ORIGINAL RESEARCH

Next Best View Planning in a Single Glance: An Approach to Improve Object Recognition

Pourya Hoseini[1,2] · Shuvo Kumar Paul[1] · Mircea Nicolescu[1] · Monica Nicolescu[1]

Received: 11 October 2021 / Accepted: 11 October 2022 / Published online: 11 November 2022

Abstract

Many real-world mobile or robotic vision systems encounter the problem of occlusion or unfavorable viewpoint in performing their tasks. A remedy to this issue is active vision, i.e. physically moving the camera or employ another camera to provide other viewpoints that hopefully provide more information for the task at hand. In the case of object recognition, an active vision system can help by offering classification decisions from another viewpoint when the current recognition confidence is low. A natural question, however, would be which next viewpoint is more effective in improving the object recognition performance. To determine the next best view, previous approaches either need multiple captures of the same object in specified poses, training datasets of 3D objects, or construction of occupancy grids. These methods are consequently computation, data, or observation intensive. In this paper, we propose a next best view method for object recognition that does not need any information about objects in other viewpoints, their 3D shape, or multiple prior observations to function properly. The proposed approach analyzes the object's appearance and foreshortening in the current view to rapidly decide where to look next. Test results show its efficacy in correctly selecting the viewpoints that improve the object recognition performance more.

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

Introduction

Mobile intelligent systems depend on sensing their environment to act informed. Among the sensing modalities, vision plays a crucial role. However, for a vision system there can be difficulties in real-world situations where capturing the most useful information is deterred in the absence of a good viewpoint. Insufficient discriminative features due to poor lighting, insufficient image resolution, or unfavorable viewpoints of the object, as well as presence of occlusion are some of the characteristics of a suboptimal viewpoint. A workaround to this issue is active vision, wherein new visual sensory information are obtained to enhance the performance of a vision system. More details about the idea of active vision can be found in [1]. Active vision has two major domains of application: three-dimensional (3D) object reconstruction and object recognition. Active object recognition (AOR) routines, which are the focus of our work, generally encompass uncertainty evaluation, physical camera repositioning, and information fusion [2]. In the case of an uncertain recognition of an object in the current viewpoint, an AOR system moves the camera to a new position and orientation with respect to the object of interest to fuse the recognition outputs in the new perspective with the ones from the earlier viewpoint. An example of an AOR system can be found in our previous paper [3], in which a humanoid robot uses an arm-mounted camera to capture new views

This article is part of the topical collection "Computer Vision, Imaging and Computer Graphics Theory and Applications" guest edited by Jose Braz, A. Augusto Sousa, Alexis Paljic, Christophe Hurter and Giovanni Maria Farinella.

✉ 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, NV, USA

[2] Department of Ophthalmology, University of California, San Diego, CA, USA

of objects that it cannot discern initially through its head camera.

In order to move the camera, an active vision system needs to first evaluate the effectiveness of the new viewpoint from the new camera pose. Accordingly, a viewpoint that provides more useful information is preferred. Finding next best view (NBV) in a single shot, however, is an ill-posed task as an active vision system needs to decide where to look next based on only the current viewpoint, which is insufficient for deterministically inferring that. The goal of the vision defines the way a NBV method is designed. For 3D reconstruction, next viewpoint is desired to reveal unobserved surfaces of an object, thus their typical goal is not to plan for a single new viewpoint, but to plan for a sequence of NBVs to completely observe the volume of an object. In contrast, for object detection and recognition applications, obtaining new discriminative features to enhance the recognition performance, while keeping the time and energy costs of camera movements lower by fewer camera relocations, are intended. In this work, we propose and evaluate a next best view method for active object recognition that plans for only a single new frame to be captured.

As detailed in "Previous Work", earlier work related to finding the next best view can be grouped into two classes: space occupancy-based and object estimation-based techniques. Computing occupancy of 3D space through ray tracing and subsequently evaluating information gain in various possible viewpoints is inherently advantageous for 3D reconstruction purposes, because it attempts to discover more surface voxels than discriminative features for classification. On the other hand, object estimation techniques try to assume the 3D shape of the current object based on the current viewpoint, or compare the current object appearance and/or shape to the ones seen in the training time to deduce the best course of action by comparing various hypothetical viewpoints. The basic drawback to these methods, however, is that they either require several observations or rely heavily on the existence of large datasets of object volumes or images taken around the objects in predefined viewpoints. The errors induced by the inaccuracies in predicting the object shape or appearance from unseen directions also constitute another set of problems with these methods.

In this paper, a single-shot next best view method for object recognition is presented. It plans for one new viewpoint based on the shape and appearance cues of the currently visible object to enhance the object recognition performance when necessary. The proposed NBV method does not depend on a prior dataset of 3D object volumes or any particular set of images taken around the object for training. Instead, it employs conventional datasets (i.e. a collection of random images of objects) merely for the training of the classifiers. To save time and energy for camera motion, the proposed NBV also does not require a chain of camera movements toward or around the object. To accommodate these characteristics, an ensemble of shape and appearance criteria is utilized in our work to analyze the current viewpoint and suggest a new viewpoint. The criteria take into account foreshortening, classification dissimilarity between the current view and a part of it, and texture variance. Additionally, we gathered a dataset to test the proposed method in a systematic and reproducible fashion. In our tests, the proposed NBV method proves to be effective in predicting next best view among a set of pre-selected test-time poses around the object.

Here, we describe an extension of our earlier work [4, 5] with enhancements in the employed techniques and improvements in the performance. The main contributions of the current work are:

1. A novel next best system is proposed exclusively for the task of object recognition.
2. The proposed NBV depends only on the current object shape and appearance, hence no prior knowledge of objects is necessary.
3. No specially designed datasets are needed for the next best view determination. Only traditional object classifiers are trained.
4. A small test dataset was gathered to efficiently and in a reproducible way test the proposed next best view system. It contains images captured around various objects in different lighting and background situations and can be used by other researchers as a benchmark.
5. Experimental results verify the performance improvement after fusion of views among a pre-defined set of possible camera poses.

Compared to our earlier work [4, 5], the current paper incorporates the following new features:

1. We investigated the contribution of various criteria to the overall results in order to achieve the best trade-off between accuracy and computation time. In the current work, we reassessed the viewpoint criteria being used by our method and optimized the number of components in the current ensemble to three. For example, in the previous work the ensemble of criteria contained the third moment of histogram as a texture analysis criterion, which we found having least contribution to the overall results, partially due to the correlation to the texture variance criterion.
2. The classification dissimilarity criterion is updated in the new method to use Kullback–Leibler divergence instead of sum of squared differences, which we found to improve performance.

