The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Interactive Learning of Visually Symmetric Objects

Wai Ho Li and Lindsay Kleeman
Intelligent Robotics Research Centre
Department of Electrical and Computer Systems Engineering
Monash University, Clayton, Victoria 3800, Australia
waiholi@gmail.com, Lindsay.Kleeman@eng.monash.edu.au

Abstract— **This paper describes a robotic system that learns visual models of symmetric objects autonomously. Our robot learns by physically interacting with an object using its end effector. This departs from eye-in-hand systems that move the camera while keeping the scene static. Our robot leverages a simple nudge action to obtain the motion segmentation of an object in stereo. The robot uses the segmentation results to pick up the object. The robot collects training images by rotating the grasped object in front of a camera. Robotic experiments show that this interactive object learning approach can deal with top-heavy and fragile objects. Trials confirm that the robot-learned object models allow robust object recognition.**

I. Introduction

Autonomous object learning is an inherently interesting concept as humans use it regularly to adapt to new environments. A robot with the ability to learn new objects on its own can adapt to different operating environments while shifting the burden of training data collection and model construction away from human users. By doing so, the robot may now be able to operate in environments such as the household where the large number of unique objects make exhaustive modelling and training intractable. Given the increasing ratio of workers versus retirees in developed nations [1] and positive public opinion towards domestic robots [2], the case for autonomous object learning has never been stronger.

There are many bilaterally symmetric objects in the household, including container objects such a cups and bottles. As such, the ability to autonomously learn symmetric objects is a useful addition to any domestic robot performing tasks such as cleaning or setting the table. In a previous paper [3], the authors demonstrated a robotic system that segments objects autonomously. Segmentation is performed by observing the object motion induced using a controlled pushing action called the *robotic nudge*. The robotic nudge removed the need for object models, allowing the robot to segment new objects autonomously, including near-symmetric objects such as a mug with a handle. We suggested that it maybe possible to use robot-obtained segmentations to perform further interactions and object learning.

This paper confirms these suggestions by demonstrating a robotic system that learns visual models of new symmetric objects via robotic interaction. The learning process is autonomous and model-free, which frees our robot from having to rely on training data and prior object models. Experiments on beverage bottles show that models learned by the robot allow reliable and robust object recognition.

II. Contributions

Contributions are made in the areas of interactive object learning and object recognition. The autonomous nature of the entire robotic system, from object segmentation to grasping to modelling, also contributes to current research.

A. From Simple to Advanced Interactions

Fitzpatrick suspected that it maybe possible to leverage simple object interactions to perform *advanced* interactions such as object grasping [4]. This paper confirms Fitzpatrick's suspicion experimentally. Our robotic system investigates objects by moving them a very short distance across the table using a robotic nudge. The information gained from this simple interaction is then used by the robot to pick up the object. The robot's ability to move autonomously from a nudge to a grasp is novel and useful in situations where the robot has to deal with new objects.

B. Object Learning using Robot-collected Training Images

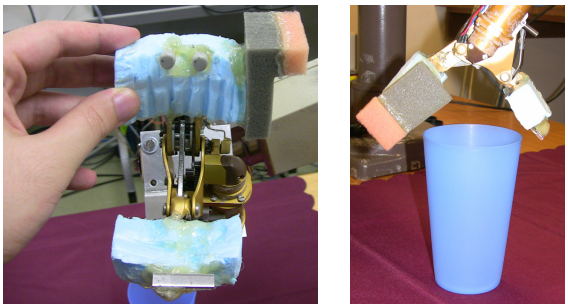
In our previous paper [3], we suggested that object segmentations obtained autonomously by our robot can be used as training data for an object recognition system. While these segmentations are accurate, nudging an object on a table only provides a single view of the moved object. The robot presented here grasps the nudged object and rotates it to collect training images over the entire 360 degrees of the grasped object. Object models are constructed using these robot-collected training images.

The proposed approach differs from the traditional approach of offline image collection and feature detection using a turntable-camera rig as surveyed in [5]. Our approach also differs from semi-autonomous systems, such as [6], that require a human user to provide the robot with different views of test objects. Instead, our robot autonomously *learns* new objects by modelling them online. Object recognition experiments suggests that the robot is able to learn useful visual models of new objects.

