Scene Interpretation for Lifelong Robot Learning

Mustafa Ersen and **Melodi Deniz Ozturk** and **Mehmet Biberci** and **Sanem Sariel** and **Hulya Yalcin** [1]

Abstract. Integration of continual planning, monitoring, reasoning and lifelong experimental learning is necessary for a robot to deal with failures by gaining experience through incremental learning and using this experience in its future tasks. In this paper, we propose a scene interpretation system combining 3D object recognition and scene segmentation in order to maintain a consistent world model involving relevant attributes of the objects and spatial relations among them. This system provides the required information for our lifelong experimental learning framework based on Inductive Logic Programming (ILP) for framing hypotheses. Necessary attributes are determined for both known and unknown objects. For known objects, size, shape, color and material attributes are provided beforehand in the model. For unknown objects, observable attributes are determined based on similarities with modeled objects. Using segmentation masks, new models are automatically extracted for these objects. The experiments to evaluate our system include both on-ground and tabletop scenarios on our Pioneer 3-AT robot and Pioneer 3-DX based humanoid robot. The results of these experiments show that the required inputs for learning can be extracted successfully from the environment by using our system for interpreting the scenes based on the data acquired through the onboard RGB-D sensors of our robots.

1 Introduction

While a cognitive robot executes its plan in the real-world, different types of failures may occur due to the gap between the real-world facts and their symbolic counterparts (Figure 1), unexpected events affecting the current state of the world or internal problems [11]. To deal with these types of failures, a consistent world modeling [16] is essential for efficient execution monitoring. Symbolic representations of world facts should continually be updated through the sequence of states. Furthermore, the robot also needs to gain experience during its lifelong operation and use its experience for different contexts to guide its future tasks. This requires having onboard learning abilities to frame hypotheses through action execution experiences. Lifelong experimental learning based on Inductive Logic Programming (ILP) can be considered for this purpose [12]. Hypotheses are framed for mapping execution contexts to action failures. Execution contexts including the observable attributes of and the relations among the objects are represented in first-order logic sentences or probabilistic graphical models. The derived hypotheses can further be generalized by incorporating background knowledge and are to be used for guidance in planning [22].

Consider in a 3-block (a, b, c) stacking tabletop scenario, after $on(b, c)$ is achieved, execution of $stack(a, b)$ fails. The hypothesis

[1] Artificial Intelligence and Robotics Laboratory, Istanbul Technical University, Turkey, email: ersenm@itu.edu.tr

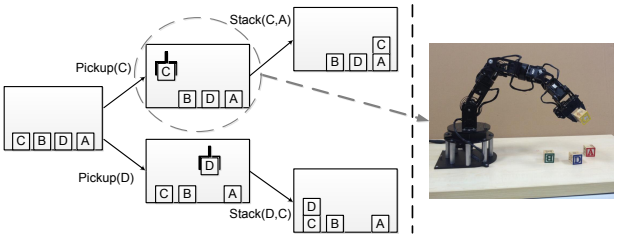


Figure 1. Planning operators are symbolic representations of actions in the real world. The scenes should be continually interpreted and symbolically represented to plan, monitor action execution and apply further reasoning.

explaining this relation is given as follows:

$$\begin{aligned} holding(a) \wedge clear(b) \wedge on\_table(c) \wedge on(b,c) \wedge \\ red(a) \wedge green(b) \wedge yellow(c) \wedge \\ cube(a) \wedge cube(b) \wedge cube(c) \\ \Rightarrow fails(stack(a,b)) \end{aligned} \quad (1)$$

where the predicates in the premise part of this implication should be maintained in the knowledge base (KB) by temporal interpretation of the scenes to date. Efficient and consistent scene interpretation is a prerequisite for determining predicates related to both the attributes of the objects (e.g., $red(a)$, $cube(a)$) and the spatial relations (e.g., $on\_table(c)$, $on(b, c)$) among them. These predicates should be continually monitored for detecting scene anomalies.

