Active Visual Object Search Using Affordance-Map in Real World : A Human-Centric Approach

Lasitha Piyathilaka and Sarath Kodagoda
Center for Autonomous Systems (CAS)
University of Technology, Sydney
Email: lasitha@ieee.org

***Abstract*—Human context is the most natural explanation why objects are placed and arranged in a particular order in an indoor environment. Usually, humans arrange objects in order to support their intended activities in a given environment. However, most of the common approaches for robotic object search involve modelling object-object relationships. In this paper, we hypothesize such relationships are centered around humans and bring human context to object search by modelling human-objects relationships through affordance-map. It identifies locations in a 3D map which support a particular affordance using virtual human models. Therefore, our approach does not require to observe real humans in the scene. The affordance-map and object-human-robot relationship are then used to infer the object search strategy. We tested our algorithm using a mobile robot that actively searched for the object "computer monitors" in an office environment with promising results.**

I. Introduction

Ability to recognize objects plays a major role in robot's understanding of its environment. To execute various tasks such as pick and carry or object manipulation require effective interaction with objects. Therefore, a service robot that operates in an indoor environment is required to localize and map objects in its operating environment. This is challenging due to several reasons. First, the operating environment of an indoor robot could extend beyond robot's sensory range. Therefore the robot should have an understanding of where to look for objects when they are located beyond its perception range. Secondly, real time object recognition is still largely an open problem. This makes object search and recognition a challenging task even though the objects are visible to the robot's sensors.

Active object search involves executing series of sensing actions in order to bring the object to the field of view of the sensor. When vision sensors like cameras are involved, this is called as 'active visual object search'. In order to increase object detection efficiency, the robot should move in a path that maximizes the object detection probability. Solving this problem is far from trivial because factors such as occlusions and poor illumination affect the object recognition capabilities. On the other hand, 3D objects can be viewed from a number of different view points. Therefore, often objects recognition techniques involve training models for multiple view points. However, even such a greedy approach may often fail for most of the unsymmetrical objects, if the object is viewed from a previously untrained view point.

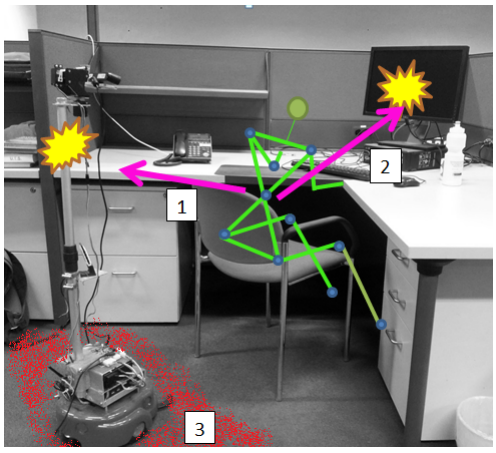

Fig. 1. Affordance-Map models the relationship between virtual human skeleton model and the 3D environment. 1.Robot- Human relationship 2. Object -Human relationship 3. Probable robot's locations that give best view of the object

However, complexities associated with object search can be minimised by understanding the relationships between objects and humans. This is because most of the objects play major role in humans' perception of their environment [1]. Further, humans arrange objects in their environment in a specific order to cater their requirements and activities [2]. For example the 'computer monitor' is placed on a tabletop so it can be easily seen, and the 'keyboard' is placed within arm reach of the human sitting beside the table. The 'Mouse' is placed near the keyboard in a place that is convenient to handle. Therefore, these strong relationships between objects and humans can be used to localize probable locations of a given object.

Our approach for active object search is based on the concept of affordance-map. This brings the human context to the active object search which has not previously explored in robotic object search research. Affordance is defined as all 'action possibilities' latent in the environment. In other terms, affordance is a property of an object, or an environment, which allows an individual to perform an action [3]. Although many affordances are possible in a given environment very few are practically probable. For example, in an office environment most frequently observed affordance is sitting and working beside an office desk. Office desk supports this action and other objects such as the monitor, keyboard, mouse and the telephone are arranged around the sitting human. Therefore,

by identifying the locations of the 3D map that support the affordance 'sitting and working', we can easily model most probable locations that the searched object can be found.