C. Robust Object Recognition by Pruning SIFT Descriptors

The robot's gripper has two wide foam fingers, which can be seen in the photos of Figure 1. The foam-padded gripper ensures a stable grasp but does not allow an accurate pose

estimate of the grasped object. This means that foreground-background segmentation is not available during training. As such, multi-view object recognition methods such as [7] are unsuitable because they rely on well segmented training images.



(a) Front view of gripper (b) Side view of gripper

Fig. 1. Photos of foam-padded robot gripper. The L-shaped protrusion is used to perform the robotic nudge

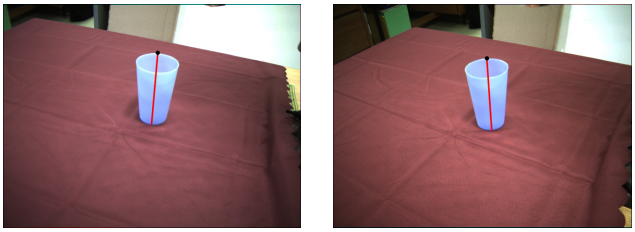
Our approach extracts SIFT descriptors [8] from the robot-collected training images to model objects. As the inclusion of background SIFT descriptors in the object model can produce false positives during recognition, we developed an automatic descriptor pruning method. The pruning method crossexamines descriptors between the images within a robot-collected training set to reject background descriptors.

III. Picking up a Nudged Object

The object learning process begins with a robotic nudge as described in [3], except that motion segmentation is now performed in stereo. This results in two object segmentations, one for each camera view. These segmentations allows the robot to crudely triangulate the top of the object. The location of the object on the table is found by looking for the intersection of its 3D axis of symmetry with the table plane. The object's axis of symmetry is found through fast symmetry triangulation as detailed in [9].

A. Determining the Height of a Nudged Object

In each camera view, the top of the nudged object is determined by following the object's symmetry line upwards. The top of the object is where its symmetry line intersects with the object-background boundary of its segmentation. Figure 2 visualizes an object's symmetry line and the top of the object as detected by our robot.



(a) Left camera image (b) Right camera image

Fig. 2. The top of a blue cup's symmetry line as detected by the robot

After locating the top of the object in each camera view, standard stereo triangulation is performed to determine the height of the object. As the top of the object is a projection of the rear of the object, there is an inherent error to the object's triangulated height. Figure 3 illustrates the pertinent geometry of the error for a single camera. The blue line joins the camera's focal point and the top of the object as detected in the camera view. The height returned by stereo triangulation is marked as a black dot. Notice that the triangulated height will always be greater than or equal to the actual height of the object.

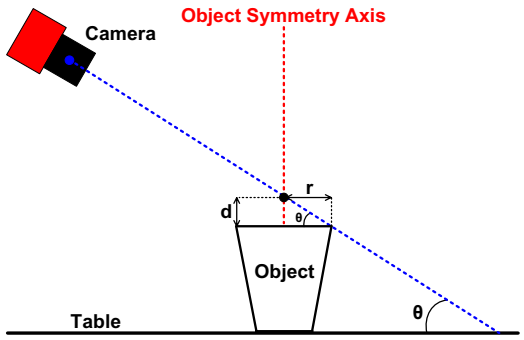


Fig. 3. Error of stereo triangulated object height.

In Figure 3, r represents the object radius and d is the error of the triangulated height. In cases where the object deviates from a surface of revolution, r represents the horizontal distance between the object's symmetry axis and the point on the top of the object that is furthest from the camera. The angle between the camera's viewing direction and the table plane is labelled as θ. Using similar triangles, the height error d is described by the following equation.

$$d = r \tan \theta \tag{1}$$

To simulate humanoid robots dealing with objects at arm's length, our experimental rig has a θ of 30 degrees. This results in a d error of roughly $0.6 \times r$. We assume objects with radii ranging from 30mm to 90mm, which produces values of d between 18mm and 54mm. To compensate for the height error d, the gripper vertical coordinates are offset downwards by 36mm. As the vertical tolerance of the robot's two-fingered end effector is well over $\pm$18mm, object grasping remains stable and reliable as demonstrated by the experiments detailed in Section V.