Figure 2 illustrates a classification for properties of physical objects to be manipulated by robots. Among these attributes, size, shape, color, texture, location, orientation and grasp positions can be extracted from a scene by applying 3D vision algorithms on the data obtained using an RGB-D sensor. However, further sensory modal-

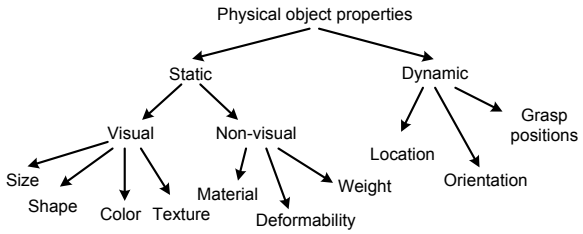


Figure 2. A general classification for properties of physical objects to be manipulated by robots. If the robot does not have a predefined model for the values of these attributes, some of them can be acquired by observations.

ities are needed to extract the other properties. When the robot interacts with known objects, it can use their predefined models (i.e., templates) and different properties encoded in the model. If the object models are not known in advance, the robot is expected to extract objects' some of the observable visual features (e.g., color features, texture features such as SIFT [15], 3D shape features [2], etc.) as well as their spatial relations.

We propose a scene interpretation system combining 3D object recognition and scene segmentation results in order to maintain a consistent world model involving relevant attributes of objects and spatial relations among them. This system provides the required information for our lifelong experimental learning framework based on Inductive Logic Programming (ILP) for framing hypotheses. Necessary attributes are determined for both known and unknown objects. For known objects, size, shape, color and material attributes are provided beforehand in the model. For unknown objects, the visual attributes are determined based on similarities with modeled objects. Our system can also automatically create templates for unknown objects by using masks created by segmentation. These templates are then stored along with the extracted attributes.

Throughout the paper, we first present earlier work related to scene interpretation. Then, we describe our system for determining attributes related to known objects based on 3D recognition and extracting these attributes from the scene for unknown objects based on segmentation. We also show how the spatial relations among all the objects are maintained as well as the attributes of the objects in a consistent world model. We then give empirical results of our approach followed by the conclusions.

2 Background

A qualitative description of the scene could be extracted by using scene interpretation based on the quantitative data obtained using visual sensors [16]. Various approaches exist for interpreting the scene. In some of these works, qualitative spatial relations are extracted from the scene by using 2D visual data [6, 20]. In another work, topological spatial relations *on* and *in* are determined among the recognized objects using SIFT keypoints [15] and these relations are used for guiding the search of the robot [19]. There is another study with a similar purpose of improving the search of a robot by considering qualitative spatial relations based on a Gaussian Mixture Model (GMM) [14]. Another work is based on extracting proximity-based high-level relations (e.g., relative object positions to find objects that are generally placed together) by considering 3D Euclidean distances between pairs of recognized objects in the scene [13]. In another study, a probability model for the symbolic world representation is constructed based on observations from the environment and this model is used for manipulation planning [3]. Another relevant study is based on learning symbolic representations of manipulation actions by extracting relevant features from a few teacher demonstrations [1]. Moreover, there are some studies using semantic knowledge for scene interpretation [8, 9]. Hawes et al. have proposed a system for reasoning about spatial relations based on context (e.g., nearby objects, functional use etc.) in previously unseen scenes [9]. To maintain a consistent model of the world, Elfring et al. have proposed a system based on probabilistic multiple hypothesis anchoring for associating data from different sources [4].

Different from these studies, our system combines 3D recognition and segmentation results to create and maintain a consistent world model involving attributes of the objects and spatial relations among them. These attributes are extracted from the scene to be used in our lifelong learning framework based on Inductive Logic Programming (ILP) [12]. Unknown objects are modeled by using the segmentation output to determine their sizes and considering similarities with existing models to determine their shapes and colors. Then, these models are also stored as templates to be used for recognition along with the extracted attributes.

3 Extracting Predicates from Scenes

In this work, we propose a temporal scene interpretation system for cognitive robots to represent their world models symbolically in their KBs including the attributes of and the relations among the objects in the environment. We focus on the extraction of *size*, *shape* and *color* attributes as well as the following unary and binary spatial relations: *on*, *on_ground/on_table*, *clear* and *near* for object manipulation scenarios. These predicates are generated and updated over time to maintain a consistent world model. Later, these predicates are to be used for monitoring execution and framing hypotheses for failure cases. The raw data taken from the environment by the sensors of the robot are processed through a series of different operations (e.g., 3D object recognition, scene segmentation etc.) running in parallel to maintain a more enhanced interpretation of the world. The system is designed to recognize modeled objects with manually fed templates and attribute values. Furthermore, a segmentation based automatic template and attribute extraction method is proposed for unknown objects.

