Active Object Detection Through Dynamic Incorporation of Dempster-Shafer Fusion for Robotic Applications

S. Pourya Hoseini A.
Department of Computer Science and Engineering
University of Nevada, Reno
Reno, USA
hoseini@nevada.unr.edu

Mircea Nicolescu
Department of Computer Science and Engineering
University of Nevada, Reno
Reno, USA
mircea@cse.unr.edu

Monica Nicolescu
Department of Computer Science and Engineering
University of Nevada, Reno
Reno, USA
monica@cse.unr.edu

ABSTRACT
Employing multiple sensing capabilities in a robotic platform offers significant advantages in increasing the recognition abilities of robots. Specifically, for vision-based object detection in a real-world environment, acquiring information from different viewpoints might be decisive for correct classifications in the presence of occlusions or to disambiguate between similar objects. For this reason, an active vision object detection system is proposed in this paper. It is implemented on a robotic environment that uses a 3D camera mounted on the robot head and an RGB camera on its hand. The system tries to detect and recognize objects being seen from the head camera, while computing a confidence score on the classification. In the case of an unreliable classification, another stage of object recognition is dynamically requested, but this time from the viewpoint of the hand camera. The objects detected from the two cameras are matched and their classification decisions are fused through a novel fusion approach based on the Dempster-Shafer evidence theory. Experimental results show sizable improvements in object recognition performance compared to a traditional single-camera configuration, as well as applicability to handling partial occlusions.

CCS Concepts
• Computing methodologies → Vision for robotics

Keywords
Object Detection; Active Vision; Distance Matching; Dempster-Shafer Fusion; Transferable Belief Mode; PR2; Robotics

1. INTRODUCTION
In the context of robotic applications, active vision can be effectively employed to address traditional limitations of static/single camera configurations. We are interested in investigating novel computer vision techniques that dynamically manipulate cameras (mounted on autonomous robots) in order to better explore and understand the environment, as compared to static camera solutions. In general, such techniques are well suited for (1) detecting and avoiding occlusion by direct camera manipulation, (2) achieving a dynamic, wide field of view for tracking, and (3) recognizing objects of interest, human postures and gestures at finer levels of camera resolution.

Current generations of robotic systems usually have a multitude of sensory inputs, including RGB and depth cameras, thus offering the prospect of utilizing more than a single viewpoint for vision-related tasks. However, it is not practical to continuously process input from all sensors, at all times. This would increase the computational burden on the system, to the point where real-time functionality becomes difficult to achieve. One approach to address this problem is sensor management by dynamically selecting the most appropriate information from the cameras. In the realm of computer vision, this strategy belongs to the family of active vision methods, which we employed in a robotic system that is able to detect and recognize objects of interest in a scene.

In [12], an active vision system uses a few fixed cameras and dynamically decides how it can make best use of information from those cameras. An active camera location planning and pose estimation method is presented in [11]. Through dynamically rotating a pan-controlled camera, a leader-follower robotic arrangement with active tracking capability of the leader is realized in [18]. A reinforcement learning approach for selectively focusing on part of the input image in a vision system is presented in [5]. Improvements in learning rate and processing speed were observed as a result of the active selective attention mechanism. Furthermore, the work in [6] proposes an attention selection technique for event recognition. A comprehensive survey of active vision literature can be found in [17].

In our work, we designed an active object recognition system and implemented it on a PR2 robot. The proposed vision system has access to a 3D Kinect v1 sensor (primary camera) mounted on the robot's head and an RGB sensor (secondary camera) mounted on the left hand of the robot. In the proposed method, objects in the scene viewed by the main camera are detected and classified first, and a confidence measure for each object is computed. Based on the level of confidence, the active vision system decides whether the classification is reliable, and dynamically requests the input from the secondary camera for unreliably detected objects. After a stage of matching, the two decisions obtained from the two classifiers are combined via a novel transferable belief model, a variant of the Dempster-Shafer evidence theory. Once the final classifications are determined, the system infers the 3D position of each object and sends it along with the object label to other nodes in the robot that are fed by the vision module.

The contributions of the proposed method are (1) an active object detection system realized on the PR2 robot with dynamic confidence-based switching capability between the head camera and the eye-in-hand camera, (2) a distance-based object matching with efficient use of available information in the robotic platform, and (3) fusing the

classification decisions with a novel Dempster-Shafer fusion technique.

