Pushing and Grasping for Autonomous Learning of Object Models with Foveated Vision

Robert Bevec and Aleš Ude
Humanoid and Cognitive Robotics Lab
Department of Automatics, Biocybernetics and Robotics
Jožef Stefan Institute, Ljubljana, Slovenia
Email: robert.bevec@ijs.si, ales.ude@ijs.si

***Abstract*—In this paper we address the problem of autonomous learning of visual appearance of unknown objects. We propose a method that integrates foveated vision on a humanoid robot with autonomous object discovery and explorative manipulation actions such as pushing, grasping, and in-hand rotation. The humanoid robot starts by searching for objects in a visual scene and generating hypotheses about which parts of the visual scene could constitute an object. The hypothetical objects are verified by applying pushing actions, where the existence of an object is considered confirmed if the visual features exhibit rigid body motion. In our previous work we showed that partial object models can be learnt by a sequential application of several robot pushes, which generates the views of object appearance from different viewpoints. However, with this approach it is not possible to guarantee that the object will be seen from all relevant viewpoints even after a large number of pushes have been carried out. Instead, in this paper we show that confirmed object hypotheses contain enough information to enable grasping and that object models can be acquired more effectively by sequentially rotating the object. We show the effectiveness of our new system by comparing object recognition results after the robot learns object models by two different approaches: 1. learning from images acquired by several pushes and 2. learning from images acquired by an initial push followed by several grasp-rotate-release action cycles.**

I. Introduction

An autonomous robot operating in home and other natural environments must be able to deal with new objects as they are encountered for the first time. Robust methods are needed so that a robot can detect a new object, explore it, and learn a complete model of its appearance and other characteristics. An active agent, e .g. a humanoid robot, has the facilities to affect the environment with its manipulation actions. It can change its point of view to see an object from a different angle [1], push an object to segment it from the background [2], [3] or even try to grasp it [4]. Feedback from manipulations has proved vital in grounding knowledge for object learning [2].

Objects can be sensed in different ways, but the most important sense is vision. In our work we exploit the advantages of foveated vision, which mimics the properties of human vision. Besides being an active sensor, human eye also differs from standard digital cameras in how photoreceptors are distributed across the retina. The highest density occupies the center of the retina called fovea, where the best acuity is achieved. In order to inspect an area of interest in greater detail, we need to shift our eyes so that the object image falls onto the fovea. At the same time, the peripheral part of the retina allows us to observe a wider field of view. It has been shown that foveal vision can be used effectively to improve the accuracy of object recognition [5]. In our previous work [6], we proposed to improve visual object learning by exploiting the advantages of this arrangement, here denoted as foveated vision, through interactive manipulation. In our system, foveated vision is realised by using two cameras per eye with different focal lengths [7]. This enables capturing wide-angle peripheral and narrow-angle foveal images at the same time using of-the-shelf equipment. The robot can simultaneously find and track objects in the peripheral cameras and recognize them in the foveal cameras, where images have the highest resolution.

978-1-4673-7509-2/15/$31.00

In this paper we extend these ideas by proposing a method that supplements foveated vision and active exploration by pushing with grasping, in-hand rotation, and release action cycles. As in our earlier work [6], the robot detects objects by processing the peripheral camera images. It then pushes an object hypothesis in order to induce a change in the scene that grounds the visual features that belong to the hypothetical object. An object hypothesis is then verified and a higher resolution object image is acquired in the foveal cameras. In order to learn a complete object representation, the robot must see the object from different viewpoints. In our new approach, the robot generates new object snapshots by first grasping the hypothetical object, rotating it slightly to change the viewpoint and then retracting the hand to make the object fully visible. Unlike in the case of pushing, where it is difficult to predict how an object will move, in our new system the robot can explore the object in a systematic manner. This results in quicker and more reliable acquisition of object models than in the case of sequential application of pushing actions.

We validated the developed system by comparing object recognition results after learning object models with the sequential application of pushing actions and with the new method, where the object is sequentially grasped, rotated, and released. The developed approach retains all the features of our previous system. It requires no prior knowledge about the objects or the environment except for the assumption that the objects of interest move as rigid bodies.

