An Active Robotic Vision System with a Pair of Moving and Stationary Cameras

S. Pourya Hoseini A.[1[0000-0003-3473-9906]], Janelle Blankenburg[1], Mircea Nicolescu[1], Monica Nicolescu[1], and David Feil-Seifer[1]

[1] University of Nevada, Reno, NV 89557, USA
hoseini@nevada.unr.edu

Abstract. Vision is one of the main potential sources of information for robots to understand their surroundings. For a vision system, a clear and close enough view of objects or events, as well as the viewpoint angle can be decisive in obtaining useful features for the vision task. In order to prevent performance drops caused by inefficient camera orientations and positions, manipulating cameras, which falls under the domain of active perception, can be a viable option in a robotic environment.
In this paper, a robotic object detection system is proposed that is capable of determining the confidence of recognition after detecting objects in a camera view. In the event of a low confidence, a secondary camera is moved toward the object and performs an independent detection round. After matching the objects in the two camera views and fusing their classification decisions through a novel transferable belief model, the final detection results are obtained. Real world experiments show the efficacy of the proposed approach in improving the object detection performance, especially in the presence of occlusion.

Keywords: Active Perception, Active Vision, Robotics, PR2, Dual-Camera, Transferable Belief Model, Dempster-Shafer, Occlusion.

1 Introduction

An important component of robotic platforms is their sensing capability. Through the sensory data, robots can interpret their surroundings. Analyzing the inputs from visual sensors, i.e. cameras, can specifically be used in determining the type of objects or events around the robot. Gaining knowledge about the existing objects or events happening around a robot can be cornerstone of many robotic applications, such as robotic manipulation, vision-based simultaneous localization and mapping (SLAM), rescue robots, social robots, etc.

In a vision system, a correct detection, among other factors, depends on obtaining distinctive features for the specific task at hand. Inadequate distinguishable features or occlusions may deteriorate a vision system's detection performance. To alleviate such problems in real world robotic platforms, trying to get a new viewpoint or moving the camera closer or farther from the event or object of interest can be advantageous.

However, in real world conditions it is not practical to capture data in all possible orientations and positions around the robot. It would be energy and time inefficient to move the cameras to all possible poses around the robot. Moreover, adding a lot of cameras on robots to process their inputs at the same time to accomplish only one task would be infeasible, due to physical, computational, and energy limits of robots. Active perception, or active vision as it is known in the computer vision community, can be the answer to these issues. Active vision is a technique to manipulate cameras to help in better performing the vision-related tasks.

Relocating cameras based on a bioinspired approach is discussed in [1], where authors analyzed the head movement of barn owls and adopted it to actuate a depth camera installed on a robot. An active object detection and pose estimation method with dynamic camera location planning is presented in [2]. The sensor used was an Asus Xtion RGB-D camera mounted on the PR2 robot's wrist. This method tries to balance the amount of energy needed to move the camera and the added chance of getting a better object detection. In [3] a relative navigation system for space robotics by means of a monocular camera is presented, which benefits from an active leader robot tracking by means of a Dual Quaternion-based controller. An active head (and camera) rotation is also implemented in [4] for humanoid soccer robots to obtain useful data for each player robot. Also, a leader-follower robotic arrangement is realized in [5] with active tracking capability of the leader through dynamically rotating a pan-controlled camera. In another work [6], an active vision system is employed on a quadrotor to detect gaps. As the quadrotor moves, optical flow is computed by considering different captures of the same scene. Subsequently, contours of the gaps are detected from the resulting optical flow.

In our work, we designed an object detection system that works with two cameras. It is implemented on a PR2 humanoid robot, with a camera mounted on the robot's head (main camera) and the other one (secondary camera) installed in the robot's arm. The main camera is a Kinect v1 3D camera, while the secondary camera is an ordinary RGB camera. In the beginning, the proposed vision system detects objects in the view seen by the main camera only. In addition to detecting objects, it also computes the confidence on recognition of the detected objects. In the case of any uncertainty in the detections, it dynamically asks for another round of detection by the secondary camera. Before getting the assistive detections from the secondary camera, it moves the arm of the robot, and with it the secondary camera, to a pose suitable for capturing another viewpoint of the object with an uncertain detection. The detections made in the two scenes viewed by the dual-camera system are matched together and then combined via a novel transferable belief model, a decision fusion method based on the Dempster-Shafer evidence theory.