In the remainder of this paper, we describe the proposed active vision system in Section 2 with more details regarding the matching and fusion techniques. Section 3 presents the experimental results. Finally, concluding remarks are provided in Section 4.

2. THE PROPOSED ACTIVE OBJECT DETETION SYSTEM

The flowchart of the proposed method is depicted in Figure 1, with the left vertical bar showing the main phases. In the beginning, raw images are captured, followed by a stage of denoising with a median filter to eliminate impulsive noise and a Gaussian filter to remove additive noise. Next, potential objects of interest in the scene are detected by using a mixture of Gaussians [2] background modeling and foreground segmentation. We used background-foreground segmentation as a simple and yet effective method in our case. Works like [12] and [10] may be mentioned as examples of detecting objects with static vision sensor setups.

Since the mixture of Gaussians background-foreground segmentation works at the pixel level, there can be some small noisy areas in the foreground map; hence, the resultant binary foreground map is then cleared from noise by morphological opening followed by morphological closing. The former ensures removal of small noisy foreground segments and the latter works for eliminating small background patches. The potential objects of interest are determined by applying a connected components technique and the subsequent removal of very small objects. From this point, operations continue for the main camera image only.

For each of the candidate objects we extract features based on a histogram of oriented gradients (HOG) [13] and a color histogram. The color histogram, which is the result of concatenating histograms of the three RGB color planes, describes the overall color information for the object, while the HOG is responsible for capturing its edge-based appearance. These features are fed into a non-linear multi-class Support Vector Machine (SVM) classifier with a Radial Basis Function kernel and one-versus-rest strategy. In coordination with the Dempster-Shafer decision fusion method mentioned above, the classifier outputs mass values for each trained object category. 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 object categories. From the mass values obtained for each trained object category a confidence measure is calculated through dividing the maximum mass value of all object categories by the second highest mass value. In this way, the confidence metric checks for large enough peaks in mass values. A low confidence typically corresponds to two close competitor categories, which makes selecting either of them unreliable for the object recognition system. Accordingly, if the confidence is greater than a threshold value, it is considered reliable and the category with the largest mass value is selected as the recognition result. Otherwise, the active vision system will dynamically request additional evidence from the secondary camera in order to improve the reliability of the recognition process.

In order to keep a reasonable computation load, it is essential to only classify the objects in the secondary view that correspond to objects unreliably recognized in the main view. To this end, a matching stage is performed between the objects found in the main view and the secondary view. Matching is also indispensable for the fusion of the classification decisions. The Euclidean distance-based matching procedure is discussed in the next section, followed by a description of the decision fusion technique, a novel variation of a transferable belief model, which in turn is a type of Dempster-Shafer fusion method. After fusing masses from the two classifiers, a single probability vector is obtained. At this point, the category with the highest probability is chosen as the winner class.