II. Related Work

The beginning of interactive perception reaches way back to its first suggestion by Tsikos and Bajcsy [8] and in the context of cognitive robotics by Fitzpatrick and Metta [2]. Recently the doctrine of interactive perception has been

Fig. 1. JSI humanoid robot consisting of Karsluhe head, Kuka LWR arms, and Barrett hands. In front of the robot we can see the objects used in the experiments.

adopted by many research groups and has seen great advances. Many researchers including Fitzpatrick and Metta [2] employed pushing to ground visual information. For example, Chang et al. [9] proposed a method for discovering objects in a pile and removing a singulated object. Evidence is gathered by matching visual features of the hypothesis before and after the perturbation, finding rigid transformations using RANSAC [10], which is similar to the system developed in [3], [11]. Similar methods using RGB-D images have been proposed in [12], where more emphasis is given to where to perturb the pile, and in [13], where depth and normal discontinuities are used to create object hypotheses. Pushing objects can also lead to the discovery of kinematic models of articulated objects [14], [15], or object affordances [16].

Pushing an object has limited capabilities in manipulation of the object. It is possible to predict how the object will move on a plane [17], [18] and in three dimensions [19]. Some methods require extensive knowledge about the environment and Newtonian physics equations, while others learn and generalize this knowledge, however it is difficult to predict motion without an accurate object model.

In-hand rotation of objects have been used by Ude et al. [20], Krainin et al. [21], Kraft et al. [22], and Browatzki et al. [23]. In these works information is acquired by observing an object while it is being rotated in the hand. The major problems with these approaches is that the object needs to be separated from the hand, which partially occludes the grasped object. The holes in the model that are left due to occlusions can be solved by grasping the object in a different way and filling the missing parts, but this requires accurate registration of object features across different views.

The problem of grasping novel objects is addressed in many works including Saxena et al. [4] who learn a classifier that determines possible grasping points of a novel object from two or more images. Kraft et al. [22] also look for possible grasping points to grasp an unknown object. The work of Li and Kleeman [24] is the closest to our present work in a sense that they use a robotic manipulator to first nudge an object to segment it from the background. The robot then grasps the object from the top and acquires different views by in-hand rotations. Unlike these works [20]–[24], we found out that we can learn better models by successfully grasping, rotating and releasing the object to gain an unobstructed object snapshots. In addition, we use foveated vision to improve the overall performance of the system.

III. System Overview

Our robot applies the following procedure to learn a new object representation:

- **Generate object hypotheses in peripheral views:** Look for surface regularity and feature proximity in the point cloud of stereo matched visual features acquired from peripheral cameras.
- **Turn the head and eyes toward one hypothesis:** The centroid of the hypothesis should lie in the middle of the foveal images.
- **Generate an object hypothesis in foveal view:** The object takes up a large portion of the foveal images, therefore all visual features represent a hypothesis.
- **Validate the hypothesis in peripheral view:** The robot pushes the object through the mean position of the hypothesis' points. It validates which features belong to the object due to the resulting change in the scene. Additional features are added if they move concurrently with the object.
- **Turn the head and eyes toward the confirmed hypothesis:** The centroid of the manipulated object should lie in the middle of foveal images.
- **Validate the hypothesis in foveal view:** The robot validates which features belong to the object due to the resulting change in the scene.
- **Object grasping:** The robot approaches the object from above until it touches the object and closes the hand. Force feedback is exploited to stop the downward robot hand motion.
- **Acquire a new view:** The robot turns the object by approximately 30°, releases it and retracts its hand.
- **Validate object features:** The robot validates which features belong to the object in the foveal and peripheral view due to the resulting change in the scene.
- **Learn the complete model:** The robot repeats the last three steps above to learn object features from a different viewpoints until a complete model is acquired.

IV. Discovering Object Candidates

The detection of object candidates starts by processing the peripheral stereo image pair. The peripheral cameras cover a much wider area than foveal cameras (see Fig. 2). Typical objects in household environments are large enough to be detected in the peripheral view of a humanoid head and still fit into the foveal view at the reaching distance, i.e. 0.5 - 1 meter. It is therefore sensible to use peripheral views when looking for object hypotheses, i.e. in the object detection phase. In this way we significantly reduce the detection time since the robot does not need to actively search for the objects by turning its head and eyes. For unusually small objects, a different, active

Fig. 2. Initial object hypotheses found in a typical scene with household objects. In the peripheral view each hypothesis is represented with the points of the same color. After the head has been turned toward the hypothesis "1", the robot generates the initial hypothesis in the foveal view, where all feature points belong to the hypothesis.

Fig. 3. The robot pushes the object to induce a change in the scene and confirm object existence. The red points represent the validated visual features of the initial hypothesis in the peripheral and foveal view.

strategy could be designed, but this is beyond the scope of this paper.