The contributions of this work can be summarized in three parts, which are: (1) dynamic allocation of cameras to improve detection performance while trying to keep number of detection efforts minimum, (2) a distance-based matching scheme to associate the detection between the two camera views, and (3) a decision fusion technique to combine the classification results of the two cameras. In the rest of this paper, the proposed active perception-based robotic vision system is described in Section 2. In

Section 3, the experimental results are presented and analyzed, while Section 4 concludes the discussion.

2 The Proposed Vision System

Fig. 1 shows the main components of the proposed vision system. They are clustered in the a few main phases that are demonstrated in the left vertical bar of Fig. 1. The flowchart of Fig. 1 shows that after an initial preprocessing and denoising stage to eliminate impulsive noise through a median filter, objects in the scene viewed by the main camera are detected by means of a sliding window technique. Since the main goal of this project is to enhance the detection performance by incorporating active perception, we chose to use a simple, yet effective, detection method in our case. For every candidate object, a feature vector containing a color histogram and a histogram of oriented gradients (HOG) [7] is constructed. HOG captures edge-based appearance, whereas the color histogram is generated by merging two flattened 2D histograms in the CIELUV [8] and HSV (i.e. Hue, Saturation, Value) color spaces. The 2D histogram of the CIELUV color space is computed from the u and v channels, while the 2D histogram of the HSV color space only considers the hue and saturation channels. In contrast to the two discarded channels, value (V) and L, which contain brightness information, these channels encode the color information of pixel.

After feature extraction, a stage of feature reduction based on the Principal Component Analysis (PCA) method is used to prevent curse of dimensionality. In other words, it ensures the subsequent classifier sees a moderate number of features given the number of available training samples. To classify the input features a non-linear multi-class Support Vector Machine (SVM) classifier with one-versus-rest strategy and a Radial Basis Function kernel is used in our method. For every trained object category, the classifier outputs mass values. It will be explained later that mass values are counterparts of probabilities in terms of the Dempster-Shafer theory, and represent the belief of the classifier concerning the similarity to the trained object categories.

In the proposed method, first the detection procedure is applied to the main camera frames. Subsequently, a confidence measure is computed by dividing the maximum mass value in the output mass vector to the second largest mass value. A low confidence value occurs when there are at least two strong candidate categories, which typically is the result of lack of discriminating features in the appearance of the candidate object. By comparing the confidence measure with a threshold value, the classification of detected objects in the main view is deemed “reliable” or “unreliable”. In the case of an unreliable classification, the secondary camera is directed to perform object detection on its input frames, otherwise the reliable detection result is regarded final.

Prior to starting the second round of object detection, the secondary camera should be moved to a pose with respect to the object with unreliable detection to have it in its perspective. After the secondary camera’s detections are done, the detections in the two camera views are matched and their classification results are fused and converted