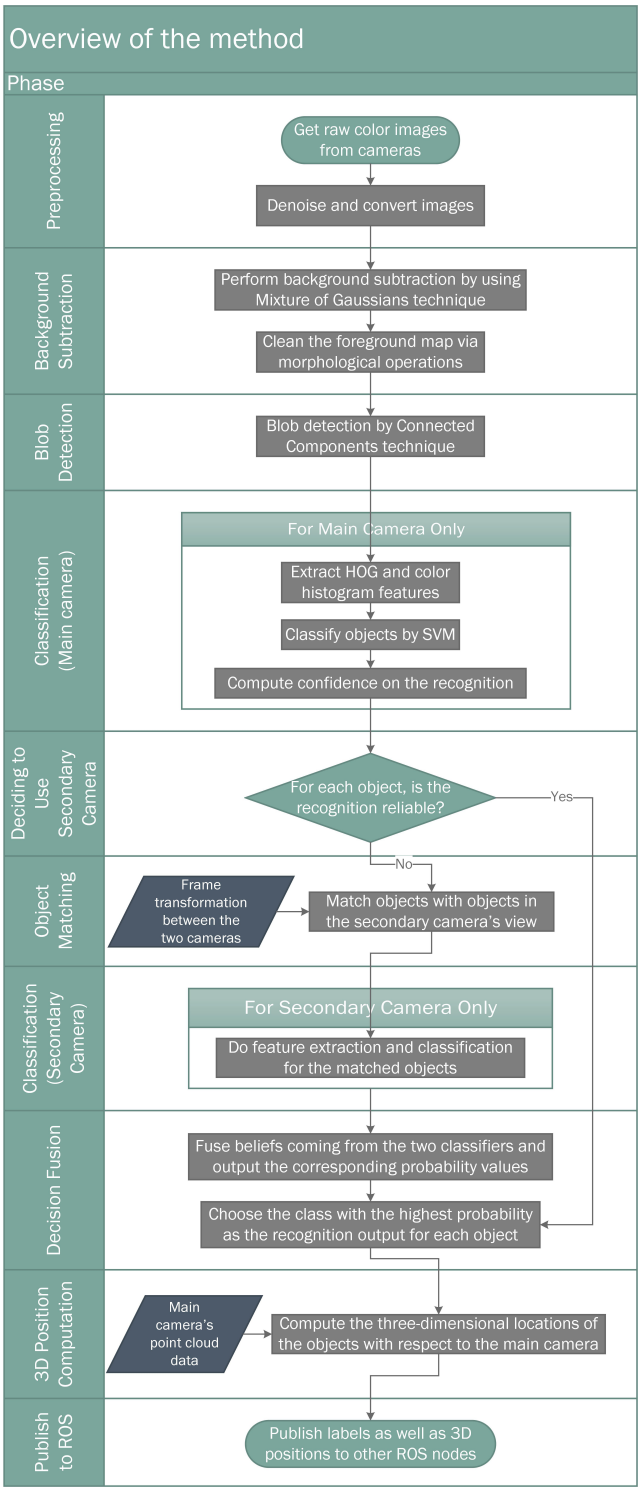

Figure 1. Overview of the proposed method.

Our active vision approach is employed in a larger robotic application for multi-agent collaboration that involves the manipulation of the recognized objects, for which the location of the detected objects in 3D space is also required. Therefore, the 3D main camera's point cloud data is utilized to infer the 3D position of object centroids. The object labels in conjunction with their 3D locations with respect to the main camera are then published to other robotic nodes over a Robot Operating System (ROS) network.

In the next two subsections, we detail the object matching and decision fusion approaches in the proposed vision system.

2.1 The Object Matching Module

As stated before, a process of matching objects in the two camera views is necessary for the later stage of decision fusion. This can be achieved by using various techniques, such as shape, keypoint, and appearance matching [16]. Nonetheless, their application in our case can be limited. Due to the physical placement of the robot cameras (on the head and on the hand, in our case), there is no guarantee that the two viewpoints are close to each other and the object appearances could be very different – in fact, it is actually desirable to have widely different viewpoints in order to allow classifications that complement each other. On the other hand, the proposed approach is implemented on a robot with access to at least one 3D camera, the transformation between the two camera coordinates, and intrinsic calibrations of both cameras. Thereby, it is feasible to transform an object's position in the pixel coordinates of the 3D main camera to the pixel coordinates of the secondary camera, making it possible to perform matching based on distances. This also provides for a fast matching procedure, because no feature extraction and correlation are required for it to operate.

Figure 2. illustrates the flowchart of the matching module. For any object in the main view being queried for a match in the secondary view, we first compute its centroid. By using the point cloud information from the 3D camera, we convert the centroid from the pixel coordinate to the camera coordinate of the main camera. The 3D position of the object with respect to the main camera is transformed to the camera coordinate of the secondary camera by using the extrinsic transformation available between the two robot frames. The intrinsic calibration of the secondary camera is then utilized for obtaining the pixel coordinates in that camera view. By transforming just the centroid point, we keep computations low. In addition, the centroid of a window around an object has a better chance for belonging to the actual object, thus avoiding the use of a point in 3D space that does not lie upon the object surface.

