A one-shot next best view system for active object recognition

Pourya Hoseini[1] · **Shuvo Kumar Paul**[1] · **Mircea Nicolescu**[1] · **Monica Nicolescu**[1]

Accepted: 29 June 2021

Abstract

Active vision is the ability of intelligent agents to dynamically gather more information about their surroundings by physical motion of the camera. In the case of object recognition, active vision enables improved performance by incorporating classification decisions from new viewpoints when there is some degree of uncertainty in the current recognition result. A natural question in an autonomous active vision system is, nonetheless, how to determine the new viewpoint, i.e. in what pose should the camera be moved? This is the traditional question of next best view in active perception systems. Current approaches to the next best view problem either need construction of occupancy grids or require training datasets of 3D objects or multiple captures of the same object in specified poses. Occupancy grid methods are usually dependent on multiple camera movements to perform well, which make them more useful for 3D reconstruction applications than object recognition. In this paper, a next best view method for active object recognition based on object appearance and surface direction is proposed that decides on the next cameras pose without requiring any specifically structured training datasets of 3D objects. It is also designed for single-shot deductions of next viewpoint and is able to determine next best views without the need for substantial knowledge of 3D voxels in the environment around the camera. The experimental results illustrate the efficiency of the proposed method, while showing large improvements in accuracy and F_1 score.

Keywords Object recognition · Active vision · Next best view · View planning · Robotics

1 Introduction

Autonomous mobile intelligent systems, such as robots or unmanned aerial vehicles (UAVs), rely heavily on sensing the surrounding environment to execute their missions. One of the major requirements of perception mechanisms is their ability in gathering useful information to accomplish their tasks. For a vision system, however, sometimes it is not possible to capture an input that is good enough for further processing. There can be several reasons for such a situation. Among them, presence of occlusion in the line of sight of the camera, insufficient resolution of the object in the captured image, and lack of discriminative features in the current viewpoint of an object being observed can be mentioned. Active vision is a solution to such issues by dynamically incorporating new information sources in the hope of improving the performance of the vision system. For more information about the concept of active vision, refer to [3]. Two major application areas of active vision are three-dimensional (3D) object reconstruction and object recognition. An active object recognition routine, which is the focus of our work, typically is comprised of an uncertainty evaluation procedure, an information fusion algorithm, and a mechanism to achieve physical camera movement.

An essential question regarding the camera movement is how to determine the most appropriate pose - the problem of finding the Next Best View (NBV) for the camera. Finding the next best view is not a trivial task. Since in many situations the observer has no prior knowledge of what the shape and appearance of an object from other camera poses would be, the current input data are not sufficient for

✉ Pourya Hoseini
hoseini@nevada.unr.edu

Shuvo Kumar Paul
shuvo.k.paul@nevada.unr.edu

Mircea Nicolescu
mircea@cse.unr.edu

Monica Nicolescu
monica@cse.unr.edu

[1] Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA

deterministically deciding for the next best camera pose. Hence, NBV is an ill-posed problem by definition.

Usually, the specific task of a vision system directly impacts the way NBV is acquired. If the goal is 3D reconstruction of objects, it is a good idea to seek for next best views that reveal the most unexplored areas of the objects. In addition, the typical goal in 3D reconstruction applications is not to plan for a single new viewpoint, but to plan for a chain of NBVs to fully observe the object's volume. In contrary, in the case of object detection and recognition applications, the least number of alternative viewpoints that provide a fresh number of discriminative features is desired to improve the recognition performance, while keeping the energy and time costs of moving the cameras as low as possible. In this work, we propose and evaluate a method to achieve the next best view, suitable for active object recognition tasks with just a single NBV frame capture necessary for its operation.

It is evident that the earlier work [1, 2, 4–6, 8, 9, 17–21, 23, 24, 27–29], which will be discussed in Section 2, in determining next best view is concentrated on two camps: information gain-based and object estimation-based techniques. Measuring occupancy of 3D space through ray tracing and information gain is inherently useful for 3D reconstruction applications, because it strives for exploring more surface voxels than discriminative features for classification. That is why it has been preferred almost constantly in previous work for 3D reconstruction, rather than in object recognition applications. On the other hand, object shape and appearance estimation methods rely on either comparing the current object shape and/or appearance to the ones in the training time, or on hallucinating the 3D shape of the current object. Afterwards, those methods try to infer the best course of action by comparing different suppositional viewpoints. The problem here is in the inaccuracies that can arise from presumptive object shapes/appearances as well as in the necessity of large datasets of object images taken from various points of view.

In our work, we propose a next best view method for object recognition. It is aimed at a single camera relocation to improve the object recognition rate based on cues from the visible object shape and appearance only. It does not require a series of camera motions toward or around the object. Our NBV method is also neither dependent on a dataset of specifically designated images from around the object, nor on 3D object volumes for training. It uses conventional datasets, a collection of random images of objects, merely for the training of the classifiers. To accommodate the aforementioned specifications in our work, an ensemble of appearance and shape criteria is used to evaluate different areas of the object viewed, in order to suggest a new camera pose. Examples of such criteria are uncertainty measurements in classification of a region of the object image, parallelism of object surface to image plane, and various statistical texture metrics. In order to test the proposed method in an organized way, a test dataset was also created. In the tests, the proposed method proves to be effective in predicting the next best camera view among a set of pre-selected test-time poses around the object.

This work represents a novel approach for a NBV module which is part of an active object recognition (AOR) system discussed in [15] and [16]. The significant contributions of the proposed work extend along several directions:

1. A novel next best system is proposed specifically for the task of object recognition.
2. The proposed NBV depends only on the current object shape and appearance, therefore no prior knowledge of objects is necessary.
3. There is no dedicated training stage for the NBV itself, thus eliminating the need to create specially designed datasets for next best view determination. The only training happening is for the object classifiers.
4. A small test dataset, containing the snapshots captured around various objects, has been created to effectively test the proposed NBV system. It can be used by other researchers as a benchmark.
5. Experimental validation shows good results in terms of the performance improvement after fusion of views, by comparing the fusion results among a pre-defined set of possible camera poses.

The remainder of the paper is organized as follows. Earlier work in the literature around the field of next best view is reviewed in Section 2. Section 3 contains an overview of the steps in an active vision system for object recognition. The single-shot next best view system is presented in Section 4. Section 5 demonstrates the results obtained in the benchmarks and analyzes them. Finally, concluding remarks are discussed in Section 6.

2 Previous work

By reviewing the literature, a few different directions can be distinguished. In [27], a deep belief network is employed to hallucinate the complete 3D shape of objects in the presence of occlusion. For any hallucinated 3D shape, the uncertainty of recognition is computed in different predefined camera poses via the conditional entropy of output classifier probabilities. The viewpoint that gives the least uncertainty is chosen as the winner for the next-best-view system. The idea of examining various hypothetical viewpoints of an object in [27] seems interesting for an NBV system, but the proposed approach depends heavily on the imagination of the object shape in the unobserved occluded areas. Therefore, its performance is prone to the

mistakes stemming from the 3D hallucination part of the algorithm. Another deep learning-based method is proposed in [29], which takes raw point cloud data and current view selection states as input to subsequently predict the information gain of all candidate views.

Relocating cameras based on a bio-inspired approach is discussed in [4], where authors analyzed the head movement of barn owls and adopted it to actuate a depth camera installed on a robot for 3D reconstruction of objects. The idea of adopting the movement of biological systems, related to their active perception attempts, looks promising, but it should be noted that the approach of replicating motions in [4] is insensitive to the object shape and appearance, thus it probably does not target the best next view for many objects. An active object detection and pose estimation method with dynamic camera location planning is presented in [2]. An Asus Xtion RGB-D camera mounted on the PR2 robot's wrist was used as the sensor. This method tries to balance the amount of energy needed to move the camera and the added chance of getting a better object detection. It plans a sequence of camera movements, whereas every movement is a step in the fastest route to get close to the object. Consequently, the method in [2], although multi-capture, does not have any intelligent NBV component.

A next best view method is proposed in [23] for 3D reconstruction applications on the basis of predicting information gain from prospective viewpoints. In order to predict the information gain in unobserved areas, an occupancy grid is constructed out of all the observations so far and a Hidden Markov Model (HMM) is employed to estimate the observation probability of unobserved cells in the grid. An information gain mechanism to estimate preferability of any potential viewpoint is reported in [8]. To determine the information gain, instead of rendering hypothetical object appearances, the next best view system directly estimates the classification probabilities. This approach overcomes the problem of computationally expensive renderings of hypothetical 3D objects. However, it necessitates 3D training data for every object and doing classification and confidence estimation for every viewpoint of the 3D objects in the training. This requirement seriously affects the applicability of the method due to the shortage of such training data for many real-world objects.

In [24], the NBV algorithm simply chooses the viewpoint with most unknown voxels as the best one to explore for 3D scanning. A path planning algorithm is used in [5] to construct a path tree to completely explore the area around an aerial vehicle. The nodes in the tree are poses in the free space. In each step, only the best node under the root of the tree is chosen for the movement. After any move, a new tree is constructed. The preference in selecting a node is based on the number of unobserved 3D volume that can be observed in the corresponding camera pose. An eye-in-hand vision system is proposed in [19] that uses multiple simultaneously-captured views, scene segmentation, and an objective function applied to each perspective to estimate a gradient, representing the direction of the next best view. Relevantly, a multi-sensor NBV method is presented in [6], which was tested for both 3D reconstruction and weld seam inspection.

In [17], a boosting technique to combine three criteria for determining the NBV is used for actively selecting a viewpoint around an object. One of the three techniques is a similarity check of a detected object with prerecorded object appearances in different views. The viewing angle with the least similarity to the current detection is then selected as the NBV. The other two criteria for choosing NBV are the prior probability of a viewpoint in successfully determining the object class given either a currently detected object pose or a currently detected object category. Aside from the priors, which are application data specific, using a similarity measure between the current viewpoint of a presumed object and its other viewpoints is appealing. However, it requires a dataset comprised of images around the training objects with their known pose. This can be cumbersome to use in general applications as there is a need to capture appearances and poses all around the objects that we want to detect at test time.

