Information Driven Sensor Placement for Robust Active Object Recognition based on Multiple Views

Dennis Stampfer, Matthias Lutz and Christian Schlegel
University of Applied Sciences Ulm
Department of Computer Science, Prittwitzstr. 10, 89075 Ulm, Germany
{stampfer, lutz, schlegel}@hs-ulm.de

***Abstract*—Robust object recognition is a mandatory prerequisite for many applications in Service Robotics. A common approach is to capture a single image of the scene from a fixed position and to recognize all objects at once. This challenging task is even more demanding in everyday environments.**

We propose an approach for object recognition which makes use of mobile manipulation. An initial object belief is enhanced by systematically inspecting objects from different views with a second camera on a manipulator. Knowledge about an object is used to generate possible viewpoints that are directed at specific features. Viewpoint selection considers the expected recognition probability and costs to optimize the recognition performance.

The approach is demonstrated in real-world experiments with a service robot. Almost identically appearing objects are reliably classified by systematic inspection of additional distinguishing features like barcodes and text labels.

I. Introduction

Many applications in Service Robotics use manipulators to interact with the environment to grasp and place objects. Accurate pose estimation and reliable object recognition is a mandatory prerequisite for such real world applications. A very common approach in object recognition for mobile manipulation applications is to recognize all objects in the scene at once from a single scene image, captured from a fixed position (as e.g. in [1]). However, bad perspectives, occlusions, insufficient image data and single algorithms that can only partially interpret the information result in insufficient recognition. This results in insufficient recognition rates in real-world scenarios. For example, a 3D model matcher cannot distinguish between two flavours of a product as in fig. 1 (left). Nevertheless, it is not always necessary to identify objects in one shot as one can take another look and easily identify an object.

In active vision for object recognition, alternative views (zoom cameras or other perspectives) are used to gather more data of an object or scene for more reliable recognition. The challenge is to find the view that allows the best object recognition.

We propose an approach that combines mobile manipulation and object recognition. Object recognition itself is done by probabilistically fusing the results of several algorithms from several views which makes it robust for everyday environments. As a manipulator is present in manipulation tasks, we use a second camera mounted on the manipulator (eye-in-hand camera) and move it around objects to collect information from alternative views (fig. 1).

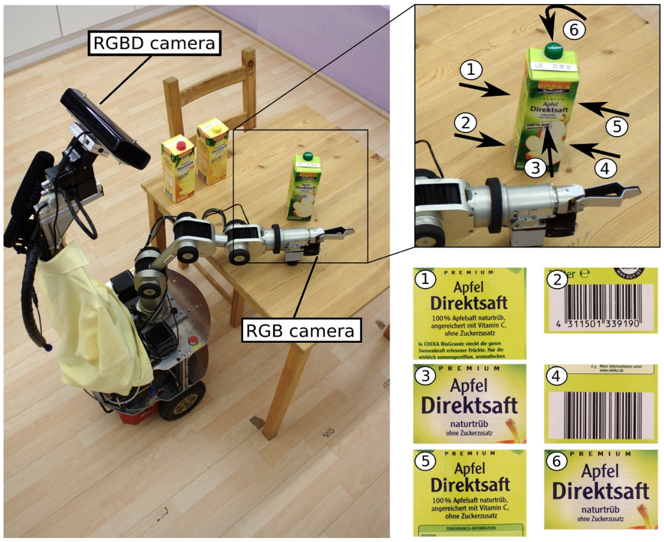
RGBD camera
RGB camera
1
2
3
4
5
6
PREMIUM
Apfel
Direktsaft
100% Apfelsaft naturtrüb, angereichert mit Vitamin C, ohne Zuckerzusatz
In EDEKA RioGrande steckt die ganze Sonnenkraft erlesener Früchte. Nur die wirklich sonnengereiften, aromatischen
4 311501 339190
naturtrüb
ohne Zuckerzusatz
Mehr Informationen unter www.edeka.de
ERNÄHRUNGS-INFORMATION

Fig. 1. A service robot in an everyday environment inspecting one of three almost identical objects on a table. A second camera mounted on the manipulator is used to acquire images from several perspectives. Six possible viewpoints and identifying features such as barcode and text are illustrated.

