2013 13th IEEE-RAS International Conference on
Humanoid Robots (Humanoids).
October 15 - 17, 2013. Atlanta, GA

Object Learning through Interactive Manipulation and Foveated Vision

Robert Bevec and Aleš Ude

Abstract— **Autonomous robots that operate in unstructured environments must be able to seamlessly expand their knowledge base. To identify and manipulate previously unknown objects, a robot should continuously acquire new object knowledge even when no prior information about the objects or the environment is available. In this paper we propose to improve visual object learning and recognition by exploiting the advantages of foveated vision. The proposed approach first creates object hypotheses in peripheral stereo cameras. Next the robot directs its view towards one of the hypotheses to acquire images of the hypothetical object by foveal cameras. This enables a more thorough investigation of a smaller area of the scene, which is seen in higher resolution. Additional information that is needed to verify the hypothesis comes through interactive manipulation. A teacher or the robot itself induces a change in the scene by manipulating the hypothetical object. We compare two methods for validating the hypotheses in the foveal view and experimentally show the advantage of foveated vision compared to standard active stereo vision that relies on camera systems with a fixed field of view.**

I. Introduction

To be able to successfully work in unstructured and uncontrolled environments, autonomous robots must have the ability to expand their library of known objects. Such robots must therefore be able to detect and learn new objects when no prior knowledge about them and the environment is available. Segmenting objects using only visual information has proved very difficult [1], [2]. However, perturbing the scene by for example pushing a hypothetical object introduces additional information that makes this task more feasible [3], [4], [5], [6], [7], [8], [9], [10], [11], [12].

In this paper we propose to improve object recognition in autonomous robots by learning and recognizing objects using foveated vision. In biological systems, the fovea is a part of the retina with a very high density of cone cells. It is responsible for color vision and color sensitivity. The density of cones slowly decreases toward the peripheral part of the retina. This layout provides sharp central vision and a relatively low average resolution over the entire field of view, therefore reducing the need for computational resources, but still achieving high precision vision in the fovea. Foveated stereo vision in robots can be accomplished using two cameras per eye with different focal lengths [13], [14], [15], [16]. This arrangement enables capturing wide-angle peripheral and narrow-angle foveal images at the same time, but requires gaze control in order to acquire the area of interest in the foveal view. A practical advantage of such an arrangement is that a robot can simultaneously analyze the wide field of view of peripheral cameras – where it is easier to find and track objects – and the narrower field of view of foveal cameras – where objects images have higher resolution and are therefore more suitable for recognition.

Some of the recently proposed methods rely on accurate depth sensors to segment objects from the scene [7], [9], [17], [11], [12], [18]. We chose to rely solely on stereo foveated vision to learn object representations (Fig. 1) because such systems are more generally applicable and are also closer to human vision and depth perception.

In our previous work [10], [19] the robot learns and recognizes objects using standard active stereo vision. It generates hypotheses about the existence of objects and tries pushing them to look for changes in the scene and validate the object hypotheses. Object representations were obtained by accumulating the confirmed features over several snapshots and a bag-of-features type models [20] have been acquired. Here we propose to extend this object learning process by making use of foveated vision, thus adding the confirmed object features acquired in the foveal view. In our current experiments, the manipulation of objects has been realized through human interaction. However, the robot

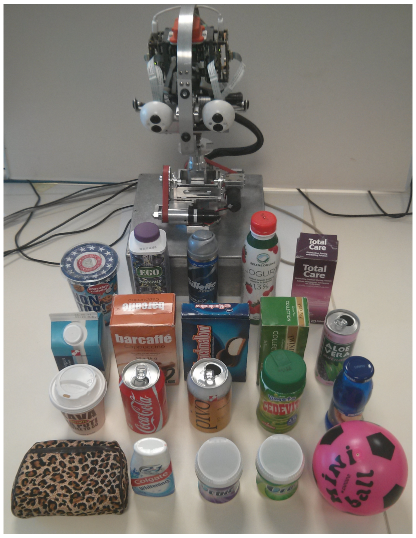

Fig. 1. Karlsruhe Humanoid Head [13] and the object test set used in the experiments. The head is equipped with two cameras in each eye. One pair of cameras models human peripheral vision, the other pair foveal vision.

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, and from the Slovenian Research Agency under grant agreement no J2-4284.