3. In contrast to the hard voting mechanism of the previous method, where each criterion casts the most preferred tile as its single vote, the presented work employs soft voting to take into account the preference order of all tiles for each viewpoint criterion.

The remainder of the paper is organized as follows. Previous work in the literature on the topic of next best view is reviewed in the subsequent section. An overview of the steps in an active vision system for object recognition is discussed in "Active Vision for Object Recognition". The proposed next best view system is presented in "The Next Best View Method". "Experimental Results" describes the test benchmarks and demonstrates the obtained results. Lastly, concluding remarks are discussed in "Conclusion".

Previous Work

Studying the literature portrays a few directions in the area of next best view. A deep belief network is presented in [6] to "hallucinate" the whole object shape and appearance 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 with the least uncertainty is then selected for the next view. Although interesting, the idea of examining various hypothetical viewpoints of an object in [6] has a major flaw in depending heavily on the estimation of the object shape and appearance in the unobserved occluded areas, which can be a large source of errors. In another deep learning-based work [7], raw point cloud data and current view selection states are taken as input to subsequently predict the information gain of all candidate views. The work in [8] presents a meta-learning based few-shot learning model to decide on a set of glimpses around an object to determine the object category. It is an answer to the issues, such as large amount of data needed in methods like [6].

Rearranging depth camera positions based on a bio-inspired approach by imitating barn owls' head motions to actuate a depth camera installed on a robot is examined in [9] for 3D reconstruction of objects. Mimicking motions of active perception attempts of a biological being in [9] looks promising. Yet it finds NBV regardless of the object shape and appearance, which can cause missing some important clues in determining the next best view. In [10], an active pose estimation and object detection framework with dynamic camera location planning is described to balance the odds of object detection enhancement and the energy needed to move the camera. An Asus Xtion RGB-D camera mounted on the PR2 robot's wrist was used as the sensor, while a sequence of captures are planned along the fastest way the camera is moved toward the object. The active vision system of [10], although multi-capture, does not consider the object shape and appearance and merely moves the camera towards an object, hence it does not have any intelligent viewpoint selection component.

A trajectory planning technique for an eye-in-hand vision system on a PR2 robot is presented in [11] to boost the expected number of voxel observations by searching for maximum local information gain. In continuation to the work of [11], a next best view method for 3D reconstruction applications on the basis of predicting information gain from prospective viewpoints is proposed in [12]. To predict the information gain in unobserved areas, an occupancy grid is formed out of all the observations so far, and a Hidden Markov Model (HMM) is used to estimate the observation probability of unobserved cells in the grid. The NBV in [13] estimates desirability of any potential viewpoint by directly estimating the classification probabilities of different views instead of rendering their hypothetical object appearances to compute the information gain in each view. Despite overcoming the problem of computationally expensive renderings of hypothetical 3D objects, this approach requires 3D training data for every test object and performing classification and confidence estimation for every viewpoint of the 3D objects in the training. This prerequisite significantly affects the functionality of the technique due to the scarcity of such training data for many real-world objects.

In [14], 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 [15] 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 [16] 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 [17], which was tested for both 3D reconstruction and weld seam inspection.

A boosting technique is proposed in [18] to combine three criteria for determining the NBV around objects. The first criterion compares the similarity of the current object with prerecorded object appearances in different views and selects the one with the least similarity. 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 an object and its other viewpoints requires a dataset made of images around the training objects with their known pose. This can be practically burdensome as there is a need to capture poses and appearances all around the objects that are to be detected at test time.

In [19], next best view is used for calibrating and operating a multi-camera 3D hyperspectral scanner. In another work [20], NBV is incorporated for sketch shape retrieval to select the candidate projection of 3D shapes to extract their features and compare them to a sketch. In order to construct 3D models of objects using depth images, captured by a team of robots, the NBV presented in [21] employs a utility function that scores sets of viewpoints and avoids overlap between multiple sensors. Improvement in robotic grasp detection by using NBV to provide informative viewpoints in cluttered scenes is reported in [22]. In [23] an unmanned aerial vehicle (UAV) equipped with NBV was reported for 3D reconstruction of large structures. Likewise, in [24] an autonomous underwater vehicle (AUV) is programmed to choose its next viewpoint optimally to map or inspect complex underwater structures. The utility function in [23], considers four criterion categories to compute NBV: traveled distance, information theory, model density, and predictive measures based on symmetries in the structure.

Active Vision for Object Recognition

Active object recognition has been adopted in robotic, vision-based surveillance, and several other applications. Figure 1 demonstrates the flowchart of a typical AOR system. After preprocessing operations, such as denoising, the process begins with an object recognition stage using the current point of view. Later, the confidence of the classification results of the initial recognition round is evaluated. Whenever the recognition is deemed uncertain (i.e. a low confidence value), the active vision mechanism is set off. 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 camera being moved 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 now in place in the pose determined by the NBV system, the object recognition is performed once again. Henceforth, detected objects in the two camera views are matched to form pairs of object classification decisions for a later fusion step. Depending on the availability of frame transformations and the application, the matching can be achieved through a purely vision-based object/keypoint/pixel correspondence operation or through the 3D geometric transformation between the camera poses in a sensor-equipped robotic system.

To combine the classification results and obtain the final probability vector of the object categories, each matched object classification pair is proceeded into a decision fusion module. The fusion module then takes the class probabilities of the two classifiers and fuse them to yield the output class probabilities. For more information regarding active object recognition in a robotic platform, refer to our previous paper [3].

The Next Best View Method

The proposed method only assumes the availability of color and depth information of the initial camera view to find a candidate viewpoint in a single try after the initial capture. For rigorous testing purposes, the NBV method can select among a set of pre-specified poses that are ordinarily reachable for UAV or eye-in-hand platforms. In the current implementation, the poses around any object are grouped into eight clusters, all of them on the plane that passes through the object and is parallel to the image plane of the camera at the initial viewpoint. A camera in the possible poses in each group generally views the same part of the object. For instance, a next viewpoint 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 is worth noting that the number of pose groups around an object can be increased if it is necessary depending on the application.

