Active segmentation of cluttered scenes

1st Pedro Rodriguez de Ledesma Jimenez
Rijksuniversiteit Groningen
Faculty of Science & Engineering
Madrid, Spain
p.rodriguez.de.ledesma.jimenez@student.rug.nl

2nd Sarantis Doulgeris
Rijksuniversiteit Groningen
Faculty of Science & Engineering
Serres, Greece
s.doulgeris@student.rug.nl

***Abstract*—It is essential for robots that operate in human centric environments to recognize the objects correctly and be able to grasp and manipulate them. Therefore, identifying objects correctly in complex surroundings is an important task that a robot needs to perform so they can assist humans with their daily tasks. In this, segmentation of cluttered environment is a crucial operation to proceed with grasping and manipulation of objects. Segmentation can be done passively or actively depending on the interaction planning. In this work we are going to focus in active segmentation procedures where robot actively interacts with the scene to gain different insights about the environment, and use this knowledge to segment it.**

I. Introduction

Robots are becoming important members in our society by performing a diverse range of tasks to serve humans, many times replacing them. These tasks require a robot to recognize, grasp and place items in different locations. Although robots have extended their capacity in the last few years, they are still unable to perform this task with a high performance in complex environment where objects overlap making it challenging for robots to distinguish individual objects. Recognizing individual objects is essential for a robot. By perceiving the environment, the agent can correctly plan a template for performing a specific task. This template should differ from object to object.

While the are accurate methods in object recognition and categorization that can detect object instances and categories accurately, this models require large training set and even do the accuracy drops when it has to recognize objects in complex system [1].

As a result, new procedures are created to handle the recognition of objects in cluttered environments and make robots more reliable in their tasks by previously segmenting the objects of the scene. The goal of these approaches is to singulated objects of an complex environment by separating them from one another or separate them up to a point where an action has higher probability to be executed. All these approaches include some kind of robot interaction with the environment. This interaction plays a key role in the segmentation of the scene. The idea of interacting with the environment to segment the scene was first stated by Fitzpatrick [16]. He states the information which is gained by that cannot be compared with approaches that depend on moving camera to segmentate the scene.

These procedures can be passive or active depending on the robot planing. In this work we are going to focus on active interactions in a cluttered environment, where the robot creates motion by pushing or grasping an uncertain object to evaluate the singularity of the object before performing other tasks with it. Both procedures, passive and active are intended to improve object recognition facilitating the object modeling and therefore reduce future task errors thanks to the spatial segmentation.

In the following sections we will look into the two categories of scene segmentation, as mentioned in literature and on section III we will solely focus on the topic of the assignment which is active segmentation. Finally will discuss and state our conclusions for the state of the art techniques.

II. CATEGORIES OF SCENE SEGMENTATION

Scene segmentation can be divided into categories. Authors in [2] propose two main categories:

- Standard model-based methods.
- Interactive model-free methods.

This categorization is also seen in [3] but different names. Standard-base model approaches are labeled as *Passives* and free-model approaches as *Actives*. These categories derive from the planning of actions.

A. Passive scene segmentation

A passive or standard model-based scene segmentation approach, is regarded as one with a preplanned set of actions for a segmentation task. The robot's actions are derived from object kinematics and/or any anticipated movement of object. Moreover, in these approaches, additional information is needed for planning motion and choose according actions (object shape, color).

For example in [4] the motion of the robot end-effector is planned in advance. After the objects are separated in the scene the robot calculates the connection type between pair of objects (revolute joint, prismatic joint or disconnected) to identify the kinematic relationship between objects. In [5], from a video stream, motion features of objects and rigid body motion for multiple rigid bodies, are extracted using recursive filtering (recursive Bayesian filter) to estimate the next state of motion of the object after comparing three different recursive filters (different prior probability for prismatic, revolute and rigid joint).