R. Bevec and A. Ude are with Jožef Stefan Institute, Department of Automatics, Biocybernetics and Robotics, Humanoid and Cognitive Robotics Lab, Jamova 39, Ljubljana, Slovenia robert.bevec@ijs.si, ales.ude@ijs.si

could also use it's own manipulation capabilities to achieve the same result. The developed approach requires no prior knowledge about the objects or the environment and retains the assumptions that the objects contain some distinctive visual features and move as rigid bodies.

II. Overview

The following procedure is applied to learn new objects and generate their representations for recognition:

- **Generate object hypotheses in peripheral view**: Find smooth surfaces in the point cloud of stereo matched visual features.
- **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.
- **Generate data for hypothesis verification**: The robot requests a human teacher to move the object and 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.
- **Feature accumulation**: The last three steps above can be repeated several times to accumulate object features from different viewpoints.

Note that in this procedure, the object manipulation step by a human teacher could be replaced by robot manipulation as in our other work [10].

III. Object hypotheses

A. Peripheral view

Object hypotheses in the peripheral view are created within a point cloud generated by the peripheral stereo vision. Initial point correspondences are found by matching Harris interest points [21] and maximally stable extremal regions (MSER) [22] in each eye. The Harris interest points are found mostly on textured parts of the image. The MSER detector balances that by finding salient points in areas with less texture. The correct correspondences and precise 3-D point positions are obtained by using epipolar geometry and stereo calibration on an active camera system [23]. At each 3-D interest point a SIFT feature descriptor [24] is calculated, which has shown robustness to scale, rotation, translation and illumination changes.

The hypotheses are created by searching for smooth surface patches within this point cloud. As in our previous work [10], [19], the robot looks for planar, spherical and cylindrical surface patches using RANSAC [25]. This iterative, nondeterministic model fitting algorithm chooses a random subset of points, fits models of the possible surface types and returns the parameters of the fitted surface that includes the largest number of points within a tolerance of that surface out of the entire point cloud. All of the points belonging to the fitted surface are then excluded from the point cloud and a search for a new hypothesis is started again. When no good fits are found anymore, the remaining points are clustered into groups of points lying close to each other using X-means algorithm [26]. Each of these clusters also represents a hypothesis if it retains enough points and has a sufficient point per volume ratio. This allows the robot to create a hypothesis for an object, even if no part of that object corresponds to a smooth surface.

In order to prevent surface hypotheses spanning over several objects, X-means clustering is applied to each hypothesis. The hypothesis is divided into several hypotheses if that is deemed appropriate by the algorithm. Erroneous splitting of hypotheses is not a problem since all features that move concurrently after interactive manipulation are later joined together as a validated object. An example of initial object hypotheses in the peripheral view is seen in Fig. 2. A detailed description about the detection of planar, spherical and cylindrical surface patches is given in [19].

B. Foveal view

Since a foveal image covers a much narrower field of view, the object of interest will cover a larger portion of the foveal than peripheral image. We can therefore assume that there is no need to search for smooth surface patches, like in the peripheral views, to create hypotheses. Instead, the entire point cloud is considered as an object hypothesis. The foveal camera pair naturally requires it's own camera calibration to find interest point correspondences and calculate the 3-D position of points. An example object hypothesis in the foveal view is seen in Fig. 2.

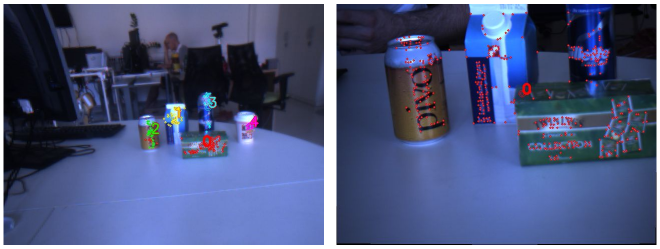

Fig. 2. Initial hypotheses in a typical scene with household objects. In the peripheral view, each hypothesis is represented by points of the same color. After the head has turned towards hypothesis ”0”, we see the initial hypothesis in the foveal view, where a hypothesis consists of all features.