Next best view is incorporated in the work of [9] for calibrating and operating a multi-camera 3D hyperspectral scanner. Recently, NBV is used in [28] for sketch shape retrieval to select the candidate projection of 3D shapes to extract their features and compare them to a sketch. The problem of choosing a set of next best viewpoints for constructing 3D models of objects using depth images, captured by a team of multiple robots, is tackled in [18], where a utility function that scores sets of viewpoints and avoids overlap between multiple sensors is employed. In another work [20], NBV helps in improving robotic grasp detection by providing informative viewpoints in cluttered scenes. An interesting application of NBV is also proposed in [21], in which an autonomous underwater vehicle (AUV) is programmed to choose its next viewpoint optimally to map or inspect complex underwater structures. Similarly, in [1] an unmanned aerial vehicle (UAV) equipped with NBV was reported for 3D reconstruction of large structures. The utility function in [1], considers four criterion categories: information theory, model density, traveled distance, and predictive measures based on symmetries in the structure.

3 Active vision for object recognition

Active Object Recognition has many uses in robotics, vision-based surveillance, and many other applications.

Figure 1 shows the flowchart of a typical AOR system. After several preprocessing operations, such as denoising, the process starts with an object recognition stage using the current point of view. Any object recognition or object detection method can be used at this stage, depending on the nature and limitations of the application. Some useful examples are [7, 10, 11, 13, 22, 26, 30]. Subsequently, the classification result of the initial recognition round is evaluated to determine the confidence of the recognition. In the event of a low confidence value (uncertain recognition), the active vision mechanism is triggered. First, it plans the new camera pose based on the principles of a next best view method, which is the subject of the work proposed here. A camera is then moved to the specified position and orientation. The moving camera can be the same camera that captured the initial view, or it can be a secondary camera employed to achieve a concurrent capture mechanism.

With a camera in the pose determined by the NBV system, the object recognition is performed again, this time by using the new camera view. Thereafter, if there are multiple detected objects in each camera view, the objects in the two camera views are matched to form pairs of object classifications for a later decision fusion step. Depending on the application and the availability of frame transformations, the matching procedure can be done in a purely vision-based style, via pixel/keypoint/object correspondence, or through a mechanical submodule in a sensor-equipped robotic system that provides the 3D geometric transformation between the camera poses.

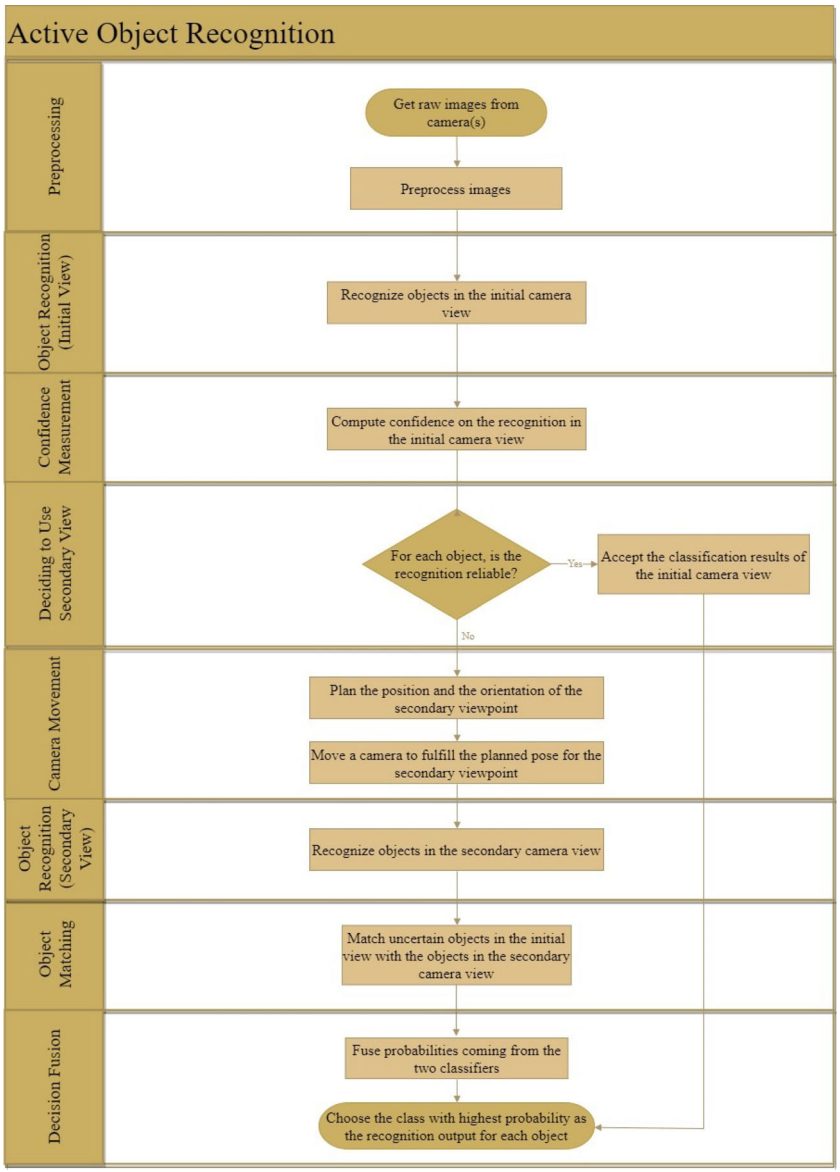

Fig. 1 Active object recognition steps

Each associated classified object pair is passed into a decision fusion module, which combines the classification results to obtain the final probability vector of the object categories. The fusion module should take the class probabilities of the two classifiers and fuse them to yield the output class probabilities. For more information about the described active object recognition method, refer to our prior work [15, 16].

4 The proposed next best view system

The aim of the proposed NBV system is to find a candidate viewpoint in a single try after the initial capture. For this reason, the only assumption is the availability of the color and depth information of the object being seen in the initial camera view. For rigorous testing purposes, the NBV poses are also limited to a number of pre-specified positions and orientations that are usually reachable for eye-in-hand or UAV platforms. There are eight groups of poses around the object, on the plane that passes through the object and is parallel to the image plane of the camera at the initial viewpoint. Each group is the set of poses that are generally viewing the same part of the object. For example, a NBV can be one of the poses that are looking at the top left of an object, which means a camera on the aforementioned plane and in the top left of the object is looking at it. It should be noted that the number of poses around an object can be extended if needed, depending on the application.

The viewpoints are chosen to be at the same depth as the object in the camera coordinate of the initial view, because while they are reasonably accessible for many eye-in-hand configurations, they can provide substantially new information from a view direction perpendicular to the initial one. Any pose from a depth less than the object's depth will probably have an overlapping view with the frontal initial view (they will see common parts of the object). In contrast, any pose with a depth farther than the object will see behind the object, which can be desirable, but it has two disadvantages. First, it is hard to reach by a robotic system. Assuming an object is in front of a robot, many robotic arms do not have degree of freedom required to move an arm-mounted camera to a pose facing back of the object and, thereby, facing the robot itself too, at a large distance from the robot. For an UAV, or any other freely moving unit with a camera, it is also difficult to plan for a pose behind the object in a single shot as there is no information accessible of the object's thickness (depth in the initial camera coordinate). A second reason is that for a single NBV based on the current frontal view of an object, it is difficult to derive any pose that is located behind the object, because the initial view has no indication of what is behind the object. Therefore, as we do not know if the self-occluded area behind the object is really valuable for active object recognition, it will not be considered as a candidate for a NBV.

4.1 Tiling the initial view

In the proposed method, the object bounding box, coming from any object detection system, is divided into different regions. The tiled regions are disjoint, and they cover the entire area of the bounding box. Figure 2 demonstrates the tiling of regions in the proposed method. In the current implementation, each bounding box is divided into nine regions. The peripheral tiles represent one of the pose groups of a camera in the corresponding areas around the object. For instance, the top left region represents a new point of view when the camera is viewing the object from the object's top left with the same depth to the camera as the object itself in the camera coordinate of the initial view. The distance of the camera in the new pose to the object can be set arbitrarily close to the object considering the pose feasibility for the camera setup and the image resolution of the camera. Figure 3 illustrates this example situation.

The rationale behind this tiling scheme is that by analyzing each region of the current view, there can be clues to find a more informative NBV corresponding to the side of the object it is representing and hence suggesting where to move the camera next. Compared to methods that simply attempt to look at unobserved voxels [5, 8, 23, 24], the proposed approach tries to further qualify that decision by choosing perpendicular views, with an informed move on the basis of what is currently being seen. In addition, the proposed method contrasts with approaches that hypothesize the object shape [8, 27] due to the fact that we do not require such a step in our work. The proposed

Fig. 2 Tiling routine in the proposed next best view system

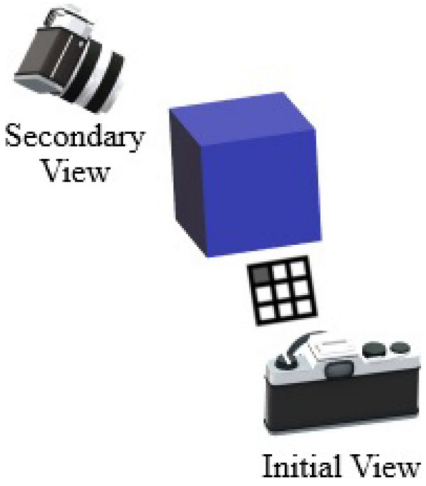

Fig. 3 An example next viewpoint selection situation, where the top left tile is selected and consequently the secondary viewpoint is looking at the object from its top left

approach only utilizes limited cues directly available in the initial view, instead of requiring the inference of explicit information about the entire shape, appearance, and relative pose of the object.

4.2 Tile voting system

The proposed method implements a poll among four different criteria to select the peripheral tile with the highest votes. The four criteria cast a single vote, only for the tile they score the most. The ensemble method is a simple voting procedure with equally weighted votes for the four criteria in the ensemble. Two of the criteria are based on statistical metrics that analyze the texture of a tile. Another one evaluates the angle between the object surface normal and the normal to the image plane in the initial view to estimate how visible the surface was in the initial view. The last criterion takes into consideration the classification dissimilarity of a tile relative to the classification of the whole object. In the following three subsections, we explain the four voting metrics in more detail.