An initial object recognition of the whole scene using a Microsoft Kinect camera brings up first hypotheses of the objects (as described in [1]). The hypotheses are then systematically enriched with information from more descriptive features by a closer look at the object (systematic object inspection) if the recognition probability is not yet high enough. Object recognition from several views achieves a significant advantage over recognition on the full scene from a single perspective. It significantly increases the reliability of the object recognition in a wide variety of applications as it combines the strengths of different algorithms and views. Sensor placement and object recognition is information driven: the robot uses knowledge about the objects and its features to evaluate views and to derive the next best view position to maximize the recognition result by evaluating the benefit of each viewpoint. While active vision often considers geometrical properties such as distance or angular offsets in finding a viewpoint, we also consider feature properties such as the expected recognition quality of different features. Viewpoint selection takes the action costs and the estimated object recognition probability into account.

We demonstrate the approach (video at [2]) with household

978-1-4673-0856-4/12/$31.00 ©2012 IEEE

goods. The initial object recognition of a scene is based on a simple 3D model matcher for the object shape, a color detector and SURF feature detection [3] for textured objects. Additional features used for systematic object inspection are text labels and barcode labels. The latter are highly descriptive, present on most household goods and reliable algorithms are available [4]–[6]. The experiments show how the recognition is significantly improved by information driven object inspection.

II. Related Work

The selection of viewpoints in object recognition has been intensively researched. For example, [7] find viewpoints such that they maximize the visible information and define the use of a viewpoint by the number of visible surfaces. Their approach is based on surfaces only, not on features.

Similar, the approach in [8] selects a viewpoint such that it covers the most surfaces. A fixed camera is directed at an object rotating on a turntable. Viewpoints are evaluated by their angular offset to the surface normal. There is no information about the benefit of the surfaces itself. In [9], a camera moves around an object. The system learns an optimal sequence of views that needs to be performed to recognize an object.

In [10], a manipulator and viewpoint planning is used for autonomous model generation. They answer the question of how to move a grasped object in front of a depth camera to generate a model. The problem addressed in our work is the other way around. We know the object model and find the next best view for recognition by taking the object properties and environment constraints (e.g. occlusions) into account.

An approach using a fixed foveal camera that can be pointed towards objects is presented in [11]. The authors run object recognition on a scene image and look at prospective objects using a foveal camera. Object recognition results are integrated for robust object search. However, a foveal camera brings data of higher quality (close-up, high-res image) but no new additional data since the perspective stays the same.

The work of [12] is the one most similar to the presented approach. A robot recognizes objects placed on a table by driving around it on a circular path. Object hypotheses from different viewpoints are integrated. The best robot position is chosen by estimating a recognition reward by the number of expected features and object occlusions. Costs are defined as the change in angle and distance of the robot movement. The costs consider only the robot movement and are not related to the objects themselves. They do not consider different kinds of features or algorithms for viewpoint selection. Turning the camera on a circle brings new perspectives but is still limited as object occlusions cannot be handled completely. Driving around the scene is only possible if the area is drivable, which is a limitation (e.g. object in shelf). Recognition is always run on the whole scene – we only focus on one individual object in question.

In [13], objects are grasped and a barcode is tried to read by turning the hand without knowledge where to look for the barcode. Only a single algorithm is used for object recognition.