Viewpoints at the same depth as an object in the camera coordinate of the initial view, which are considered in the proposed NBV, are reasonably accessible for many eye-in-hand arrangements, while at the same time can provide substantially new information from a view direction perpendicular to the initial one. A camera in a pose from a depth less than the object's depth will probably have overlapping field of view on the object's surface with the initial frontal view and see the same features of the object. In contrary, any pose with a depth farther than the object will see behind the object, hence offering a perspective with new features. Notwithstanding the desirability of that, this approach has two disadvantages. First, it is difficult to reach by a robotic system. Many robotic arms do not have the degrees of freedom to move an arm-mounted camera to an orientation facing back of an object, thus facing the robot itself, at a large distance from the robot. Since the object's thickness is unknown in a single frontal shot, it is also challenging to plan for a pose behind an object for an unmanned aerial vehicle or a freely moving camera unit. The second reason is that for a single NBV based on the

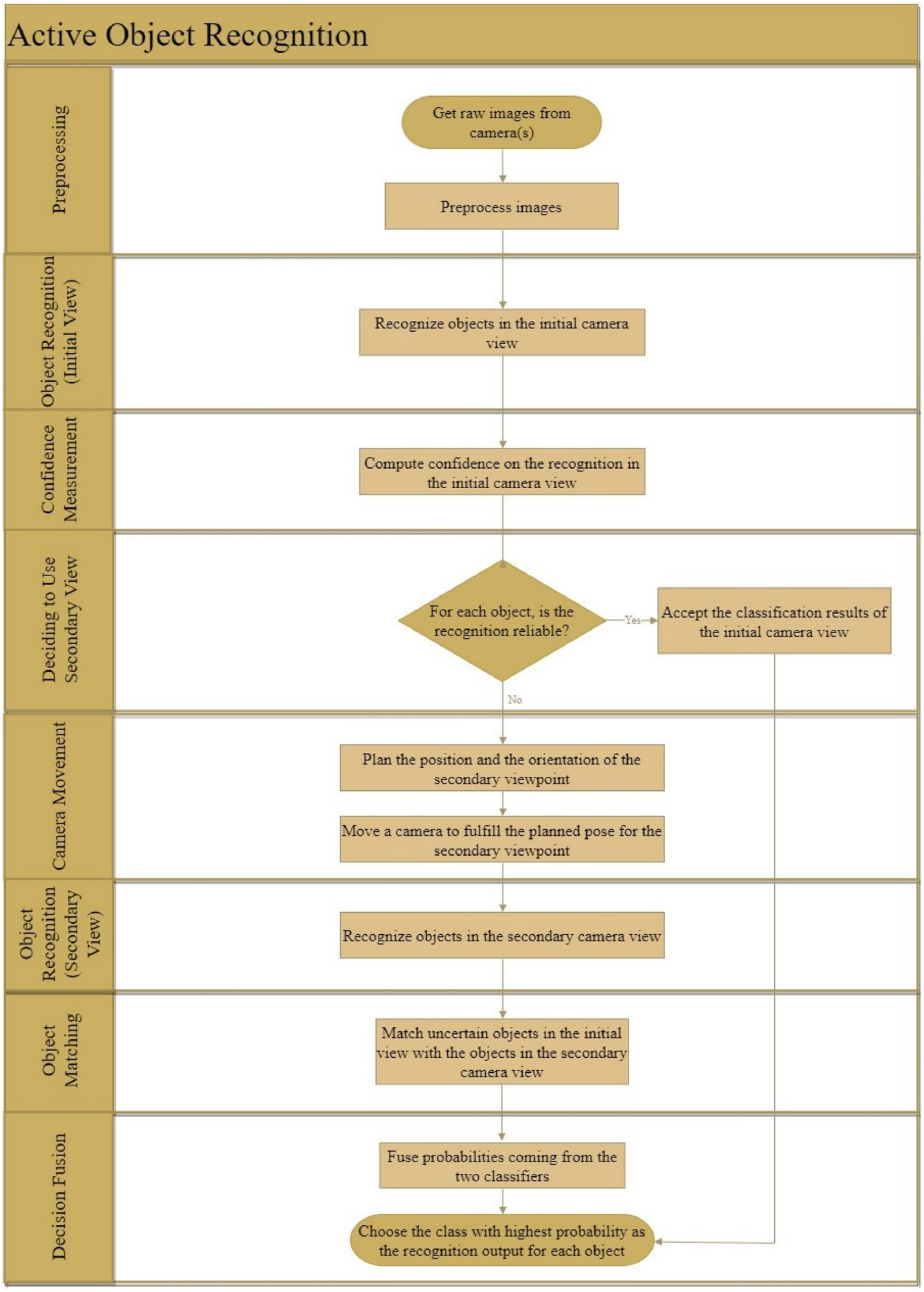

Active Object Recognition
Preprocessing
Get raw images from camera(s)
Preprocess images
Object Recognition (Initial View)
Recognize objects in the initial camera view
Confidence Measurement
Compute confidence on the recognition in the initial camera view
Deciding to Use Secondary View
For each object, is the recognition reliable?
Yes
Accept the classification results of the initial camera view
No
Camera Movement
Plan the position and the orientation of the secondary viewpoint
Move a camera to fulfill the planned pose for the secondary viewpoint
Object Recognition (Secondary View)
Recognize objects in the secondary camera view
Object Matching
Match uncertain objects in the initial view with the objects in the secondary camera view
Decision Fusion
Fuse probabilities coming from the two classifiers
Choose the class with highest probability as the recognition output for each object

Fig. 1 Active object recognition steps

current frontal view of an object, the viewpoint quality of the self-occluded area behind the object for active object recognition is not known. As a result of the aforementioned issues, viewpoints that try to observe behind an object are not considered as the next viewpoint candidates.

Pose Representation by Splitting the Initial Viewpoint

In the proposed method, the object bounding box, being generated by any desirable object detection procedure, is divided into different regions, *tiles*, that cover the entire area of the bounding box in a non-overlapping style. The splitting scheme in the proposed method is shown in Fig. 2. Every bounding box is divided into nine tiles in the current implementation, where each of the eight peripheral tiles corresponds to one of the pose groups of a camera around the object. For example, the top left tile represents a point of view when the camera is viewing the object from the object's top left with the same depth with respect to the camera as the object itself in the camera coordinate of the initial view. The camera in the new