In order to build an affordance map, human actions need to be observed in the environment. However, observing real humans could be time consuming and a robot could easily miss a possible affordance if a real human is never observed in the environment. Therefore in this research, we used virtual human models to learn affordances in a given 3d environment. First, the 3D map of the environment is obtained and converted to a 3d distance map. Then the virtual human models are trained using distance measures by manually placing human models on 3D cad models of furniture downloaded from a public dataset [6]. Later, these models are placed across the 3d environment in search of locations that support the given affordance. Finally, affordance map is created by calculating an affordance value for each and every grid cell which represents how likely that locations support the given affordance. In addition, each grid cell consists of orientation information of human skeleton model. This information embedded in the affordance map is later used to model the relationship between the human,object and the robot's position which gives the best view of the object that is being searched. Therefore our approach for active object search requires a fewer number of training samples that are taken from a limited number of view points. In other terms, our method looks at the objects only from similar viewpoints that were used to train the object detector.

II. Related Works

Active object search has been a popular research area among robotic research community in recent years. Bulk of the previous works have relied on the fact that the object is already located within the field of views of robot's sensors [4]. These aproaches have used object features such as color to guide the robot towards the object. However, it is unclear how the same approach can be used, if the object is located outside the sensory range of the robot.

Some other researchers have used object-object relations to search objects in large search areas [5], [6]. Common approach is to model object-object relationships using probabilistic graphical models such Markov Random Fields (MRF). However to this to work, atleast few objects need to be recognized before the object search step or strong prior assumptions have to be made about object's locations. On the other hand, these approaches do not model the human context in the environment, which has a strong relationship with objects that are being used in the environment.

In recent years, human context has been introduced to the object recognition research field in the form of affordance. In [7] researchers used virtual human models to recognize objects that have "sittable" affordance without using common approaches such as 3D features for object recognition. They used a human model with sitting pose to detect locations that support sitting, and tested their approach in synthetic datasets. Although they achieved good recognition accuracies with synthetic datasets, they failed to record good results when tested on a real 3D environment.

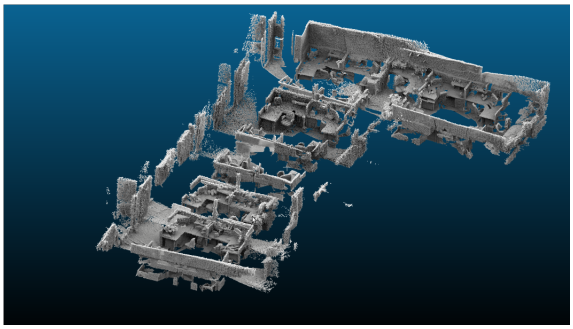

Fig. 2. Built 3D map of the office environment

Recently, few researchers introduced virtual humans models to learn human context in 3D environments [2]. They hallucinated human models in a 3D environment to learn human context in the environment and used them for object detection. However, they did not explore the possibilities of using virtual human models for active object search in a large environment.

III. Our Approach

In this research, our main goal is to identify 'human-working' areas in an office space and use it for object search. However, our approach does not require real humans to be seen. Instead, we modelled the relationship between the virtual human and the environment. The 3D environment is modelled as a 3D distance field and humans are modelled as 3D skeletons, obtained from a real human activity detection dataset [2].