IV. Gaze control

In order to acquire the object hypothesis in the foveal view, the robot must rotate the head and eyes so that the center of the hypothesis appears in the foveal cameras. Hypothesis i is first created in the peripheral view. It contains N_i points corresponding to a smooth surface patch. The 3-D centroid $\mathbf{P}_i$ of the hypothesis in the peripheral view is calculated as follows

$$\mathbf{P}_i = \frac{\sum_{n=1}^{N_i} \mathbf{x}_n}{N_i} \tag{1}$$

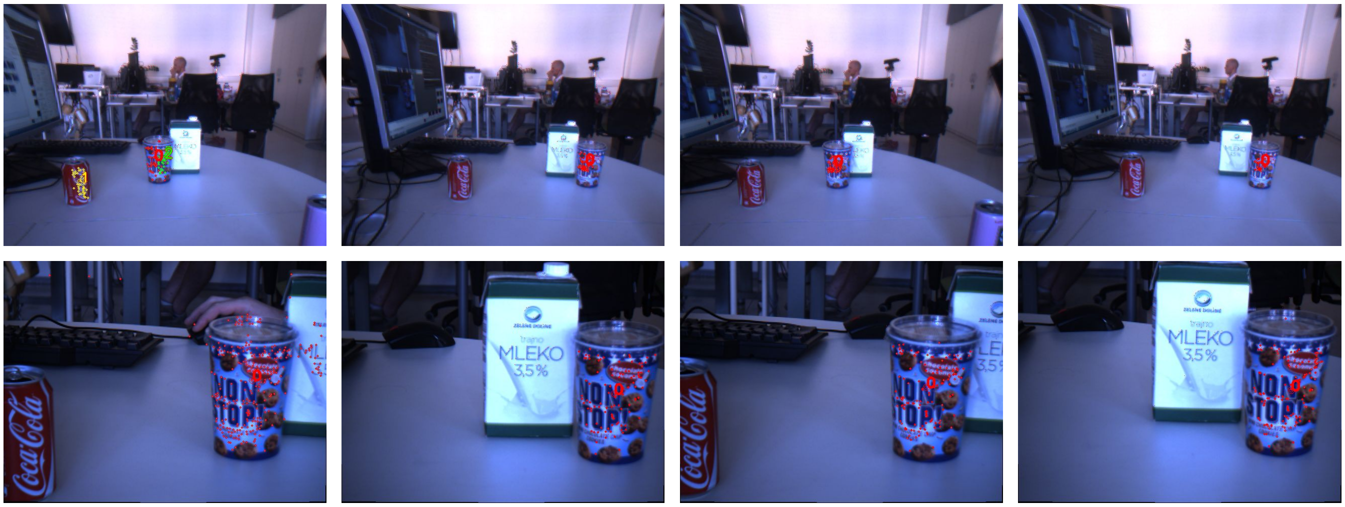

Fig. 3. A typical object learning / recognition procedure. The upper row respectively shows the peripheral and the bottom row the foveal view. The images in the first column contain the initial object hypotheses as the head is turned towards hypothesis ”0”. Each of the following columns shows the scene after moving the object, validating the initial hypothesis and accumulating the verified feature points for learning or recognition.

Using direct kinematics equations, the robot calculates the positions of all the hypotheses in the global coordinate system. It then uses a virtual mechanism approach to calculate the proper joint configuration to center the chosen hypothesis in the foveal cameras and moves the head and eyes accordingly [27].

V. Object validation

After the robot has identified the object hypotheses and turned its view towards one of them, it needs additional information to validate or discard the hypothesis. This additional information is provided by inducing motion on the object. Changes in the scene can then be analyzed for simultaneous feature motion, which is a very strong indicator of object existence in the object definition given by Gibson [28]. In our system, the robot requests a human to move the object hypothesis the robot is looking at. This is indicated by displaying the hypothetical object feature points in the acquired stereo image pair. Alternatively, the robot could also use it’s own manipulation capabilities to try and push the object hypothesis or perform some other manipulation. A typical learning or recognition procedure is seen in Fig. 3, where the initial hypotheses in the first column are validated after changes in the scene in each of the following columns.

After the manipulation action is completed, the robot recomputes the point cloud using Harris interest points and MSERs as described in Section III.