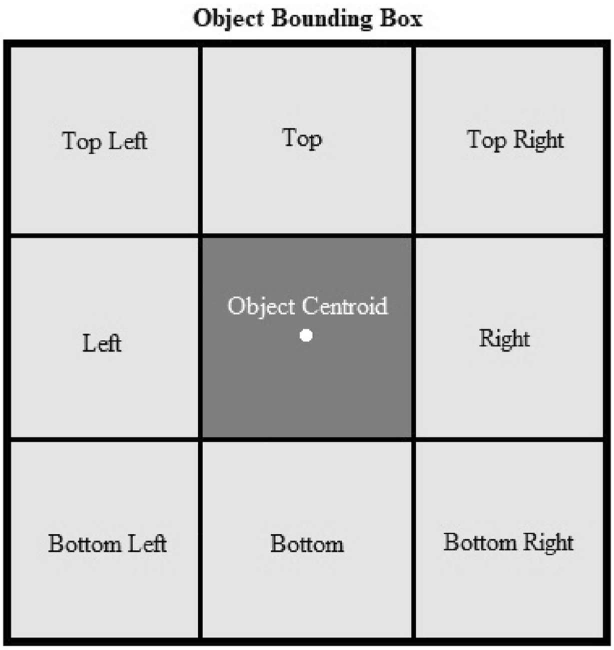


Fig. 2 Tiling pattern in the proposed next best view system

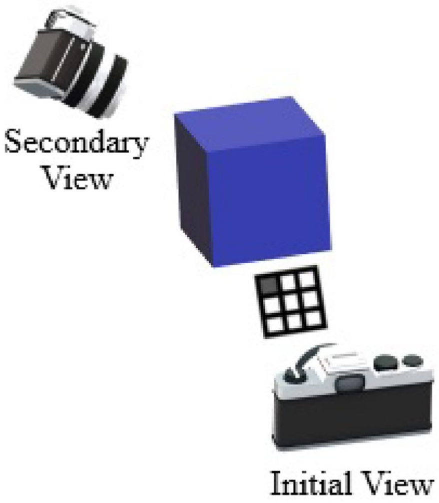


Fig. 3 An example next viewpoint selection situation, where the top left tile is selected and consequently the secondary viewpoint is looking at the object from its top left

orientation can be placed at any arbitrary distance from the object surface it is observing, given that the pose is feasible for the camera setup, object surface is entirely visible from that distance, and the image resolution of the camera is satisfactory for capturing a clear image of the object. Figure 3 illustrates this example situation.

With the splitting scheme comes the opportunity of analyzing each tile of the current view separately in the hope of revealing clues to a more informative NBV, corresponding to the same side of the object it is representing, that consequently contributes to a better overall recognition performance. In contrast to methods of [12–15] that simply try to look at unobserved voxels, the proposed approach attempts to further qualify its decision based on what is currently being observed. Moreover, the proposed method differs from techniques of [6, 13] that hypothesize the object shape as a part of their NBV prediction process, because instead of requiring inference of explicit information about the entire shape, appearance, and relative pose of the objects, it only utilizes limited cues directly available in the initial view.

The Ensemble of Viewpoint Criteria

A weighted voting mechanism among three criteria in the proposed method selects the peripheral tile with the highest votes. Each of the three criteria cast their votes according to the preference they give to the tiles. The lowest ranked tile gets no votes, while others get one more vote than their less preferred one. One criterion statistically analyzes the texture of a tile. Another one considers the classification dissimilarity between a tile and the entire object, whereas the third criterion evaluates the foreshortening of the object to estimate how visible its surface is in the initial view. The three voting criteria are explained in detail in the following three subsections.

Second Moment (Variance) of Histogram

Among other cases, active object recognition may prove helpful when the object being observed is not clearly recognizable due to occlusion, lighting conditions, object shape, etc. One way to confront these situations can be to shift the view toward poses that are likely to provide better quality images. To this end, the second moment (variance) texture analysis tool is utilized. In this work, it is computed from the intensity histograms of each tile's image to facilitate its faster processing. A high-contrast image has a higher chance of containing more features than a uniform one. The second moment or variance of intensity histogram is a measure of contrast of an image [25]. The variance of an intensity histogram is defined in (1) [25].

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

In the equation, $\sigma^2(z)$ is the variance of intensity levels (z), which is identical to the second moment, $\mu_2(z)$. In Addition, L represents the total number of intensity levels in the histogram, i is the index of the current intensity level, $p(z_i)$ is the probability of an intensity level, and m is the mean of intensities, computed as follows:

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

The second moment should be ideally high, because a greater $\mu_2(z)$ means higher contrast and perhaps more features, with which a tile can be a cue to a feature-rich sideways surface for a good next viewpoint. It is also worth mentioning that, because the second moment and the third moment-based criteria result in similar scoring of tiles in a bounding box, in contrast to the earlier version of the proposed method [4, 5] we only use second moment of histogram to prevent overemphasizing statistical texture analysis in the voting ensemble of criteria.

Foreshortening Score

By taking into consideration that we examine all criteria on the periphery tiles of an object's bounding box, an object surface that is nearly parallel to the image plane of the camera in the initial viewpoint will probably be easily visible to the sensor. Oppositely, a peripheral surface with a perspective to the current view, exhibits foreshortening and is likely to be less visible in the current view as its surface is tilted. Based on this idea, the foreshortening score measures how much foreshortening is present, or in other words how parallel the object surface being seen in a tile is to the image plane of the 3D camera observing the object. Assuming the depth map of a tile is segmented, and the object surface constitutes the foreground pixels, the foreshortening score is defined in the following:

$$\begin{cases} P = 1 - \dfrac{\sum_{p \in F} N\left(\overrightarrow{\left(\frac{\mathrm{d}z}{\mathrm{d}x_p}, \frac{\mathrm{d}z}{\mathrm{d}y_p}, 1\right)}\right) . \vec{z}}{|F|} & |F| \neq 0 \\ P = 0 & |F| = 0 \end{cases} \tag{3}$$

where P is the foreshortening score, p is a pixel in the current tile, F is the set of foreground pixels, $|F|$ shows the number of pixels in the tile that are designated as foreground, $N()$ is a vector normalization function, and $\vec{z}$ is the 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:

$$\frac{\mathrm{d}z}{\mathrm{d}x}\bigg|_{x=x_p} = z(x_p + 1, y_p) - z(x_p - 1, y_p) \tag{4}$$

$$\frac{\mathrm{d}z}{\mathrm{d}y}\bigg|_{y=y_p} = z(x_p, y_p + 1) - z(x_p, y_p - 1) \tag{5}$$

In (4) and (5), $z(., .)$ denotes the depth at a pixel of 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. It is common for ordinary 3D cameras to produce small spots of unknown values spread over their generated depth map, for which the camera is not able to compute the depth. To overcome this issue, the unknown values are replaced with the maximum known depth in the current tile. Because we are presuming the background has more depth than the object surface, the substitution of unknown values with maximum depth effectively marks them as background. Usually, actual object surfaces do not completely fill their bounding boxes, which means object bounding boxes contain areas showing other unintended entities, i.e. background. To prevent the background areas from affecting the foreshortening score, an input depth map passes through a segmentation step before being considered for its surface foreshortening. In our work, we opted for Otsu's segmentation [25], but any well-performing binary segmentation method might be used. The foreground areas (F) that are obtained through Otsu's segmentation are then assumed to represent the object surface and are used in (3).