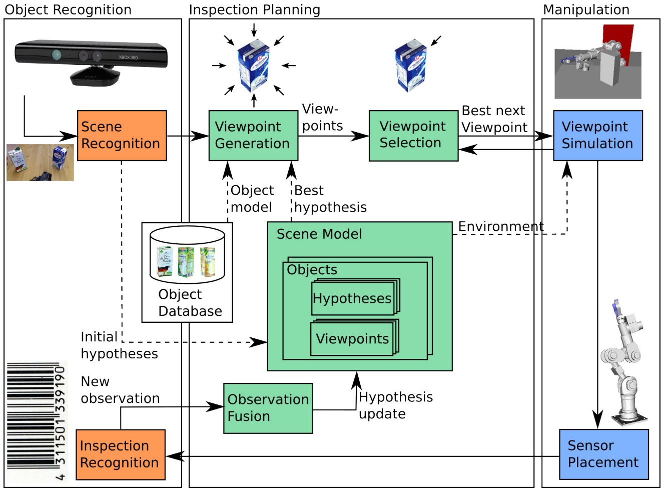

Fig. 2. Overall structure of the approach which consists of three parts: object recognition, inspection planning and manipulation. First, an initial recognition is performed on the full scene. Based on the object type of the best hypothesis, viewpoints are generated, selected and simulated to confirm that the manipulator is able to reach the desired position for the eye-in-hand camera mounted on the manipulator. Object recognition performed on the eye-in-hand camera images generates new observations which update the object in the scene model.

However, object manipulation is not always necessary or possible for object recognition. It is expensive and has to deal with further difficulties such as object occlusion by the gripper and (re-)grasping.

III. Informed Active Object Recognition

A. Structure / Architecture

The overall structure of the object recognition process is illustrated in figure 2. An initial object recognition (as described in [1]) is run on the full scene color and depth image to get poses and an initial hypotheses of prospective objects in the scene including their 3D model for later manipulation planning. The scene model holds all obstacles and objects in the scene. In the following steps, each object is systematically inspected one after another if the initial recognition probability is too low.

Based on the object type of the best hypothesis, viewpoints are generated for each feature on the object (fig. 3). The benefit of each viewpoint is calculated based on its cost and utility. Among all viewpoints of an object hypothesis, the best viewpoint is selected as the next pose for the camera. The strategy is to confirm the best hypothesis. The manipulator movement in order to place the camera at the desired viewpoint is calculated and simulated using OpenRave [14]. OpenRave uses the scene model which includes the current object hypotheses and obstacles. This enables safe and collision free movement to the desired viewpoint and allows to reject viewpoints where the manipulator would collide with objects. In this case, the next best viewpoint will be tried. The path to the viewpoint is planned in OpenRave and executed on the real manipulator. The object recognition is run again with the eye-in-hand camera image as input. The result is a new hypothesis from an independent observation which is fused with existing ones and updates the scene model. The inspection of one object continues until the best object hypothesis reaches a threshold, and is thus sufficiently identified, or until no further unvisited viewpoints exist.

B. *Object Recognition*

The object recognition used in this approach works on RGB and depth images and is described in [1]. It can be configured to recognize only a set of expected objects depending on the current situation (e.g. few but fast algorithms when time is an issue). Several algorithms for object recognition can be used: simple 3D model matcher for the shape of the object, color detector, SURF [3] feature detection for textured objects, OCR software and barcode scanning. Each algorithm works on different features.

There are two modes of recognition: *scene recognition* and *inspection*. In *scene recognition*, the object recognition is run on the whole scene (e.g. all objects on a table) using Kinect depth and RGB image. In *inspection* mode, the depth image is not processed and the RGB image of the eye-in-hand camera is used.

In all steps and modes, the object recognition crops the RGB input images to only contain one object. This relieves the algorithms from finding objects in the scene image and simplifies the object recognition from recognition in a cluttered scene to recognition of a single object in full view.

The results of the algorithms heavily depend on the type of object. Not all algorithms perform equally well on all object types. Thus, the object recognition uses a quality for each algorithm $Q(x)$, which tells the probability that an algorithm will identify an object x.

The final object recognition output is a list of objects in the scene indexed by c. Each of the objects comes with a list of classification probabilities $P(x_c)$ to each object type x in the database. For example, a milk box in a scene might get index $c = 1$ and the object recognition reports probabilities for matching each of the object types in the database: $P(\text{MILK}_1) = 0.8$, $P(\text{CUP}_1) = 0.1$, etc.

C. *Viewpoint Generation*

Systematic object inspection requires the determination of possible viewpoints before the best one can be chosen. Therefore, possible viewpoints around an object are generated. A viewpoint is the 6D pose at which the camera is placed to look at the object. Since the object is identified by features, the viewpoint generation is based on the location of these features on the object (fig. 3).

