Calculation of Attention Points Using 3D Cues

Ekaterina Potapova, Michael Zillich, Markus Vincze

Automation and Control Institute
Vienna University of Technology
{potapova,zillich,vincze}@acin.tuwien.ac.at

Abstract
Attention points are an effective means to tackle scene complexity, for example for grasping objects in a table scene. One way to obtain attention points is from a saliency map. Inspired by findings from preattentive human vision we investigated 3D cues to build a new type of saliency map. We implemented two 3D cues - one based on surface height and the other based on relative surface orientation. To evaluate the approach we built up an RGB-D (colour and depth) image database with table scenes of different complexity. We compared results of our algorithm to the classical 2D Itti-Koch-Niebur saliency map. In all types of scenes 3D cues showed better results than using 2D cues only.

Keywords - preattentive vision, 3D cues, attention points

1. Introduction

Imagine a table with a box full of different objects and a robot that is asked to empty the box. While a trivial task form humans, it remains a tough challenge for robots and still cannot be successfully solved in general.

The reason is that the problem of selecting objects in a table scene for grasping is quite complex. Objects come in a large variety of appearances and shapes, some may not be known, partially occluded, casting shadows onto each other, or cause inter-reflections. However, to grasp an object we have to segment it from the other objects. One way is to use one of the many segmentation algorithms [7, 3, 4, 21, 26], but none of them has been successfully applied to autonomously segment jumbled objects in a box. These algorithms typically lead to either an over- or undersegmentation of the scene. Neither of the outcomes is appropriate to enable grasping an individual object.

To overcome the problems of generic segmentation several recent segmentation algorithms use seed points to guide segmentation [22], object-of-interest segmentation described in [18], and [25] in which points of attention are seeds for the growing of regions.

The theory of human vision [17] claims that there are different stages of vision: preattentive, attentive vision, postattentive. Preattentive processing is the first stage of visual processing, during which basic features of the scene are extracted. The basic features include colors, closure, curvature, size, depth, and 3D orientation. This stage of visual processing takes no more than 200 to 250 msec [11]. Preattentive vision determines significant positions that are considered as the starting points for later
One way of implementing preattentive vision is to calculate attention points from a saliency map. Saliency maps have a long history in computer vision and there are several algorithms that calculate saliency maps, e.g., [10, 13, 14, 16, 15, 12, 1]. Existing approaches for saliency map calculation can be divided into two classes. The first class includes approaches based on feature integration theory [27] and the second class includes informational approaches [12]. One of the most popular algorithms which employ feature integration theory has been published by Itti, Koch and Niebur (IKN) [16]. This is one of the most relevant algorithms for saliency map calculation and used as a benchmark for evaluating new algorithms. The main idea of IKN is to combine color, intensity and orientation features. However most of these approaches are of limited use in robotics applications, where it is the 3D scene structure that is of interest and not 2D image saliency.

In Fig. 1 you can see an example image of a scene with objects in a box and the first five attention points selected by a repeated application of the Winner-Take-All [19] algorithm (Fig. 1.a) and seven areas which would be interesting from a robotic grasping perspective (Fig. 1.b). While IKN presents reasonable choices they are not necessarily the same that we consider relevant for grasping.

Despite of the fact that 3D cues are seen to be basic features during early vision [28], they are still not widely used in preattentive vision. There is a number of 3D preattentive cues that have been found to be relevant in human vision, e.g., 3D orientation [6], depth and stereoscopic depth [5, 23], and lighting direction [6].

Several algorithms for saliency map calculation use 3D information, obtained from point clouds or laser-scan data [9]. There is a number of algorithms that use disparity maps [20, 24], but they do not use spatial features that can be extracted from this 3D information. The work most similar to our approach is the algorithm in [2], which computes a saliency map from spatial information in point cloud data.

In this paper we propose to use attention cues derived from 3D point clouds, which are more suitable for the robotic applications we envision. The main idea and the novelty of our paper is to show how attention points can be calculated for table scenes in terms of preattentive vision using 3D information. For our work we employ a combination of two selected 3D pre-attentive cues: relative surface orientation (RSO) and surface height (SH). Saliency maps based on these two cues are combined with the saliency map calculated by IKN algorithm [16]. We evaluate these cues on labelled training data against the basic IKN algorithm, and show improved performance for scenes containing complex 3D structures.
The paper presents in the next Section the scheme to integrate the cues and details on cue calculation and Section 3. presents the results of the evaluation.

2. Attention Points

In our work we combine results derived from classical IKN algorithm [16] with relative surface orientation (RSO) based on the angle estimation between surface normals and support plane normal and surface height (SH) based on the height of the scene. A block-scheme of the algorithm is shown in Fig. 2.

![Figure 2. A block-scheme of proposed algorithm.](image)

The IKN saliency map is calculated from color images. An example of the IKN saliency map along with original image is presented in Fig. 3.

![Figure 3. Example of the original image (a) and IKN saliency map (b).](image)

For SH and RSO saliency maps we use point clouds derived from depth images, which are captured by a Kinect RGB-D sensor.