A. Peripheral view

Our basic assumption is that the object moves as a rigid body. The feature points contained in the hypothesis are matched in the peripheral images after the change using a SIFT descriptor. Due to occlusions or large rotations, some of the features may not be matched at all. If enough matches are found, we can test our assumption and find a rigid body motion that corresponds to the motion induced by manipulation. Since there will be false feature matches and matches of points that didn’t belong to the object in the first place, we use the RANSAC algorithm for finding the most probable object rotation and translation, thereby filtering out all feature correspondence matches not within the tolerance of the transformation. The parameters of the transformation can be estimated from three pairs of points before and after the change. Let $\mathbf{x}_n$ be the position of a feature point before the change and $\mathbf{x}'_n$ the position of the matched point after. If the new feature position corresponds to the transformation

$$\mathbf{x}'_n = \mathbf{R} \cdot \mathbf{x}_n + \mathbf{t}, \tag{2}$$

where $\mathbf{R}$ is the rotation matrix and $\mathbf{t}$ is the translation vector, it moved concurrently with this transformation. If more than a minimum amount of features correspond to this transformation, the hypothesis is considered as validated. All of the features that move according to the estimated rigid body transformation are considered confirmed object features. In the following we call this process a rigid body motion filter (RBMF).

All other feature matches from the point cloud, i. e. features from the point cloud that do not belong to the initial hypothesis, are also checked regarding the estimated rigid body transformation. If they move as the object features, these features are added as candidate features of the validated hypothesis. If they are matched and move together with the object also after the next manipulation, they are considered confirmed. The first row in Fig. 3 shows an example of successful object validation through several manipulations.

Even though in this paper we propose learning based on foveal views, it is still necessary to track the motion of the hypotheses in peripheral view as well. This is due to the fact that after manipulation the object usually disappears from the foveal view. The only way to get it back into the foveal view is to perform a saccade towards the new object position, which can be done by estimating the new object pose in the peripheral view.

B. Foveal view

The same process of hypothesis validation described for the peripheral views can be used in the foveal view as well. After the successful validation of the hypothesis in the peripheral view, the head and eyes are turned toward the hypothesis' centroid. The new point cloud is computed in the foveal view and matched to the one before the change using SIFT descriptors. Matches are verified with the rigid body motion filter described in Section V-A and the object hypothesis is validated or discarded. The second row in Fig. 3 shows an example of successful object validation through several manipulations.

We also suggest a simpler solution for use in the foveal views to reduce the computational complexity of the validation procedure. Since RANSAC is a statistical method, it needs many repetitions in order to provide a good solution with high certainty [25]. Being a nondeterministic algorithm, it does not guarantee the best solution even after an arbitrary number of repetitions. Instead, we propose to make use of the known movement of the head and eyes and assume that the surroundings of the hypothesis does not move when the human moves the object. We can therefore filter all static features in the peripheral views, i. e. features that moved according to the motion of the head-eye system, and confirm all other features as features belonging to the object. The assumption of static surrounding is much more justified in foveal than peripheral views because foveal views contain only a small portion of the scene.

Let $\mathbf{x}_n$ be the position of a feature point before the change and $\mathbf{x}'_n$ the position of the matched feature point after manipulation and head-eye rotation. Let $\mathbf{R}_\mathrm{V}$ be the rotation matrix and $\mathbf{t}_\mathrm{V}$ the translation vector describing the change of viewpoint from the previous gaze direction. We define threshold ϵ as a minimum displacement that implies motion. If statement (3) is true, the feature point has moved and belongs to the validated hypothesis:

$$\|\mathbf{x}'_n - \mathbf{R}_\mathrm{V} \cdot \mathbf{x}_n + \mathbf{t}_\mathrm{V}\| \geq \epsilon. \tag{3}$$

We call this method a static feature filter (SFF). Although SFF requires a partially static scene, it provides a valid alternative to the first approach. In Fig. 4 we can see how successful these two methods are at filtering feature matches and validating the hypotheses. In these examples, the initial object hypotheses in foveal views (recall that our assumption is that all feature points detected in foveal views belong to the object) contained a large number of features found in the background. The first row shows validation of the initial hypothesis with rigid body motion filter (RBMF) and the second row validation with the static feature filter (SFF). Both approaches succeeded in eliminating most of the spurious features, although SFF failed to discard a small number of false matches on the box in the background.

VI. Object learning and recognition

As in our previous work, the visual appearance of objects is learned using a bag of features (BoF) model [20]. Firstly, a visual vocabulary is created by clustering SIFT feature descriptors extracted from training images. To compute the vocabulary, we use the SIFT feature descriptors extracted while learning different objects from different viewpoints. To represent an object, each SIFT descriptor that is confirmed to belong to the object in the current view, is matched with the closest descriptor in the vocabulary. A histogram of descriptors from the vocabulary corresponding to object features is built and later used for recognition.