3.1 Associating the Attributes of Known Objects

In our system, objects are modeled using LINE-MOD templates [10] to recognize them in the environment. LINE-MOD is a linearized multi-modal template matching approach considering both the color gradients around the borders of the object and the surface normals inside the template of the object. This method is suitable for object recognition tasks in object manipulation scenarios for cognitive robots as it is fast enough to be processed in real-time using an ordinary computer. We further involve color histograms in the HSV color space to verify the recognition results obtained from LINE-MOD in order to have more reliable outcomes [5]. V(value) channel is not considered while creating these histograms as it is strongly dependent on the illumination conditions.

In order to maintain a consistent world model by eliminating wrong recognition results due to noisy data obtained using the RGB-D camera, a temporal filtering method is used [16]. This system generates and maintains objects in the world model and their attributes depending on the cumulative recognitions made by LINE-MOD. Before determining an object's attributes, first, a predicate $object(o_i)$ is created in the KB representing the existence of an object and is associated with a temporal confidence value (c_i^t) for representing the belief on the existence of the object at time step t. This value is computed with the following formula:

$$c_i^t = c_i^{t-1} + \begin{cases} -0.1, & \text{if } \sum\limits_{s_j} c_{ij}^t = 0 \wedge \sum\limits_{s_j} f_{ij}^t > 0 \\ \sum\limits_{s_j} w_j \cdot f_{ij}^t \cdot c_{ij}^t, & \text{if } \sum\limits_{s_j} c_{ij}^t > 0 \wedge c_i^{t-1} < 1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

When an object is detected in the scene, the confidence value for the existence of this object is increased based on similarity values (Formula 2 Equation 2). c_{ij}^t denotes the match similarity of the recognition of object o_i by source s_j (LINE-MOD, LINE-MOD&HS or segmentation), which has an empirically determined weight w_j that

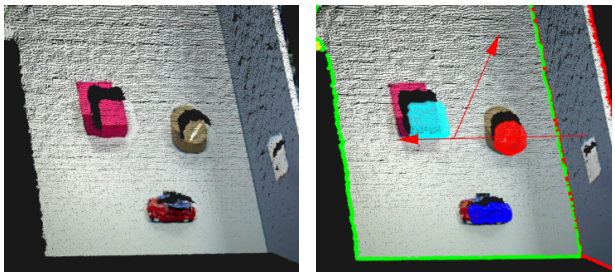
specifies the reliability of the source. Similarity value for segmentation is taken as 1 as it does not involve matching based on similarity. According to our previous experiments, we have observed that using LINE-MOD along with color histograms is more reliable than only using LINE-MOD [5]. Depending on the robot's location and orientation, objects in some portions of the environment are expected to be recognized while those that are outside the robot's field of view (f_{ij}^t) cannot be detected. This field of view is determined by the range of the RGB-D sensor and the recognition performance [16]. When an object o_i is not in the robot's field of view, its confidence value is preserved (Formula 2 Equation 3).

Recognition results on intersecting regions are treated as belonging to the same object even the recognized types and colors are different. For this case of conflicting assignments, the possible types and colors for an object are held in weighted lists to find the likely values. The type and color with the maximum likelihoods are then predicted to be the real type and color of the object in the final decision. Deletion of an object from the KB is performed gradually by considering the noise in the obtained sensor data as the robot moves in the environment. If the robot cannot recognize an object that is expected to be recognized in a location within its field of view in a time step t, the confidence value for the object is decreased by 0.1 (Formula 2 Equation 1). When the confidence value of an object becomes zero, the object is deleted from the KB.