A. Environment Model

State of the art 3D point cloud maps generated from a depth camera usually contain large amount of 3D points, and searching for areas with 'workable' affordance in this large space is computationally infeasible. Therefore in order to generate 'workable' affordance-map, we preprocessed bulky 3D point clouds in to much lighter point clouds that only contain horizontal surfaces and vertical surfaces. Our heuristic here is that human workable areas are always need to be supported by horizontal flat surfaces like table tops. On the other hand, vertical flat surfaces like 'wall partitions' and 'table drawers' oppose the existence of 'sitting' virtual human skeletons. In other terms, given an office table as shown in the Fig. 1, human prefers to sit in a area where there is a free space under the table top rather than sitting facing a drawer. In order to segment the horizontal and vertical surfaces the original point cloud is first voxelised. Then We segment the horizontal surfaces by calculating 3D correlation across all points in the map, using a point cloud template extracted from a table top surface. Vertical surfaces are extracted with the same procedure but point cloud templates extracted from vertical surfaces like partitioning walls and table drawers are used.

The environment model is consisted of 3d Distance Transform Map $DT(\mathbf{x})$ and 3D Occupancy Map $OC(\mathbf{x})$, where

x is any 3D position of the environment. The 3D Distance Transform (DT) is a shape representation which indicates the minimum distance from a point in the environment to the closet occupied voxel. In our approach, we calculate 3D Distance Transform from the occupied voxels of horizontal surfaces, OC_h. The distance transform map $DT(\mathbf{x})$ of the occupancy grid map OC_h can be generated using an unsigned distance function given by (1), which represents Euclidean distance from each location **x** of the environment to the nearest occupied voxel in OC_h.

$$DT(\mathbf{x}) = \min_{O_j \in OC_h} |O_j - \mathbf{x}| \tag{1}$$

Finally, 3d occupancy map for the environment can be obtained by combining occupancy map of the horizontal surfaces OC_h and the occupancy map the vertical surfaces OC_v.

$$OC(\mathbf{x}) = OC_v(\mathbf{x}) + OC_h(\mathbf{x}) \tag{2}$$

B. Human Model

In this research, virtual human models that can effectively model interaction between the environment and the human is used to build the affordance-map. Therefore selected human models should have a direct relationship between the affordance 'workable' and the given environment. Although many human poses can be observed in an office environment, very few of them directly relate to 'workable area'. Most frequently observed human pose in an office environment is sitting beside an office desk and working with the computer. Therefore human pose model are selected from the activity 'working with the computer' from a human activity detection dataset [8]. Fig. 1 shows the human pose model that is used in this experiment. It is a human skeleton that consist of 3D joint positions in 3D. Given these 3D points of the human skeleton H_l, we can move it across the test environment using the rigid transformations of translation and rotation. Then we can effectively map each human skeleton model to the coordinate system of the environment using (3), where $X_k = (x_k, y_k, z_k, \theta_k)$ is the position and orientation of the skeleton's Torso in the world coordinate system and $R_z(\theta_k)$ is the rotational matrix about Z axis. Since we only move skeleton model in a plane parallel to the ground plane only rotation about Z axis is considered.

$$H_w(X_k) = [x_k, y_k, z_k]^T + R_z(\theta_k).H_l \tag{3}$$

C. Human Environment Relationship

Once the models for the environment and the human have been built, the next step is to model the relationship between them. This is achieved using two geometric features, namely distance features and collision features. Selection of these features are motivated by two factors. First the human needs to be close enough to the object for effective interaction, and the second is to prevent collisions with occupied voxels of the environment.

Distance features are obtained by moving the human model across the voxels in the environment and calculating distance measure for each and every skeleton points in the human model. Once the environment is modelled by (1) and (2), we can effectively calculate distance features for a human skeleton with location and orientation $X_k = (x_k, y_k, z_k, \theta_k)$ by (4), where n is the number of 3D points in the skeleton.

$$\begin{vmatrix} d_1 \\ d_2 \\ . \\ d_n \end{vmatrix} = DT(H_w(X_k)) \tag{4}$$

In the same way, we can check for any collision for a skeleton at location and orientation, X_k by (5). In case of a collision c_i becomes 1 and 0 otherwise.

$$\begin{vmatrix} c_1 \\ c_2 \\ . \\ c_i \\ c_n \end{vmatrix} = OC(H_w(X_k)) \tag{5}$$