The basic idea of our approach for the generation of object hypotheses is to search for surface regularity and feature proximity to find the candidates in peripheral views. Most common household objects consist, at least partially, of regular geometric shapes such as planes, spheres and cylinders. Thus, the detection of such a shape is a strong indication about the existence of an object. A detailed description of how to look for hypotheses in peripheral views is provided in [3], [25]. The result of this procedure is shown in Fig. 2, left.

The object candidate is then inspected in detail in the foveal view by turning the head and eyes toward it. Since foveal cameras have a narrower field of view, the object covers a much larger portion of the foveal than peripheral images. It is therefore not necessary to search for object cues like in the peripheral views. Instead all visual features are included in a foveal object hypothesis. A detailed description of determining foveal view hypotheses and how to control the robot to acquire the object in foveal views is provided in [6], [26] and an example in Fig. 2, right.

V. Confirming Object Existence

Since feature proximity or a smooth surface is not sufficient evidence for existence of a physical object, the robot acquires additional information to validate the object hypothesis. This information is provided by the robot itself by pushing the hypothetical object. A push represents an appropriate manipulation at this stage, since we don't know a lot about the object candidate. Any interaction that causes movement of the object can help to resolve ambiguities about the object's extent. Our assumption is that objects move as rigid bodies. Changes in the scene can be analyzed for simultaneous feature motion, which is a very strong indicator of object existence [3].

The robot calculates the mean position of the hypothesis' feature points. It then executes a straight line push from the outside of the object toward a central point on the table, where grasping can be performed. The details of this manipulation are described in [25]. As the robot moves the object, it disappears from the narrow angle foveal view. Therefore, the robot detects rigid motion in the peripheral view and validates object existence. If an object is found, the robot turns its head and eyes toward the object and reacquires it in the foveal view. It then verifies the object features in foveal views and stores them for later model training. The verification procedure is done using a rigid body motion filter in the peripheral view and a static feature filter in the foveal view as described in [6]. The result of this procedure is shown in Fig. 3.

VI. Building the Complete Object Representation

When object existence is confirmed, the robot has learned one view of that particular object. In order to successfully recognize the object in the future, it must acquire a complete object representation, which is only possible if the robot sees the object from different views. Since the robot confirmed feature points that belong to the object in the previous stage after pushing it, it can now try to grasp the object. This task is much more feasible now than grasping the object before pushing when its extent was not known reliably. The idea is to grasp and rotate the object by 30° and then release it. This action is repeated several times until the object has been seen from all sides.

A. Grasping of Hypothetical Objects

The robot system is calibrated, which is why the robot knows the global positions of the visual features belonging to the object. We do not have a 3-D surface model to plan a precise grasp, but we can plan a grasp based on the estimation of the object size from the validated feature points. This is done by first calculating the mean position $\mathbf{p}$ and the dominant horizontal principal axis $\mathbf{a}$ of the object from the foveal confirmed visual features. Let $\mathbf{x}_i = [x_i, y_i, z_i]$ be the position of N visual features confirmed in the foveal views

$$\mathbf{p} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\,. \tag{1}$$

The gripper approaches the object in the configuration seen in Figure 4, with one finger directly opposite to the other two. It therefore make sense to grasp the object on the narrow side. By calculating the principal axes of the object feature points we estimate the object's greatest extent in each direction. First the covariance matrix $\mathbf{\Sigma}$ is calculated as

$$\mathbf{\Sigma} = \mathrm{cov}\left(\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N\}\right) = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{p})(\mathbf{x}_i - \mathbf{p})^T. \tag{2}$$

Next we calculate the three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ and eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ of the covariance matrix Σ, which is

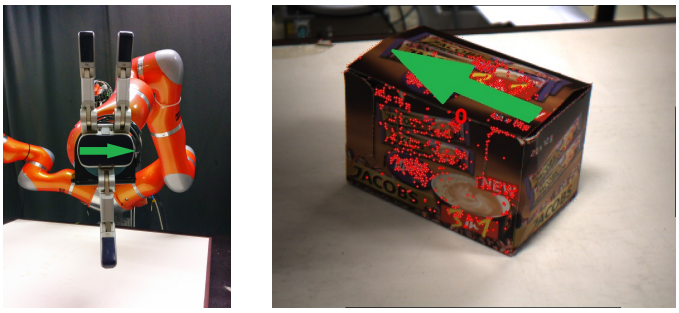

Fig. 4. The configuration of the gripper used for the grasping reflex. The green arrow shows how the dominant axis of the object should be aligned in the hand. On the right, an example of the dominant axis calculated from the object's confirmed visual features.

done by solving the equation

$$(\mathbf{\Sigma} - \lambda \mathbf{I})\,\mathbf{v} = 0 \tag{3}$$

The eigenvector associated with the biggest eigenvalue represents the axis of the greatest extent of the object.