In theory, for each feature one perfect viewpoint exists where the distance to the feature is the focal distance and the optical axis matches the surface normal of the feature. In the real world however, there are several reasons why this viewpoint is not feasible, because the camera may not be positioned at the viewpoint: the configuration may be within an obstacle, no path exists to reach the configuration or due to limitations of the manipulator. We therefore generate several viewpoints for each feature (fig. 3, left). Features are point-shaped with a 6D pose relative to the object reference point and stored in an offline generated database for each object.

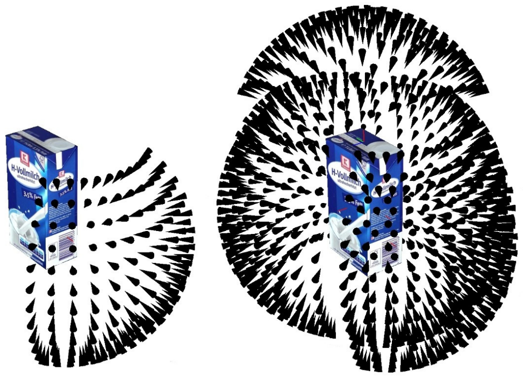

Fig. 3. Generated viewpoints for the barcode feature (left) and all viewpoints of a milk box including viewpoints of all text and barcode features (right). Viewpoints for text features on the top are clearly visible against the otherwise spherical appearing viewpoints to keep the focal distance to the surface.

The first hypothesis of an object comes from the full scene recognition. In every subsequent step, an object is assumed to be of the type of the likeliest hypothesis. Viewpoints are generated by sampling a half sphere around the surface normal vector of the feature. They are pointed at the feature so that it is in the center of the optical axis of the viewpoint/camera.

In the experiments, we set a deviation of ϕ degrees to the feature normal and a sampling of ψ degrees for viewpoint generation. A low sampling rate ψ changes the view only slightly and a higher deviation ϕ leads to a high perspective distortion and unrecognizable features.

The viewpoints are regenerated as soon as the best hypothesis changes to another object type, since the features that the viewpoints are based on have changed.

D. *Viewpoint Evaluation and Selection*

The viewpoints generated to look at features on the object are evaluated with the strategy to confirm an object hypothesis, not to disprove it. The hypotheses are enhanced by looking at the feature that promises the best recognition result. Every viewpoint is evaluated concerning its benefit and its cost resulting in an utility value. The viewpoint with the highest utility is considered the “best viewpoint”, promising the most probability gain.

The basis for the utility is the quality of a feature $Q(x)$ which represents the probability how well the selected feature is able to classify the object in general. The value $Q(x)$ from the object recognition defined as the quality of an algorithm can be taken as the quality of the feature, since the algorithm works on it. $Q(x)$ is determined empirically for each object type. The general classification probability $Q(x)$ depends on the object only, but not on any constraints decreasing the recognition probability, such as image perspective distortion. Therefore the angle to the normal vector of the feature is used as demotion to the utility. Another measure could be

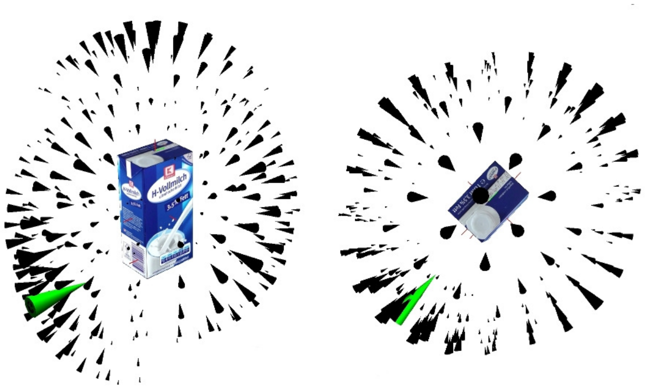

Fig. 4. Isometric- and top-view of evaluated viewpoints for a milk box. Larger cones represent viewpoints with a higher utility. In this example, viewpoints with the highest utility can be found around the surface normals of the object. The most and best viewpoints are in the lower left where the barcode is located. The best viewpoint (green, lower left) is directed at the barcode.

the distance to the focal point. In everyday object recognition, one wants to detect objects robust and fast. This implies to look at features near the current sensor position first. Therefore short travel distances of the tool center point are preferred over longer ones. Further measures could be related to the cost or execution time of an algorithm.