3.2 Reasoning About the Attributes of Unknown Objects

It is likely to encounter unknown objects as the robot explores its environment (perception in the wild). To reason about these objects, besides a template-based recognition method, a depth-based segmentation algorithm is applied to the point cloud acquired from the RGB-D sensor. Organized Point Cloud Segmentation [21] is used for this purpose. The point cloud frame gathered from the sensor is first filtered such that the points farther than a given threshold (empirically determined as 2 m) are removed. Then, planar regions above a certain size (empirically determined as 10,000 points) are segmented out since they represent large planes such as ground, table, and walls. On the points of the remaining cloud, Connected Component Labeling is applied using Euclidean distance as a comparison metric [21]. Here, each label represents a cluster. The clusters having less points than the lower threshold and more points above the upper threshold (500-2000 points for ground robot scenarios and 300-1000 points for tabletop scenarios) are discarded. The others are defined as object clusters/segments, and their sizes and center locations in the global map are calculated. A sample output from the segmentation algorithm depicting three objects can be seen in Figure 3.

In order to extract only the target objects in the scene, segments that are not on the ground/table or larger than a certain threshold are further discarded by the scene interpreter. As the objects detected only by segmentation have no type information in the KB, it is not possible to associate encoded attributes with these objects. However, values for the observable attributes can be determined by further reasoning based on similarities with the modeled objects. To be able to extract visual features for unknown objects detected as segments, binary masks are passed to the automated template generation procedure for creating multi-modal LINE-MOD templates and HS histograms to reason about the shape and color of unknown objects. When an unknown object is added to the KB, the automatically created LINE-MOD template for this object is compared with the templates of the known objects in order to find the object with the most



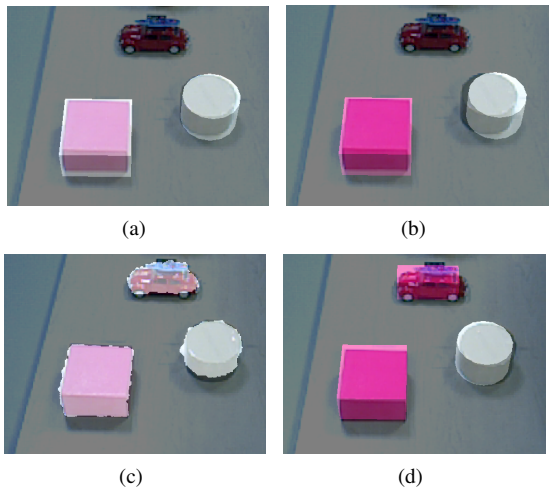
(a) (b)

Figure 3. (a) Point cloud of the scene and (b) output of segmentation where the two planes are determined and the three objects are segmented.

similar shape. Similarity threshold is taken as half of the one used in recognition. To determine the color of an unknown object, HS color histogram comparison is applied with the histograms of the modeled objects and the result with the best correlation above 0.2 is selected. In Figure 4, an example case is given to illustrate the results of LINE-MOD, LINE-MOD&HS, segmentation and the fused information in the KB. In the given case, the red toy car is not modeled beforehand. However as it can be seen from the figure, its shape and color attributes can be determined by using automatically created templates. For its shape, a similar sized prism is selected as the most similar template which is the one that belongs to the pink box. Similarly, its color is determined to be pink. Note that the color histogram of the cylindrical covered tape is not provided. Since a similar color object is not modeled, the color of the object cannot be determined by recognition, and only its shape information is taken from the output of only LINE-MOD.

3.3 Determining Spatial Relations Among Objects

In object manipulation tasks, extracting spatial relations among the objects is of great importance as well as the attributes describing these objects. Two spatial relations are considered in our system: *on* and *near* [16]. Moreover, spatial predicates associated with the



(a) (b)

(c) (d)

Figure 4. Construction of the KB: (a) LINE-MOD results, (b) LINE-MOD &HS results, (c) segmentation results and (d) KB fusing all this information.

on relation in the blocks world domain (i.e., *on_table* and *clear*) are determined along with the *stability* for the *on* relation.