Thereafter, the collision check at X_k in the map can be converted into a probability value using (6).

$$P(C|X_k) = 1 - \frac{\sum_{i=1}^{n} c_i}{n} \tag{6}$$

Then the affordance-map for 'workable area' in the given environment can be calculated by using the distance features, d_i and collision features, c_i. Each cell in the affordance-map represents the likelihood of that place being a 'workable area'. Finally, given a virtual skeleton with location and orientation X_k, the underlying Likelihood, $P(A_k|X_k, \lambda)$ of that location being a 'worakable area' can be calculated by (7).

$$P(A_k|X_k, \lambda) = P(C|X_k).\prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(d_i-\mu_i)^2/2\sigma_i^2} \tag{7}$$

The Parameters of the above likelihood function $\lambda = \{\mu_i, \sigma_i\}$, for each distance measure can be estimated via training as explained in the next section.

IV. Mapping Affordance

In order to build the affordance-map, probable working areas in the given map need to be identified. This involves training and detection steps.

A. Data set

In order to test the proposed approach, first a 3D point cloud map of the environment needs to be built. Recent advancements in RGBD SLAM algorithms allow us to build 3d map of the environment in real-time. We mounted a depth camera on a mobile robot and used CCNY−RGBD [9], a ROS tool for fast visual odometry and mapping RGBD data, to map an office environment. The mapped area covers 30m X 30m and includes several office cubicals and corridors. A snapshot of the mapped area is shown in Fig. 2.

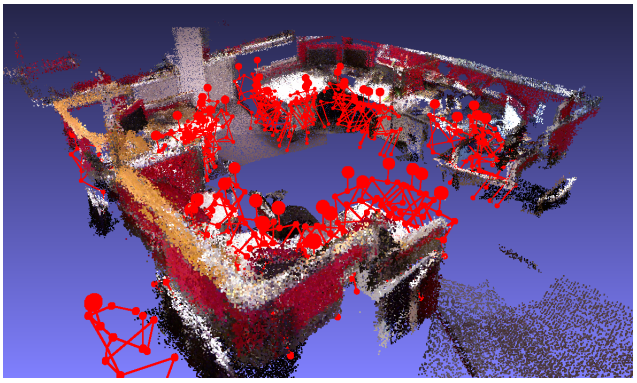

Fig. 3. Affordance map with highest probabilities

B. Training Human Model

The parameters of the human models are trained using a simple process that calculates (μ_i, σ_i) in (7). We downloaded 3d model of office furniture that support affordance "workable" from "Google 3d Ware House" and manually placed the human skeleton model in areas that support the affordance "workable". Then we recorded distance measure for each point in the human skeleton model and used them as training data to estimate μ_i and σ_i of the normal distribution.

C. Detection

In order to build the affordance-map for the affordance 'workable', the virtual human model needs to be moved across the environment while searching for most probable locations that support affordance, 'workable'. First, the map is voxelized into 10cm x 10cm x 10cm grid and distance fields and occupancy map are built. Then $P(A_k|X_k, \lambda)$ for each grid location is densely calculated. For each grid cell, rotation angle θ is set to 10 discrete values to calculate affordance $P(A_k|X_k, \lambda)$. The z axis position of the virtual human model is set to a preselected value in order to increase the calculation efficiency. Fig. (3) shows the human skeleton samples with high affordance likelihood, mapped on a section of the office environment.

V. Affordance map for Object search

Once the affordance-map is built, it can be used to effectively search for objects. The main challenges that need to be solved in any robotic object search algorithm are " where to look for a specific object" and "what is the camera angle need to be?". These challenges can be addressed effectively by using affordance map as it already embodies human object relationships.

A. Object, Human and Robot Relationship

The affordance map which is built for the office environment, models the relationship between the environment and the 'working' human pose. Therefore, the position of the robot can be inferred by modelling the relationships between the object, human and the robot. More specifically once the human pose position is given, we can infer the best position and the viewing angle of the robot that gives clear view of the object to be recognised.