After transforming object centroids to the secondary camera view, we have two groups of components to match: centroids of the main view objects and bounding boxes of the objects in the secondary view. The group with a higher number of components is selected as the queried component group, while the other one will be the searching component group. The next step is calculating the L2 distance between any centroid and eight surrounding points of all bounding boxes of the secondary view objects. Four corners of a bounding box plus four middle points of its edges constitute the eight points of an object's surrounding window. The minimum of the eight distances between the centroid and the eight points is regarded as the distance of that centroid from the bounding box for every centroid-bounding box pair. By considering eight points of a window instead of its centroid to match, we prevent problems that may arise with objects that appear elongated in the secondary view. From the secondary camera viewpoint, the centroid of an elongated object may be too far from the viewable surface from the main viewpoint, thus the centroid from the secondary viewpoint will be far from the centroid of the viewable surface from the main camera.

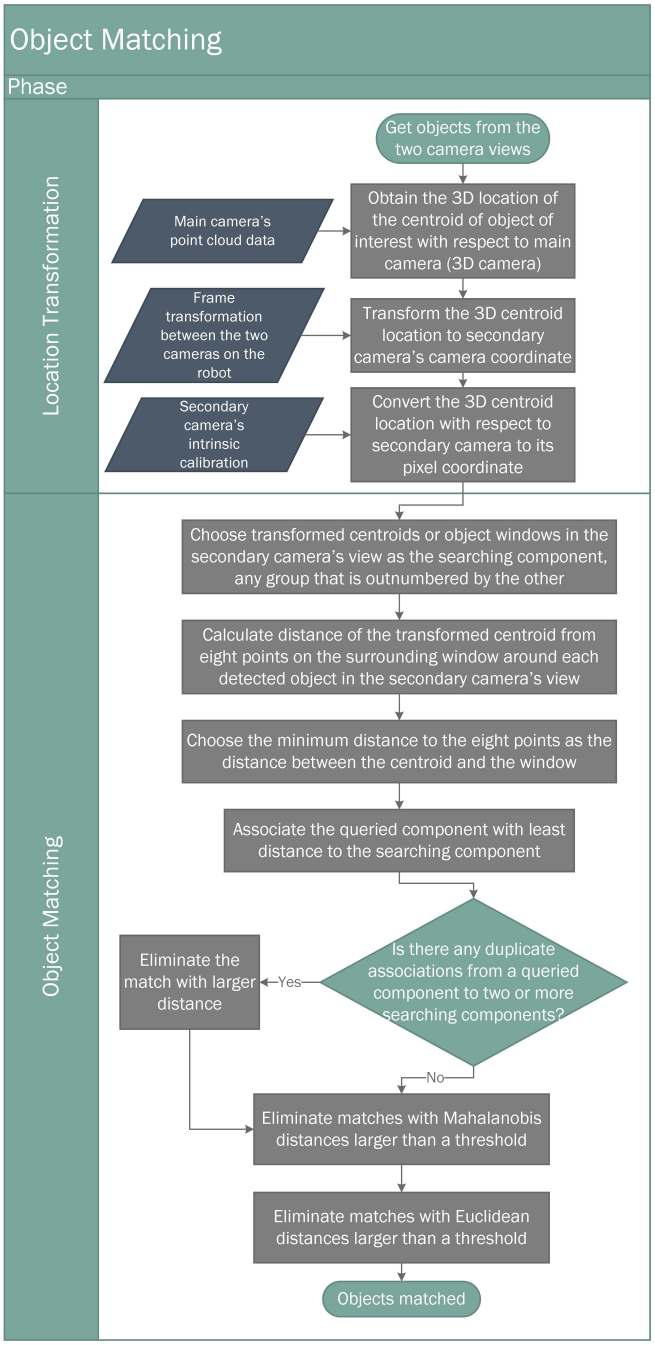

Figure 2. Flowchart of the object matching module.

Matches are determined by associating a queried component with the smallest distance to any searching component. However, it is possible for a queried component to get associated with more than one searching component. This issue is resolved by keeping only the association with a smaller distance. Later, we remove matches with Mahalanobis distances more than a threshold in order to prune associations with irregular distances compared to all the others. Additionally, associations with Euclidean distances more than a specified threshold are cancelled to avoid matches with very large absolute distances.

2.2 The Dempster-Shafer Decision Fusion Module

The Dempster-Shafer evidence theory [9] is an information fusion approach that takes into consideration uncertainty and inaccuracy [1]. It is in contrast to Bayesian fusion, as it does not deal with singleton probabilities only. As a substitute, there can be alternative units of belief with non-empty intersections [4]. The works in [3,7,8] are a few examples of the application of the Dempster-Shafer fusion, used in airborne object identification, human activity recognition, and vehicle location verification, respectively.