The term $N\left(\overrightarrow{\left(\frac{\mathrm{d}z}{\mathrm{d}x_p}, \frac{\mathrm{d}z}{\mathrm{d}y_p}, 1\right)}\right)$ in (3) computes the surface normal for every foreground pixel in the depth map. The inner product of the surface normal and the z axis of the camera coordinate measures how parallel are the object surface and the image plane of the camera in the initial view, effectively quantifying inverse foreshortening of the object. Ultimately, to find the average foreshortening of the object surface to the camera, the results of the inner products are averaged over all the foreground pixels and later negated. The proposed score prefers a tile when its score is higher. In the case of no foreground pixels in a tile (i.e. no object surface), the score is set to the least possible value, zero, as the foreshortening criterion would not have any clue of the object surface in the tile.

Classification Dissimilarity

In the event of an uncertain recognition by an AOR system, locating the tiles that contribute more to the uncertainty of the initial view by not confirming the initial view's recognition can constitute a promising NBV strategy. Incidentally, if a tile's recognition is in agreement with the recognition results of the whole initial view, the prospect of finding new information is less compared to an opposing classification. Therefore, the classification dissimilarity opts for the contradicting tiles to take a new look from their respective direction. Accordingly, we adopted Kullback–Leibler divergence of class probabilities to calculate the classification dissimilarity (6).

$$D_{\mathrm{KL}_j} = \sum_{i \in G} p_o^c(i) \times \log p_o^c(i) - \sum_{i \in G} p_o^c(i) \times \log p_{tj}^{c_j}(i) \tag{6}$$

where D_{KL_j} is the dissimilarity score (Kullback–Leibler divergence) between the tile j and the complete object image, G represents the set of object classes, and $p_{tj}^{c_j}(i)$ and $p_o^c(i)$ are probabilities of a 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 c and c_j are conventional classifiers, trained with color images of objects, with the difference that c_j only uses the portion of images related to tile j, while c considers the whole object images in the training. 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.

The proposed dissimilarity score is an improvement over our previous work [4, 5], where sum of squared differences (SAD) of the classification probabilities of a tile and the whole object image was used for scoring the classification dissimilarity. SAD is defined in the following.

$$\mathrm{SAD}_j = \sum_{i \in G} |p_{tj}^{c_j}(i) - p_o^c(i)| \tag{7}$$

Experimental Results

To test next best view techniques, and more generally active object recognition methods, standard test benchmarks are necessary for comparison purposes. However, due to the robotic nature of these approaches, many methods resort to situation-specific test environments with different objects, backgrounds, viewpoint choices, lighting conditions, etc. To address this problem and to set a standard way of testing NBV methods for object recognition, we gathered a dataset, particularly for benchmarking of active object recognition techniques, with which the proposed next best view method was evaluated. It should be noted that there are other useful datasets, such as those presented in [6] and [26] in the literature that are not suitable to test our method since such evaluation necessitates availability of both color and depth images of objects taken from the front and sides. For example, the dataset in [6] does not offer real world colors of objects to provide the texture clues needed in our work. To imitate real-world conditions, the initial views of objects were intentionally distorted in parts of the tests. This ensures occurrence of uncertain initial recognition situations that cause an AOR system to trigger. In addition, the tests were performed on AOR systems with different classifiers and fusion methods to ensure the results are not biased for a specific type of AOR system.

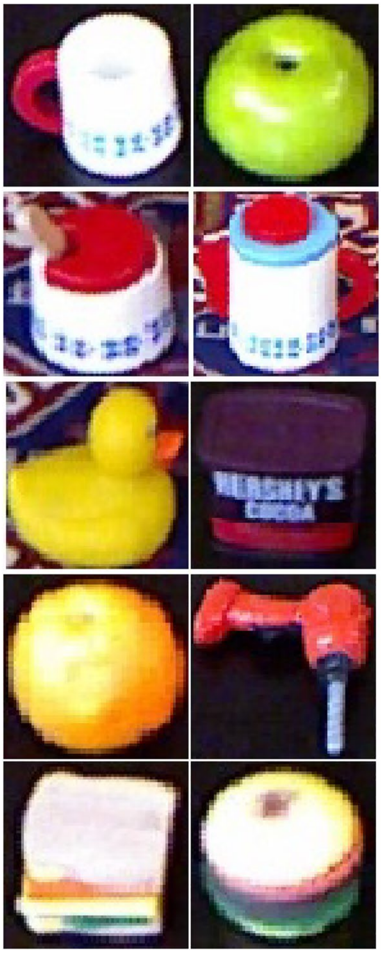


Fig. 4 The 10 objects in the dataset

Test Dataset for Active Object Recognition

We collected 240 test situations, generally for evaluating active recognition systems, especially next best view methods. 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). There are seven images and their corresponding depth maps in each situation: one for a frontal initial view, another for an initial view with a slightly higher altitude, and five others for the images/depth maps taken from the sides of objects as follows: left, top left, top, top right, and right. Figure 5 illustrates a sample situation for one of the objects in the dataset. A calibrated Microsoft Kinect v1 3D camera, with a resolution of 640 × 480 pixels was used to capture both the color images and depth maps of objects and their immediate surrounding. The objects were then

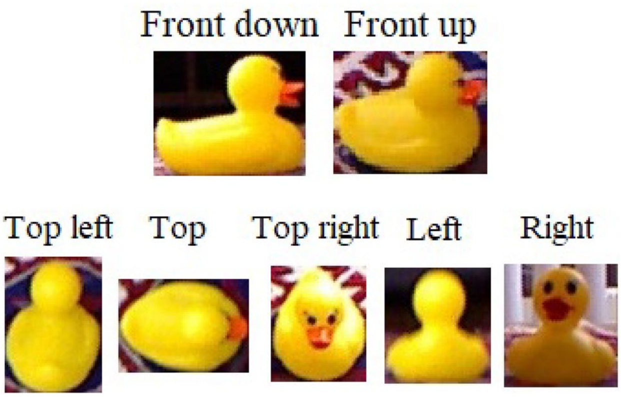


Fig. 5 A sample situation in the dataset

labeled and cropped from the scenes. The quality of all the captures, their cropping and labeling, and heterogeneity of test samples according to the above-mentioned distribution of pose, lighting condition, and background texture were separately verified. Because the objects were placed on the ground or on opaque hard surfaces during the photo shoot 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 the test. The dataset is published along with the current paper[1].