B. Object Grasping, Rotation and Training Data Collection

A PUMA260 6-DOF robot arm is used to perform all object manipulations. Grasping is performed by lowering the opened gripper along an object's symmetry axis. When the gripper arrives at the top of the object, offset downwards by the height triangulation error d, grasping is performed by closing the gripper. The gripper is then raised until the majority of the gripper is no longer visible in the camera image. This helps prevent the inclusion of gripper features in the object model.

Training images are collected by rotating the grasped object about the vertical axis of the robot manipulator. Right camera images are taken at 30-degree intervals over 360 degrees to produce 12 training images per object. The 30-degree angle increment is chosen according to the ± 15 degrees viewing orientation tolerance reported for SIFT descriptors [8]. The first two images of the twelve-image training set collected by the robot is shown in Figure 4. Each training image is 640×480 pixels in size.

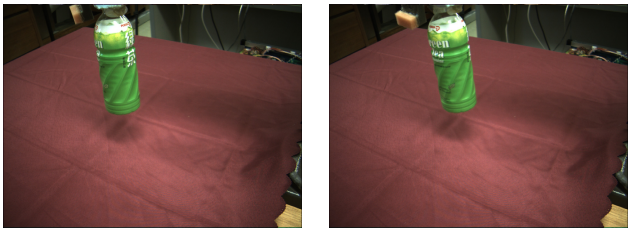


Fig. 4. First two images of the green bottle training set. The robot has rotated the grasped bottle by 30 degrees.

IV. Building Object Models using SIFT

The *scale invariant feature transform* (SIFT) [8] is a multi-scale method that extracts highly unique *descriptors* from *affine regions* on an image. The descriptors are robust against translation, rotation and illumination changes as well as small changes in viewing orientation and minor perspective effects. This makes the SIFT descriptor a viable feature for modelling objects visually.

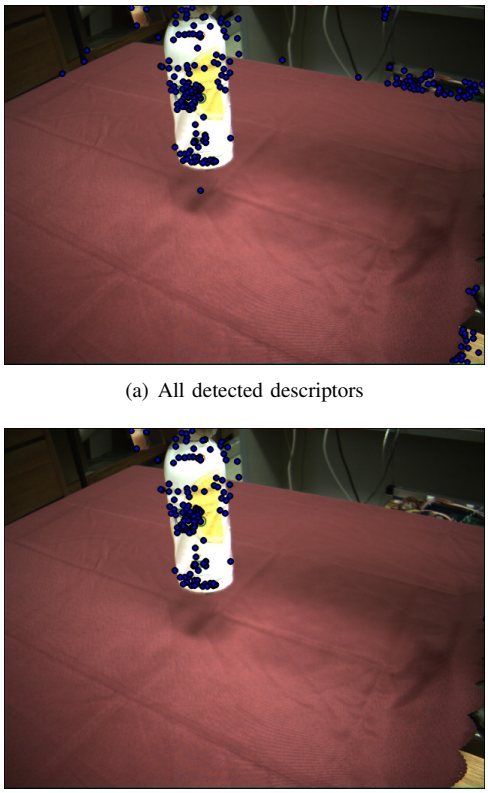
A. SIFT Detection

As shown in Figure 4, the robot rotates a grasped object to collect training images at 30-degree increments. SIFT detection is performed on each of the twelve images in a training set. David Lowe's SIFT binary is used to extract descriptors from the image. Our own C++ code is used to match and visualize the detected descriptors. The location of SIFT descriptors detected from a training image of a white bottle are visualized as blue dots in Figure 5(a). Note the dense coverage of descriptors over the object. Object modelling is performed offline after object interaction.

B. Pruning Background Descriptors

Figure 5(a) highlights the need to prune away non-object descriptors before building object models. The inclusion of background descriptors, such as those in the upper right and lower right of Figure 5(a), will probably lead to false positives during object recognition. This is especially worrying when the robot is operating on objects set against similar backgrounds.