Assume there exists a set of singleton probabilities **Ω**, which is called frame of discernment. Here, singleton means that any two probabilities in **Ω** are mutually exclusive. In a classification task, **Ω** characterizes the set of probabilities for every object category. Instead of merely using singleton probabilities, the Dempster-Shafer fusion allows for the power set of **Ω**, which is a broader set of probabilities. All combinations of singleton probabilities from an empty set to the set universe of **Ω** are included in the power set of **Ω**. In the context of the Dempster-Shafer fusion, a value called mass, in the range of [0, 1], is attributed to any subset element in the power set. Any subset of **Ω** with a mass value greater than zero is termed a focal element. By the above definition, we may consider masses in Dempster-Shafer fusion as a replacement for probabilities in Bayesian fusion. As shown in (1), the sum of all masses must be equal to 1. Here Ψ represents an input subset of **Ω**, while m(.) is a mass value for it.

$$\sum_{\Psi \subseteq \Omega} m(\Psi) = 1 \qquad (1)$$

In the proposed method, a probability vector coming from one of the classifiers stands for the frame of discernment. Therefore, we have two frames of discernment for our two classifiers. Given n object categories, there will be n elements in every probability vector, thus in each frame of discernment. The n elements of the frame of discernment plus an additional "universal" element constitute the focal elements of the power set for each classifier output. The universal element is actually the universe of the probability vector elements. By defining the power set this way, there is a mass value for every object class. Additionally, the extra universal mass value determines how similar the object of interest is to the whole training set. Because the universal element does not point to any specific category, its mass is analogous to the probability of an "unknown" object.

In order to obtain $(n + 1)$ masses in the case of the proposed Dempster-Shafer fusion, we shall have $(n + 1)$ output classes in each classifier. Out of them, n classes (related to each object category), are trained similar to a normal training routine. The last one, the class "unknown", is trained with a training set created by merging half of each object categories training images. Only half of the training set is used in order to reduce the training time considerably. To counterbalance the effect of a class with much larger training data than others, categories are balanced during training by using weights relative to the training set size in the optimization formulae of the SVM classifier.

As stated before, a variant of the Dempster-Shafer fusion, the transferable belief model [15], is adopted in our work. The transferable belief model accomplishes fusion via an un-normalized rule of combination. For our dual-classifier fusion case, it is shown in (2):

$$m(\Psi) = \sum_{\alpha \cap \beta = \Psi} m_A(\alpha) * m_B(\beta), \ \forall \Psi \subseteq \Omega \qquad (2)$$

where sets A and B are mass vectors of the main view and the secondary view, respectively and $m(\Psi)$ is the mass of category Ψ. α and β are each a category from the mass vector of the main view and the secondary view classifiers, respectively. Considering the above explanations, except for the class "unknown", a category α in the mass vector of the main view (A) has intersection with two βs in the mass vector B. The first one is the "unknown" category, and the second one is the same category as α in B. The same thing is true in the opposite way for a β.

To convert the mass values back to the probability domain, we utilize the pignistic transformation described in [14], as illustrated in (3):

$$P(\omega) = \sum_{\omega \in \Psi} \frac{m(\Psi)}{|\Psi|}, \ \forall \omega \in \Omega \text{ and } \forall \Psi \subseteq \Omega \qquad (3)$$

where Ψ is any non-empty element (focal element) of the set of object categories and $|\Psi|$ designates the number of object classes in the subset Ψ. Moreover, $P(\omega)$ is the probability of an object category ω, excluding the "unknown" class. Equation (3) expresses that any mass belief is distributed among its comprising class probabilities. After fusing mass beliefs to form probabilities of each object class, probabilities are normalized to sum to one.

As mentioned before, besides the actual classes of objects present in the training, the two classifiers provide a mass value for an "unknown" category. The "unknown" category can be considered as a way to indicate to the fusion module the opinion of the classifier about its uncertainty in detecting an object. Due to the fact that the sum of all masses in a mass vector is equal to 1, an increase in the mass of the unknown category causes the other categories in the same vector to receive lower mass values. From equation (2) we observe that an increase in the mass of the unknown category of a classifier not only decreases other masses of that classifier, but also through multiplying it weighs more toward the masses of the other classifier, with which it has a non-empty intersection. In the opposite way, when the unknown mass of a classifier is low it means a more resolute classifier which has a higher contribution to the final fused decision.