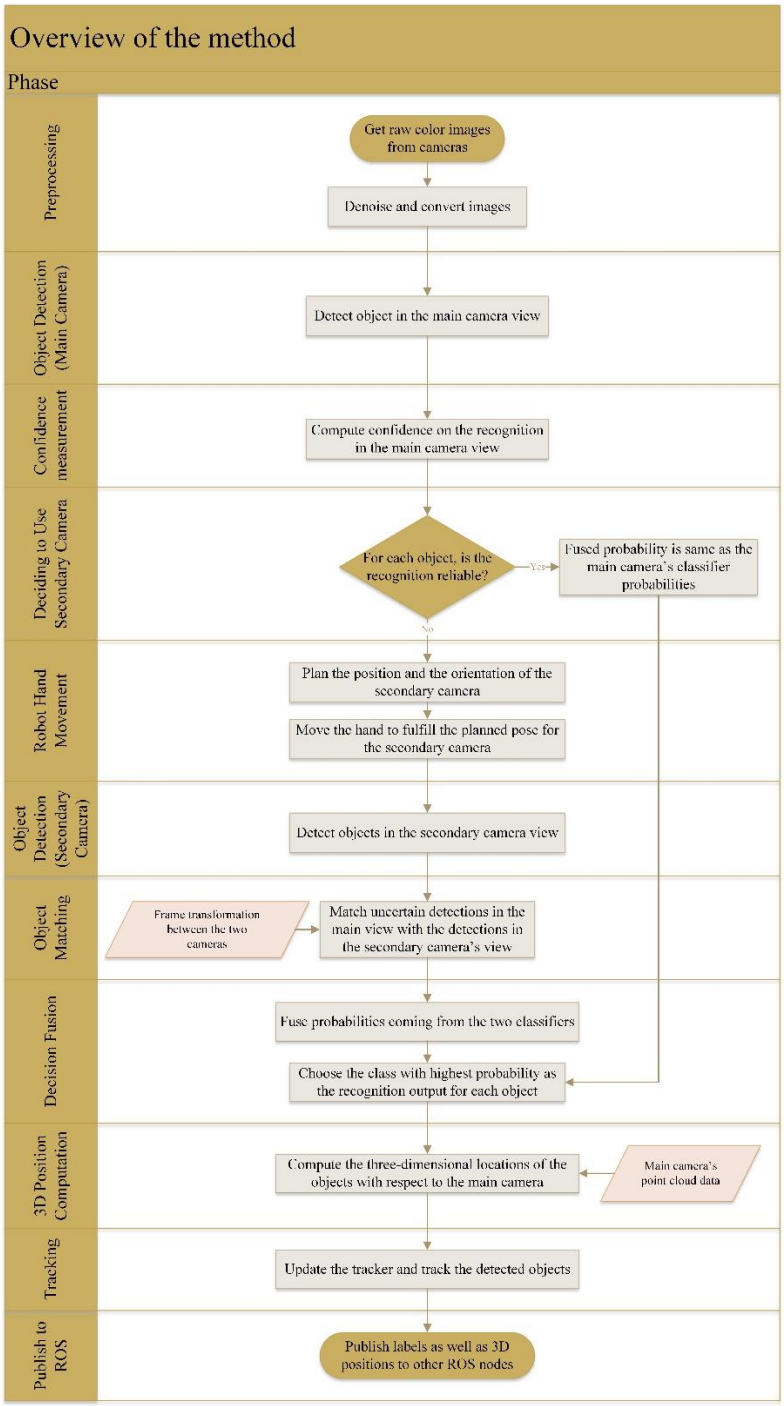

Fig. 1. Main steps of the proposed vision system.

to probability values to form the final probability vectors. The classification output of each object is then the winner category with the highest probability.

In order to keep the computational load of the proposed vision method suitable for real-time applications, median flow tracker [9] is embedded in the system. It handles the tracking of the objects between every two successive object detections.

By completing the active perception-based object detection, the proposed system obtains the 3D location of the objects with respect to the robot by means of the point cloud data of the 3D camera and publishes them to other robotic nodes in the robot operating system (ROS). The proposed vision system has been incorporated into a multi-robot control architecture for collaborative task execution [15]. This real-time distributed architecture enables collaborative execution of tasks with hierarchical representations and multiple types of execution constraints. In the following we focus more on the matching, camera pose planning, and decision fusion modules.

2.1 Matching

To enable fusing the decisions of the two classifiers, it is required to associate objects in both views to form object pairs. For any unreliable detection in the main view, if there is an association in the secondary view, the pair of matched objects is formed, else the classification of the main view object is considered final without any fusion.