Initially all the present objects in the world model are assumed to be *on_table* (*on_ground* for ground robot scenarios) and *clear*. The *on* relation is determined as follows for each pair of objects (o_i, o_k) by checking whether their projections on the ground plane (i.e., xy plane) overlap and the bounding box of one of the objects lie on the bounding box of the other one.

$$\forall o_i, o_k, \neg(DC_{xy}(o_i, o_k) \vee EC_{xy}(o_i, o_k)) \wedge UP(o_i, o_k) \Rightarrow on(o_i, o_k) \quad (3)$$

where *DC(disconnected)* and *EC(externally connected)* are topological predicates of RCC8 [18] and *UP* is a directional predicate which can be considered as the 3D expansion of *N(north)* from cardinal direction calculus [7]. *stability* is determined for stacks of objects as inversely proportional to the size of the unsupported region of the top object. For the objects involved in the *on* relation, *on_table* (or *on_ground*) and *clear* predicates are updated as follows,

$$\forall o_i, o_k, on(o_i, o_k) \Rightarrow \neg on\_table(o_i) \wedge \neg clear(o_k) \quad (4)$$

For each pair of localized objects in the environment, the *near* relation is determined by comparing the distance between the centers of these objects in each dimension (i.e., x, y and z) with the sum of the corresponding sizes in that dimension.

4 Experimental Evaluation

To evaluate the proposed system, we have conducted both ground and tabletop experiments. In ground experiments, we have used our Pioneer 3-AT robot and the objects shown in Figure 5. Tabletop experiments have been made on our humanoid robot with a Pioneer 3-DX base and the object set shown in Figure 6. Both robots are equipped with front sonar sensors to detect obstacles and Hokuyo laser rangefinders on top of the bases of them facing forward for mapping and localization in the environment. We also placed ASUS Xtion Pro Live RGB-D cameras on top of the laser rangefinders for 3D object recognition and segmentation. Humanoid robot has an additional RGB-D camera mounted on its head to be used for interpreting the scene in tabletop object manipulation scenarios. The robots have 2-DOF grippers to manipulate objects. Intel i5 laptops with Ubuntu 12.04 are used to control the robots and all the system is implemented under ROS (Robot Operating System) framework [17].

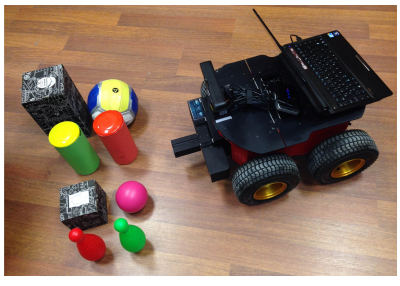


Figure 5. Pioneer 3-AT robot and the objects used in the ground exp.

Figure 6. Pioneer 3-DX and the objects used in the tabletop exp.

4.1 Ground Experiments

In these sets of experiments, we have evaluated the performance of our proposed system using our ground robot on the objects shown in Figure 5. These objects are a green plastic bowling pin, a red plastic bowling pin, a green cylindrical box, a red cylindrical box, two black boxes with different sizes, a small pink ball and a big beach ball. In these experiments, threshold for matching LINE-MOD templates is set to 80% considering the noise in the acquired data from the RGB-D sensor and threshold for HS histogram correlation is determined to be 0.3 to deal with changing lighting conditions.

First, we have evaluated the validity of the automatically extracted attributes for unknown objects in the physical world. In this set of experiments (involving 20 trials by 5 trials per each object), the models of the large black box, beach ball and the cylinders have been provided beforehand. The shape and color attributes of the remaining objects are tested. The results are given in Table 1. As seen from these results, the proposed similarity-based reasoning approach generally gives reasonable values for the shape and color attributes. The shapes of the plastic bowling pins cannot be determined in half of the trials as there is not a similar shaped object in the stored models and in the other half of the trials these pins are determined to be cylindrical as the curvature of the modeled cylinders is somewhat similar to that of the pins. Similarly, the pink ball has been observed to be classified as a cylinder in one of the trials. For the small box, its shape has been determined correctly as a prism in all 5 trials. It seems in general, the most similar color has been observed to be assigned for each unknown object. In some trials (e.g., the black box), wrong values are determined for the color due to changing illumination conditions.

Table 1. Confusion matrix for unknown objects in the ground exp.