We have developed an automatic pruning method to remove background descriptors. The pruned result is shown in Figure 5(b). Notice that the majority of background descriptors, including the descriptor extracted from the object's shadow, have been successfully removed. Our experiments suggests that the remaining non-object descriptors have negligible effect on object recognition performance.



(a) All detected descriptors

(b) Background descriptors pruned

Fig. 5. Removing background SIFT descriptors.

The descriptor pruning method is a two-step process. Firstly, a loose bounding box is placed around the grasped object to remove non-object descriptors. The bounding box is large enough to accommodate the object tilt and displacement incurred as a result of grasping and rotation. This removes a large portion of background descriptors.

The second pruning step exploits the repetitions within a training image set to remove background descriptors. As the grasped object is rotated in front of a static background, background descriptors will occur much more frequently within a training image set. We assume that an object descriptor in the current training image may also be detected in the image collected at the previous object rotation as well as the next object rotation. This means that an object descriptor can at most match with descriptors from two other images within the training set. As such, the second pruning step rejects any descriptor in a training image that has three or more descriptor matches with other images in the training set.

The proposed pruning method can be applied to other training data collection situations where an object is rotated in front of a static background. The three descriptor threshold can be altered depending on the angular increment used for object rotation. Apart from increasing object recognition robustness by reducing the probability of false positives, the reduction in the number of descriptors also reduces the computational cost of recognition. In the example shown in Figure 5, the number of descriptors is reduced from 268 to

163.

C. Object Recognition

The robot's object recognition system is described in Figure 6. The object database is created automatically from robot-collected training images of grasped objects. Each object in the database is represented by the descriptors detected from a twelve-image training set, with background descriptors pruned using the method described in Section IV-B. Note that SIFT detection is carried out on grayscale images, so colour information is not used by our object recognition system. The names of the object labels are specified manually by the user.

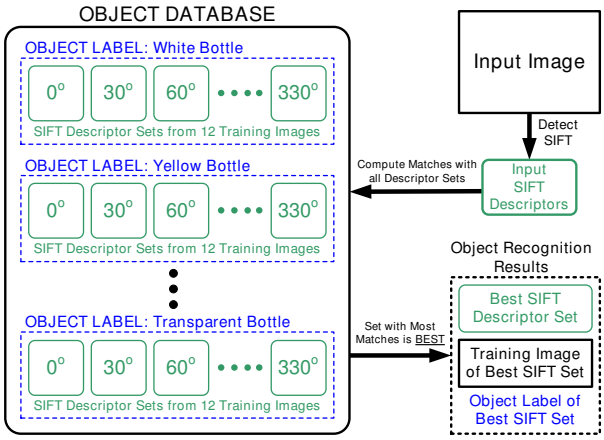


Fig. 6. Object recognition using learned SIFT descriptors.

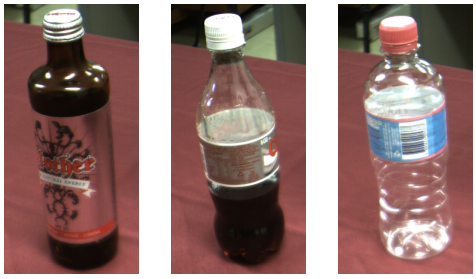
Object recognition is performed by matching descriptors in the input image against each descriptor set in the object database, shown as green squares with angles in Figure 6. As there are twelve descriptor sets per object learned, twelve set matches are performed for each object in the database. The descriptor set with the most matches is considered *best* and its object label is returned as the recognized object. The *best* descriptor set and the corresponding training image are also returned by the system. As pose estimation requires three correct matches [8], the recognition system will only return a result when three or more matches are found.

Descriptor matches are found using the ratio-based method described in [8]. Each descriptor of one set is compared with all descriptors of the other set. The best matching descriptor pair is the one with the shortest Euclidean distance. Additionally, the nearest pair is only considered a valid match if its distance is shorter than 0.6 times the second nearest pair.