2.1. Saliency Map Based on the Surface Height

For the task of picking an object from a cluttered box objects or parts of objects that stick out from the clutter are good candidates for initial grasping attempts, and they should therefore be considered more interesting than the rest.
The first cue we propose is therefore based on a height map. As input we have a point cloud of the table scene. First of all the point cloud has to be filtered to delete outliers. Because height is a rather relative measure, we have to determine the ground base or absolute zero the supporting plane, which is the plane on which the box rests. We use RANSAC [8] to determine plane coefficients $Ax + By + Cz + D = 0$. For every point in the point cloud distance to the main plane is calculated. We set $d_{\text{max}}$ to be the distance between the main plane and the most remote point in the point cloud. Values of the SH saliency map are calculated according to:

$$H(i, j) = s_{\text{max}} f(D(i, j))$$  \hspace{1cm} (1)

where $s_{\text{max}} = 255$, $D(i, j)$ is a distance from correspondence point in the point cloud to the main plane. We furthermore scale height values non-linearly according to

$$f(x) = ax^2$$  \hspace{1cm} (2)

to obtain more pronounced salient regions, where $a$ is determined such that $f(d_{\text{max}}) = 1$ An example of the SH saliency map along with original image are presented in Fig. 4.

![Figure 4. Example of the original image (a) and height map (b).](image)

2.2. Saliency Map Based on the Relative Surface Orientation

The top surfaces of objects standing up-right often present good candidates for grasping positions. So our second 3D cue aims to identify top-surfaces based on surface orientation. To this end we calculate local surface orientation from the 3D point cloud relative to the orientation of the supporting plane.

We use the same supporting plane coefficients $Ax + By + Cz + D = 0$ as during calculation of the SH saliency map. Vector with coordinates $\vec{N} = \{A, B, C\}$ represents the normalized plane normal. For every point $p(i, j)$ in the point cloud $P$ we calculate local surface normal vectors by fitting a local plane and estimating plane normal:

$$\forall p(i, j) \in P : \vec{n}(i, j) = \{n_x(i, j), n_y(i, j), n_z(i, j)\}$$

Values of the RSO saliency map are calculated according to:

$$A(i, j) = s_{\text{max}} |g(n(i, j))|$$  \hspace{1cm} (3)

where $s_{\text{max}} = 255$ and $g(\cdot)$ calculates the cosine between two vectors:

$$g(\vec{n}) = (\vec{n}, \vec{N})$$  \hspace{1cm} (4)

An example of the RSO saliency map along with original image is presented in Fig. 5.
2.3. Combining IKN, SH and RSO saliency maps

Assuming that we have an IKN saliency map \( S(i, j) \), SH saliency map \( H(i, j) \) and RSO saliency map \( A(i, j) \), we multiply IKN saliency map \( S(i, j) \) with SH saliency map \( H(i, j) \) and RSO saliency map \( A(i, j) \) and then normalize final map to the range \([0, 255]\) to get combined map \( SM(i, j) \):

\[
SM(i, j) = N(S(i, j)H(i, j)A(i, j))
\]

(5)

We calculate attention points from combined map using several iterations of Winner-Take-All algorithm. An example of original image with first five attention points and corresponding combined map are shown in Fig. 6.

3. Results and Evaluation

To evaluate our results we built up an RGB-D (colour and depth) image database with ground truth ROIs signifying scene areas we consider relevant for robotic grasping. These regions are labeled with rectangular bounding boxes. The database consists of a total of 82 scenes and corresponding point clouds and contains four different types of scenes: single standing objects (SSO), occluded objects (OO), objects in a box (OB) and a box with objects which is situated among other objects (BOSO). The amount of ROIs differs from five rectangles to ten per one scene.

Examples of different types of scenes along with first five attention points calculated on classical IKN saliency map and on introduced combined map are shown in Fig. 8.

To evaluate results we find first five attention points for every scene using several iterations of Winner-Take-All algorithm and calculate the hit ratio of these points being situated inside different labeled...
Figure 7. The hit ratio of five first attention points situated inside different ROIs (SSO - single standing objects, OO - occluded objects, OB - objects in a box, BOSO - a box with objects which is situated among other objects. ROIs of the image. Averaged results are presented in Fig. 7.

For SSO IKN cue and RSO cue show the best hit ratio. While for OO, OB and BOSO SH cue is the most valuable. Furthermore for all types of scenes the results show that a combination of cues outperforms any single cue.

4. Conclusion and Future Work

In this paper we investigated the use of 3D cues to obtain attention points that are suited for determining objects for robotic grasping tasks. Inspired by findings from human vision, we implemented two 3D cues to augment the standard IKN model [16]. We then evaluated the approach in scenes with growing complexity: single standing objects, occluded objects, objects in a box, and object in a box with additional objects. We could show that height and relative surface orientation cues considerably improve picking out potential objects for grasping over the standard IKN model [16]. While single standing objects are always captured well, the 3D cues improve performance up to about 85% for all the other cases. In the most complex case the combination of both 3D cues gives clearly the best results. This indicates that 3D cues need to be considered more when moving out into the real world with robots.

Our future work will lay in the area of investigating more 3D preattentive cues and finding the best way to combine them. We also going to improve labeling of interesting parts of scenes in the database.

References


Figure 8. Examples of images from data base. a), b) - example of a scene with single-standing objects. c), d) - example of a scene with occluded objects. e), f) - example of a scene with box full of objects. g), h) - example of a scene with box full of objects surrounded by other objects. a), c), e), g) - attention points calculated on classical IKN saliency map, yellow circles - attention regions, red dots - attention points. b), d), f), h) - attention points calculated on combined map, green rectangles - labeled regions, yellow dots - attention points.