4.3 Statistical texture criteria

One of the circumstances where active vision is particularly useful is when the object being seen is not clearly recognizable. It can be due to occlusion, lighting conditions, object shape, etc. One approach to tackle these situations can be to change the view toward poses that bring a better perspective. Analyzing the object from a side that is likely to be well-lit and provide better quality images is therefore desirable. To this end, the second and third moment texture analysis approaches are used. The measures are chosen to be extracted from the intensity histograms of each image patch to accommodate faster processing speeds.

4.3.1 Second moment (variance) of histogram

The second moment or variance of intensity histogram is a measure of contrast of an image [12]. A uniform surface is not the best area in an image to find features for object recognition. The higher the contrast, the more feature-rich an image can be in many cases. The variance of an intensity histogram is defined in (1) [12].

$$\sigma^2(z) = \mu_2(z) = \sum_{i=0}^{L-1}(z_i - m)^2 p(z_i) \tag{1}$$

In the equation, $\sigma^2(z)$ is the variance of intensity levels (z) and is equal to the second moment, $\mu_2(z)$. Additionally, L is the number of bins in the histogram, i is the index of the current histogram bin, $p(z_i)$ is the probability of occurrence of a bin, and m is the mean of intensities, calculated as follows:

$$m = \sum_{i=0}^{L-1} z_i p(z_i) \tag{2}$$

In order to scale the metric to the range of [0, 1), the contrast score $V(z)$ is calculated via (3).

$$V(z) = 1 - \frac{1}{1 + \sigma^2(z)} \tag{3}$$

With a larger $V(z)$ value, thus higher contrast and possibly more features, a tile can be a cue to a sideways surface with plenty of features for a good next viewpoint. Hence, the scaled second moment, $V(z)$, is preferred to be high.

4.3.2 Third moment of histogram

The third moment of histogram is a measure of its skewness [12], thus it indicates if an image histogram is inclined towards dark or bright levels. It is calculated similar to the second moment in the earlier subsection:

$$\mu_3(z) = \sum_{i=0}^{L-1}(z_i - m)^3 p(z_i) \tag{4}$$

If $\mu_3(z)$ is negative the histogram is skewed toward darker intensities, and if it is positive the histogram is inclined toward brighter ones. Hence, a third moment that is close to zero signifies a balanced histogram, which in turn means the image is neither dark nor bright. This intuition in many cases translates to good lighting when the raw unprocessed image is being considered.

Consequently, if the third moment of a tile is close to zero, the respective side probably would provide a well-lit

perspective of the object. To convert the third moment to a score usable for the proposed NBV, the following formula is used.

$$L(z) = \frac{1}{1 + |\mu_3(z)|} \tag{5}$$

The equation in (5) transforms large negative (very dark) and positive (very bright) values of the third moment to small scores and balanced lightings to larger scores. Therefore, higher $L(z)$ values are preferable.

4.4 Surface parallelism score

Because the image plane of a camera is flat, any object surface, parallel to the image plane on average, is approximately a flat plane facing the camera. Considering that we examine all criteria on the periphery tiles of an object's bounding box, an object surface parallel to the camera probably means that the peripheral tile should be easily visible to the sensor. However, in that case, there can be other faces of the object that are not being seen by the camera completely, since the peripheral surface is parallel to the camera and has no angle to the camera's image plane. On the other hand, a peripheral surface with a perspective to the current view, is likely not clearly visible in the current view as its surface is tilted and exhibits foreshortening too. Based on this idea, the surface parallelism score takes into account how much the object surface being seen in a tile is parallel to a 3D camera observing the object. Assuming depth map of a tile is segmented, and the object surface constitute the foreground pixels, the parallelism score is defined in the following:

$$P = -\frac{\sum_{p \in F} N\left(\overrightarrow{\left(\frac{dz}{dx_p}, \frac{dz}{dy_p}, 1\right)}\right) . \overrightarrow{z}}{|F|} \tag{6}$$

where P is the parallelism score, p is a pixel in the current image patch being processed (current tile), F is the set of foreground pixels, $|F|$ means the number of foreground pixels in the tile, $N()$ is a vector normalization function, and $\overrightarrow{z}$ is the z (depth) axis in the camera coordinate of the initial view. The derivatives of depth (z) with respect to x and y axes for a certain pixel (x_p, y_p) in the pixel coordinate of the initial view are calculated in the following way:

$$\left.\frac{dz}{dx}\right|_{x=x_p} z(x_p + 1, y_p) - z(x_p - 1, y_p) \tag{7}$$

$$\left.\frac{dz}{dy}\right|_{y=y_p} z(x_p, y_p + 1) - z(x_p, y_p - 1) \tag{8}$$

In (7) and (8), the $z(.,.)$ is the depth at a pixel location in the depth map. To compute the score, the camera capturing the initial view should provide 3D depth data to make it possible to obtain a depth map. In many ordinary 3D cameras, there are multiple small spots of unknown values spread over the depth map, for which the camera is not able to compute the depth. In order to handle those areas, any unknown values are replaced with the maximum depth in the depth map being considered. Since we are assuming that the background in an object's image corresponds to points with depth larger than the depth of the object, this substitution implies that any unknown value spots in the depth map are converted to background pixels that have no effect in the computations of the surface parallelism score.

In addition, the actual objects do not completely fill their bounding boxes – they usually contain areas showing other unintended entities, i.e. background. To exclude the background from computations, an input depth map firstly passes through a segmentation step. In our work, we opted for Otsu's segmentation [12], but any well-performing binary segmentation method, such as [25] can be used. The foreground areas (F) are then assumed to be representing the object and are used in (6).

In the next step, the normalized surface normal is calculated for every foreground pixel in the depth map, as shown by the term $N\left(\overrightarrow{\left(\frac{dz}{dx_p}, \frac{dz}{dy_p}, 1\right)}\right)$ in (6). The inner product of the surface normal with the z axis of the camera coordinate measures how parallel the two are. Since the z axis is effectively the surface normal to the image plane of the camera in the initial view, the inner product also measures how parallel the camera image plane and the object surface are, in effect measuring the foreshortening of the object. Lastly, the results of the inner products are averaged over all the foreground pixels. The averaging operation gives an overall insight of how the surface is parallel to the camera on average. The proposed parallelism score favors a tile when its score is higher.

4.5 Tile classification dissimilarity

If after the classification of a tile's image, the output class probability vector is in disagreement to the class probability vector of the whole initial view of the object, it is probably one interesting region to view next. Such differing classification results potentially reveal where in the object image (i.e. which tile) is contributing more to the ambiguity of the initial recognition. Therefore, to resolve the uncertainty in classification, it is a promising approach to choose that direction in order to achieve a new view. Moreover, if a tile of the initial view is confirming the initial classification output, the prospect of finding some new information is less compared to an opposing classification. Accordingly, measurement of dissimilarity of class probabilities in a tile to the whole object image provides another opportunity to find the NBV.

The score computed by this measure is the sum of absolute differences (SAD) of class probabilities of the object image and each tile, as shown in (9),

$$S_j = \sum_{i \in G} \left| p_{tj}^{cj}(i) - p_O^c(i) \right| \tag{9}$$

where S_j is the score of the dissimilarity measure between the tile j and the complete object image, G is the set of object classes, i is an object class, and $p_{tj}^{cj}(i)$ and $p_O^c(i)$ are probabilities of class i after classifying the tile j and the whole object image by the classifiers trained for tile j (c_j) and the whole object (c), respectively. The classifier c is a conventional one, trained with the color images of objects. In contrast, the classifier c_j is trained by the patches cropped from the original training data in the areas specified for tile j. The existence of separate classifiers for every tile makes the tile classification more accurate compared to the case of only using a single classifier for all the tiles. It is also noteworthy that tile-specific classifiers do not impose new data requirements for the training phase - they just use portions of the same traditional dataset.

5 Experimental validation

Works on next best view, and in general active vision methods, face the challenge of lacking a common test benchmark. Due to the robotic nature of these systems, many of the experimental verifications are performed in situation-specific test environments with different objects, available viewpoint choices, lighting conditions, backgrounds, etc. To answer this issue and to set a standard way of testing NBV methods for object recognition, we gathered a dataset, specifically for benchmarking active object recognition techniques, with which the proposed next best view method was tested. In the tests, the initial views of objects were intentionally distorted in order to bring about the conditions where an active object recognition system would be triggered. The tests were performed with different classifiers and fusion methods as a part of the complete vision system.

5.1 The active object recognition dataset

We gathered 240 test situations, generally for evaluating active recognition systems. There are 10 objects in the dataset, each one being shown in 24 situations. Figure 4 shows the 10 objects in the dataset. The objects in each of their 24 test situations were placed in various poses (4 random faces of the object), lighting conditions (2 modes: darker and brighter), and background textures (3 modes: dark tabletop, light carpet, and colorful rug). The dataset is especially useful for testing next best view systems. In each

Fig. 4 The 10 objects in the dataset

situation, there are seven images and their corresponding depth maps: one for a frontal initial view, another for an initial view with a slightly higher altitude initial view, and five others for the images/depth maps taken from the sides of objects as follows: from left, top left, top, top right, and right. Figure 5 demonstrates a sample situation for one of the objects in the dataset. Because the objects were placed on the ground or an opaque hard surface during the photoshoot it was not possible to take images from the lower views. It is, nevertheless, not a significant limitation as in many real-world conditions, objects are placed on opaque surfaces. Furthermore, the existence of five choices for each of the two frontal views offers enough range of options in

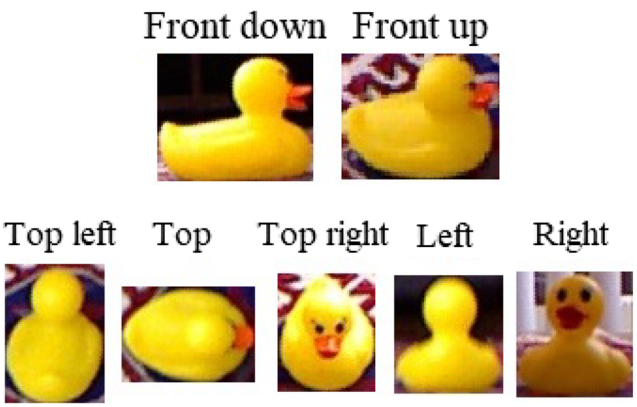

Fig. 5 A sample situation in the dataset

the test. The dataset is published along with the current paper[1].

