^{-1}. Since the metric dimensions of the calibration pattern are known, H^{-1} can easily be converted into a mapping from image plane to a metric ground-plane view. Then, the high resolution camera is directed to each of those regions (coordinates) in such a way that the camera field of view captures the entire region.

5 Recognition

The object recognition is performed on high resolution images by using scale invariant features. Both SIFT [13] and SURF [3] features are tested to decide on their performance in our application. Although, SIFT has a slightly better recognition performance than SURF, its high computational load makes SURF the better choice for our system.

SURF-128, which has a 128 dimensional descrip-

tor vector, is used for recognition. Keypoint matching is done by using the method explained in Lowe's work [13]: for each SURF 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 by matching, they must be segmented in order to cluster them into individual objects. The only available information is the position, orientation and scale of every features. Feature clustering is performed by using a *minimal spanning tree* [8] of the graph resulting from the features, where features are defined as nodes (or vertices) of this undirected graph. Affinity-based graph theoretic clustering is avoided due to its computational burden. In order to further decrease the computational cost, only one node is chosen in the $N \times N$ neighborhood of that node. Assuming that the segments of the same object are close to each other and in most of the cases will have similar orientation and scale values, the weight of a link or an edge between the *nodes-i* and *-j* can be defined as

$$w_{ij} = \alpha O_{diff} + \beta S_{diff} + \theta D_{diff} \qquad (2)$$

where O_{diff} and S_{diff} are the difference between the orientation and scale values of the *nodes-i* and *-j*, D_{diff} is the Euclidean distance between these nodes, whereas α, β and θ are the weights between these measures. The edges, whose weights are larger than a certain threshold value in the minimal spanning tree, are cut and a minimal spanning forests which represent individual objects are generated.

False matches and noise result in errors in model fitting and object loss or false object recognition. Therefore, a verification step is necessary to reject outliers (false matches). For this purpose RANSAC is utilized to find the homographic mapping between the minimal spanning forests and the actual item models, assuming that the objects are planar. Calculated mapping and real (metric) dimensions of the object are used to find its boundary/position. Rotation and scale of an object are calculated by taking the average of corresponding values of inliers for each object. Afterwards, the arm can be positioned on top of the recognised object and picking can occur.

6 Experimental results

Various scenes are tested by the proposed approach. The peripheral views are estimated by using approximately 20 frames. The required calibration is obtained by means of a checker board pattern in the starting position. A typical result is given in Fig.2. The proposed approach has good results on selecting candidate regions. Due to the textureless scenes, the peripheral view generation can sometimes

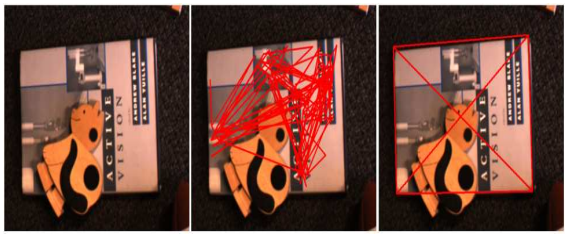

Figure 3: Recognition and localization results left to right, input image, the generated minimal spanning tree via the matched features and the bounding box

fail but this drawback can be overcome by investigating images individually instead of stitching them first. In this case, the displacement of the robot arm between frames should be known. The noise in the range sensor might also yield to undesired oversegmented or undersegmented regions in the mask, especially for the pixels close to the corners of the image. But this problem can be solved by weighting corner pixels with small values during saliency map calculation or simply by ignoring them. During the stitching process small projection errors appear and they are accumulated in each step. However, the precision is not the main concern in the exploration phase since the detailed analysis is done in the recognition phase. Finally, our algorithm achieves real time performance (on a 2,2 GHz Intel dual core CPU computer running under Linux) due to the low computational load of the homographic projections.

The recognition algorithm is tested on the segmented regions and a typical result is given in Fig.3. The proposed clustering algorithm has promising results on segmenting features into objects. Features belonging to the different objects from the same class can easily be segmented if they have enough spatial distance or scale/orientation difference.