V. Autonomous Object Learning Experiments

Autonomous learning experiments were performed by the robot on the seven test objects displayed in Figure 7. The test objects are beverage bottles, including two transparent bottles and a glass bottle. Apart from the cola bottle, all the other bottles are empty. This raises the object's center of gravity and makes object interactions more difficult. As beverage bottles are surfaces of revolution, they are visually symmetric from multiple views and therefore easily segmented using the robotic nudge. The robot was successful in collecting training images autonomously for all seven test objects.

(a) White (b) Yellow (c) Green (d) Brown



(e) Glass (f) Cola (g) Transparent

Fig. 7. Bottles used in object learning and recognition experiments.

The object database and recognition system is tested using 28 images, four for each test object. Each quartet of images show the test object at different orientations and set against varying amounts of object clutter. The object recognition results are shown beside their corresponding object learning videos at:

www.ecse.monash.edu.au/centres/irrc/iros09li.php

The long pause after the nudge in the object learning videos is due to the saving of data to document the experiment. This includes the saving of several hundred 640×480 tracking images collected during the nudge, which takes a considerable amount of time. Without the logging of experiment data, object grasping can occur 160ms after the robotic nudge. Two videos where experiment data saving has been disabled are available under the *Video Walkthroughs* heading on the website. These videos are a part of the first author's thesis documentation and includes a verbal narration of the autonomous learning process.

The recognition system returned the correct object label for all 28 test images. This result implies that the SIFT descriptors models built autonomously by our robot is sufficient for reliable object recognition. Statistics of the descriptor matches obtained for the test images are shown in Table I. *Good* descriptor matches are tabulated under the columns labelled with a $\surd$. *Bad* matches are shown under the $\times$ columns. We define a good match as one where the descriptor in the input image appears at a similar location on the object in the matching training image. Bad matches are those with descriptors that belong to different parts of the object.

Good and bad matches are judged by manual observation. As only three correct SIFT matches are needed for pose estimation [8], the results in Table I suggests that our robot-learned object models can be used for pose estimation.

TABLE I
OBJECT RECOGNITION RESULTS – SIFT DESCRIPTOR MATCHES

Bottle	**Image number**							
	00		01		02		03	
	√	×	√	×	√	×	√	×
White	16	0	6	0	17	0	7	0
Yellow	14	0	11	0	24	0	4	0
Green	23	1	21	1	11	0	9	1
Brown	15	0	16	0	16	0	8	0
Glass	5	0	6	1	4	1	4	1
Cola	7	0	4	0	9	0	11	0
Transparent	6	0	7	1	11	0	6	0

The recognition results for four test images are shown in Figures 8, 9, 10 and 11. The bracketed number in the captions is the same as the image numbers in Table I and the website. Each figure shows the input image above the matching training image returned by the recognition system. The recognized object label is shown as green text at the bottom of the figure. The red lines linking the two images represent the SIFT descriptor matches between the input image and the matching training image.

Figure 8 shows the affine invariant nature of our SIFT-based recognition system as the white bottle is successfully recognized despite an inversion of orientation. Figure 9 contains a difficult object recognition scenario. The green bottle is heavily occluded by objects and also has several specular reflections on its shiny surface. Despite these challenges, the correct object is recognized and the figure also shows numerous descriptor matches.

Figure 10 shows the small number of SIFT descriptor matches found for the glass bottle. This can be attributed to the bottle's reflective label and its low texture surface. Note also that a bad descriptor match is found between the bottle cap in the input image and the bottle's label in the training image. However, as four correct matches are found, object recognition remained successful. Additionally, as pose estimation only requires three correct matches, the noisy match is of minor concern.

Figure 11 shows the object recognition result for a cola bottle with large patches of transparency. Note also that the cola bottle in the input image is empty. The recognition system remains robust against the change in object appearance and the presence of background clutter.

VI. CONCLUSION AND FUTURE WORK

This paper has demonstrated a robot that can autonomously learn new objects through the intelligent use of object interaction. Our robot is able leverage a simple nudge action to pick up and rotate new objects. Experiments show that our robot is able learn beverage bottles autonomously, including transparent plastic bottles and a fragile glass bottle.