	Shape				Color				
	prism	cylinder	sphere	none	green	red	black	mixed	none
green pin	0	2	0	3	5	0	0	0	0
red pin	0	3	0	2	0	4	1	0	0
pink ball	0	1	3	1	0	5	0	0	0
black box	5	0	0	0	0	1	3	1	0

Second, the determined spatial relations among the objects in the world model have been evaluated using *precision*, *recall* and *F-score* measures. This set of experiments have been conducted on 20 different scenes for each case with the scenes involving objects having the *on* relation, the scenes involving objects having the *near* relation and the scenes involving no pairwise object relations. As seen from the results presented in Table 2, our system can successfully determine relations for/among the objects in most of the trials. The highest error rates are observed for the *on* relation and determining its stability where the former is mainly caused by object recognition and segmentation problems, and the latter is due to alignment problems. The success in determining the *on* relation also accounts for the errors in the *clear* and *on_ground* relations. *near* relation is observed to be determined correctly in all the scenes.

Table 2. Success in extracting spatial relations in the ground exp.

	Precision	Recall	F-Score
on	94.74%	90.00%	92.31%
stable	88.89%	100.00%	94.12%
on_ground	97.44%	98.70%	98.06%
clear	97.44%	98.70%	98.06%
near	100.00%	100.00%	100.00%

4.2 Tabletop Experiments

These sets of experiments have been conducted to see whether the results for the ground experiments are reproduced for smaller objects suitable for tabletop object manipulation tasks. The used objects are a pink paper box, a brown cylindrical covered tape, a yellow toy car, a red toy car, a wooden toy wagon, a beige world globe, a toy pumpkin and a white half spherical squeezed plastic bag (See Figure 6). Different from the ground experiments, similarity threshold for LINE-MOD is taken as a higher value (i.e., 90%) to have more reliable recognition results for these smaller objects.

First, we have evaluated the validity of the determined attributes for unknown objects. In this set of experiments (involving 20 trials by 5 trials per each object), the models of the pink box, the brown cylindrical tape, the yellow toy car and the beige world globe have been provided beforehand. The shape and color attributes of the remaining objects are validated. Similar results are obtained as the ground experiments (See Table 3). Note that since the shape of the yellow car is encoded in the KB as a prism, the shape of the red car is automatically determined as a prism in all the trials. The validity of color values are better than that of the ground experiments as the lighting conditions are more stable.

Table 3. Confusion matrix for unknown objects in the tabletop exp.

	Shape				Color				
	prism	cylinder	sphere	none	pink	yellow	brown	beige	none
red car	5	0	0	0	5	0	0	0	0
pumpkin	0	4	1	0	0	1	0	0	4
wagon	5	0	0	0	0	1	0	4	0
plastic bag	0	1	4	0	0	0	0	5	0

Second, the determined spatial relations among the objects in the world model have been evaluated using the same measures as in the ground experiments with the same number of trials for each case (See Table 4). Slightly better results are obtained for the tabletop objects as the recognition performance is better due to more stable illumination conditions.

Table 4. Success in extracting spatial relations in the tabletop exp.

	Precision	Recall	F-Score
on	100.00%	95.00%	97.44%
stable	94.44%	94.44%	94.44%
on_ground	98.04%	100.00%	99.01%
clear	98.04%	100.00%	99.01%
near	100.00%	100.00%	100.00%

4.3 Scene Interpretation Experiments

Finally, we have conducted object manipulation experiments on Pioneer 3-AT robot to evaluate the performance of the scene interpreter. In this set of experiments, the robot is given the task of recognizing and collecting objects (see Figure 7 for the objects and how they are placed in the environment) from the environment and moving them to a given destination. The used objects are two cylindrical boxes in different colors, three plastic bowling pins in different colors, three small balls in different colors and two big balls in different colors. These similar shaped and similar colored objects are selected to ensure confusion of types and colors for testing our scene interpreter in a challenging scenario. The detection results for the types of objects using LINE-MOD and for the colors of objects using LINE-MOD&HS histograms are given in Table 5 and Table 6 along with the interpretations for each case. The numbers in the tables show the number of detections obtained for each object as the objects are maintained in the KB. As seen from these results, the robot obtains wrong detections for the types and the colors of some of the objects. This is mainly caused by the noise in the acquired data from the RGB-D camera while the robot is moving in the environment. For example, the robot recognizes small balls on the surface of big balls or on the top of cylindrical boxes. The colors of similar shaped objects are also confused using LINEMOD&HS histograms due to changing light conditions from different angles and the noisy camera data. The most remarkable erroneous detection has been made for the type of the beach ball. Although the numbers of correct and wrong detections are the same, the interpreter concludes with the correct type in the KB for the beach ball as the detections for the type big ball have higher similarity values. When all these results are examined, it can be seen that the scene interpreter has concluded the correct type and color for each of the objects manipulated by the robot in this set of experiments.