5.2 Initial view deterioration

The initial views in the test dataset are clear and unobstructed. However, active object recognition systems are usually employed whenever the classifiers experience degraded performance due to occlusion or unfavorable perspective of objects. The initial views are, therefore, distorted to simulate the conditions to trigger AOR. In the following, these alterations are described:

1. A corner of the image is superimposed by a patch of another randomly selected object image. The depth information of the superimposed object part is also replaced in the respective location of the depth map. In the tests, we chose corner patches of size 60% of the length and width, totaling 36% of the area of the original image.
2. A half of the image is whited or blacked out. A top and a bottom whiteout plus a left and a right blackout generate four new alterations of the original image.
3. Gaussian blurring in two levels: one with a 5×5 kernel and the other with a 9×9 kernel.
4. Added noise with standard deviations of 20 and 30 in the 8-bit color images.
5. Image darkening and brightening by 150 levels.

The tests were performed on both the altered images and their corresponding depth maps as well as the original ones. This amounts for 15 test scenarios for any test situation in the dataset, which are shown in Fig. 6 for an object in a sample situation.

5.3 Test setup

Since there are two initial images in each test situation in the dataset, two trials can be performed for a single situation. As mentioned in the former section, for each initial image 15 test scenarios are possible. Hence, 30 trials are conducted for any test situation. With the availability of 240 test situations, 7200 situations were evaluated for any vision system in the tests.

To make sure that the proposed NBV is independent of the classifier and the fusion algorithms in the AOR system, five different classifiers and three fusion techniques were examined to take their average results. No matter which classifier or fusion technique is selected, each of them may be used in the context of the proposed next best view system through the flow described in Fig. 1. Averaging,

[1]Dataset available at https://github.com/pouryahoseini/Next-Best-View-Dataset.

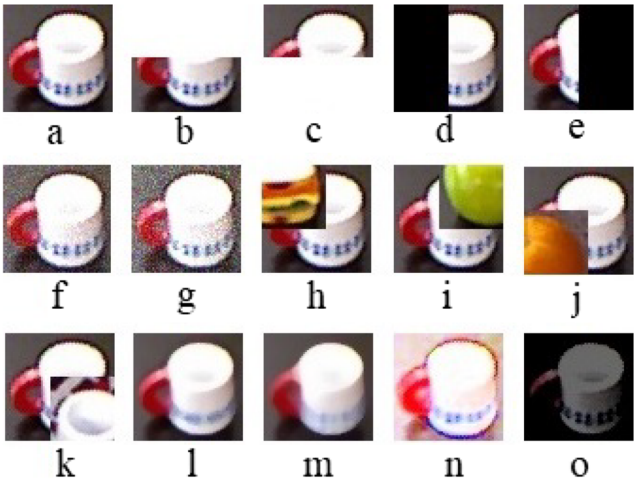

Fig. 6 Initial view distortions. a) Original image, b) Top whiteout, c) Bottom whiteout, d) Left blackout, e) Right blackout, f) Lighter noise, g) Heavier noise, h) Top left superimpose, i) Top right superimpose, j) Bottom left superimpose, k) Bottom right superimpose, l) Lighter blur, m) Heavier blur, n) Bright, o) Dark

Naïve Bayes [15], and Dempster-Shafer (DS) [16] fusion algorithms are used in the tests. The classifiers are:

- CNN 1: A convolutional neural network (CNN). By naming the convolution, dropout, fully connected, and pooling layers as C, D, F, and P respectively, the network structure is written as $(C \rightarrow D \rightarrow C \rightarrow P \rightarrow D \rightarrow C \rightarrow D \rightarrow C \rightarrow P \rightarrow D \rightarrow F \rightarrow D \rightarrow F \rightarrow F)$. All the activation functions, except the last layer, are Rectified Linear Unit (ReLU). The activation function of the last layer is Softmax. The pooling layers take the maximum of the inputs (max pooling). The dropout rate is set to 0.1. All the layers, with the exception of the last one, have an ensuing L2 activity regularization function. The learning rate is 0.01 in the beginning of training and is reduced over the epochs. The number of epochs is 200, while batch size is 50. The loss function is categorical cross-entropy, which is optimized by the Adam optimizer [18].
- CNN 2: A similar neural network to CNN 1, but with average pooling in place of max pooling and hyperbolic tangent instead of ReLU activation functions.
- CNN 3: This network is also close to CNN 1. The only difference is that the network does not have the last convolutional layer and its following dropout and pooling layers.
- A one-versus-rest non-linear Support Vector Machine (SVM) classifier with the feature vector comprised of Hu moments of the three RGB (red-green-blue) planes, besides the reduced Histogram of Oriented Gradients (HOG) of the gray level image of the input. The SVM kernel was selected to the Radial Basis Function (RBF), while the feature reduction for the HOG component

of the feature vector is Principal Component Analysis (PCA) with a reduced feature number of 60. The regularization parameter and the kernel coefficient are determined through a five-fold cross-validation grid search.

– A random forest with 150 decision trees and a split criterion of the Gini impurity that uses a bag of 150 visual words of Speeded Up Robust Features (SURF) keypoint descriptors. The bag of words uses L2 distance and k-means algorithm in its clustering procedure. A five-fold cross-validation grid search is also utilized to decide the max depth of a tree, minimum samples for a split to happen, and minimum samples in a leaf node.

Considering the possible combinations of the classification and fusion approaches, 15 benchmarks were evaluated, each with 7200 situations tested. In the tests, the confidence threshold of the AOR system was set to 20, except for those tests needing a sweep of the threshold value. That is, the initiation of the active new viewpoint is triggered if the maximum probability in the initial classification probability vector is less than 20 times of the second highest probability.

5.4 Test results

The proposed NBV is evaluated by comparing its suggested selections with the other viewing poses in each test case. The voting system is also compared to a few measures to probe the efficacy of the ensemble method. Further, the energy-efficiency of the proposed method and changes to the Receiver Operating Characteristic (ROC) curves are evaluated in the following. The dedicated classification dissimilarity is the one used in the ensemble, which incorporates dedicated classifiers for any specific tile, although we reported the same measure with using a single classifier for all the classifications. There are also other methods reported for comparison purposes. They are histogram uniformity and entropy [12] as defined below.

$$U(z) = \sum_{i=0}^{L-1} p^2(z_i) \tag{10}$$

$$e(z) = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i) \tag{11}$$

In the last two equations, z is a random variable denoting intensity and $p(z_i)$ is the corresponding histogram with L bins.

5.4.1 Tile-ranked improvements

In every test situation, the five prospective next viewpoints are examined for the scores they get from every criterion. In Fig. 7, the tiles are sorted on the horizontal axis in an

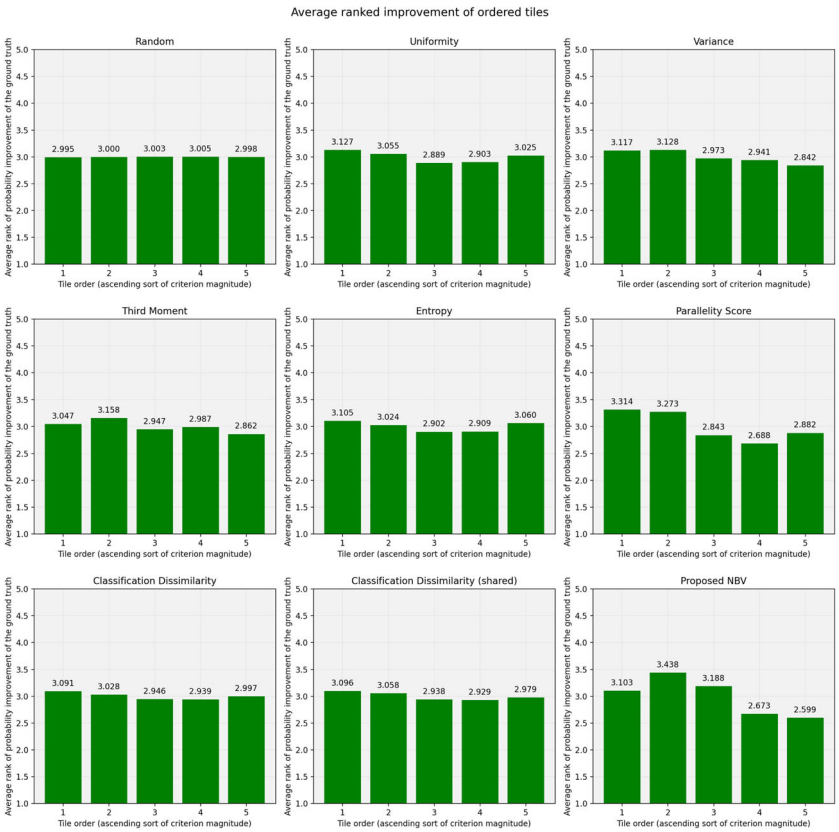

Fig. 7 Average ranked improvement of tiles in ascending order of scores

ascending order of the scores of each designated criterion. The height of the bars for any tile shows the average rank of the tile in attaining better probability for the ground truth classes after the decision fusion stage. The lower the rank and the closer it is to 1, the better it is. Therefore, lower height of the bars in the right sides of the plots in Fig. 7 is desirable. For example, for the proposed method, the mean rank of the third highest scoring tiles (represented by the middle bar) is 3.188 and the average rank of the highest scoring tiles (the rightmost bar) is 2.599, which is comparatively lower and better.

From the results, it is obvious that the proposed NBV method achieves better ranks for the tiles it scores higher. It means that it is able to find the viewpoints that improve the probability of the true class in the AOR system's output. In addition, Fig. 7 demonstrates the performance of the individual metrics, some of which are part of the ensemble. It can be observed that all the individual proposed metrics in the ensemble generally find the better tiles with their scoring, though not as well as the combination of them. The member criteria of the ensemble tend to bring the height of their very right bar down. This decline is sharper for the ensemble itself.