The robot position is modelled as a multivariate normal distribution relative to the human skeleton model using (8). Here μ_s and Σ_s are the mean and covariances of position and orientation of robot in human skeleton co-ordinate system.

$$R_s \sim \mathcal{N}(\mu_s, \Sigma_s) \tag{8}$$

This probability distribution can be converted to world coordinate system for a skeleton at $X_k = (x_k, y_k, z_k, \theta_k)$ by,

$$R_k \sim \mathcal{N}(X_k + B\mu_s, B\Sigma_s B^T) \tag{9}$$

where B is the rotation matrix given by (10)

$$B = \begin{vmatrix} cos(\theta_k) & -sin(\theta_k) & 0 & 0 \\ sin(\theta_k) & cos(\theta_k) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} \tag{10}$$

Finally, likelihood of a robot at location $Y_j = (x_j, y_j, z_j, \theta_j)$ can see an object is given by

$$P(z|Y_j) \propto \sum_{k=1}^{n} P(A_k|X_k, \lambda).P(Y_j|\eta_k) \tag{11}$$

where η_k is the parameters of the probability distribution given by (9) and n is the number of skeletons in the map. Note that only skeletons with high affordance likelihood contributes for (11).

VI. Results and Discussion

To test the effectiveness of the proposed approach series of experiments were carried out in an office environment. Our objective was to find 'computer monitors' in the given office space. Learned affordance-map for the affordance 'working' is shown in the Fig. 4. It only shows 2d locations of the map with likelihood values related to the rotation angle θ that maximize affordance likelihood. The ground truth map was generated with ROS Gmapping using a laser range finder. High 'workable' affordances were recorded near work benches as can be seen from the Fig. 4. A 3D representation of the affordance map with skeletons is shown in Fig. 3. It is clear from these results that our algorithm is sufficiently capable of learning tested affordance in the given environment. More importantly learned affordances are more or less realistic to real human behaviours as can be seen from the 3D affordance-map in Fig. 3. Few high affordance probabilities can be observed outside normal working areas due to the lack of sufficient 3D data.

The affordance-map is then used for active object search. 'Computer monitor' was selected as the object to be searched in the experiments. Fig. 5 and Fig. 6 show position information of the robot that gives the best view of the object 'computer monitor'. These positions are calculated using object, human and Robot relationships explained in Section VI. It is clear from these results, most of the time robot positions itself

Fig. 4. Affordance Probability shown on a 2D Laser map

TABLE I
OBJECT DETECTION RESULTS SUMMARY

Number of Monitors in the environment	33
Total Number of Snaps taken	49
Number of Snaps with Monitors in the image	43
True Positive Detections	37
False positive Detections	14
True Negative Detections	6
False Negative Detections	8
Precision	0.72
Recall	0.87

near work benches so it can obtain clear views of "computer monitors".

Fig. 7 shows the path planning results of the robot for object search. It uses the Probabilistic Road Map (PRM) based path planner [10]. To move the robot across all possible search locations efficiently, we formulated this problem as a 'Travelling Salesman problem'. Since finding exact solution for the Travelling salesman problem is NP hard, Nearest Neighbour Algorithm [11] is used to estimate the possible path of the robot. As depicted in the Fig. 7, calculated path of the robot covers all possible locations that "computer monitors" can be found.

Once the robot reaches the predicted location, it captures RGB images of the possible areas that the computer monitors can be found. Then a simple object classifier based on boosting [12] is used to recognize monitors in the scene. Samples form the object detection experiments are shown in Fig. 8, and object recognition results are summarized in the Table 1.