Matching of objects between the two views might be fulfilled in a variety of ways. Among them we can enumerate matching of shape, keypoints, appearance, etc. Although they may work very well in some situations, but since in our work there is no guarantee that the two camera views are close to each other, there can be large variations in objects shape, size, appearance or keypoints, which in turn may hamper proper functionality of the matching [10]. On the other hand, the proposed active vision system is implemented on a robot with access to transformations between its coordinate frames and depth information on at least one camera. Thus, given the transformation between the two corresponding camera coordinate frames it is possible to transform the relative 3D position of an object with respect to the depth camera to another camera. After this transformation, the 3D location with respect to the secondary camera can be converted to its pixel coordinate using its intrinsic calibration data. The accessibility to detections of the main view in the pixel coordinate of the secondary view, facilitates distance matching between the detections of the secondary view and the detections of the main view mapped to the secondary camera's frame. Not only this technique avoids the aforementioned problems encountered by other types of matching techniques, but also by taking into account their computational complexity, distance matching should be faster than other methods, since it avoids the extraction of image features.

Fig. 2 delineates the steps in the proposed matching algorithm. It is evident in the flowchart that instead of transforming all the pixels of an object in the main view to the secondary view, centroid of a window around an object is converted only. Assuming that centroids of objects have high chances of being from the actual object surface this strategy keeps computations low by avoiding segmentation and transformations in pixel granularity.

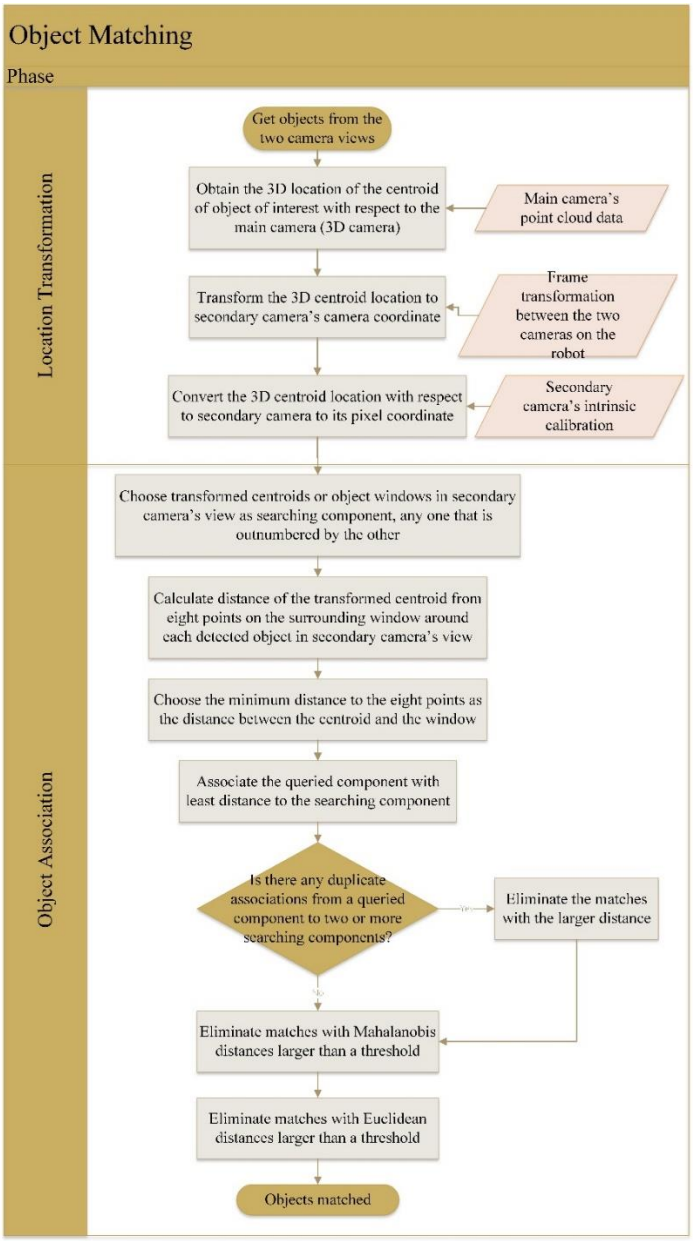

Fig. 2. Flowchart of the proposed object matching

Fig. 2 also shows that the distances of any transformed centroid to eight points around the bounding boxes of original detections in the secondary view are evaluated. By considering eight points of a window instead of its centroid to match, we prevent problems that may arise with objects that look long in the secondary viewpoint. From

a secondary perspective, centroid of a long object may be too far from the viewable surface of that object in the main viewpoint (and its centroid).