3. EXPERINMENTAL RESULTS

In this section we present the results obtained in nine real-world benchmarks. We implemented the proposed active vision system on a PR2 robot and tested it in different object and camera placements. All the tests were performed in table-top settings, i.e. objects were placed on a table in different positions in front of the robot. Figure 3. shows an example robot gesture in one of the tests. In addition, a sample visual output of the system is illustrated in Figure 4. We observe that there are two errors in classification in the main view (bottom left window): Tea Pot is incorrectly classified as Tea Can and Sugar is erroneously labelled as Tape Measure. Also, we see those two objects along with another one having red bounding boxes, which signals unreliable recognitions. All the three objects with unreliable classifications in the main view are then selected for reclassification in the secondary view (bottom right window). After the matching and fusion stages, the final detection results are all correct and reliable, as illustrated in the top left window of Figure 4.

The confusion matrices resulted from our tests with eight objects for the main view detections, as well as the actively fused detections, are shown in order in Figure 5 and Figure 6. In the confusion matrices, the Background column indicates target objects not detected at all, and the Background row counts any undefined entity being falsely

detected as a target object. The intersection of the Background row and column is also intuitively void.

Figure 3. An example robot gesture in the table-top benchmarks.

Figure 4. A sample visual output of the proposed active vision system. (Top left) The actively detected objects, (Bottom left) Main view, (Bottom right) Secondary view. Note: The red bounding boxes in the main view denote unreliable classifications.

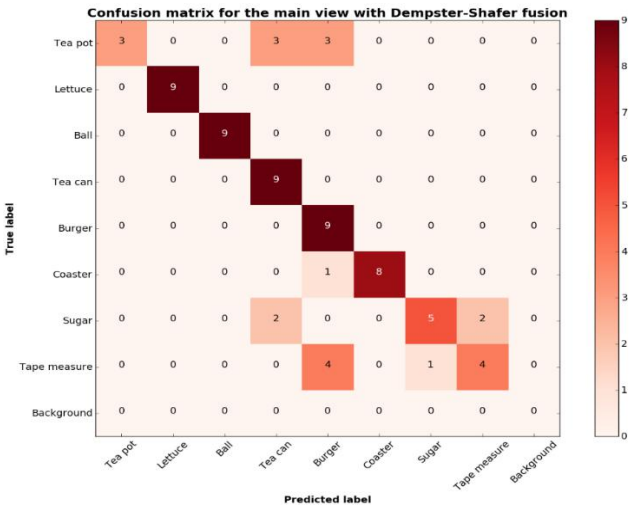

Figure 5. Confusion matrix of the detections made by the main view only classifier.

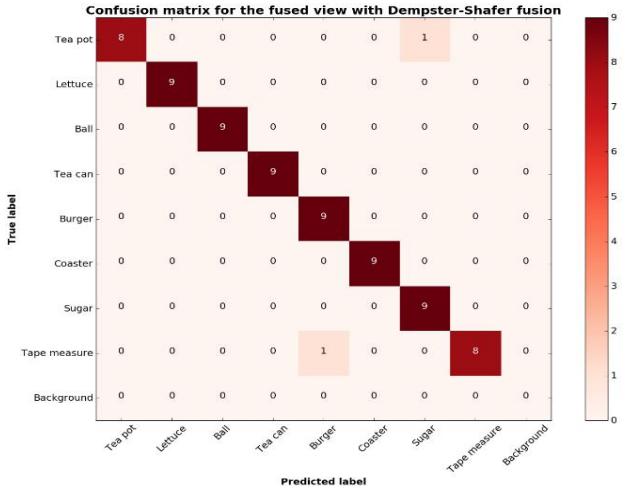

Figure 6. Confusion matrix of the detections made by the complete active vision system.