According to the test results, only 49 snaps are taken across the entire office environment and 43 of those images contain one or more monitors. This shows, our approach can effectively infer the possible locations of monitors in a large environment. On the other hand, object detector performed well giving acceptable recall and precision values. Note here that, none of the monitors present in the environment are used to train the object detector. Although the object detector is neither scale invariant nor view point invariant our algorithm was able to detect 27 monitors out of 33 monitors present in the given

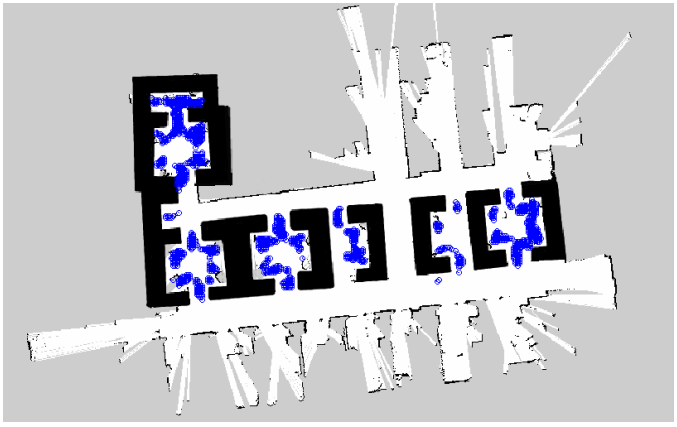

Fig. 5. Positions of the Robot that give the best view of the object

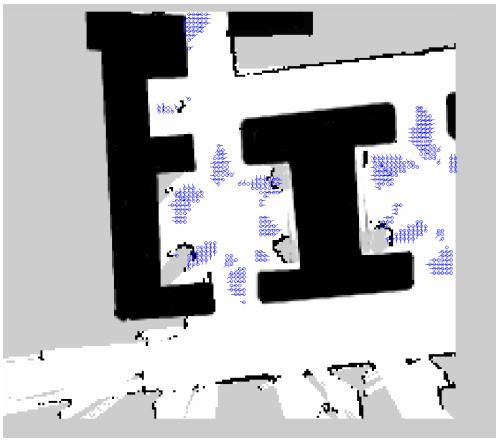

Fig. 6. Scaled view of the robot's position and orientation for object search

search area. This proves robot is able to position its camera correctly towards the object. It is clear from these results that the proposed algorithm can solve the object search problem efficiently by modelling the human context in the environment through affordance-map.