2.2 Camera Pose Planning

Whenever there is an unreliable detection, the secondary camera should be planned to be moved to a proper pose to have a clear view of the object. Fig. 3 shows a schematic of the PR2 robot, in which the main parameters for planning the robot's arm are shown in red. Distance of the object to the shoulder joint is computed by using the point cloud data from the 3D camera and the robot's inner frame transformations. Camera angle, forearm length, and upper arm length are also known in advance. Through the geometric computations, it is possible to find the shoulder lift and flex joints to have the secondary camera toward the object. The forearm roll (rotation) is determined relative to the computed shoulder joint values to keep the forearm camera (secondary camera) facing the object. The elbow flex angle and the upper arm rotation are set to a fixed value to simplify calculations by reducing the number of degrees of freedom in the planner. Detailed discussion about the geometric computations is beyond the scope of this paper.

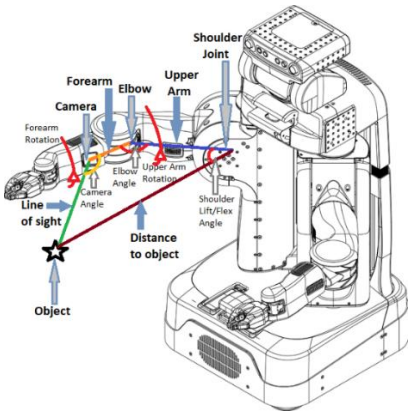

Fig. 3. Schematic of the parameters needed to plan the movement of the secondary camera.

2.3 Decision Fusion

Typically, a multiclass classifier outputs a one-dimensional probability array. It constitutes a set of singleton (mutually exclusive) probabilities, which in terms of Dempster-Shafer evidence theory (DST), is called frame of discernment (Ω). Instead of merely relying on Ω, DST decision fusion alternatively allows for the power set of Ω and attributes a value called mass in the range of [0, 1] to any element of the power set. Here, masses can be compared to probabilities in Bayesian fusion. Any element of the power set with mass value greater than zero is called a focal element. Additionally, the sum of all masses in the power set must be equal to 1, as shown in (1):

$$\sum_{\Psi \in 2^{\Omega}} m(\Psi) = 1 \tag{1}$$

where Ψ is an element of the power set of Ω and *m(.)* is a mass value for it. Because in our work there are two sets of classification results from the two classifiers, we have two frames of discernments to combine. In the proposed fusion method, the power set is designed to have *n + 1* focal elements, in which *n* elements are the singleton object categories and the last one is the universe of object categories. By defining the power set this way, we have masses for each object class that work like probabilities of that object class, while there is a mass value for the set of all the training samples [11]. We call this extra category, *unknown* class, because it actually represents how a classifier believes an object of interest is similar to all of the objects in its training set. Due to the larger size of the training samples for the *unknown* class, a balancing scheme by using weights relative to the training set size of each object category is used in the optimization formulae of the SVM classifier.

To fuse the two mass vectors from the main and secondary view classifiers, the unnormalized rule of combination [12] is used, as shown in the following:

$$m(\Psi) = \sum_{\alpha \cap \beta = \Psi} m_A(\alpha) * m_B(\beta), \quad \forall\, \Psi \subseteq \Omega, \alpha \in A, \beta \in B \tag{2}$$

where $m(\Psi)$ is mass of category Ψ, and sets A and B are mass vectors of the main view and the secondary view, respectively. Focal elements α and β are each a category from the mass vector of the main view and the secondary view classifiers in order. As an example, by considering (2) in our application, a category α in the mass vector A, except the class *unknown*, has intersection with two βs, namely the *unknown* category of the mass vector B and the corresponding element in B (the same label in B).