5.4.2 Performance per NBV measure

The obtained accuracy, precision, recall, and F_1 score of the proposed system is compared to its constituting measures and the other three measures in Fig. 8. It is evident that the proposed method is capable of suggesting next views to enhance the performance of the object recognition system. Figure 9 illustrates absolute improvements achieved by the

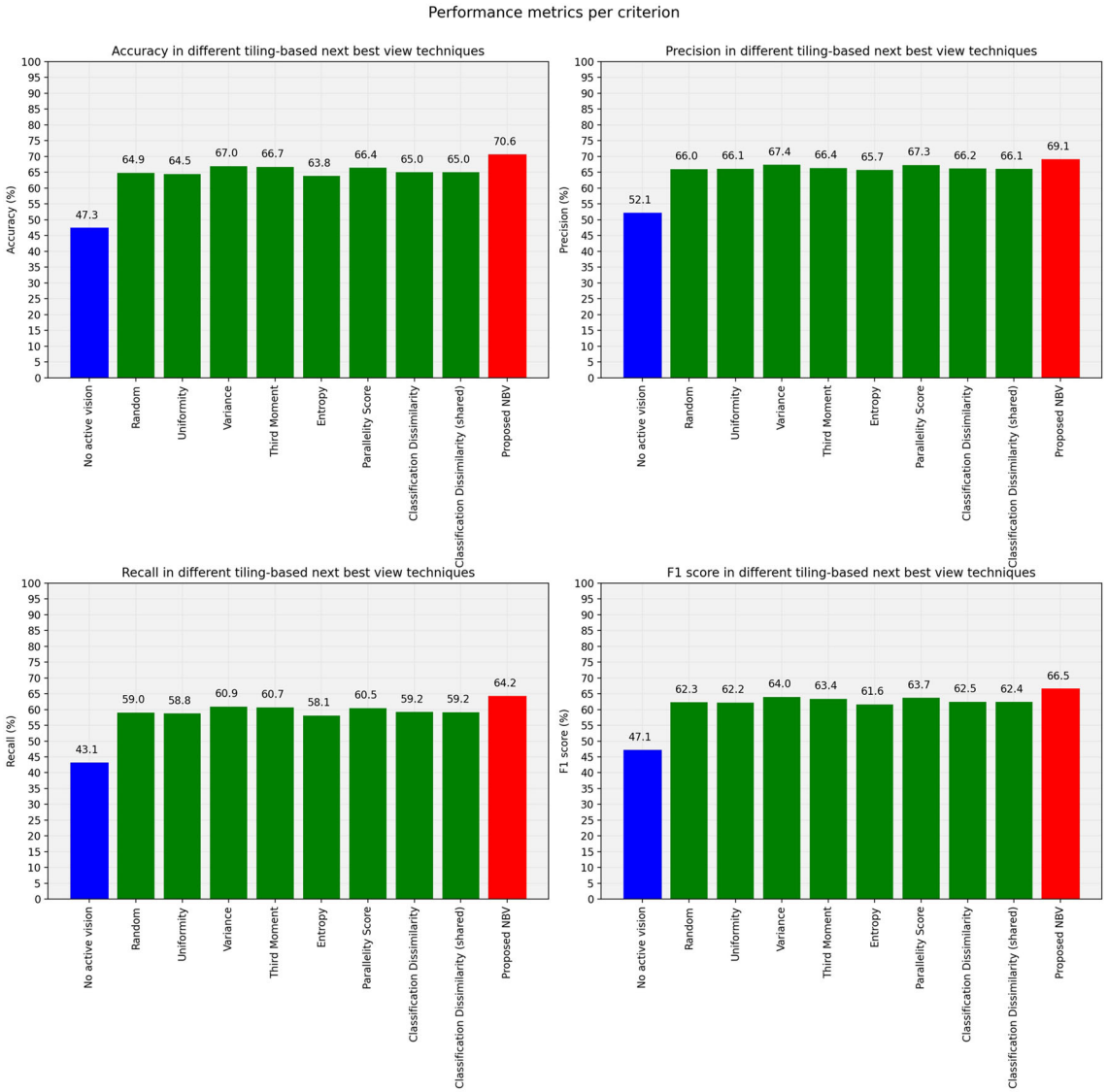

Fig. 8 Performance metrics per measure

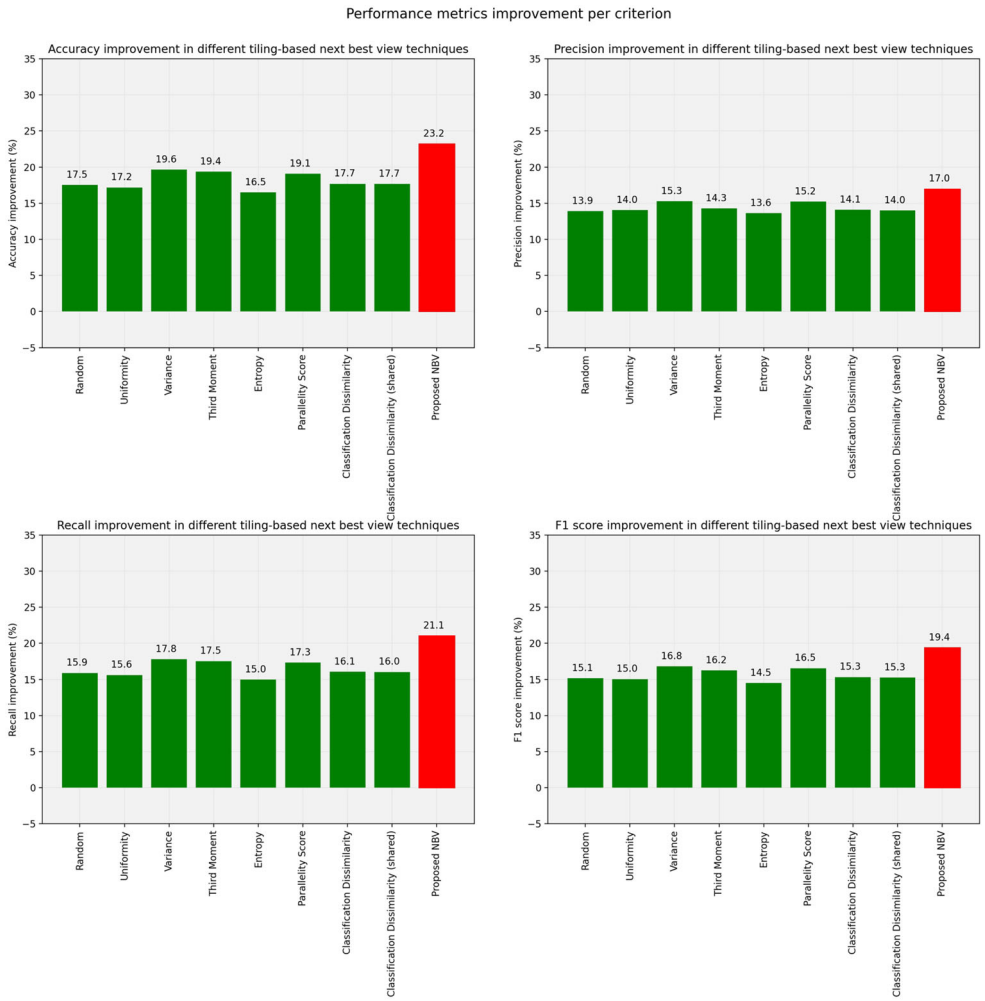

Fig. 9 Performance metrics improvement per measure

four aforementioned performance metrics, compared to the single view recognition without any NBV-enhanced active vision. We see that the proposed system is able to achieve high improvements and is better than other criteria in the figure, including the random selection of next viewpoint.

5.4.3 Performance per tile

Figure 10 shows the accuracy, precision, recall, and F_1 score improvement of the AOR system by using any of the five possible tiles in the tests. The tiles are sorted in the horizontal axis based on the scores they receive from each measure. It is desired that the NBV system performs better in the tiles it puts more emphasis on, i.e. the ones with higher scores in the right side of each plot. Here, we want to see higher bars on the right side of each plot. The results prove that the proposed NBV is successful in obtaining higher performance indices in its best choices. Additionally, the individual measures participating in the ensemble show a trend of increasing accuracy, precision, recall, and F_1 score with the higher scores they generate. They are in contrast to the random selection or other monitored measures, such as uniformity and entropy.

5.4.4 Receiver operating characteristic (ROC) curve

The ROC curves obtained through micro-averaging for all the samples in the 15 benchmarks are shown in Fig. 11. The blue curves in the figure, show the results for the initial view recognitions only, while the green curves indicate the effect of fusing with the results of a randomly selected view. The red curves, on the other hand, show the results of utilizing the proposed method. Comparing the three sets of the curves, verifies the effectiveness of the AOR system in ameliorating the ROC curve and the NBV method in further improving it.