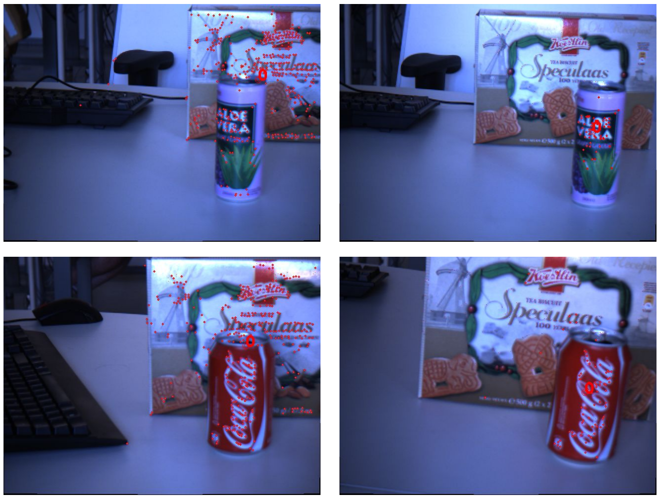

Fig. 4. In the first row, the initial hypothesis in the foveal view is validated with the rigid body motion filter and in the second row with the static feature filter. The rigid body motion filter discards all the background features and all false matches. The static feature filter requires a static background and cannot filter false SIFT matches.

To include color information, the robot also calculates a saturation-weighted hue histogram [29] within the ellipse spanned by the principal axes of the confirmed features. Both the BoF and hue histograms are calculated for individual viewpoints and for the accumulated representation after each successful validation in foveal and peripheral views. Combined, the histograms represent global color information and descriptors of salient area in all relevant views.

Object recognition is realized by calculating the corresponding histograms of the initial or validated object hypothesis as described in the previous paragraphs. A k-nearest neighbors decision is based on the distance measure between known objects in the database and the hypothesis. The distance measure is a weighted sum of normalized χ^2 histogram distances of the BoF and hue histograms as described in [10].

Some examples of object recognition for the initial hypothesis in the foveal view are shown in Fig. 5. Objects rich with features can be quite successfully recognized in this stage already, even though there are a lot of features belonging to the background included in the hypothesis (upper left image). Smaller objects with fewer features are sometimes classified as incorrect objects (unknown object in the background, upper right image) or as the object in the background itself (lower left image). When initial hypotheses do not include many features from the background, they are recognized rather successfully as shown later in the experimental evaluation (lower right image).

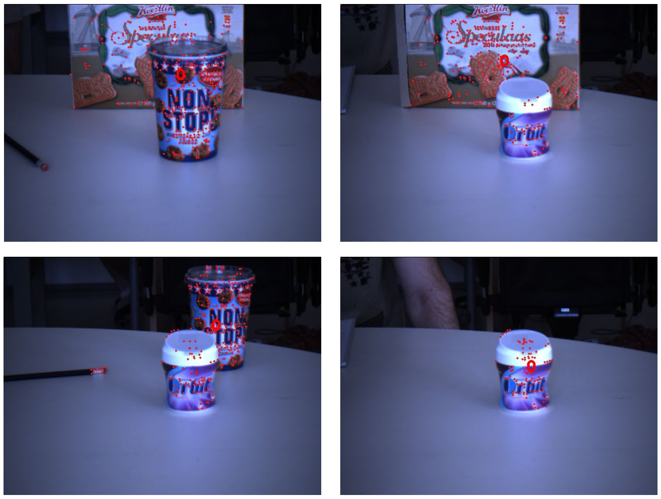

Fig. 5. Recognition of initial hypotheses in the foveal view depends greatly on the amount of features in the background. In the upper left image we have an unknown object in the background, but the object in the foreground is correctly recognized, since it is very rich with features. In the upper right image the object in the foreground has much fewer features and is therefore incorrectly recognized. For the same reason, the object in the lower left image is falsely recognized as the known object in the background. If there are few features in the background, as seen in the lower right image, the recognition of objects in the initial hypotheses can reach 93%.

VII. Experimental evaluation

We preformed several experiments to evaluate the gain of using foveated vision for object learning and recognition. We also compared the proposed filters (RBMF and SFF) for validating hypotheses in the foveal images.