By using the confusion matrices, four performance metrics for object detection are calculated, namely, precision, recall, F1 score, and accuracy. Precision, defined as the number of true positives over the total number of positives, is a measure of how well the classifier differentiates true objects of interest from false positives. In contrast, recall (the ratio of true positives over ground truth positives) assesses the ability of the classifier in finding objects of interest. The F1 score balances the two metrics mentioned above by taking their harmonic mean. Accuracy, on the other hand, evaluates the ability of the classifier in correctly performing the classification task by means of dividing the total number of true detections to all the existing objects in the experiment. Table 1 shows the computed performance measures. Macro-averaging in Table 1 means the measure is calculated separately and is averaged over the results. In contrary, micro-averaging is the process of computing the measures for all the object categories collectively. We did not include the micro-averaging results in Table 1, because micro-averaging precision and recall are equal to accuracy in the case of multi-class classifiers.

Table 1 shows large improvements in all the four measures compared to the traditional single camera setup. The proposed active vision system achieved 19.5%, 19.8%, 14.1%, and 17.1% increase in accuracy, recall, precision, and F1 score, respectively. This enhancement in performance brings up the accuracy of the experiments to 97.2%. Other metrics are also over 97%.

Table 1. Performance measures of the proposed vision system.

Performance Measure	Main Camera	Actively Fused View
Macro-Averaging Precision	0.834	0.975
Macro-Averaging Recall	0.777	0.975
Accuracy	0.777	0.972
F_1 Score	0.804	0.975

We also performed another set of experiments in order to assess the applicability of the proposed system to dealing with object detection uncertainties, this time by intentionally adding partial occlusions. Figure 7. shows a sample situation with three objects being partially occluded by obstacles not defined as objects in our training set. Due to the significant occlusions in the main camera viewpoint, all three object detections are deemed unreliable. By dynamically employing the secondary camera, better views of the objects in the scene are

obtained, leading to correct fused classifications for all objects in the scene. Table 2 illustrates the computed metrics for this test benchmark, showing the advantage of the proposed method in dealing with partial occlusions from the viewpoint of the main camera. Precision, recall, F1 score, and accuracy are enhanced by 37.8%, 38.3%, 38.1%, and 36.1%, respectively. Since the secondary camera had a better view of the objects in the test, its recognitions were both more accurate and more confident, hence the fusion results frequently leaned toward the correct secondary camera recognitions.

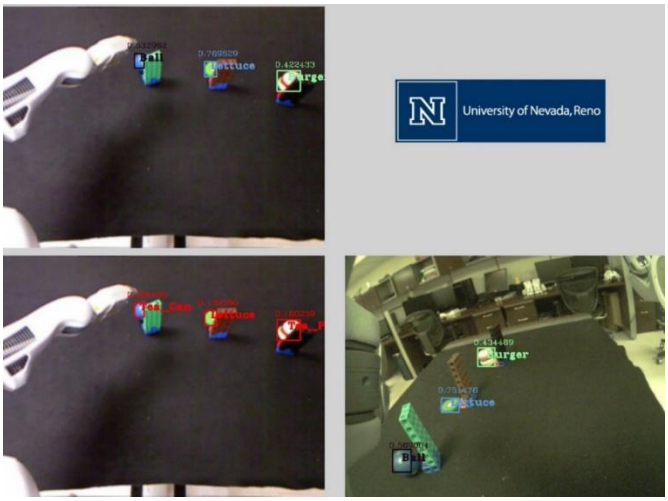

Figure 7. A sample situation with added partial occlusions. (Top left) The actively detected objects, (Bottom left) Main view, (Bottom right) Secondary view. Note: The red bounding boxes in the main view denote unreliable classifications.

Table 2. Performance measures of the proposed vision system in benchmarks with partial occlusions.

Performance Measure	Main Camera	Actively Fused View
Macro-Averaging Precision	0.555	0.933
Macro-Averaging Recall	0.533	0.916
Accuracy	0.555	0. 916
F_1 Score	0.543	0.924

4. CONCLUSION

In this paper we presented an active object detection system for robotic environments, implemented on a PR2 robot. The vision system uses an RGB-D (RGB-depth) camera on the robot head and an RGB camera mounted on the robot's hand. The contributions of the presented work are the design of an active visual sensor management approach in a robotic platform with dynamic confidence assessment, a fast distance-based object matching algorithm that utilizes the internal information available in the robotic system, and a novel variation of the Dempster-Shafer decision fusion to weight classification decisions.