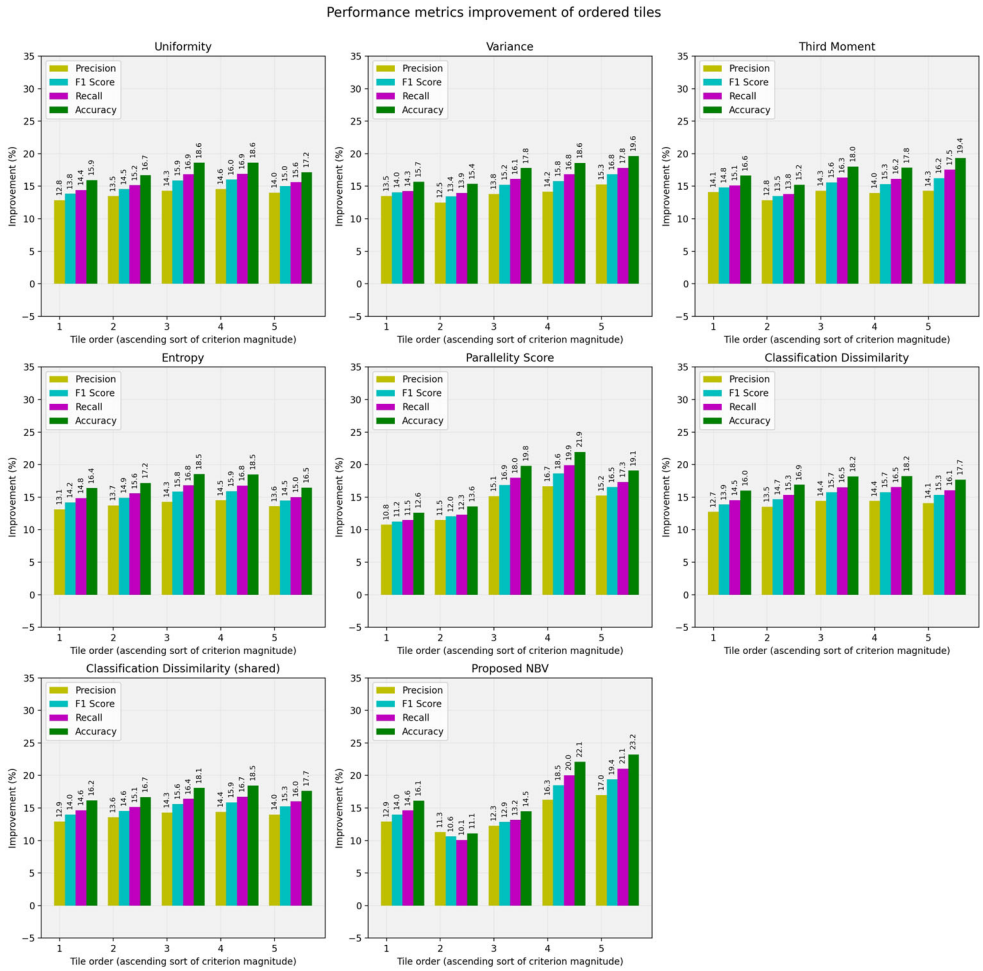

Fig. 10 Performance metrics improvement of tiles in ascending order of scores

5.4.5 Performance over confidence threshold

One of the influencing parameters of the AOR system in general is the confidence threshold. It defines the tendency of the active vision system to retrieve new viewpoints. Figure 12 illustrates the performance metrics over different confidence thresholds. Likewise, performance metrics improvements are swept in various confidence thresholds in Fig. 13. From the two figures, it is observed that the performance does not improve by increasing over an already high confidence threshold. This can be explained that due to the high confidence threshold, even good recognitions require a decision fusion after an active vision procedure, which probably does not contribute to improved classifications as they were performing well from the beginning. In contrast, in the lower confidence thresholds, we observe larger enhancements in the performance. This illustrates the fact that for the smaller group of classifications with high uncertainty, active object recognition and next best view are essential improvement steps.

5.5 Discussion

The experimental results clearly show the applicability of the proposed NBV in improving accuracy, recall, precision, and thus F_1 score of the active object recognition systems. The AOR themselves are effective in empowering the object recognition systems. For example, the SVM classifier with the Averaging fusion experienced improvements of 28.3% and 22.4% in accuracy and F_1 score with the help of the proposed NBV, while the random forest accompanied by the Dempster-Shafer fusion secured increases of 21.8% and 20.8% in accuracy and F_1 score respectively, again by utilizing the viewpoints suggested by the proposed system. These can be compared to the case of randomly selecting the next viewpoint where the former classifier gets accuracy and F_1 score improvements of 23.5% and 18.8%, whereas the

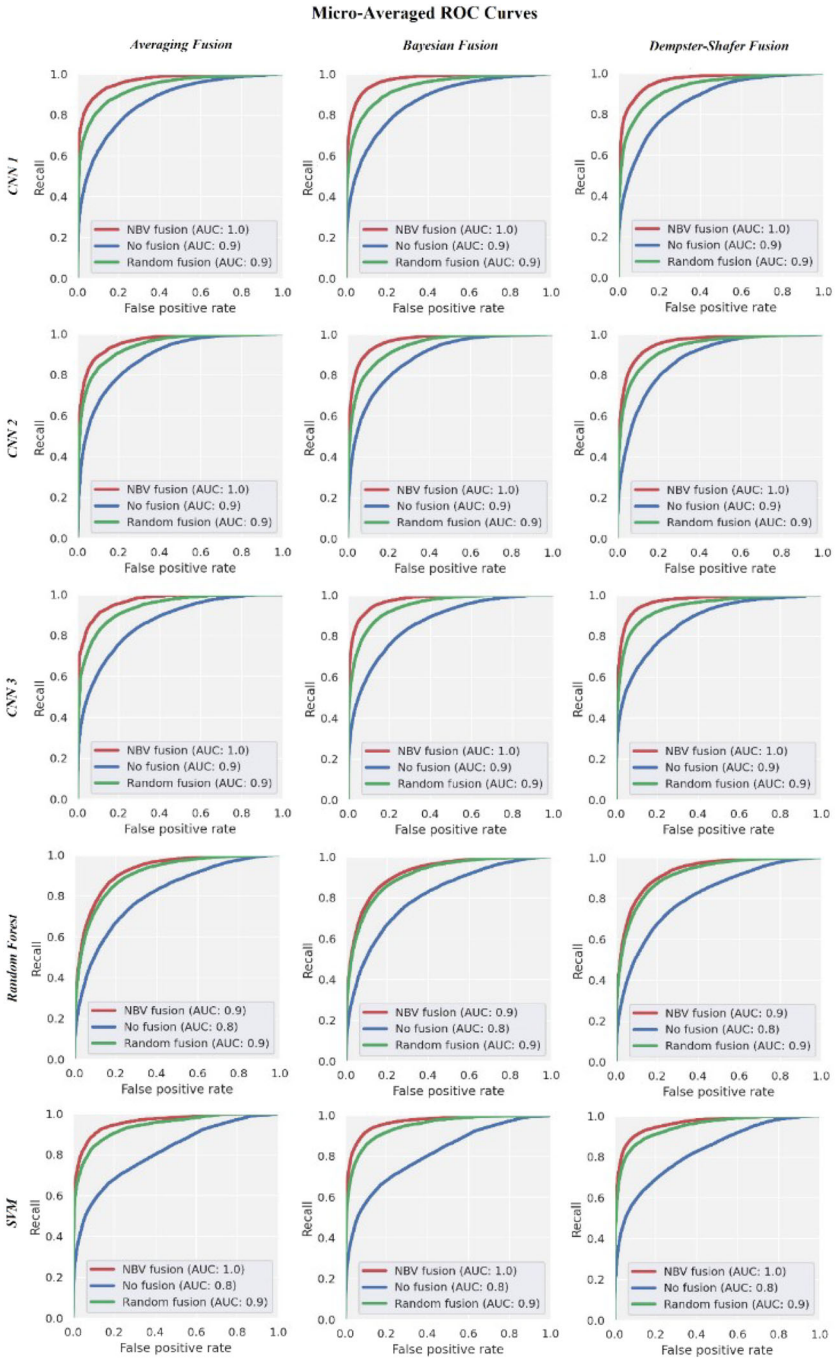

Fig. 11 ROC curves of different test benchmarks

latter gained 18.5% and 17.3% increase in the same metrics, respectively. As another example, CNN 1 with the Naïve Bayes fusion obtained 24.4% and 19.5% rise in accuracy and F_1 score through the described NBV method. It is in contrast to random selection of a viewpoint for the same classifier, where accuracy and F_1 score enhancements were 16.3% and 13.9%.

The proposed NBV method extends the applicability of many modern object recognition systems in real-world conditions. In the presented test benchmark with a variety of added image lighting and quality changes, as well as occlusions, for example, ResNet-101 [14] obtains accuracy and F_1 score of 59.5% and 58.2%, as shown in Fig. 14, while the same classifier as a component of

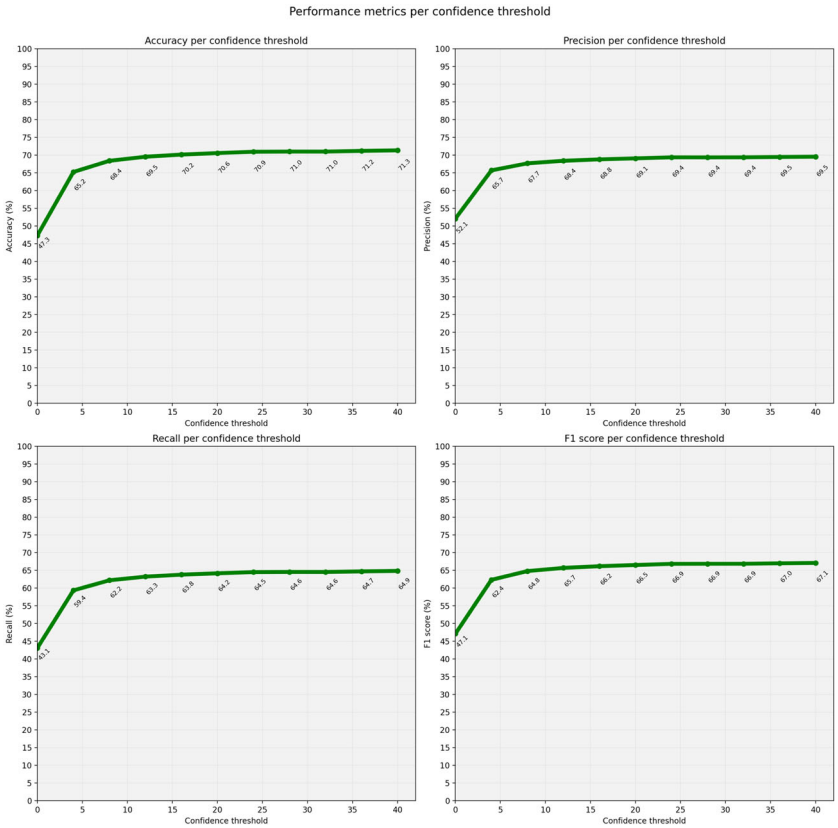

Fig. 12 Performance metrics over different confidence thresholds

the proposed NBV mechanism achieves 68.3% and 66.2%, respectively.

It should be noted that although the classification dissimilarity technique uses dedicated classifiers for each tile, it is also possible to employ only one classifier to perform all the classification tasks, thus making the training stage simpler. Examining Figs. 7 and 10 reveals that it is possible to alternatively use a single-classifier classification dissimilarity, but at the expense of a slightly reduced performance. In addition, we tested the proposed classification dissimilarity against a closely related idea based on information gain in a tile. It is defined in the following:

$$score_t(i) = e(v_i) - e(v_t) \tag{12}$$

where i is an initial view, t is a tile in the initial view, v is a probability vector after classifying an initial view or a tile in the initial view, and $e(.)$ is the entropy operator of an array. A comparison of score-based tile order between the two classification-based methods is presented in Fig. 15 for 3 of the benchmarks in the tests. The figure compares the tile ordering of the classification dissimilarity in Fig. 7 with the one represented by (12). From the figure, it can be inferred that the proposed one works better compared to the one based on the information gain in an area of an initial view. However, it is evident from the results (Figs. 7 and 10) that the classification dissimilarity has lesser impact in the ensemble in deciding the better views. Even though the proposed NBV technique is fairly simple and lightweight, if in any case a faster NBV system is needed it is recommended to drop the classification dissimilarity from the ensemble as it is less influential in determining the more successful views than other member methods and is probably the heaviest one computationally.