7 Conclusions

In this paper we propose an active vision approach for the vision system of a pick/place robotic framework. We employed a range camera as a peripheral view sensor and a high resolution color camera as a central vision sensor. Initially, the environment is explored via the range camera and important regions are selected. Afterwards, selected regions are analysed in more detail to recognise and localize the targeted object. The presented method has a good performance on finding and localizing candidate items in the scene. There are many directions for further research to improve the robustness of our system. For a better recognition and localization performance, local affine features can be combined with features that are more geometric in nature, such as geometric blur and shape context. The proposed system can be tested and

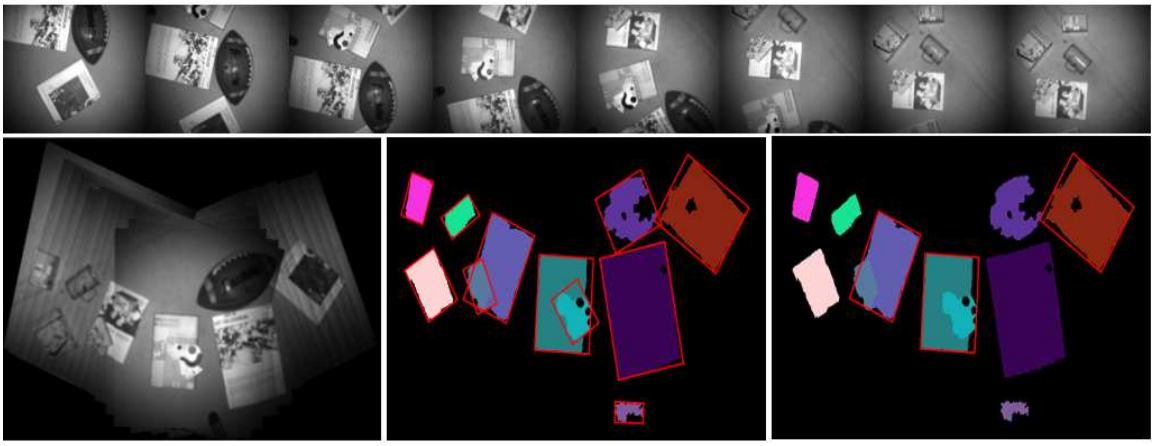

Figure 2: Results of the peripheral view generation from left to right and top to bottom, some examples of the input images, the generated mosaic view, the saliency map and the segmented regions, the candidate regions for the book situated in the middle

improved on different robot arms with more complex motion paths. Moreover, a gripper can be added to the system to observe its performance in real gripping scenarios.

8 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.

References

[1] E. Al-Hujazi and A. Sood. Range image segmentation combining edge-detection and region-growing techniques with applications sto robot bin-picking using vacuum gripper. *Systems, Man and Cybernetics, IEEE Transactions on*, 20(6):1313–1325, Nov/Dec 1990.

[2] N. Ayache and O. Faugeras. Hyper: a new approach for the recognition and positioning to two-dimensional objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8:44–54, 1986.

[3] H. Bay, T. Tuytelaars, V. Gool, and L. Surf: Speeded up robust features. *9th European Conference on Computer Vision*, 2006.

[4] M. Berger, G. Bachler, and S. Scherer. Vision guided bin picking and mounting in a flexible assembly cell. *IEA/AIE*, pages 109–118, 2000.

[5] M. Bjorkman and J. Eklundh. Vision in the real world: Finding, attending and recognizing objects. 16(5):189–208, 2006.

[6] P. S. Churchland and T. J. Sejnowski. *The Computational Brain*. MIT Press, 1992.

[7] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[8] D. Eppstein. Spanning trees and spanners. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, chapter 9, pages 425–461. Elsevier, 2000.

[9] J. M. Findlay and I. D. Gilchrist. *Active Vision: The psychology of looking and seeing*. Oxford University Press, 2003.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[11] D. Kragic and M. Bjorkman. Strategies for object manipulation using foveal and peripheral vision. *ICVS '06: Proceedings of the Fourth IEEE International Conference on Computer Vision Systems*, page 50, 2006.

[12] D. Kragic and H. I. Christensen. Robust visual servoing. *International Journal of Robotics Research*, 22, 2003.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[14] A. Maki, P. Nordlund, and J.-O. Eklundh. A computational model of depth-based attention. *Proceedings of the 13th International Conference on Pattern Recognition*, 4:734–739 vol.4, Aug 1996.

[15] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Computer Vision*, 2(1), 2006.

[16] T. Vieville, E. Clergue, R. Enciso, and H. Mathieu. Experimenting with 3d vision on a robotic head. *Robotics and Autonomous Systems*, 14(1):1–27, 1995.

[17] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330–1334, 2000.