All values are summed up (equally weighted, costs negative) to a single utility value rating the viewpoint. The utilities of all viewpoints are recalculated for each object inspection. Fig. 4 visualizes utility values of viewpoints.

E. Sensor Manipulation

OpenRave [14] is used for manipulation planning to position the camera. As the inspection is in progress, OpenRave works with sensor data in case of unidentified objects (treated as obstacles) and with the exact 3D model as soon as the object was identified (cf. fig. 10). This adds to the robustness as the depth image is only 2.5D and OpenRave can then work with exact and complete 3D models. If OpenRave is not able to simulate the manipulation (e.g. configuration collides with obstacle), the viewpoint is discarded and the next best is tried.

As soon as the inspection process for a certain feature is finished, the remaining viewpoints of the feature are discarded. It can be assumed that other views of this feature will not result in a better recognition since it already was the best viewpoint. If the object was not recognized satisfactorily (below a threshold δ), a different feature of the object is inspected. Other strategies and decisions which viewpoints can be discarded are part of further work. The threshold δ depends on the requirements of the application (e.g. high for critical applications / pharmaceuticals).

F. Observation Fusion

The results of the object recognition $P(x_{c_t})$ of object c in inspection t are fused probabilistically. Assuming independent measurements, the observation results are combined using $P(X_{c_{t=1}} \cup X_{c_{t=2}} \cup ...) = P(\bigcup_{t=1}^{n} X_{c_t})$ where x_{c_t} is the result of one object recognition run for one object type. The fusion of observations increases the probability monotonically. For example, when looking at the barcode but the suspected barcode is not there, the probability is not decreased. Instead, the probability of another object type may be increased if its barcode is recognized. This does not limit the recognition performance as the probability of the correct hypothesis will outrun the wrong one.

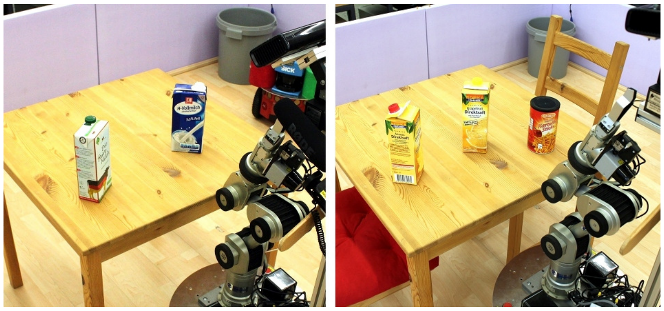

Fig. 5. Scene images of experiment 1 (left) with two milk tetra packs and experiment 2 (right) with two almost identical but different flavoured juice tetra packs and a chips can.

IV. Experiments and Results

The experiments are run on the service robot “Kate”. It is placed in front of a table in a dining-room environment (fig. 5). The purpose is to recognize five objects in two experiments with the focus on few but similar objects.

A. Experimental Setup

Kate is based on a Pioneer P3DX platform and equipped with a Microsoft Kinect RGBD camera mounted on a pan-tilt-unit. A Neuronics Katana manipulator is used for object manipulation and inspection using a small high-resolution (2560x1920 pixel) RGB iDS imaging uEye camera. The camera is mounted near the tool center point (fig. 1).

The used Katana manipulator has only 5 DOF, so possible camera poses are limited. The viewpoints are thus filtered to only contain reasonable viewpoints. This is not a general limitation of the approach but of the manipulator used in the experiments.

Viewpoints are generated about $\phi = 120$ degree around a feature with a sampling of $\psi = 15$ degree. With these parameters and the manipulator limitations, there are about 20 to 30 viewpoints per feature (depending on the feature pose in relation to the manipulator and viewpoint sampling).

The initial object recognition uses algorithms for color, texture features and a simple custom model matcher. Inspection recognition uses two different OCR algorithms [4], [5] and one barcode algorithm [6]. The $Q(x)$ for OCR is 0.5, for barcode 0.6 as given by the object recognition. The classification threshold for further inspection is $\delta = 0.6$.