Interestingly, the tile ranking using the parallelism score shows that sometimes the tiles with the penultimate score reach better ranks than the highest scoring ones. Those cases probably happen when the higher scoring tile possesses a very steep object surface with respect to the image plane of the camera in the initial view. Too steep of a surface may inhibit the proper view of the respective object side from the standpoint of the initial view, compared to less steep ones.

The tiling scheme of proposed NBV affects the granularity of the next viewpoint options. The method is also based on the assumption that any object surface is visible in the tiles around the object bounding boxes.

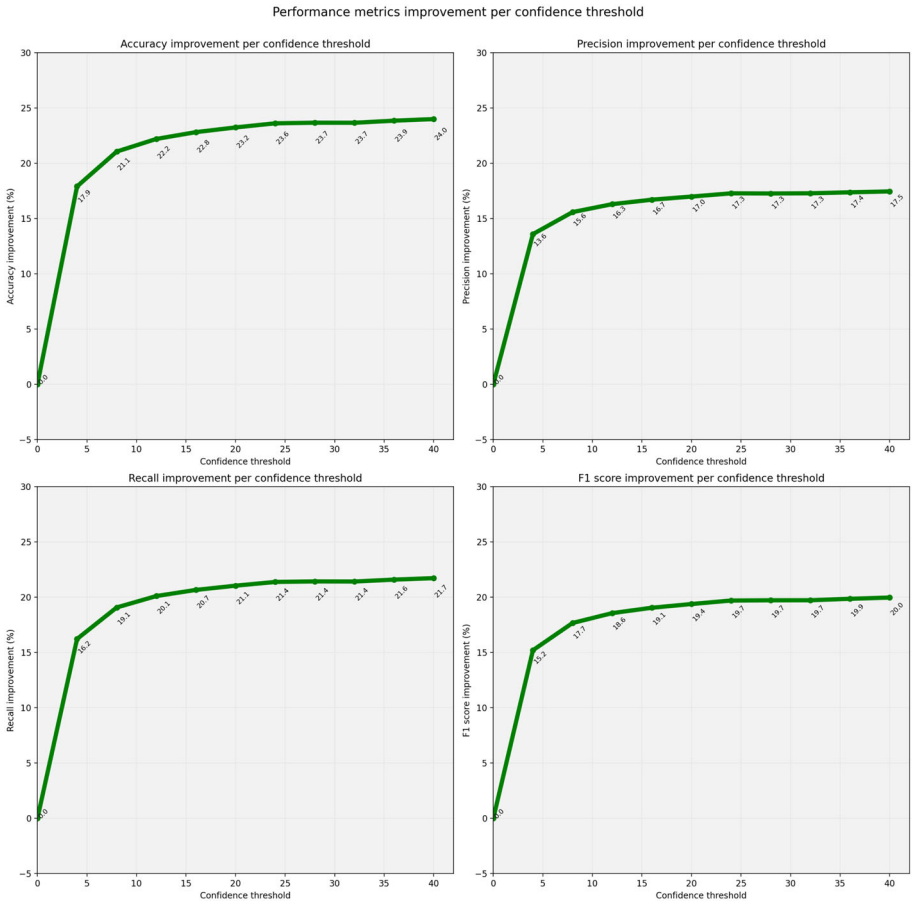

Fig. 13 Performance metrics improvement over different confidence thresholds

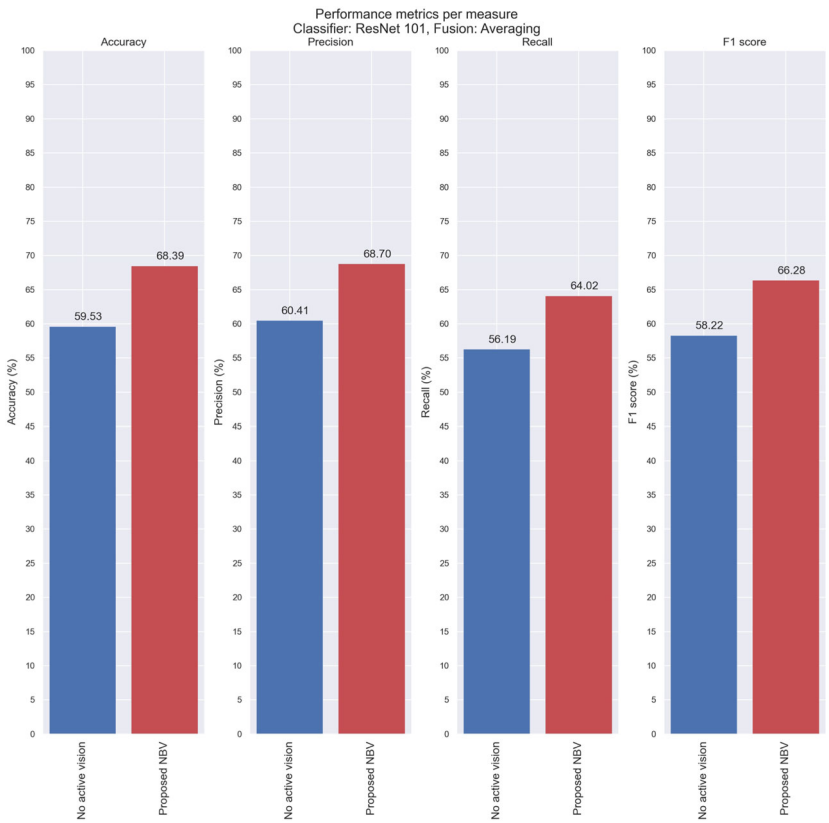

Fig. 14 Performance improvement of ResNet-101, working as a component of the proposed next best view method

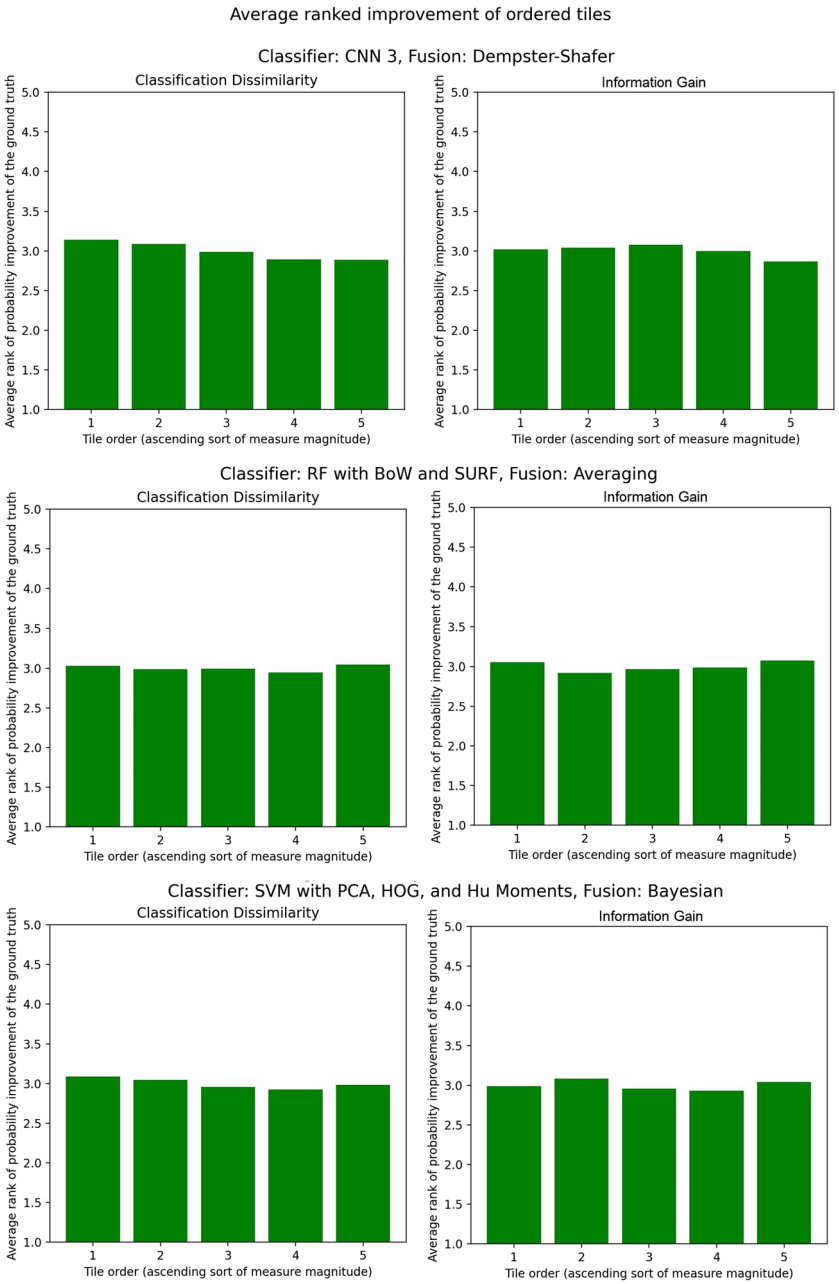

Fig. 15 Comparison of the classification dissimilarity measure with a related information gain-based technique

Although, this is a fairly reasonable supposition, there is no guarantee for that in every tile.

With the recognition performance improvements of the proposed NBV method, there comes some computational overhead, though, compared to a single recognition stage. This is common among all active object recognition systems generally, since an active vision system conducts more observations than a traditional single-frame vision system. The proposed NBV method is designed to capture and process a second frame. Nevertheless, it is computationally light compared to many other active vision systems [2, 5, 8, 17, 23, 27], because it decides on the next best move by using just the current view (thus not needing a series of images taken before making a decision) and performs merely one more move to complete the task of object recognition. In order to determine the NBV from the current viewpoint, the four presented criteria of the ensemble method are also not computation intensive comparatively. Two of the criteria (histogram second and third moments) are histogram processing algorithms, which

are intrinsically faster than local image processing methods. Surface parallelism criterion is also a combination of simple geometric surface normal computation and gray-scale depth map segmentation. As mentioned before, classification dissimilarity is slightly heavier than the other three criteria and is made of a combination of five classifications that can be done very fast in a modern computer. It should be noted that the most time-consuming part of an active vision system is probably the physical camera movement, which depends mostly on the robotic system being employed and the application.