On the other hand [6], in order to place an object in a cluttered environment, the robot has to calculate all possible scenarios that will result from an action and follow a reverse plan to execute the placing of object. For that, objects on the workspace are separated from the table using binary pixel labeling (0 for table 1 for object) and the object image with each orientation is convolved with the workspace 2D representation. The result is a image with possible free-space for object placing. If there is enough space no further actions are needed. If the space is not sufficient then object interaction in the form push is examined. All possible pushing actions are searched based on a fixed pushing distance. If the object can be moved without colliding with other objects or exiting the workspace the push is considered viable. If no pushing action is considered, the next object in tested. When we find a pair of object, viable-push, the action planning module backtracks to the previous object and tests again for a viable push.(same) This is continued until all objects are moved from the designated area. [7] propose a different planning framework for rearranging cluttered environments. When an action is defined, the planner determines the object spaces that are invaded by this action and the objects inside need to be move. The robot has four actions, *Push-grasp, Sweep, GoTo, Pickup*. This is not the only difference between those two approaches. In the latter framework, uncertainty of the object state is used and measured before and after each action to better calculate the free space on the workspace.

B. Active scene segmentation

In an active or model-free scene segmentation approach, the robot interacts with the scene with push and grasp primitives to directly declutter a scene. By that interaction, information about object specifics is gained by accumulating evidence over time. The criterion for picking an action is for the scene to be more disambiguated than before that action. That is very helpful in a human-centered environment which has no structure, because the robot does not require previous knowledge, but just tries to ”explore” its surroundings. Most approaches contain push and grasp primitive actions. Research has been made about different techniques, that will provide robot with that perception[reference?]. The main focus of this assignment is to investigate different methods of understanding the scene and algorithms that decluttered it.

III. ACTIVE SEGMENTATION TECHNIQUES

According to [8] object segmentation algorithm by themselves are not enough to solve the the segmentation problem in unstructured environments. Based on that, they propose that the best way to approach this problem is to combine robot's perception with learning. Basically, the robot should not just interact with the environment but also to learn if each interaction provided the expected results and learn from experience. This would be very beneficial for service robots. The first step for a robot to interact with the environment is first to perceive it. Then, based on the task the robot can take the appropriate action to achieve it. In the following section we analyze different perception and action techniques for actively segmenting a scene.

A. Visual and geometrical features

Many active segmentation algorithms depend on visual and geometrical elements to implement a task. Geometrical features are constructed by geometrical elements such as lines, points, curves or surfaces. Typical geometrical features include corners, edges, textures, etc. This kind of information can help the robot understand the shape and size of each object or object cluster in the scenery. Katz [8] attempted to manipulate unknown objects from a pile. For that, he uses a facet segmentation algorithm to segment the scene. This algorithm is based on the fact that a facet is an area that keeps its consistency in depth and on normals orientation. Depth discontinuities are extracted by a 3×3 convolution filter. Surface normal computed by applying a least-square plane fitting for the local planes that are defined for each two point neighbors. By combining these two features the algorithm can understand each facet of any object in the scene and can act on them to segment the image. One of five possible actions (actionable, push, pull, grasp along the principal axis and grasp along the secondary axis which are derived from principal component analysis of the region's cloud point) is assigned to every facet from a user through a GUI that was also implemented. This method has two major drawbacks. It cannot perceive reflective material and also cannot distinguish between two object that are touching (no depth discontinuity)

Gupta [9], on the other hand based on known geometrical features (Duplo bricks) but also visual features (every block has no color variations) applies a modified euclidean clustering algorithm which also includes color. The task is to singulate every object in three different scenarios. In each scenario *uncluttered, cluttered and piled*, the action is different but the task remains the same. The robot decomposes a pile, spreads clusters of bricks in order to declutter the scene with two primitive actions *push, grasp*. The robots understand single objects when it notices no color changes in a cluster. This approach as we can see has a lot of drawbacks in a scenario that multi-color objects exist in the scene.