However, we are interested in the extent of the object in the plane perpendicular to the approach direction. Because we approach the object from above and the z-axis of our global coordinate system is vertical, we have to project all the eigenvectors to the x-y plane and choose the biggest. The projected principal axes $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are:

$$[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left[\sqrt{\lambda_1}\mathbf{v}_1, \sqrt{\lambda_2}\mathbf{v}_2, \sqrt{\lambda_3}\mathbf{v}_3\right] \tag{4}$$

The dominant axis $\mathbf{a}$ is then given by

$$\mathbf{a} = \underset{\mathbf{y} \in \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}}{\arg\max}\ (\text{norm}(\mathbf{y})). \tag{5}$$

An example of the calculated dominant axis is seen in Figure 4, right.

The robot approaches the object from the top at the centre point $\mathbf{p}$ with the hand oriented along the detected dominant axis. It then moves downward until it detects a contact using force-torque sensor in the hand. When the force in the wrist exceeds a predetermined threshold, the robot stops the movement and starts closing the fingers. We evaluate the success of the grasp by reading the finger joints values. If the fingers reached the end position, the object was likely not grasped successfully. Otherwise the robot continues and rotates the object by thirty degrees. The reason why we chose that amount is explained in Section VI-B2. After that the robot slowly releases the grip and retracts the gripper back above the object, out of sight for the foveal view cameras. An example of this procedure is shown in Fig. 5. Each time the robot attempts to grasp the object, it recalculates the dominant axis. With successive manipulation it learns more information about the object, so the object's extent is calculated more reliably and consequently the successive grasping action also succeed provided the first grasp action has succeeded.

B. Feature Registration

After the first push of the object, the object disappears from the foveal views, so the robot has to locate it and confirm object existence in the peripheral view. Afterwards, the robot turns the head and reacquires the object in foveal cameras. Because the exploration by grasping only induces a rotation on the object, the object remains in the view of the foveal cameras and the robot does not need to look for it in peripheral cameras again.

1) Foveal View: The robot generates the foveal view point cloud from point correspondences, found by matching interest points in each eye. The entire point cloud represents candidate points that might belong to the object and is calculated after each grasp. There are many interest point detectors to choose from and many of them or a combination of them work with our method. In this paper we use a combination of Harris interest points [27], which are found mostly in textured parts of the image, and maximally stable extremal regions (MSER) [28] found in the areas with less texture. The two detectors balance each other by finding salient points in different image areas.

After the push the robot validates the features using a static feature filter described in [6]. We propose using a similar filtering process after the grasping, which also works under the assumption that the surrounding of the object in foveal images didn't move. The difference is, we don't take the head motion in account, because there is none and we propagate the new validated features to the previous snapshot representation.

Let $\mathbf{x}_m^k$ be a candidate visual feature from the point cloud before the grasp and $\mathbf{x}_m^{k+1}$ a matched visual feature after the rotation. A tolerance ϵ is defined as a maximum displacement of static features to account for small changes in their positions due to noise in the active stereo system. A feature has moved if it satisfies the following equation:

$$\|\mathbf{x}_m^{k+1} - \mathbf{x}_m^k\| \geq \epsilon. \tag{6}$$

After the grasp, the robot retracts the hand out of the foveal cameras' view. Because the only thing that changed in the foveal view is the position of object, all the features that have moved belong to the object.

The robot observes the object in a snapshot-like fashion - only between manipulations. Therefore, it sees a new part of the object for the first time after each rotation. The new features seen after the k-th grasp can only be successfully validated after the $k+1$-th grasp. That indicates that the set of visual features belonging to the object after the k-th grasp is the union of the features satisfying Eq. (6) after the k-th and $k+1$-th grasp. We can see an example of how the propagation of confirmed features to the previous view complements the object model in Figure 6.