Emulating AOR Triggering Conditions

The initial views in the test dataset are clear and unobstructed. Nonetheless, active object recognition systems are typically employed when the classifiers suffer reduced performance due to occlusion or unfavorable perspective of objects. For this reason, to emulate such conditions that trigger AOR and mix it with the cases that the object image is clear, in our test benchmarks, we used each test situation in the dataset once with severe distortions, next with milder alterations, and lastly without any changes. 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% (for severe alteration) and 40% (for mild alteration) of the length and width, totaling 36% (for severe alteration) and 16% (for mild alteration) of the area of the original image.

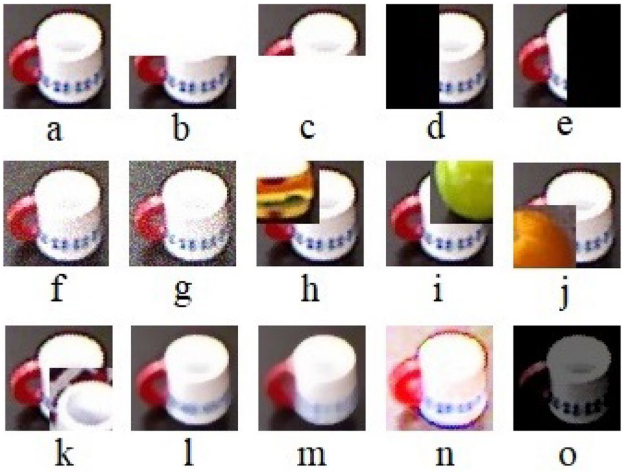


Fig. 6 Initial view distortions. **a** Original image, **b**, **c** top/bottom whiteout, **d**, **e** left/right blackout, **f**, **g** lighter/heavier noise, **h**–**k** corner superimpose, **l**, **m** lighter/heavier blur, **n** bright, **o** dark

2. A half (for severe alteration) or 30% (for mild alteration) 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 (for severe alteration) or 3×3 (for mild alteration) kernel and the other with a 9×9 (for severe alteration) or 7×7 (for mild alteration) kernel.
4. Added noise with standard deviations of 20 and 30 (for severe alteration) or 15 and 25 (for mild alteration) in the 8-bit color images.
5. Image darkening and brightening by 150 (for severe alteration) or 100 (for mild alteration) levels.

The tests were performed on the original images and their altered versions as well as their corresponding depth maps, totaling 29 test scenarios for any test situation in the dataset, of which 14 cases experience severe alterations, another 14 undergo mild changes, and the last one, with an arbitrary weight factor of 17, is unaltered. Figure 6 shows the 14 severe alterations of the initial view for an object, plus the unaltered one, in a sample situation.

Test Benchmarks

Since there are two initial images in each test situation in the dataset, two experiments can be performed for a single situation. As mentioned in the former section, for each initial image 29 test scenarios with various alterations are possible. Therefore, 58 tests are performed for any test situation. With the existence of 240 test situations in the dataset, 13,920 situations were evaluated for any single vision system in the tests.

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

To ensure that the proposed NBV is independent of any specific classifier or fusion algorithm in the AOR system, six different classifiers and three fusion techniques were examined in order 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, Naïve Bayes [3], and Dempster-Shafer (DS) [2] fusion algorithms are used in the tests. The classifiers are:

- ResNet 101, a residual neural network with 101 layers. The convolutional layers are reused by transfer learning with weights pre-trained on the ImageNet dataset. After the convolutional layers, we added a global average pooling and two dense layers that are trained with the small training dataset of objects we had gathered.
- 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 500. The loss function is categorical cross-entropy, which is optimized by the Adam optimizer.
- A one-versus-rest non-linear Support Vector Machine (SVM) classifier with the feature vector comprised of Hu moment invariants 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 be 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 similar SVM classifier as above, but instead with a sigmoid kernel and without Hu moment invariants as features.
- A Random Forest (RF) with 150 decision trees and a split criterion of the Gini impurity that uses a bag of 150 visual words of Scale-Invariant Feature Transform (SIFT) 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.
- A random forest classifier as above, but with KAZE keypoint descriptor instead of SIFT.

Considering the possible combinations of the classification and fusion approaches, 18 benchmarks were evaluated, each with 13920 situations tested. In the tests, the confidence threshold of the AOR was set to 20, except for those tests needing a sweep of the threshold value. That means the second viewpoint is retrieved if the highest class probability of the initial view is less than 20 times of the second highest one.

Obtained Results

We evaluated the proposed NBV by comparing its suggested selections with the other viewing poses in each test case. The results obtained by the ensemble method are also compared to its constituting members and a few other methods. Among the other methods being compared with, negative entropy (8) of intensity histogram and classifications can be mentioned. For negative entropy of intensity histogram, z_i in (8) represents a random variable denoting an intensity and $p(z_i)$ is the probability of that in the intensity histogram with L bins. Similarly, for negative entropy of tile classifications, L in (8) is the number of object classes, z_i means an object class, and $p(z_i)$ is the probability of that class that is generated by a classifier.

$$n(z) = \sum_{i=0}^{L-1} p(z_i) \log p(z_i) \tag{8}$$

Further, the energy-efficiency of the proposed method and changes to the Receiver Operating Characteristic (ROC) curves are evaluated in the following. As mentioned in "Classification Dissimilarity", to gauge classification dissimilarity we used dedicated classifiers for each tile. Yet, we report the same metric with a shared classifier for all the classifications.

Tile Preference Ranking

The five prospective next viewpoints are examined in every test situation for the scores they get from every criterion. In Fig. 7, the tiles are sorted on the horizontal axis in an ascending order of the scores, and therefore preference 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. Hence, it is desirable to have lower height of the bars in the right sides of the plots in Fig. 7. For example, for the