The proposed Dempster-Shafer fusion is a transferable belief model as a result of using an unnormalized rule of combination [12]. To convert the combined mass vector to a probability vector, we employ the pignistic transformation described in [13]:

$$P(\omega) = \sum_{\Psi \in \Phi} \frac{m(\Psi)}{|\Psi|}, \quad \forall\, \omega \in \Omega, \Phi = \{\Psi \mid \Psi \subseteq \Omega, \omega \in \Psi\} \tag{3}$$

where $P(\omega)$ is the probability of an object class ω, excluding the *unknown* category. In addition, Ψ is a focal element of the power set of object classes and $|\Psi|$ designates the number of object classes in Ψ. The above equation shows that any mass of belief is distributed among its comprising class probabilities [14].

Generally speaking, the two classifiers feeding the DST fusion provide a mass value for an *unknown* category besides actual the classes of objects present in the training. This specific mass value can be considered as indicating to the fusion module the uncertainty of the classifier about its recognition results. Considering the combination rule in (2) and the fact that sum of all masses resulted from a classifier is 1 (as stated in (1)), we observe not only that an increase in the mass of the *unknown* category of a classifier decreases other masses of that classifier, but also it weighs more toward the

masses of the other classifier, with which it has a non-empty intersection. In contrary, a more resolute classifier that contributes more to the final fusion result is obtained when the *unknown* mass of the classifier is low.

3 Experimental Results

The proposed active vision system was implemented on a PR2 robot. Fig. 4 shows a sample situation where there are three objects in the scene viewed by the main camera. One of the objects, though, (Tea Pot) is partially occluded. Each of the three parts of the figure, show the output of the proposed system in a different time. The bottom left image in the system output is the processed main view, while the bottom right is the secondary view, and the top left is the output of the active vision system superimposed on the scene viewed in the main view. In the beginning (Fig. 4a), there is an incorrect detection in the main view for the occluded object (Tea Pot is recognized as Sugar.) The active perception system has found it as an unreliable detection and marked it with a red bounding box. After that (Fig. 4b), the active vision system moves an arm with the secondary camera on it toward the object and detects objects in the secondary view. The detections in the secondary view are matched with those in the main view, and only the one which has been matched with an unreliable detection in the main view is shown (Tea Pot). We observe that it is correctly detected there as it has a clearer view of the object compared to the main view. The fusion result (top left of Fig. 4b) demonstrates the system was successful in correcting the initial erroneous detection. Consequently in Fig. 4c, the manipulation section receives the object detection and position information from the proposed system and based on that starts to manipulate the objects with the other arm.

Our method was tested in fifteen different real-world benchmarks. In the tests, objects were placed in front of the robot in various table-top settings in different lighting conditions. The benchmarks were also divided into two types: those with objects being partially occluded and those without any obstacles in viewpoint of the main camera. Table 1 shows the test results in terms of precision, recall, accuracy, and F_1 score (harmonic mean of precision and recall) for three scenarios: a single camera only (main camera), the proposed system without any secondary camera movement, and the system with its complete functionality. Macro-averaging in Table 1 means the measure is calculated separately for each object category and is averaged over the results. In contrary, micro-averaging is the process of computing the measures for all the object categories collectively. Micro-averaging results were not included in Table 1, because micro-averaging precision and recall are equal to accuracy in the case of multi-class classifiers.

Table 1 indicates large improvements in the performance of the proposed system in comparison to the conventional single camera configuration. In accurately detecting the non-occluded objects, the proposed method outperforms the traditional single camera setup by 12.9%, 12.2%, 13.1%, and 12.5% in precision, recall, accuracy, and F_1 score, respectively. These enhancements bring the percentage of the four