2) Peripheral View: While exploring the object, the robot can also learn the object model in the peripheral view, to be able to classify an object in the object discovery phase using just the peripheral cameras. This is useful when the robot is looking for a particular object and sees several object candidates in the peripheral view. It has to decide which one to gaze toward to inspect it in the foveal view.

Since the visual features confirmed in the foveal view include 3-D positions, the robot can calculate the rigid motion the object exhibited between manipulations. Using RANSAC

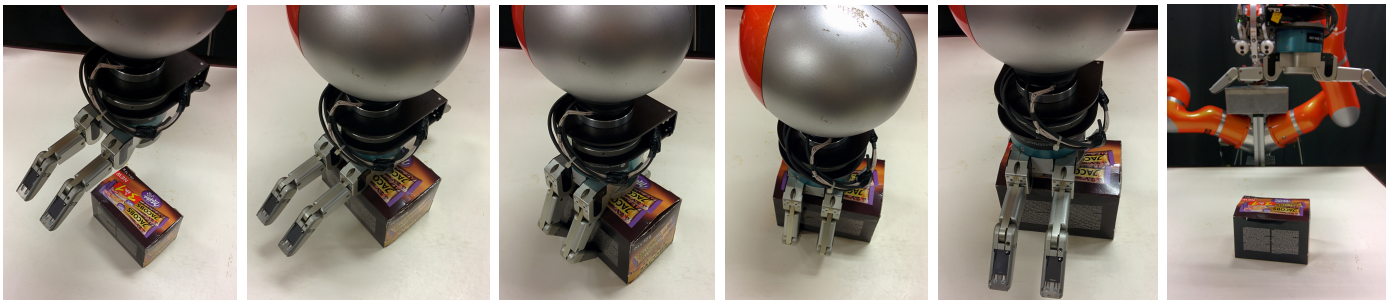

Fig. 5. The robot aligns the gripper with the dominant axis of the object, moves downward until it detects a contact, grasps the object and rotates it approximately 30°. Afterwards the object is released and the robot retracts the manipulator out of the view of the foveal cameras.

the robot finds the best rigid transformation (rotation $\mathbf{R}_{k+1}$, translation $\mathbf{t}_{k+1}$) between the corresponding confirmed visual features after the k-th and $k+1$-th grasp.

The robot can use this transformation to validate the features that belong to the object in the peripheral view. Similarly to the procedure in the foveal view, it generates the point cloud after each manipulation and checks which features correspond to the motion model with a minimum tolerance γ:

$$\|\mathbf{x}_f^{k+1} - \mathbf{R}_{k+1} \cdot \mathbf{x}_f + \mathbf{t}_{k+1}\| \leq \gamma. \quad (7)$$

Fortunately, the robot does not have to check the entire point cloud. The peripheral cameras capture a much wider area, therefore we can focus only on the volume close to the object. We limit the search area to a sphere around the mean point of the object $\mathbf{p}$. In our case a sphere with a radius of 30 cm captures all the movement that happens in the foveal view in the range of 0.5 - 1 meter from the robot.

In our system we use SIFT descriptors [29] to match the interest points, which is why we prompt the robot to try to rotate the object by 30°. It has been shown that the SIFT descriptors can be matched well enough for viewpoint changes up to 30° [30]. Since the robot rotates the object by 30° every time, it could simply execute eleven rotations and assume it has rotated it a full circle. However, the grasp of the object can cause some unintended rotation, so the rotation of the object should not be assumed. But, since we calculate the transformation of the object's pose between individual grasps, we can instead rely on this information to determine when to finish object exploration. The robot sums the transformations together and when the combined rotation around z-axis exceeds 330°, the exploration stops.

3) Object Representation: Numerous visual features define an object in snapshot-like groups from individual manipulation steps. Using these groups the visual appearance of objects is learned using a bag of features model (BoF) [31]. This method of representing objects has been shown to be distinctive and robust for object classification, even under partial occlusions.

A visual vocabulary is created by clustering SIFT feature descriptors extracted from random object training images. Each cluster is represented by a descriptor - a visual word - that replaces the descriptors defining the object, by finding their closest matches in the visual vocabulary. Since SIFT is based on gray scale data, we calculate a saturation-weighted hue histogram [32] within the ellipse spanned by the principal axes of the confirmed features projected on the image, to also include color information. Combined, the two different types of histograms represent feature vectors for recognition.