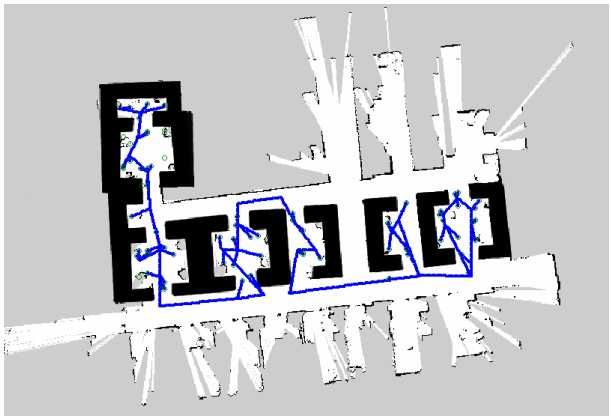

Fig. 7. Path planning for active object search

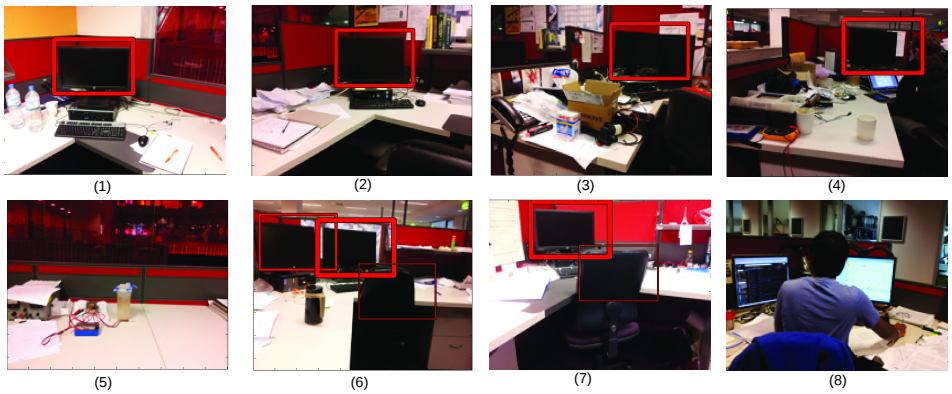

Fig. 8. Samples from object search and detection experiments. (1-4) True Positive, (5) True Negative, (6-7) False positive, (8) False Negative

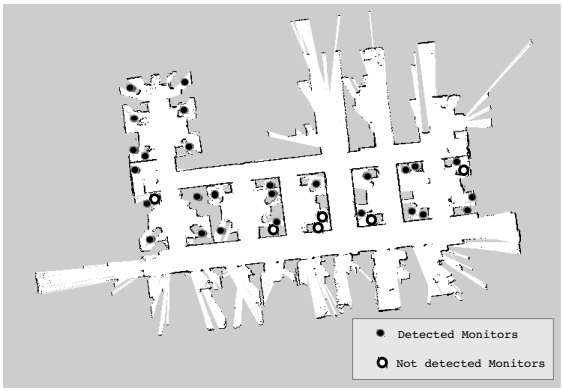

Fig. 9. Object detection results

VII. Conclusions

In this paper, we introduced novel active object search approach centered around human context in an indoor environment. We also showed how a dense 3D point cloud can be converted in to a more informative semantic map called affordance-map which consists of virtual human models. The affordance-map is then used to actively search objects in an office environment. The object search is carried out by a naive algorithm but lead to great detection accuracies. This is due to the correct pose estimation based on the object-human-robot relationship model. The experiments carried out in the large office environment proved our approach can efficiently search for given objects.

Our future works involve detecting multiple affordances in a given environment and using affordance-map as a prior for human activity detection [13] [14] .

References

[1] S. Vasudevan, S. Gchter, and R. Y. Siegwart, *Cognitive Spatial Representations for Mobile Robots: Perspectives from a User Study*.

[2] Y. Jiang, H. Koppula, and A. Saxena, "Hallucinated humans as the hidden context for labeling 3d scenes," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2993–3000.

[3] J. G. Greeno, "Gibson's affordances." 1994.

[4] J. Ma and J. W. Burdick, "A probabilistic framework for stereo-vision based 3d object search with 6d pose estimation," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2036–2042.

[5] A. Aydemir, K. Sjoo, J. Folkesson, A. Pronobis, and P. Jensfelt, "Search in the real world: Active visual object search based on spatial relations," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2818–2824.

[6] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena, "Contextually guided semantic labeling and search for three-dimensional point clouds," *The International Journal of Robotics Research*, vol. 32, no. 1, pp. 19–34, 2013.

[7] H. Grabner, J. Gall, and L. V. Gool, "What makes a chair a chair?" in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1529–1536.

[8] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 842–849.

[9] I. Dryanovski, R. G. Valenti, and J. Xiao, "Fast visual odometry and mapping from rgb-d data," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2305–2310.

[10] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *Robotics and Automation, IEEE Transactions on*, vol. 12, no. 4, pp. 566–580, 1996.

[11] D. S. Johnson and L. A. McGeoch, "The traveling salesman problem: A case study in local optimization," *Local search in combinatorial optimization*, vol. 1, pp. 215–310, 1997.

[12] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004.Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, June 2004, pp. II–762–II–769 Vol.2.

[13] L. Piyathilaka and S. Kodagoda, "Human activity recognition for domestic robots," in *Field and Service Robotics Conference*. Springer, 2013, pp. 567–572.

[14] L. Piyathilaka and S. Kodaagoda, "Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features," in *Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on*. IEEE, 2013, pp. 567–572.