B. Experiment 1

The experiment (fig. 5, left) uses two milk tetra packs which cannot be recognized by their shape. The initial hypothesis

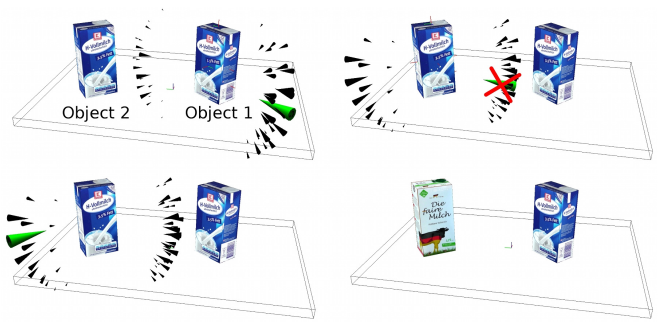

Fig. 6. Viewpoints and object hypothesis of experiment 1. From upper left to lower right: successful recognition through barcode; discarded viewpoint due to table collision detected in simulation; reading text; both objects recognized.

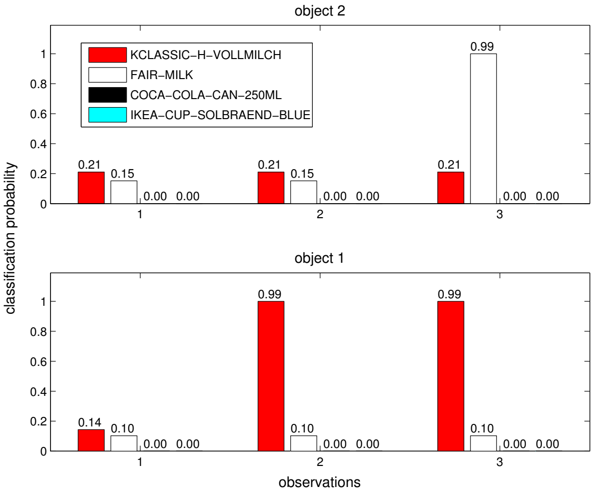

Fig. 7. Rounded classification probabilities of both objects during experiment 1. The first observation is the initial scene recognition, the others are systematic object inspections. The legend shows all object types that the object recognition is configured to recognize.

(scene recognition / first observation) of object 1 (fig. 6, upper left) is correct but recognition probability is very low (fig. 7). It is recognized at once in the first inspection (observation 2) to a very high probability through the barcode.

The initial hypothesis of object 2 is wrong. The first viewpoint (crossed out in fig. 6, upper right) is directed at the barcode as expected by the (wrong) hypothesis. This viewpoint is discarded due to collision with the table detected in simulation.

Even though there are other views at the barcode, there are viewpoints to text features in a better perspective. Thus the planner decides to look at the text on the left of object 2 as the angle to the surface is smaller (fig. 6, lower left).

As the real object has enough text on this perspective, the probability of FAIR-MILK for object 2 outruns the initial best hypothesis (K-CLASSIC) and successfully recognizes it by reading the text (fig. 6, lower right).

C. Experiment 2

The experiment (fig. 5, right) recognizes two of three possible juice tetra packs of the same shape and one chips can. The apple juice can be separated from a pineapple and grapefruit juice by its color. The latter two are impossible to distinct by shape, color and even textural features – even by humans without reading the label (cf. fig. 8, 10, 11).

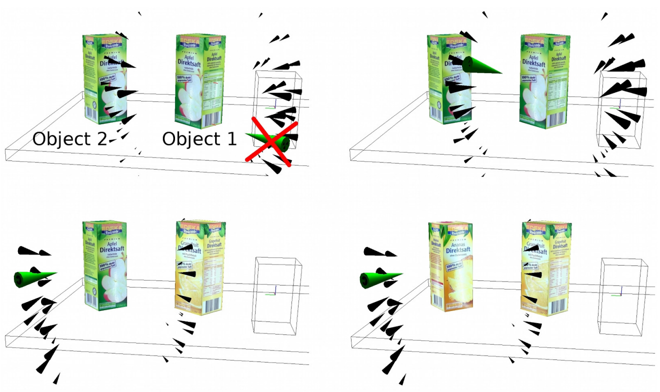

Fig. 8. Viewpoints and object hypothesis of experiment 2. From upper left to lower right: impossible viewpoint(s) due to collision with chips can detected in simulation; view to the front text of object 1; view to the side of object 2 which reflected light; changed hypothesis and new slightly different view without reflection. The box on the right represents the chips can.