VII. Experimental Evaluation

We performed several experiments to evaluate the gain of autonomous object exploration by pushing and grasping. We compared the performance of recognition using models acquired by two different approaches: 1. learning from images acquired by several pushes (our previous approach presented in [6]) and 2. learning from images acquired by an initial push followed by several grasp-rotate-release action cycles (this paper). The robot learned representations of 10 different household objects (Fig. 1). Each object was placed on the table in front of the robot at an approximate distance of 80 cm. Using the first exploring approach, the robot pushed each object 15 times. Using our proposed method the robot grasped and rotated each object until it completed a full circle as described in Section VI-B2, but was limited to using no more than 15 manipulations. The acquired object representations were used to learn a classifier. The classification was realized by Support Vector Machines (SVM) [33] with a linear kernel and a multi-class classification approach.

For recognition each object was placed in front of the robot in 6 different poses. The randomness was ensured by asking a person, who did not witness the robot's learning procedure, to place an object in front of the robot in six different poses. The objects were placed at approximately the same distance as in the learning procedure. The recognition procedure started by classifying the object hypothesis. The robot then pushed the object three times and classified it after each manipulation. Table I shows the results of object recognition.

A setup with dual KUKA LWR arms [34] and the Karlsruhe robot head [7] was used, as seen in Figure 1. We applied active calibration procedures [35] to account for the changing robot configuration and thus enable 3D vision.

The classification of the initial hypotheses was not very successful because they didn't capture the objects in their entire extent and also included some false features from the background. After the first push the recognition rate improved in both strategies, showing the benefit of interaction for segmentation. With additional pushing both rates rose, because the object representation included more visual features, that corresponded to object motion. The grasping method (G+R) was much more successful at some objects, mainly because

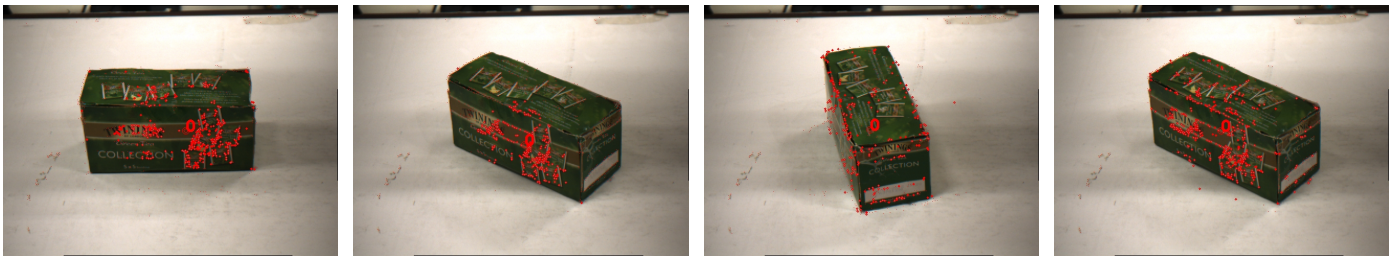

Fig. 6. As the robot grasp and rotates an object (1. image), it observes a new part of the object (2. image). The new features can only be successfully validated after the next grasp (3. image). The new validated features are then propagated back to the previous snapshot representation and joined with the existing ones (4. image).

TABLE I. OBJECT RECOGNITION RATES USING AUTONOMOUSLY LEARNED OBJECT REPRESENTATIONS BY GRASPING AND ROTATION (G+R) AND BY PUSHING ONLY (P). THE RECOGNITION RATE IS EVALUATED FOR THE INITIAL HYPOTHESES AND AFTER THE FOLLOWING THREE PUSHES.

Object	Exploration	init. hyp.	1 push	2 pushes	3 pushes
Orbit green	G+R	50 %	67 %	83 %	83 %
	P	50 %	67 %	83 %	67 %
Orbit purple	G+R	33 %	67 %	83 %	100 %
	P	33 %	33 %	50 %	67 %
Total care	G+R	67 %	83 %	67 %	83 %
	P	33 %	50 %	50 %	67 %
Light yogurt	G+R	50 %	67 %	83 %	83 %
	P	50 %	67 %	67 %	67 %
Ego yogurt	G+R	67 %	83 %	83 %	100 %
	P	33 %	83 %	67 %	83 %
Yogurt	G+R	33 %	67 %	100 %	83 %
	P	50 %	67 %	50 %	67 %
Barcaffe	G+R	50 %	67 %	83 %	100 %
	P	50 %	67 %	50 %	83 %
Leopard bag	G+R	83 %	100 %	100 %	100 %
	P	83 %	100 %	100 %	100 %
Non stop	G+R	67 %	83 %	100 %	100 %
	P	50 %	67 %	67 %	83 %
Ball	G+R	83 %	100 %	100 %	100 %
	P	83 %	100 %	100 %	100 %
Total	**G+R**	**58 %**	**78 %**	**88 %**	**93 %**
	P	**51 %**	**70 %**	**68 %**	**78 %**

our proposed systematic learning strategy managed to learn the object from more different views. When an object was presented to the robot from a view it hasn't seen before, it had a hard time classifying it, even after interacting with it. On the other hand, some objects, e.g. the ball, have a symmetrical pattern that looks similar from different views and could be successfully recognized from different views.