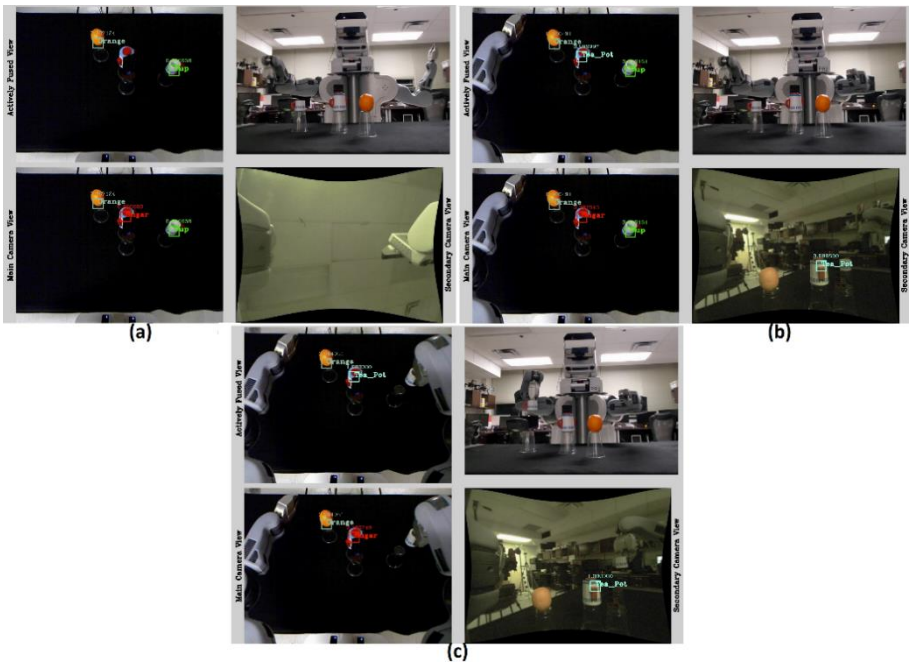

Fig. 4. A sample situation involving active perception through the proposed vision system and manipulation of objects. (a) Before using the secondary camera, (b) After completing the active detection, (c) Subsequent manipulation of objects based on the provided information from the vision system.

measures to over 97.7%. In the tests consisting of only partially occluded objects, the active vision system was even more successful in ameliorating over the traditional single camera robotic vision with 18.3%, 15.5%, 17.1%, and 16.9% increases in precision, recall, accuracy, and F_1 score, respectively. The reason is that, the active vision had access to more informative viewpoints via its use of the secondary camera. Furthermore, it is obvious in Table 1 that the proposed method with the dynamic secondary camera movement works better than the same system with fixed secondary camera. The accuracy of the vision system increased by 4.4% and 2.2% in the benchmarks with non-occluded and occluded objects, respectively, when the secondary camera pose planning and movement was added to the proposed system. This proves, not only using cameras in different viewpoints contributes to a better detection performance, but also planning and moving the extra views help in further improvements.

In order to evaluate the efficacy of the transferable belief model (TBM) decision fusion proposed in this paper, we implemented our method with the fusion module replaced with Bayesian fusion [11] instead. Table 2 shows the results obtained for the same tests we conducted for Table 1. By comparing them to those illustrated in Table 1, we observe that without any fusion taken place (single camera only) the vision system with the classifier trained for the Bayesian fusion works better. The reason is probability the added *unknown* category in the classifier for TBM, which reduces the

probability of other object categories before any fusion is taken place. In spite of a weaker single camera performance, the proposed TBM fusion, however, is superior in the final results obtained after the decision fusion are applied.

Table 1. Performance results of the proposed active vision system.

Performance Measure	Single Camera		Active Perception (no camera motion)		Active Perception (with camera motion)	
	Non-occluded	Partially occluded	Non-occluded	Partially occluded	Non-occluded	Partially occluded
Macro-Averaging Precision	0.860	0.745	0.944	0.903	0.989	0.928
Macro-Averaging Recall	0.855	0.700	0.933	0.833	0.977	0.855
Accuracy	0.846	0.684	0.933	0.833	0.977	0.855
F_1 Score	0.857	0.721	0.938	0.866	0.982	0.890

Table 2. Performance results of the proposed active vision system with Bayesian fusion.