The robot first learned representations of 20 different typical household objects (Fig. 1), where a human teacher manipulated the objects. The objects were placed on a table in sets of 5 at a distance approximately 1 meter from the robot, as shown in Fig. 2. Using KUKA LWR manipulators, this distance would be well within reach of the robot if pushing was done autonomously. There were some occlusions present at times, but eventually each object was shown in full extent. The system created initial hypotheses about the objects and then learned a model for each. Each model was learned through 6 consecutive manipulations. The human teacher (the first author) moved the objects mainly laterally to the image plane with small rotations. This ensured good feature matching, but the objects were learned mainly from one viewpoint. For the foveal view different representations were learned, once using RBMF and once SFF.

For recognition, the sets of objects were randomly mixed and placed back on the table. The robot tried recognizing the initial hypotheses and then requested the human to move the object it was facing. It followed the object through 3 consecutive manipulations in the scene and tried recognizing it after each one. Table I shows the results of object recognition.

TABLE I

Object recognition rate in the peripheral view, foveal view with rigid body motion filter (RBMF) and foveal view with static feature filter (SFF) for the initial hypotheses and the following pushes.

	init. hyp.	1 push	2 pushes	3 pushes
Peripheral	51 %	52 %	60 %	62 %
Foveal w. RBMF	60 %	95 %	95 %	95 %
Foveal w. SFF	60 %	100 %	100 %	100 %

The results of recognition in peripheral views are significantly worse than in our previous work [10], where we used the same camera pair as a standard active stereo vision system. This is due to the increased distance of the objects from the robot. In our experiments the objects were placed approximately 50 cm further away form the cameras. The recognition rate improves with each push, until it starts to converge toward approximately 65 % and doesn't improve even after more pushes. At such a distance few object features were found and even fewer matched. On average, only 32 features were accumulated in the recognition stage, describing each object after 3 pushes. A larger number of features allows for more robust recognition under partial occlusion in cluttered scenes [24].

Recognition rate using foveated vision varied a lot in the initial hypothesis. Depending on the amount of clutter in the view, features of the hypothesis might belong to the background. Singulated objects were correctly recognized 93 % of the time, while dense clutter reduced this rate down to 27 %. On average, the initial hypothesis recognition rate in our experiments was 60 %. Using the proposed approach, background features were filtered our after the first push and the recognition rate improved significantly.

Both of the proposed filters used the same initial hypotheses. Using RBMF, the validated hypothesis after the first push was successfully recognized 95 % of the time. This rate stayed stable with consequent pushes. It turned out that there was just one particular object, which was constantly recognized as another object from the database. These recognition rates are significantly higher compared to the rates in the peripheral view. On average, 181 features described each object after 3 pushes, which is significantly more than in the peripheral view.

Using SFF, the validated hypothesis after the first push was successfully recognized 100% of the time. This rate remained stable throughout the interactive recognition process. SFF is not able to discard false descriptor matches and therefore builds a richer representation than RBMF including false positives. On average, 227 points described each object after 3 pushes, but as it turns out, a richer representation including some false descriptor matches does not hurt the recognition rates, since the proportion of false features is low. Some of them were actually on the object itself and therefore benefited the representation. The results prove that both methods are valid options for hypothesis validation in foveal views. Overall, our results confirm the usefulness of foveal vision for object learning and recognition.

VIII. Conclusions

We developed a novel system for object learning and recognition by manipulation, which can exploit the advantages of foveal vision. Initial object hypotheses are generated

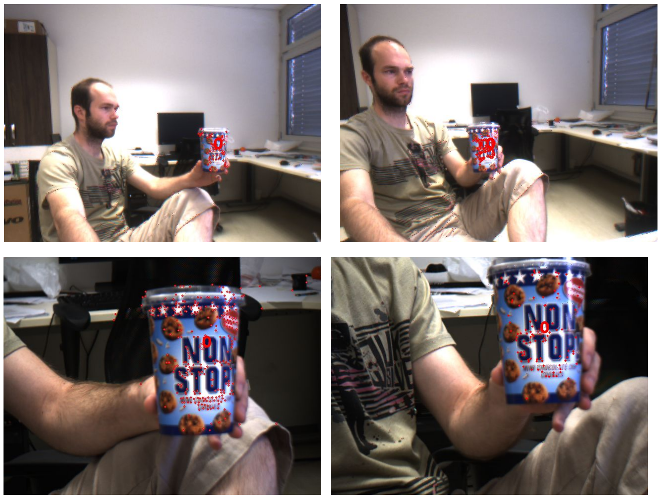

Fig. 6. Our method does not require a static tabletop scene. The system is able to learn new objects or recognize known objects in an arbitrary environment. In the pictures above, we can see the object learning through human interaction.