Compared to the results in [6] the total recognition rates were lower, because in that work the robot tried to recognize the objects only from previously seen poses. The task was to segment a single object from a group of objects. Here only one object was placed in front of the robot, but the pose of the object was random. On some rare occasions even the bottom of the object was facing the robot, however, the person placing the objects on the table subconsciously avoided these poses. That comes at no surprise since humans mostly encounter these household objects in an upright pose. The results show that using our proposed method the robot was able to learn a more complete object representation, which resulted in better object recognition.

VIII. CONCLUSION

We presented a new method for autonomous discovery of unknown objects using active perception. The main novelty compared to previous approaches is in how information about the objects is acquired; through a combination of initial pushes to verify object hypotheses followed by successive application of grasp-rotate-release action cycle. The system also exploits the advantages of foveated vision to acquire initial hypotheses faster and more reliably and to learn more accurate models of object appearance. Our approach works in arbitrary tabletop environments and relies on only two assumptions: that the object moves according to the assumed motion model, currently rigid body motion, and that it has some distinctive visual features.

The proposed approach does not require the detection, tracking and exclusion of the robot hand and the rest of the robot manipulator as some previous approaches that exploited in-hand rotation for object learning. The grasp and the object rotation do not need to be precise, it is only important that the change in the object's pose is not too large for the robot to match visual features across successive views. Our experimental results show that with the proposed approach, the robot is able to learn a more complete object representation than when only pushing the object. Since up to now we have only implemented grasping from above and rotating the object around the vertical axis, the robot cannot learn the bottom side of the object. However, this is sufficient for many practical applications because the bottom side of an object is less likely to be seen in standard interaction. It is, however, quite straightforward to implement also grasping from directions different than vertical direction, which would enable us to acquire object snapshot across the complete viewing sphere.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience. R. Bevec was also supported within the framework of the Operational Programme for Human Resources Development for the period 2007-2013, 1. development priorities Promoting entrepreneurship and adaptability, policy priorities 1.3: Scholarship Scheme. Its operation is partly financed by the European Union, European Social Fund.

REFERENCES

[1] G. Kootstra, J. Ypma, and B. de Boer, "Active exploration and keypoint clustering for object recognition," in *IEEE International Conference*

on Robotics and Automation (ICRA), (Pasadena, California), pp. 1005–1010, 2008.

[2] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2165–2185, 2003.

[3] D. Schiebener, J. Morimoto, T. Asfour, and A. Ude, "Integrating visual perception and manipulation for autonomous learning of object representations," *Adaptive Behavior*, vol. 21, no. 5, pp. 328–345, 2013.

[4] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[5] A. Ude, C. G. Atkeson, and G. Cheng, "Combining peripheral and foveal humanoid vision to detect, pursue, recognize and act," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Las Vegas, Nevada), pp. 2173–2178, 2003.

[6] R. Bevec and A. Ude, "Object learning through interactive manipulation and foveated vision," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, (Atlanta,Georgia), pp. 234–239, 2013.

[7] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The karlsruhe humanoid head," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, (Daejon, Korea), pp. 447–453, 2008.

[8] C. Tsikos and R. Bajcsy, "Segmentation via manipulation," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 3, pp. 306–319, 1991.

[9] L. Chang, J. R. Smith, and D. Fox, "Interactive singulation of objects from a pile," in *IEEE International Conference on Robotics and Automation*, (St. Paul, Minnesota), pp. 3875–3882, 2012.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[11] E. Stergaršek Kuzmič and A. Ude, "Object segmentation and learning through feature grouping and manipulation," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, (Nashville, TN), pp. 371–378, 2010.

[12] T. Hermans, J. M. Rehg, and A. Bobick, "Guided pushing for object singulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Vilamoura, Portugal), pp. 4783–4790, 2012.