The transition from simple to advanced object interactions allow the robot to collect its own training images. An object database is constructed using the robot-collected training images by applying SIFT detection with background descriptor pruning. The resulting object recognition system performs well in our trials, reliably identifying the correct object in all test images. The recognition results show that the robot-collected training images are of sufficient quantity and quality to build useful object models. Future works can increase the discriminatory power of learned object models by including colour information. Additionally, the robustness descriptor pruning with different backgrounds should be further investigated. Possible robustness improvements may include the estimation of the fundamental matrix between object views to improve pruning and model building.

Our robot is an addition to a sparse field of systems [4], [10], [11] that actuates objects in front of a static camera instead of actuating a camera around static objects. The proposed approach takes a small but important step towards greater robot autonomy by shifting the labour intensive task of object modelling from the human user to the tireless robot.

VII. ACKNOWLEDGEMENTS

We gratefully acknowledge IRRC for their financial support. We thank the reviewers for their insightful comments.

REFERENCES

[1] H. I. Christensen, "Robotics as an enabler for aging in place," in *Robot Services in Aging Society IROS 2008 Workshop*, Nice, France, September 2008.

[2] C. Ray, F. Mondada, and R. Siegwart, "What do people expect from robots?" in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France, September 2008, pp. 3816–3821.

[3] W. H. Li and L. Kleeman, "Autonomous segmentation of near-symmetric objects through vision and robotic nudging," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France, September 2008, pp. 3604–3609.

[4] P. Fitzpatrick, "First contact: an active vision approach to segmentation," in *Proceedings of Intelligent Robots and Systems (IROS)*, vol. 3. Las Vegas, Nevada: IEEE, October 2003, pp. 2161–2166.

[5] P. Moreels and P. Perona, "Evaluation of features detectors and descriptors based on 3d objects," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 800–807.

[6] H. Kim, E. Murphy-Chutorian, and J. Triesch, "Semi-autonomous learning of objects," in *Conference on Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06.*, June 2006, pp. 145–145.

[7] J. Chen and C. Chen, "Object recognition based on image sequences by using inter-feature-line consistencies," *Pattern Recognition*, vol. 37, pp. 1913–1923, 2004.

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

[9] W. H. Li and L. Kleeman, "Fast stereo triangulation using symmetry," in *Australasian Conference on Robotics and Automation*. Auckland, New Zealand: Online, December 2006,
URL: http://www.araa.asn.au/acra/acra2006/.

[10] A. Ude, D. Omrcen, and G. Cheng, "Making object learning and recognition and active process," *International Journal of Humanoid Robotics*, vol. 5, pp. 267–286, 2008, special Issue: Towards Cognitive Humanoid Robots.

[11] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," in *In Proceedings of the IEEE International Conference on Robotics and Automation*, Kobe, Japan, May 2009.

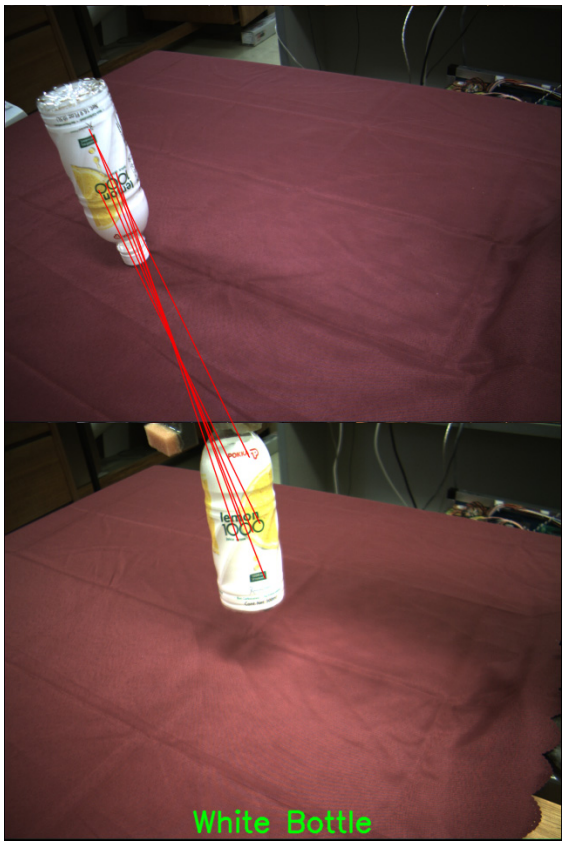


Fig. 8. Object recognition result – White bottle (Image Number 01).