Real-world experimental results indicate the high performance of the presented approach with average accuracy of 97.2% and F1 score of 97.5%, while providing considerable improvements over the static camera case.

We plan to extend our active vision approach to the domain of scene understanding with redundant or missing visual data, by providing new capabilities for automated, vision-based scene understanding, which allow both the fusion of redundant information and the selection of the most appropriate input from multiple sensors. This includes the dynamic selection of the best source of sensor data for a specific recognition task, reasoning with missing data due to partial/temporary occlusions, and the seamless integration of data from multiple sensors for tracking multiple targets in a potentially wide field of view.

ACKNOWLEDGMENTS

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

REFERENCES

[1] B. Scheuermann and B. Rosenhahn, "Feature quarrels: the Dempster-Shafer evidence theory for image segmentation using a variational framework" in Computer Vision – ACCV 2010. Lecture Notes in Computer Science, vol. 6493, R. Kimmel, R. Kletter, and A. Sugimoto (eds.), Springer, Berlin, Heidelberg, 2011.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Fort Collins, USA, June 1999.

[3] D. K. Sheet, O. Kaiwartya, A. H. Abdullah, Y. Cao, A. N. Hassan, and S. Kumar, "Location information verification using transferable belief model for geographic routing in vehicular ad hoc networks," IET Intelligent Transport Systems, vol. 11, no. 2, pp. 53-60, March 2017.

[4] D. Koks and S. Challa, "An introduction to Bayesian and Dempster-Shafer data fusion," DSTO Systems Sciences Laboratory, Edinburgh, Australia, August 2003.

[5] D. Ognibene and G. Baldassare, "Ecological active vision: four bioinspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot," IEEE Transactions on Autonomous Mental Development, vol. 7, no. 1, pp. 3-25, March 2015.

[6] D. Ognibene and Y. Demiris, "Towards active event recognition," International Joint Conference on Artificial Intelligence, pp. 2495-2501, Beijing, China, August 2013.

[7] E. Ramasso, D. Pellerin, and M. Rombaut, "Belief scheduling for recognition of human action sequence," International Conference on Information Fusion, Florence, Italy, July 2006.

[8] G. Powell, D. Marshall, P. Smets, B. Ristic, and S. Maskell, "Joint Tracking and Classification of Airbourne Objects using Particle Filters and the Continuous Transferable Belief Model," International Conference on Information Fusion, Florence, Italy, July 2006.

[9] J. B. Yang and D. L. Xu, "Evidential reasoning rule for evidence combination," Artificial Intelligence, vol. 205, pp. 1-29, December 2013.

[10] M. Rahimi and Y. Shen, "PSD microscopy: a new technique for adaptive local scanning of microscale objects," Robotics and Bioinformatics, vol. 4, no. 1, December 2017.

[11] N. Atanasov, B. Sankaran, J. Le Ny, G. J. Pappas, and K. Daniilidis, "Nonmyopic view planning for active object classification and pose estimation," IEEE Transactions on Robotics, vol. 30, no. 5, pp. 1078-1090, October 2014.

[12] N. Bellotto, B. Benfold, H. Harland, H. H. Nagel, N. Pirlo, I. Reid, E. Sommerlade, and C. Zhao, "Cognitive visual tracking and camera control," Computer Vision and Image Understanding, vol. 116, no. 3, pp. 457-471, March 2012.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, June 2005.

[14] P. Smets and R. Kennes, "The transferable belief model," Artificial Intelligence, vol. 66, no. 2, pp. 191–243, April 1994.

[15] P. Smets, “The combination of evidence in the transferable belief model,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pp. 447–458, May 1990.

[16] R. Szeliski, Computer Vision: Algorithms and Applications, 1st ed. London: Springer-Verlag, 2011.

[17] S. Chen, Y. Li, and N. M. Kwok, “Active vision in robotic systems: a survey of recent developments,” The International Journal of Robotics Research, vol. 30, no. 11, pp. 1343-1377, August 2011.

[18] X. Chen and Y. Jia, “Adaptive leader-follower formation control of non-holonomic mobile robots using active vision,” IET Control Theory and Applications, vol. 9, no. 8, pp. 1302-1311, May 2015.