Figure 7. The objects manipulated by our Pioneer 3-AT robot.

Table 5. Success in interpreting types of objects.

	cylinder	small ball	big ball	pin	*interpretation*
red cylinder	57	6	0	4	cylinder
green cylinder	64	0	0	0	cylinder
pink ball	0	123	0	0	small ball
purple ball	0	134	0	0	small ball
yellow ball	0	64	0	0	small ball
orange ball	0	9	23	0	big ball
beach ball	0	14	14	0	big ball
red pin	0	0	0	66	pin
green pin	0	0	0	79	pin
blue pin	0	1	0	68	pin

Table 6. Success in interpreting colors of objects.

	red	green	pink	purple	yellow	orange	mixed	blue	*interpretation*
red cylinder	20	0	0	0	0	0	0	0	red
green cylinder	0	26	0	0	0	0	0	0	green
pink ball	0	0	40	0	11	0	0	0	pink
purple ball	0	0	0	37	9	0	0	0	purple
yellow ball	0	0	0	0	34	0	0	0	yellow
orange ball	0	0	0	0	0	11	0	0	orange
beach ball	0	0	0	0	3	0	8	0	mixed
red pin	26	0	0	0	0	0	0	0	red
green pin	0	22	0	0	0	0	0	0	green
blue pin	0	0	0	0	0	0	0	24	blue

5 Conclusion

The temporal scene interpreter presented in this paper is designed for automated extraction of the required predicates to be used in our lifelong experimental learning framework based on ILP. Attributes of the objects and spatial relations among them can be used as the premise parts of hypotheses to reason about failure situations and to guide the future decisions of the robot. First, we have presented the attributes that are considered for hypothesis generation and which of them can be determined from visual data obtained using an RGB-D sensor. Then, we have demonstrated how the required attributes are modeled for known objects and how observable attributes of unknown objects can be derived from the scene in a consistent world modeling framework. Finally, we have shown how the spatial relations among the objects in the world model are determined to represent the states and the execution contexts during runtime. The proposed scene interpreter is evaluated on object manipulation experiments in which the robot is given the task of recognizing objects in the environment and moving them to a given destination. The results of these experiments indicate that the presented system can be used to extract the required predicates from the scene to be used as an input to our lifelong learning framework based on ILP.

Acknowledgments

This research is funded by a grant from the Scientific and Technological Research Council of Turkey (TUBITAK), Grant No. 111E-286. We thank Prof. Muhittin Gokmen for his recommendations on vision algorithms. We also thank Melis Kapotoglu, Cagatay Koc and Dogan Altan for their helpful comments.

REFERENCES

[1] N. Abdo, H. Kretzschmar, L. Spinello, and C. Stachniss, 'Learning manipulation actions from a few demonstrations', in *Proc. of 2013 IEEE International Conference on Robotics and Automation (ICRA'13)*, pp. 1268–1275, (2013).

[2] A. Aldoma, Z. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze, 'Tutorial: Point cloud library: Three-dimensional object recognition and 6 DOF pose estimation', *IEEE Robotics and Automation Magazine*, **19**(3), 80–91, (2012).

[3] R. Dearden and C. Burbridge, 'Manipulation planning using learned symbolic state abstractions', *Robotics and Autonomous Systems*, **62**(3), 355–365, (2014).

[4] J. Elfring, S. van den Dries, M. J. G. van de Molengraft, and M. Steinbuch, 'Semantic world modeling using probabilistic multiple hypothesis anchoring', *Robotics and Autonomous Systems*, **61**(2), 95–105, (2013).