[13] D. Katz, M. Kazemi, J. A. Bagnell, and A. Stentz, "Clearing a pile of unknown objects using interactive perception," in *IEEE International Conference on Robotics and Automation*, (Karlsruhe, Germany), pp. 154–161, 2013.

[14] D. Katz, M. Kazemi, J. A. Bagnell, and A. Stentz, "Interactive segmentation, tracking, and kinematic modeling of unknown 3D articulated objects," in *IEEE International Conference on Robotics and Automation*, (Karlsruhe, Germany), pp. 4988–4995, 2013.

[15] R. Gaschler, D. Katz, M. Grund, A. P. Frensch, and O. Brock, "Intelligent object exploration," in *Human Machine Interaction - Getting Closer* (M. Inaki, ed.), ch. 12, pp. 236–260, InTech, 2012.

[16] B. Ridge, D. Skocaj, and A. Leonardis, "Self-supervised cross-modal online learning of basic object affordances for developmental robotic systems," in *IEEE International Conference on Robotics and Automation*, (Anchorage, Alaska), pp. 5047–5054, 2010.

[17] M. T. Mason, "Mechanics and planning of manipulator pushing operations," *The International Journal of Robotics Research*, vol. 5, no. 3, pp. 53–71, 1986.

[18] Z. Balorda and T. Bajd, "Reducing positioning uncertainty of objects by robot pushing," *Robotics and Automation*, vol. 10, no. 4, pp. 535–541, 1994.

[19] M. Kopicki, S. Zurek, R. Stolkin, T. Morwald, and J. Wyatt, "Learning to predict how rigid objects behave under simple manipulation," in *IEEE International Conference on Robotics and Automation*, (Shanghai, China), pp. 5722–5729, 2011.

[20] A. Ude, D. Omrčen, and G. Cheng, "Making object learning and recognition an active process," *International Journal of Humanoid Robotics*, vol. 5, no. 2, pp. 267–286, 2008.

[21] M. Krainin, P. Henry, X. Ren, and D. Fox, "Manipulator and object tracking for in-hand 3D object modeling," *The International Journal of Robotics Research*, vol. 30, pp. 1311–1327, July 2011.

[22] D. Kraft, N. Pugeault, E. Baeski, M. Popović, D. Kragić, S. Kalkan, F. Wörgötter, and N. Krüger, "Birth of the object: Detection of objectness and extraction of object shape through object-action complexes," *International Journal of Humanoid Robotics*, vol. 5, pp. 247–265, June 2008.

[23] B. Browatzki, V. Tikhanoff, G. Metta, H. H. Bülthoff, and C. Wallraven, "Active in-hand object recognition on a humanoid robot," *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1260–1269, 2014.

[24] W. H. Li and L. Kleeman, "Segmentation and modeling of visually symmetric objects by robot actions," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1124 –1142, 2011.

[25] A. Ude, D. Schiebener, N. Sugimoto, and J. Morimoto, "Integrating surface-based hypotheses and manipulation for autonomous segmentation and learning of object representations," in *IEEE International Conference on Robotics and Automation*, (St. Paul, Minnesota), pp. 1709–1715, 2012.

[26] D. Omrčen and A. Ude, "Redundancy control of a humanoid head for foveation and three-dimensional object tracking: A virtual mechanism approach," *Advanced Robotics*, vol. 24, no. 15, pp. 2171–2197, 2010.

[27] C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, vol. 15, (Manchester, UK), pp. 147–151, 1988.

[28] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3D objects," in *IEEE International Conference on Computer Vision (ICCV)*, (Beijing, China), pp. 800–807, 2005.

[31] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV Workshop on statistical learning in computer vision*, (Prague, Czech Republic), 2004.

[32] K. E. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1582–96, 2010.

[33] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2002.

[34] R. Bischoff, J. Kurth, G. Schreiber, R. Koeppe, A. Albu-Schaeffer, A. Beyer, O. Eiberger, S. Haddadin, A. Stemmer, G. Grunwald, and G. Hirzinger, "The KUKA-DLR Lightweight Robot arm - a new reference platform for robotics research and manufacturing," in *International Symposium on Robotics and German Conference on Robotics*, (Munich, Germany), pp. 741–748, 2010.

[35] A. Ude and E. Oztop, "Active 3-D vision on a humanoid head," in *International Conference on Advanced Robotics*, (Munich, Germany), pp. 1–6, 2009.