Performance Measure	Single Camera		Active Perception (no camera motion)		Active Perception (with camera motion)	
	Non-occluded	Partially occluded	Non-occluded	Partially occluded	Non-occluded	Partially occluded
Macro-Averaging Precision	0.884	0.765	0.922	0.878	0.967	0.907
Macro-Averaging Recall	0.877	0.722	0.911	0.811	0.955	0.833
Accuracy	0.868	0.706	0.911	0.811	0.955	0.833
F_1 Score	0.880	0.742	0.916	0.843	0.960	0.868

4 Conclusion

In this paper, a dual-camera robotic object detection system based on the idea of active perception was presented. It is implemented on a PR2 humanoid robot. The contributions of the work are the dynamic switching capability between cameras, which is accomplished automatically, a fast matching algorithm between the cameras, and a decision fusion method established on the basis of Dempster-Shafer evidence theory. The experimental results in real-world tests prove the efficiency of the proposed method in enhancing the vision performance of robots. The accuracy of the presented vision approach is 13.1% more than a traditional single camera robotic vision system. The practicality of the proposed decision fusion is also verified in the tests. The presented method is useful in making robotic vision more robust when there is more than one camera present. It is especially applicable in dealing with partial occlusions as shown in the tests.

A future work can be adding the ability of handling complete occlusion of objects. Another future direction can be extending the functionalities of the current system to the field of activity and intent recognition.

Acknowledgment. This work has been supported by Office of Naval Research Award #N00014-16-1-2312.

References

1. Barzilay, Q., Zelnik-Manor, L., Gutfreund, Y., Wagner, H., Wolf, A.: From biokinematics to a robotic active vision system. Bioinspiration and Biomimetics 12(5): 056004 (2017).
2. Atanasov, N., Sankaran, B., Le Ny, J., Pappas, G. J., Daniilidis, K.: Nonmyopic view planning for active object classification and pose estimation. IEEE Transactions on Robotics 30(5), 1078-1090 (2014).
3. Zhang, G., Kontitsis, M., Filipe, N., Tsiotras, P., Vela, P. A.: Cooperative relative navigation for space rendezvous and proximity operations using controlled active vision. Journal of Field Robotics 33(2), 205-228 (2016).
4. Mattamala, M., Villegas, C., Yanez, J. M., Cano, P., Ruiz-del-Solar, J.: A dynamic and efficient active vision system for humanoid soccer robots. In: Almeida L., Ji J., Steinbauer G., Luke S. (eds) RoboCup 2015: Robot World Cup XIX. LNCS, vol. 9513, pp. 316-327. Springer, Cham (2015).
5. Chen, X., Jia, Y.: Adaptive leader-follower formation control of non-holonomic mobile robots using active vision. IET Control Theory and Applications 9(8), 1302-1311 (2015).
6. Sanket, N. J., Singh, C. D., Ganguly, K., Fermuller, C., Aloimonos, Y.: GapFlyt: Active vision based minimalist structure-less gap detection for quadrotor flight. IEEE Robotics and Automation Letters 3(4), 2799-2806 (2018).
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Diego (2005).
8. CIE 015:2018: Colorimetry. 4th ed., International Commission on Illumination.
9. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection and tracking failures. In: International Conference on Pattern Recognition, Istanbul (2010).
10. Hoseini A., S. P., Nicolescu, M., Nicolescu, M.: Active object detection through dynamic incorporation of Dempster-Shafer fusion for robotic applications. In: International Conference on Vision, Image and Signal Processing (ICVISP), Las Vegas (2018).
11. Hoseini. A., S.P., Nicolescu, M., Nicolescu, M.: Handling ambiguous object recognition situations in a robotic environment via dynamic information fusion. In: IEEE Conference on Cognitive and Computational Aspects of Situation Management, Boston (2018).
12. Smets, P.: The combination of evidence in the transferable belief model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990).
13. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence 66(2), 191–243 (1994).
14. Denoeux, T.: A neural network classifier based on Dempster-Shafer theory. IEEE Trans. on Systems, Man, and Cybernetics, Part A: Systems and Humans 30(2), 131-150 (2000).
15. Blankenburg, J., Banisetty, S.B., Hoseini Alinodehi, S. P., Fraser, L., Feil-Seifer, D., Nicolescu, M., Nicolescu, M.: A distributed control architecture for collaborative multi-robot task allocation. In: International Conference on Humanoid Robotics, Birmingham (2017).