[5] M. Ersen, S. Sariel-Talay, and H. Yalcin, 'Extracting spatial relations among objects for failure detection', in *Proc. of the KI 2013 Workshop on Visual and Spatial Cognition*, pp. 13–20, (2013).

[6] Z. Falomir, E. Jiménez-Ruiz, M. T. Escrig, and L. Museros, 'Describing images using qualitative models and description logics', *Spatial Cognition and Computation*, **11**(1), 45–74, (2011).

[7] A. U. Frank, 'Qualitative spatial reasoning with cardinal directions', in *Proceedings of the 7th Austrian Conference on Artificial Intelligence*, pp. 157–167, (1991).

[8] C. Gurău and A. Nüchter, 'Challenges in using semantic knowledge for 3D object classification', in *Proc. of the KI 2013 Workshop on Visual and Spatial Cognition*, pp. 29–35, (2013).

[9] N. Hawes, M. Klenk, K. Lockwood, G. S. Horn, and J. D. Kelleher, 'Towards a cognitive system that can recognize spatial regions based on context', in *Proc. of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, pp. 200–206, (2012).

[10] S. Hinterstoisser, C. Cagniart, S. Ilic, P. F. Sturm, N. Navab, P. Fua, and V. Lepetit, 'Gradient response maps for real-time detection of textureless objects', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **34**(5), 876–888, (2012).

[11] S. Karapinar, D. Altan, and S. Sariel-Talay, 'A robust planning framework for cognitive robots', in *Proc. of the AAAI-12 Workshop on Cognitive Robotics*, pp. 102–108, (2012).

[12] S. Karapinar, S. Sariel-Talay, P. Yildiz, and M. Ersen, 'Learning guided planning for robust task execution in cognitive robotics', in *Proc. of the AAAI-13 Workshop on Intelligent Robotic Systems*, pp. 26–31, (2013).

[13] A. Kasper, R. Jäkel, and R. Dillmann, 'Using spatial relations of objects in real world scenes for scene structuring and scene understanding', in *Proc. of the 15th IEEE Intl. Conference on Advanced Robotics (ICAR'11)*, pp. 421–426, (2011).

[14] L. Kunze and N. Hawes, 'Indirect object search based on qualitative spatial relations', in *Proc. of the IROS-13 Workshop on AI-based Robotics*, (2013).

[15] D. G. Lowe, 'Object recognition from local scale-invariant features', in *Proc. of the 7th IEEE Intl. Conference on Computer Vision (ICCV'99)*, pp. 1150–1157, (1999).

[16] M. D. Ozturk, M. Ersen, M. Kapotoglu, C. Koc, S. Sariel-Talay, and H. Yalcin, 'Scene interpretation for self-aware cognitive robots', in *Proc. of the AAAI Spring Symposia 2014 Qualitative Representations for Robots*, (2014).

[17] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, 'ROS: an open-source robot operating system', in *Proc. of the ICRA-09 Workshop on Open Source Software*, (2009).

[18] D. A. Randell, Z. Cui, and A. G. Cohn, 'A spatial logic based on regions and connection', in *Proc. of the 3rd Intl. Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, pp. 165–176, (1992).

[19] K. Sjöö, A. Aydemir, and P. Jensfelt, 'Topological spatial relations for active visual search', *Robotics and Autonomous Systems*, **60**(9), 1093–1107, (2012).

[20] H. S. Sokeh, S. Gould, and J. Renz, 'Efficient extraction and representation of spatial information from video data', in *Proc. of the 23rd Intl. Joint Conference on Artificial Intelligence (IJCAI'13)*, pp. 1076–1082, (2013).

[21] A. J. B. Trevor, S. Gedikli, R. B. Rusu, and H. I. Christensen, 'Efficient organized point cloud segmentation with connected components', in *3rd Workshop on Semantic Perception, Mapping and Exploration (SPME)*, (2013).

[22] P. Yildiz, S. Karapinar, and S. Sariel-Talay, 'Learning guided symbolic planning for cognitive robots', in *Proc. of the ICRA-13 Workshop on Autonomous Learning*, (2013).