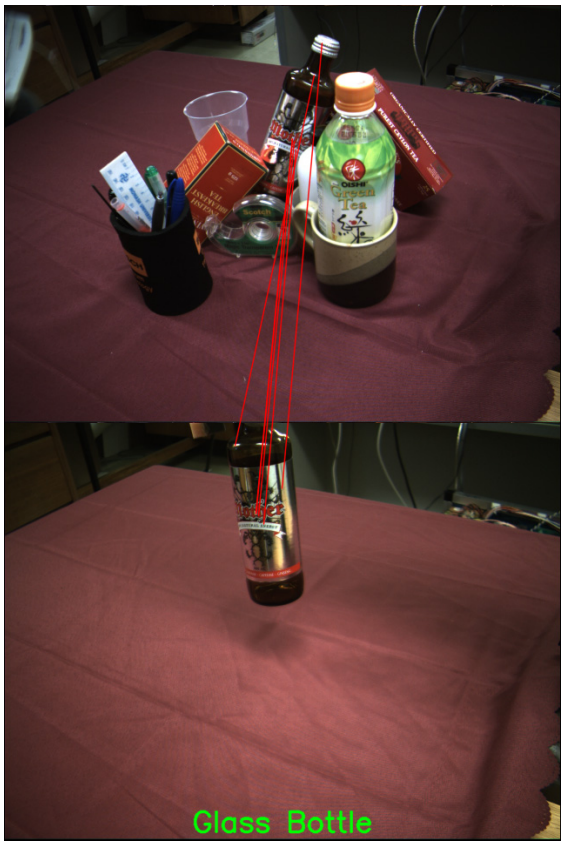


Fig. 10. Object recognition result – Glass bottle (Image Number 03).

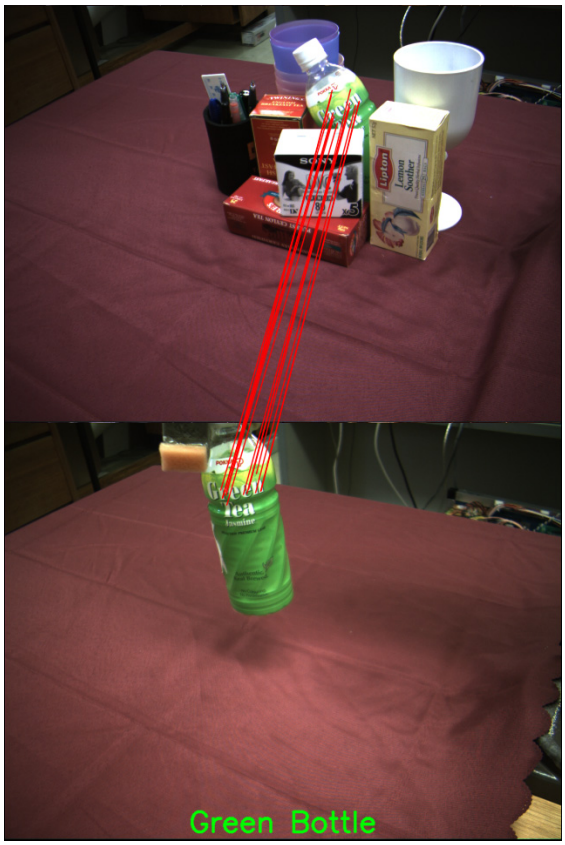


Fig. 9. Object recognition result – Green bottle (Image Number 02).

Fig. 11. Object recognition result – Cola bottle (Image Number 03).