[10] attempt to singulate objects based on depth and intensity images. The approach obtains possible edges, from these two images, which represent candidate boundaries of an object. This is done by using the derivatives images in both axis directions, for each point, therefore obtaining information of significant changes in pixel values. With this information an object can be located in the scene and a favorable push can be determined for singulation of the scene. The robot calculates four push action (fixed length, push normal) for each boundary and ranks them before performing the more suitable one. The before action scene is compared with the after one to check if new cluster are created.

B. Probabilistic presentation of the scene

Probabilistic models as part of the segmentation techniques can boost low-level algorithms (data captured from a camera)

can boost their performance [15]. This upgrade comes from integrating knowledge after a robots interaction with the environment which enables it to distinguish individual object from a clutter. This models provide the robot with high localization accuracy and high certainty of object detection.

For instance, [12] uses a probabilistic segmentation approach to determine where should the robot interact. In this approach the frames captured by the robot are stored in octrees from which a neighborhood graph is created. This is an can also be considered as encoding approach because edges are assigned as a similarity value between adjacent nodes. The encoded value comes a combination of independent probabilities that derive from geometric properties of the nodes. The probabilities are modeled as normal distributions of the Euclidean distance between point centroids, angular difference between normals and scalar difference between curvatures. During the motion the trajectory information is used to update the edge probabilities to improve the segmentation quality. The target selection is the one with the lower geometric similarity.

The [11] is a similar probabilistic approach to determine the likelihood ratio of a cluster to be one or more objects. Piles of objects are segmented in clusters after removing the surface from the input point cloud. The target pile to interact with is the one with the smaller footprint and a motion perturbation is created to collect evidence of singulation. After the interaction in the form of a push, the before motion scene is compared with the after one. Spatial units are matched between perturbation states to watch if new piles are created due to separation of objects. A sparse correspondence between feature points of the initial scene and the possible matching point after is evaluated to obtain the more suitable candidate. Then, the probabilistic framework evaluates based on the magnitude of the transform motion and the percentage of dense points matches if a cluster of points is a single object or multiple objects. More than one motion could be need to collect enough evidences of singularity of one spatial unit.

Much like the other approaches, [13] uses Gaussian mixture model to understand single object in a scene. The difference with the other approaches is that it uses motion signal, from an interaction with the environment, to calculate the probability of pixel to express movement. Pixel probability histograms are compared and translated to foreground pixels (moving object) or background pixel (table or already detected objects). Low probability of a pixel provides bigger confidence of motion. Histograms are kept to compare scene snapshots to exclude detected objects or to generate a motion signal (moving object). This approach can work pretty well at demanding situation such as multiple objects in the scene, multi-colored objects, etc.

C. Learning Approaches

As stated above a learning approach is ideal for service robots. That is because under no circumstances, a robot can understand everything that it perceives or perform impeccably in all scenarios. Learning from experiences can help in environment perception as well as action planning.

Authors in [2] use an aggregated convolutional neural network in order to rank and select push handles of objects in clutter. The first convolutional network called *Vanilla network*, is trained in a simulated environment only from robots interactions with the scene. Given an input image, the Vanilla classifier tries to mimic the labels that were provided by a expert user. The labels were provided after each random interaction of the robot with a significant object in the scene. These interactions combined with testing examples trained the aggregated network in an iterative manner. As a result, the robot can gain more insight about pushing action in different rigid objects.

Katz [4] implemented a Support Vector Machine (SVM) machine learning model for correctly classifying actions for each segmented facet of the scene. The SVM is trained in a supervised manner, with multiple features. Some features include information about the facet (cloud size, facet area, length, height etc.) but also from each surroundings (free space). The labeling is done by comparing two frames one before and one after the interaction with a facet. This active approach of the robot also helps to estimate an important feature that takes part in the training process that is facet matching [14]. Only high probability facet are considered same as [13].