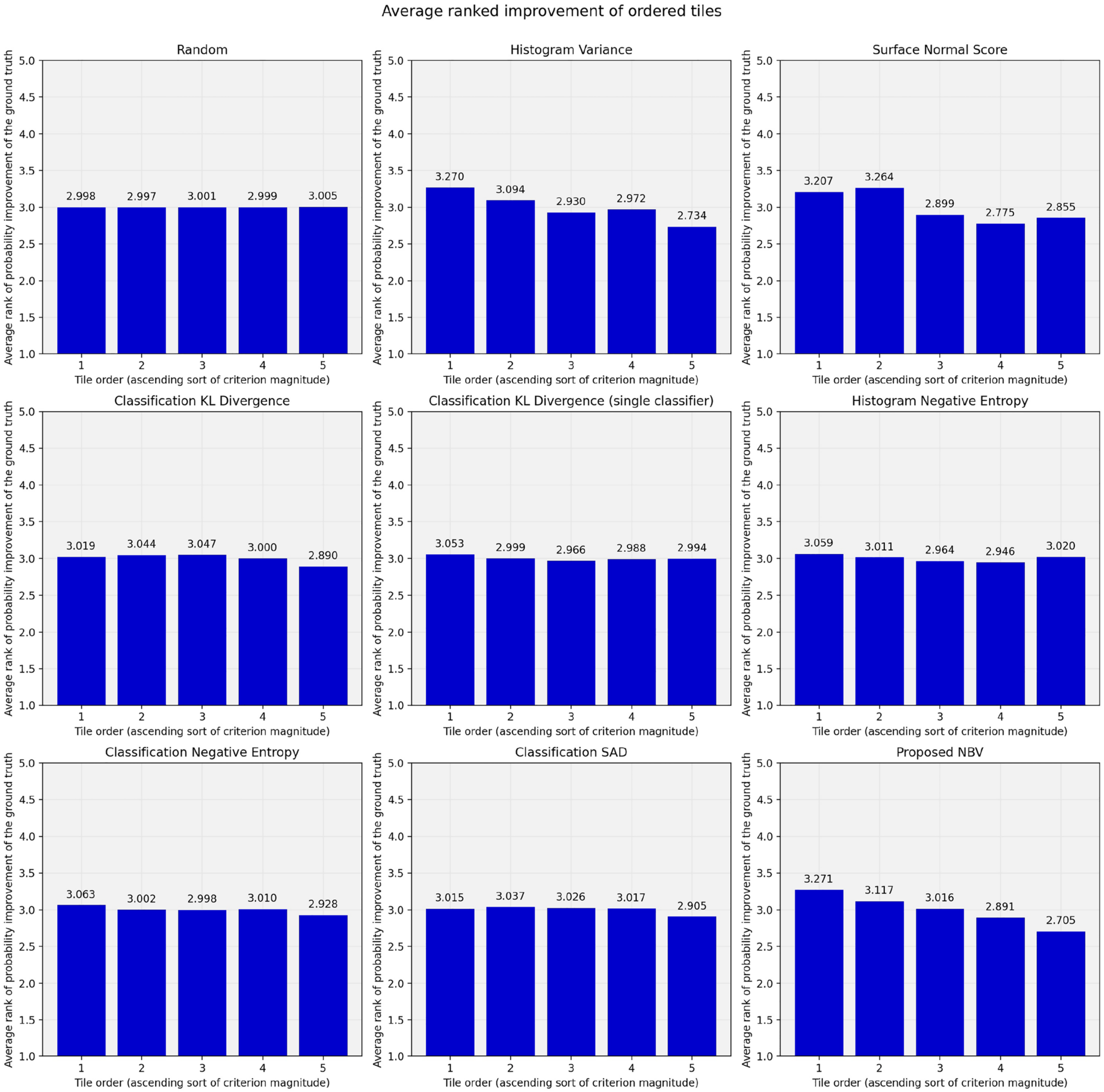


Fig. 7 Average ranked improvement of tiles in ascending order of scores

proposed method, the mean rank of the third highest scoring tiles (represented by the middle bar) is 3.016 and the average rank of the highest scoring tiles (the rightmost bar) is 2.705, which is comparatively lower and better.

It can be seen from the results that the proposed NBV method attains better ranks for the tiles it emphasizes more and scores higher. It shows that it is effectual in selecting the viewpoints that offer the best improvement in probability of the true class in the AOR system's output. Figure 7 also shows the performance of randomly selecting the next view and the individual criteria that are part of the ensemble. All the participating criteria in the ensemble almost prefer better tiles and tend to bring the height of their very right bar down. Despite the efficacy of each single proposed criterion, the combination of them still yields better outcome that causes sharper decline in the height of the right bars.