in the peripheral view and more accurately investigated in the foveal view by turning the head and eyes toward the hypothesis. Hypotheses are validated, corrected and extended after interactive manipulation by a teacher or the robot itself. We compared different methods for validating the hypotheses in the foveal view and showed the advantages of foveal vision compared to to the standard active stereo vision with a fixed field of view for object learning and recognition.

A representation of accumulated features that is built through manipulation shows a particular advantage when an object is learned from several viewpoints. As it is evident in Fig. 6, our methods works in an arbitrary environment in cooperation with a human teacher and relies on only two assumptions: that the object moves as a rigid body and that is has distinctive visual features.

References

[1] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Philosophical Transactions of the Royal Society A*, vol. 361, pp. 2165–2185, Oct. 2003.

[2] G. Kootstra, J. Ypma, and B. de Boer, "Active exploration and keypoint clustering for object recognition," in *IEEE International Conference on Robotics and Automation*, (Pasadena, CA), pp. 1005–1010, 2008.

[3] P. Fitzpatrick, "First contact: an active vision approach to segmentation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Las Vegas, Nevada), pp. 2161–2166, 2003.

[4] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," in *IEEE International Conference on Robotics and Automation*, (Kobe, Japan), pp. 1377–1382, 2009.

[5] E. Stergaršek Kuzmič and A. Ude, "Object segmentation and learning through feature grouping and manipulation," in *2010 10th IEEE-RAS International Conference on Humanoid Robots*, (Nashville, Tennessee), pp. 371–378, 2010.

[6] 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.

[7] 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.

[8] M. Rudinac, G. Kootstra, D. Kragic, and P. P. Jonker, "Learning and recognition of objects inspired by early cognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Vilamoura, Algarve, Portugal), pp. 4177–4184, 2012.

[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] 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.

[11] 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.

[12] 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.

[13] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The Karlsruhe Humanoid Head," in *2008 8th IEEE-RAS International Conference on Humanoid Robots*, (Daejeon, Korea), pp. 447–453, 2008.

[14] C. G. Atkeson, J. G. Hale, F. Pollick, M. Riley, S. Kotosaka, S. Schaul, T. Shibata, G. Tevatia, A. Ude, S. Vijayakumar, E. Kawato, and M. Kawato, "Using humanoid robots to study human behavior," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 4, pp. 46–56, 2000.

[15] H. Kozima and H. Yano, "A robot that learns to communicate with human caregivers," in *Proceedings of the First International Workshop on Epigenetic Robotics*, (Lund, Sweden), pp. 47–52, 2001.

[16] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal, "Biomimetic oculomotor control," *Adaptive Behavior*, vol. 9, pp. 189–207, 2001.

[17] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. D. Stefano, and M. Vincze, "Multimodal Cue Integration through Hypotheses Verification for RGB-D Object Recognition and 6DOF Pose Estimation," in *IEEE International Conference on Robotics and Automation*, (Karlsruhe, Germany), pp. 2096–2103, 2013.

[18] I. Lysenkov and V. Rabaud, "Pose Estimation of Rigid Transparent Objects in Transparent Clutter," in *IEEE International Conference on Robotics and Automation*, (Karlsruhe, Germany), pp. 162–169, 2013.

[19] D. Schiebener, J. Morimoto, T. Asfour, and A. Ude, "Integrating visual perception and manipulation for autonomous learning of object representations," *Adaptive Behvior*, vol. 21, no. 5, 2013.

[20] 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.

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

[22] 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.

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

[24] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] 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.

[26] D. Pelleg, A. Moore, and Others, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proceedings of the Seventeenth International Conference on Machine Learning*, (Stanford, California), pp. 727–734, 2000.

[27] 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.

[28] J. J. Gibson, "The Ecological Approach to the Visual Perception of Pictures," *Leonardo*, vol. 11, no. 3, pp. 227–235, 1978.

[29] K. E. A. 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. 1582–1596, 2010.