IV. Summary and conclusion

In this work we have presented different robot frameworks for singulating objects in a complex scene through physical interactions. Interactions are done by a mobile manipulator. Research showed that passive segmentation procedures are more time-consuming with similar performance, which makes them unsuitable for real life scenarios. We find out that many approaches face problems singulating objects who have similar textures and the are touching or same have depth level, because they rely solely on geometrical and visual information to segment the scene. This is tackled by approaches that use motion signals as a way to interpret every object in the robot's workspace. Another gap that we detected lies in approaches that use push as the only action for separating objects in a pile. Complex situation where objects are stacked on top of each other can not be solved by pushing actions. For example, objects that are similar and placed inside a boundary (bowl of apples) could not be singulated using pushing actions. In that case deep learning approaches the declutter an image for grasping such as [17], which segments scene without primitive actions, are preferable. So based on the different scenarios a robot should be able to choose the best option based on the scene but also on the action. Finally, while the learning approaches, can help the robot long-term, they need constant human supervision. It would be more preferable if the labeling and supervision was performed in a reinforcement learning fashion, where simple strategies will give the robot more autonomy to decide based on its own beliefs.

REFERENCES

[1] Breyer, Michel, et al. "Volumetric grasping network: Real-time 6 dof grasp detection in clutter." arXiv preprint arXiv:2101.01132 (2021).

[2] Eitel, Andreas, Nico Hauff, and Wolfram Burgard. "Learning to singulate objects using a push proposal network." Robotics research. Springer, Cham, 2020. 405-419.

[3] Patten, Timothy, Michael Zillich, and Markus Vincze. "Action selection for interactive object segmentation in clutter." 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2018.

[4] Katz, Dov, et al. "Interactive segmentation, tracking, and kinematic modeling of unknown 3d articulated objects." 2013 IEEE International Conference on Robotics and Automation. IEEE, 2013.

[5] Martin, Roberto Martin, and Oliver Brock. "Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors." 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2014.

[6] Cosgun, Akansel, et al. "Push planning for object placement on cluttered table surfaces." 2011 IEEE/RSJ international conference on intelligent robots and systems. IEEE, 2011.

[7] Dogar, Mehmet, and Siddhartha Srinivasa. "A framework for push-grasping in clutter." Robotics: Science and systems VII 1 (2011).

[8] Katz, Dov, et al. "Perceiving, learning, and exploiting object affordances for autonomous pile manipulation." Autonomous Robots 37.4 (2014): 369-382.

[9] Gupta, Megha, and Gaurav S. Sukhatme. "Using manipulation primitives for brick sorting in clutter." 2012 IEEE International Conference on Robotics and Automation. IEEE, 2012.

[10] Hermans, Tucker, James M. Rehg, and Aaron Bobick. "Guided pushing for object singulation." 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012.

[11] Chang, Lillian, Joshua R. Smith, and Dieter Fox. "Interactive singulation of objects from a pile." 2012 IEEE International Conference on Robotics and Automation. IEEE, 2012.

[12] Patten, Timothy, Michael Zillich, and Markus Vincze. "Action selection for interactive object segmentation in clutter." 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2018.

[13] Kenney, Jacqueline, Thomas Buckley, and Oliver Brock. "Interactive segmentation for manipulation in unstructured environments." 2009 IEEE International Conference on Robotics and Automation. IEEE, 2009.

[14] Katz, Dov Kazemi, Moslem Bagnell, J. Stentz, Anthony. (2013). Clearing a pile of unknown objects using interactive perception. Proceedings - IEEE International Conference on Robotics and Automation. 154-161. 10.1109/ICRA.2013.6630570.

[15] Daniel Beale, Pejman Iravani, Peter Hall, "Probabilistic models for robot-based object segmentation", Robotics and Autonomous Systems, Volume 59, Issue 12, 2011, Pages 1080-1089, ISSN 0921-8890

[16] P. Fitzpatrick, "First contact: an active vision approach to segmentation," Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453), 2003, pp. 2161-2166 vol.3, doi: 10.1109/IROS.2003.1249191.

[17] Kasaei, Hamidreza Kasaei, Mohammadreza. (2021). "MVGrasp: Real-Time Multi-View 3D Object Grasping in Highly Cluttered Environments".