6 Conclusion

In this paper we have presented a next best view approach for active object recognition systems. The proposed view selection divides an initial image of an object into a number of areas in order to analyze each one for clues in determining better next views. Analysis of each area is performed through an ensemble of four different techniques: histogram second moment (variance), histogram third moment, surface parallelism (measuring foreshortening), and classification dissimilarity. The proposed technique does not require a prior training set of specific views of objects or their 3D models. It is capable of suggesting the next viewpoint merely based on the information of a single initial view. This property along with the fact that the proposed method considers both the 3D shape and appearance of objects provides an inherent advantage for active object recognition tasks.

A dataset for testing active object recognition systems was developed in this work and was used to evaluate the proposed next best view technique. The results verified its efficacy in improving accuracy, recall, precision, and F_1 score over entropy, uniformity, information gain of classification, and the case of randomly choosing the next viewpoint. In the presence of heavy occlusions in the initial view, we report average accuracy and F_1 score increase of 23.2% and 19.4% compared to a non-active vision system and 5.7% and 4.3% compared to a randomly-selecting active vision. These improvements prove the effectiveness of the proposed next best view method in improving recognition performance over unsatisfactory initial viewpoints.

In continuation to this work, future efforts should be directed toward adding more selectable viewpoints around an object and possibly devising alternative tiling schemes of the initial view. Adding a mechanism to verify the existence of object surface in a tile to filter out next viewpoints that the current view does not provide any information about may be another direction to focus next. Another area of work can be experimenting with other ensemble methods to replace the voting for the winner tiles. A meta-learning approach would be a potentially interesting way to combine the tile scores.

Author Contributions Conceptualization: Pourya Hoseini, Mircea Nicolescu, Monica Nicolescu; Methodology: Pourya Hoseini, Mircea Nicolescu, Shuvo Kumar Paul; Formal Ananlysis and Investigation: Pourya Hoseini, Mircea Nicolescu, Monica Nicolescu; Writing - original draft preparation: Pourya Hoseini; Writing - review and editing: Pourya Hoseini, Shuvo Kumar Paul, Mircea Nicolescu; Funding acquisition: Monica Nicolescu, Mircea Nicolescu; Resources: Monica Nicolescu, Mircea Nicolescu, Pourya Hoseini, Shuvo Kumar Paul; Supervision: Mircea Nicolescu, Monica Nicolescu.

Funding This work has been supported in part by the Office of Naval Research award N00014-16-1-2312 and US Army Research Laboratory (ARO) award W911NF-20-2-0084.

Availability of data and material Yes. Dataset available at https://github.com/pouryahoseini/Next-Best-View-Dataset.

References

1. Almadhoun R, Abduldayem A, Taha T, Seneviratne L, Zweiri Y (2019) Guided next best view for 3d reconstruction of large complex structures. Remote Sens 11(20):2440
2. Atanasov N, Sankaran B, Le Ny J, Pappas GJ, Daniilidis K (2014) Nonmyopic view planning for active object classification and pose estimation. IEEE Trans Robot 30(5):1078–1090
3. Bajcsy R, Aloimonos Y, Tsotsos JK (2018) Revisiting active perception. Auton Robot 42(2):177–196
4. Barzilay O, Zelnik-Manor L, Gutfreund Y, Wagner H, Wolf A (2017) From biokinematics to a robotic active vision system. Bioinspir Biomim 12(5):056004
5. Bircher A, Kamel M, Alexis K, Oleynikova H, Siegwart R (2016) Receding horizon" next-best-view" planner for 3d exploration. In: 2016 IEEE international conference on robotics and automation (ICRA), IEEE, pp 1462–1468
6. Cui J, Wen JT, Trinkle J (2019) A multi-sensor next-best-view framework for geometric model-based robotics applications. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp 8769–8775
7. Das D, Lee CG (2019) A two-stage approach to few-shot learning for image recognition. IEEE Trans Image Process 29:3336–3350
8. Doumanoglou A, Kouskouridas R, Malassiotis S, Kim TK (2016) Recovering 6d object pose and predicting next-best-view in the crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3583–3592
9. Edmonds M, Yigit T, Yi J (2020) Auto-calibrated 3d hyperspectral scanning using a heterogeneous set of cameras and lights with spectrally-optimal next-best-view planning. In: 2020 IEEE 16th International conference on automation science and engineering (CASE), pp 863–868. IEEE
10. Gao P, Yuan R, Wang F, Xiao L, Fujita H, Zhang Y (2020) Siamese attentional keypoint network for high performance visual tracking. Knowl Based Syst 193:105448
11. Gao P, Zhang Q, Wang F, Xiao L, Fujita H, Zhang Y (2020) Learning reinforced attentional representation for end-to-end visual tracking. Inform Sci 517:52–67

12. Gonzalez RC, Richard E (2018) Woods digital image processing, Pearson Prentice Hall
13. Hayashi T, Fujita H (2020) Cluster-based zero-shot learning for multivariate data. J Ambient Intell Humaniz Comput 12:1–15
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
15. Hoseini P, Blankenburg J, Nicolescu M, Nicolescu M, Feil-Seifer D (2019) Active eye-in-hand data management to improve the robotic object detection performance. Computers 8(4):71
16. Hoseini P, Blankenburg J, Nicolescu M, Nicolescu M, Feil-Seifer D (2019) An active robotic vision system with a pair of moving and stationary cameras. In: International symposium on visual computing, Springer, pp 184–195
17. Jia Z, Chang YJ, Chen T (2010) A general boosting-based framework for active object recognition. In: British machine vision conference (BMVC), Citeseer, pp 1–11
18. Lauri M, Pajarinen J, Peters J, Frintrop S (2020) Multi-sensor next-best-view planning as matroid-constrained submodular maximization. IEEE Robot Autom Lett 5(4):5323–5330
19. Lehnert C, Tsai D, Eriksson A, McCool C (2019) 3d move to see: Multi-perspective visual servoing towards the next best view within unstructured and occluded environments. In: 2019 IEEE/RSJ International conference on intelligent robots and systems (IROS), IEEE, pp 3890–3897
20. Morrison D, Corke P, Leitner J (2019) Multi-view picking: Next-best-view reaching for improved grasping in clutter. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp 8762–8768
21. Palomeras N, Hurtós N, Vidal E, Carreras M (2019) Autonomous exploration of complex underwater environments using a probabilistic next-best-view planner. IEEE Robot Autom Lett 4(2):1619–1625
22. Pérez-Hernández F, Tabik S, Lamas A, Olmos R, Fujita H, Herrera F (2020) Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance. Knowl Based Syst 194: 105590
23. Potthast C, Sukhatme GS (2014) A probabilistic framework for next best view estimation in a cluttered environment. J Vis Commun Image Represent 25(1):148–164
24. Rebull Mestres J (2017) Implementation of an automated eye-in hand scanning system using best-path planning, Master's thesis, Universitat Politècnica de Catalunya
25. Wang Z, Xiong J, Yang Y, Li H (2017) A flexible and robust threshold selection method. IEEE Trans Circuits Syst Video Technol 28(9):2220–2232
26. Wu Y, Jiang X, Fang Z, Gao Y, Fujita H (2021) Multi-modal 3d object detection by 2d-guided precision anchor proposal and multi-layer fusion. Appl Soft Comput 108:107405
27. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1912–1920
28. Xu Y, Hu J, Wattanachote K, Zeng K, Gong Y (2020) Sketch-based shape retrieval via best view selection and a cross-domain similarity measure. IEEE Trans Multimed 22(11):2950–2962
29. Zeng R, Zhao W, Liu YJ (2020) Pc-nbv: A point cloud based deep network for efficient next best view planning. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 7050–7057
30. Zhu K, Jiang X, Fang Z, Gao Y, Fujita H, Hwang JN (2021) Photometric transfer for direct visual odometry. Knowl Based Syst 213:106671

Pourya Hoseini is a postdoctoral researcher at University of California, San Diego. He received his Ph.D. in 2020 and M.S. in 2017, both in computer science and engineering from University of Nevada, Reno. He also received a M.S. degree in 2011 and a B.S. in 2008 in electrical engineering from Urmia University and Azad University, Iran, respectively. His research interests are machine learning, computer vision, and evolutionary computing.

Shuvo Kumar Paul received his M.S. in 2020 in computer science and engineering from University of Nevada, Reno. He completed his B.S. from North South University, Bangladesh. Before joining UNR, he worked as a research associate at the AGENCY lab (previously CVCR). His research interests include machine learning, computer vision, computational linguistics, and robotics.

Mircea Nicolescu is a Professor of Computer Science and Engineering at the University of Nevada, Reno and co-director of the UNR Computer Vision Laboratory. He received a PhD degree from the University of Southern California in 2003, a MS degree from USC in 1999, and a BS degree from the Polytechnic University Bucharest, Romania in 1995, all in Computer Science. His research interests include visual motion analysis, perceptual organization, vision-based surveillance, and activity recognition. Dr. Nicolescu's research has been funded by the Department of Homeland Security, the Office of Naval Research, the National Science Foundation and NASA. He is a member of the IEEE Computer Society.

Dr. Monica Nicolescu is a Professor with the Computer Science and Engineering Department at the University of Nevada, Reno and is the Director of the UNR Robotics Research Lab. Dr. Nicolescu earned her PhD degree in Computer Science from the University of Southern California (2003) at the Center for Robotics and Embedded Systems. She obtained her MS degree in Computer Science from USC (1999) and a BS in Computer Science at the Polytechnic University Bucharest (Romania, 1995). Her research interests are in the areas of human-robot interaction, robot control, learning, and multi-robot systems. Dr. Nicolescu's research has been supported by the National Science Foundation, the Office of Naval Research, the Army Research Laboratory, the Department of Energy and Nevada Nanotech Systems. In 2006 she was a recipient of the NSF Early Career Development Award (CAREER) Award for her work on robot learning by demonstration.