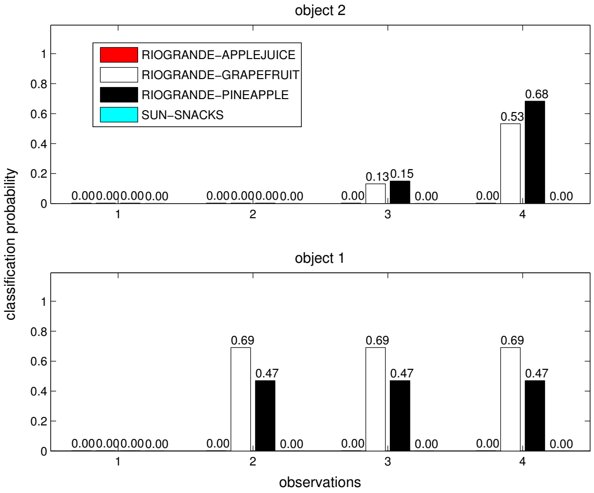

Fig. 9. Rounded classification probabilities of both objects during experiment 2. The first observation is the initial scene recognition, the others are systematic object inspections. The legend shows all object types that the object recognition is configured to recognize.

The chips can (SUN-SNACKS) is identified at once in the initial object recognition and thus not further inspected. The hypothesis for pineapple and grapefruit juice is wrong (apple juice) but the probability is still very low (fig. 9). The best viewpoints for object 1 are directed at the barcode (fig. 8, crossed out upper left) but the camera cannot be placed at that viewpoints due to collision with the chips can, which is detected in simulation. The object is successfully recognized reading the text at the front side of a successive viewpoint (fig. 8, upper right) and the hypothesis changed to the correct one (grapefruit) with a high enough probability.

Object 2 is inspected from the left (fig. 8, lower left / fig. 10). However, a bright light reflection (fig. 11) prevents the

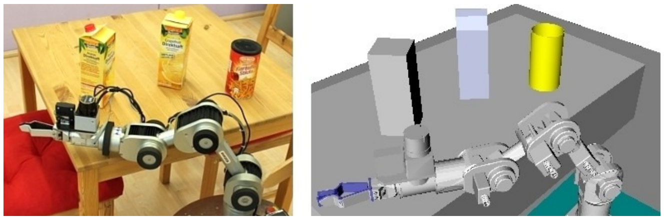

Fig. 10. Inspection of object 2 in experiment 2. Left: real scene. Right: simulation in OpenRave. Chips box and object 1 is represented by the real 3D models. Object 2 is represented as bounding box because the recognition probability is still too low. It will be exchanged with the correct 3D model after successful recognition.

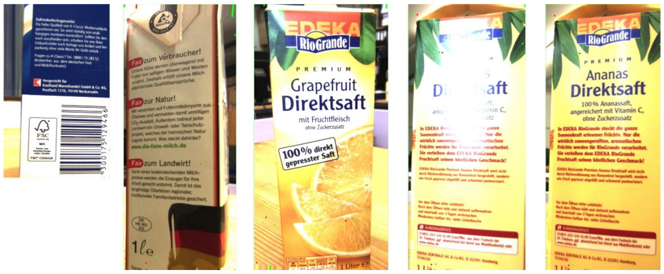

Fig. 11. Images acquired from the eye-in-hand camera. These cropped images are the input for the object recognition. The two left images are from experiment 1, the two right from experiment 2.

satisfying recognition of the object. The recognition probability after this inspection is still low but enough to change the best hypothesis. Since the type of the object has changed, new viewpoints are generated and the perspective is slightly changed (fig. 8, lower right). The reflection disappears and the recognition probability increases above the required threshold.

The fourth observation of object 2 also raises the probability of the wrong hypothesis. This is due to the almost identical words printed on the objects which differ only in the flavour. Even though, the likeliest (correct) hypothesis outruns the second highest.

D. Results

The experiments show that the initial object recognition probability, which is not always sufficient, is significantly improved by new information through systematic inspection. All objects were recognized in the first successful inspection which shows that the viewpoint selection indeed selects viewpoints that improve the recognition probability. The systematic inspection of different viewpoints is even able to handle illumination and light reflections that must be expected in everyday environments of service robots. Even individual inspection thresholds are possible: e.g. high probability for pharmaceuticals and low for juices.

The robust recognition was only achieved by inspecting objects from different viewpoints and probabilistic fusion of the results from different recognition algorithms. None of the used recognition algorithms alone was able to identify the objects reliably enough. It is possible to recognize almost identical objects. The used 5-DOF manipulator limits the results, but not the power of the approach.

V. Conclusion and Further Work

This paper proposed a method for information driven sensor placement that combines object recognition and positioning of a camera on a manipulator. Additional views on an object gained by systematic inspection significantly improve the recognition performance which makes object recognition suitable for everyday environments. It is possible to recognize almost identical objects to a high probability.

The systematic inspection using text and barcode is suitable for everyday environments as text or barcode features are printed on almost all goods. This allows object recognition of a wide variety of objects. In further work, we will evaluate more strategies for inspection as well as more recognition algorithms. A video is available at [2].

Acknowledgment

This work has been conducted within the *ZAFH Servicerobotik* (http://www.zafh-servicerobotik.de/). The authors gratefully acknowledge the research grants of the state of Baden-Württemberg and the European Union. We thank Siegfried Hochdorfer for his support in probabilistic fusion and good discussions.

References

[1] M. Lutz, D. Stampfer, S. Hochdorfer, and C. Schlegel, “Probabilistic Fusion of Multiple Algorithms for Object Recognition at Information Level,” in *IEEE Int. Conf. on Technologies for Practical Robot Applications*, Woburn, MA, USA, 2012.

[2] YouTube: Robotics@HS-Ulm, http://www.youtube.com/roboticsathsulm.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in *9th European Conf. on Computer Vision*, Graz, Austria, 2006.

[4] R. Smith, “An Overview of the Tesseract OCR Engine,” in *Int. Conf. on Document Analysis and Recognition*, vol. 2, Sept. 2007, pp. 629–633.

[5] ABBYY OCR, http://ocr4linux.com, visited: Nov. 11th 2011.

[6] ZBar bar code reader, http://zbar.sf.net, visited: Nov. 11th 2011.

[7] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich, “Viewpoint Selection using Viewpoint Entropy,” in *Proc. of the Vision Modeling and Visualization Conference 2001*, 2001, pp. 273–280.

[8] D. Roberts and A. Marshall, “Viewpoint Selection for Complete Surface Coverage of Three Dimensional Objects,” in *Proc. of the British Machine Vision Conference*, 1998, pp. 740–750.

[9] F. Deinzer, C. Derichs, J. Denzler, and H. Niemann, “Integrated Viewpoint Fusion and Viewpoint Selection for Optimal Object Recognition,” in *Proc. of the British Machine Vision Conference*, 2006, pp. 287–296.

[10] M. Krainin, B. Curless, and D. Fox, “Autonomous generation of complete 3D object models using next best view manipulation planning,” in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2011, pp. 5031–5037.

[11] K. Welke, T. Asfour, and R. Dillmann, “Active Multi-View Object Search on a Humanoid Head,” in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2009, pp. 417–423.

[12] R. Eidenberger and J. Scharinger, “Active Perception and Scene Modeling by Planning with Probabilistic 6D Object Poses,” in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Oct. 2010, pp. 1036–1043.

[13] E. Klingbeil, D. Rao, B. Carpenter, V. Ganapathi, A. Y. Ng, and O. Khatib, “Grasping with Application to an Autonomous Checkout Robot,” in *IEEE Int. Conf. on Robotics and Automation*, May 2011, pp. 2837–2844.

[14] R. Diankov and J. Kuffner, “OpenRAVE: A Planning Architecture for Autonomous Robotics,” Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-34, July 2008.