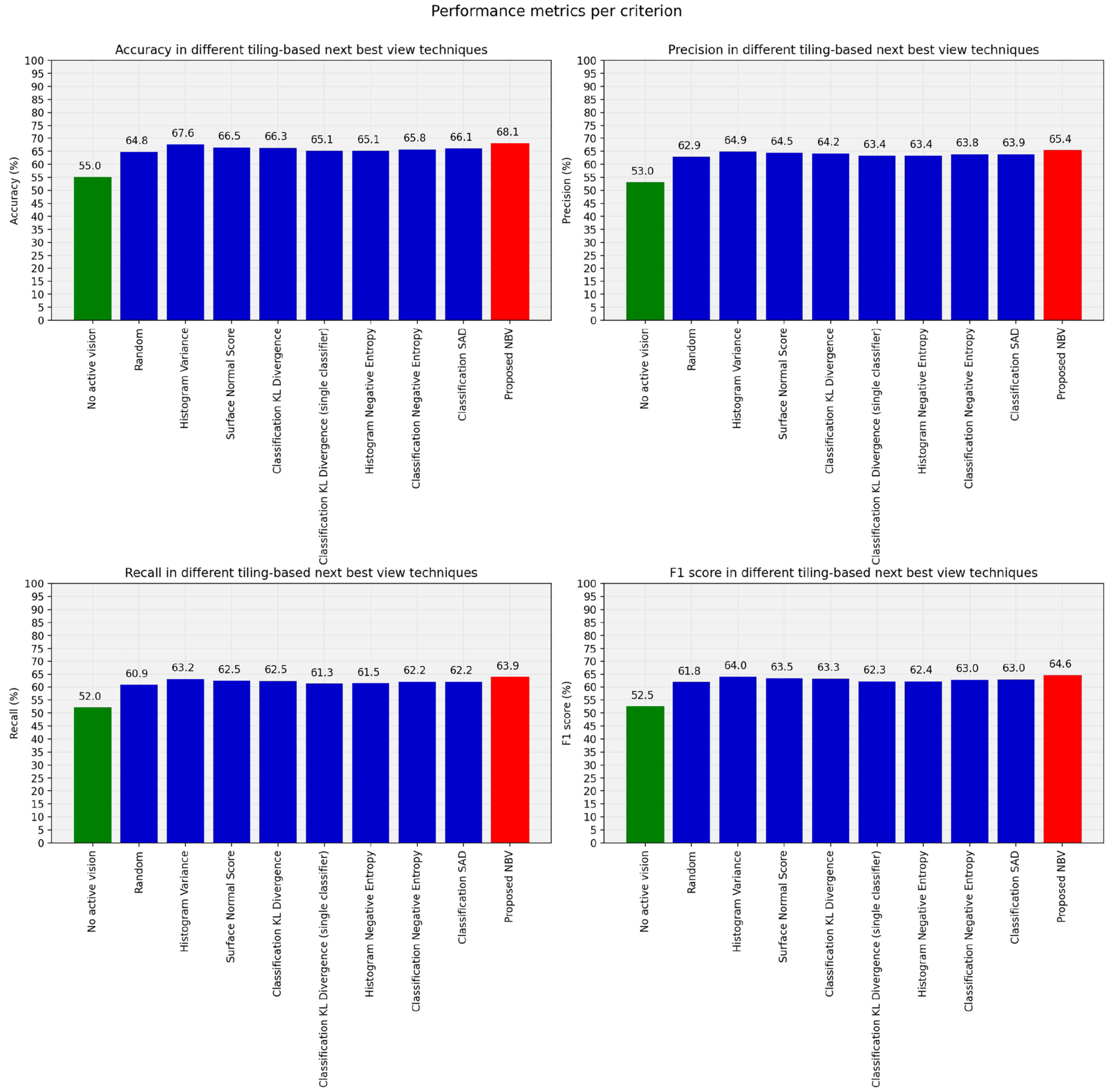


Fig. 8 Performance metrics per measure

Performance Per Criterion

We evaluated the proposed system by measuring accuracy, precision, recall, and F_1 score it obtains. Figure 8 shows the results, along with those obtained from individual criteria in the ensemble, random next view selection, and the other measures for comparison. In addition, Fig. 9 delineates the performance improvement ratio of each measure compared to their initial values with no active vision. It is obvious that the proposed method achieves higher improvements than the other methods in the figure, including random selection of next viewpoint.

Performance Improvement Per Tile

The accuracy, precision, recall, and F_1 score improvement of the AOR system by using any of the five possible tiles in the tests are shown in Fig. 10. The tiles are sorted in the horizontal axis based on the scores they receive from each

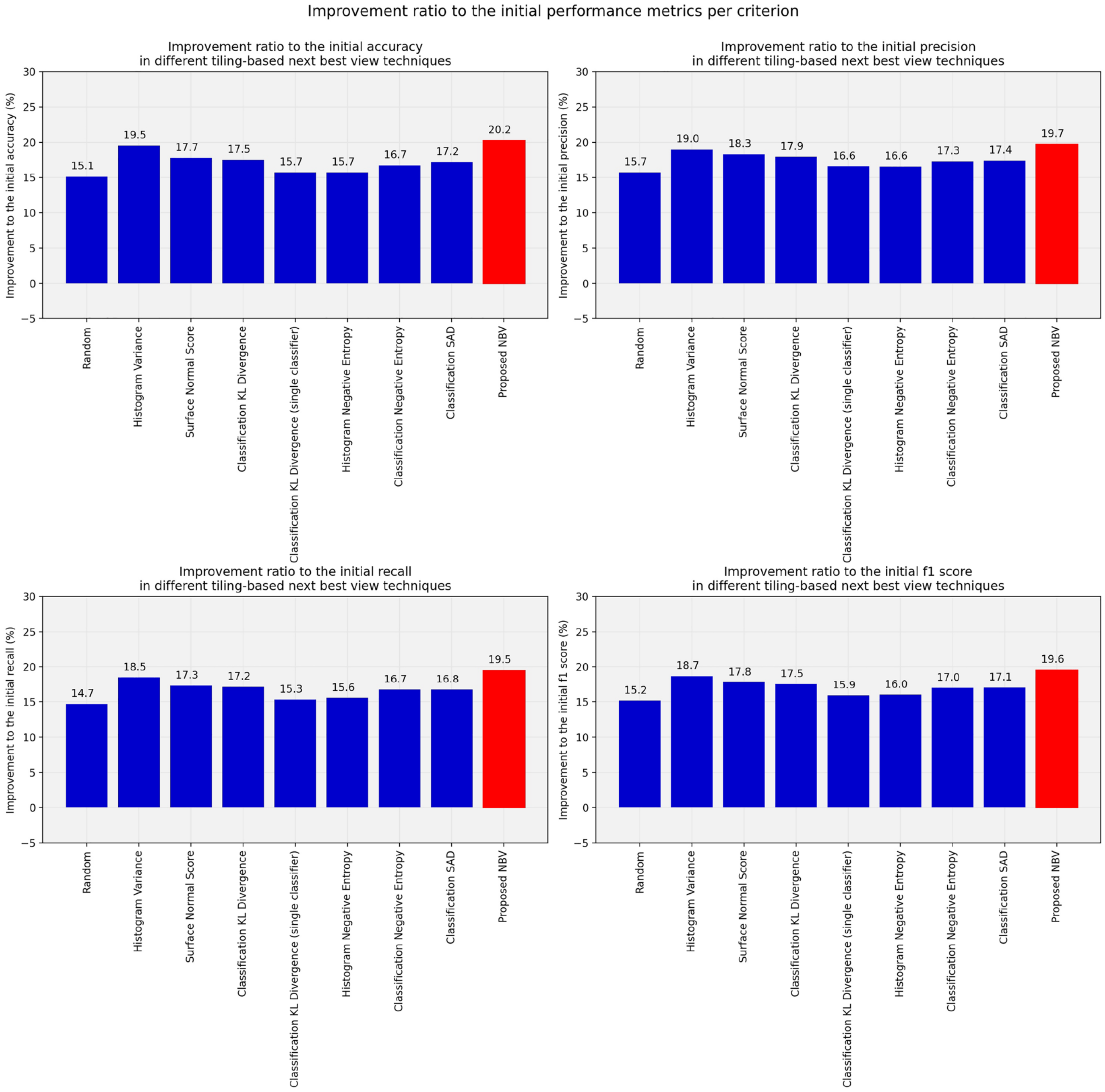


Fig. 9 Improvement ratio to the initial performance metrics per measure

measure. Greater performance improvements are expected for the tiles the NBV system emphasizes more, i.e. the ones with higher scores at the right side of each plot. The results prove that the proposed NBV is successful in obtaining higher performance indices in its top picks. The individual measures participating in the ensemble also demonstrate a trend of increasing accuracy, precision, recall, and F_1 score with the higher preference they indicate.

Receiver Operating Characteristic (ROC) Curve

For the 18 benchmarks in the tests with different classifiers and information fusion techniques, the ROC curves and their area under the curve (AUC) with micro-averaging over the mildly altered test situations are shown in Fig. 11.

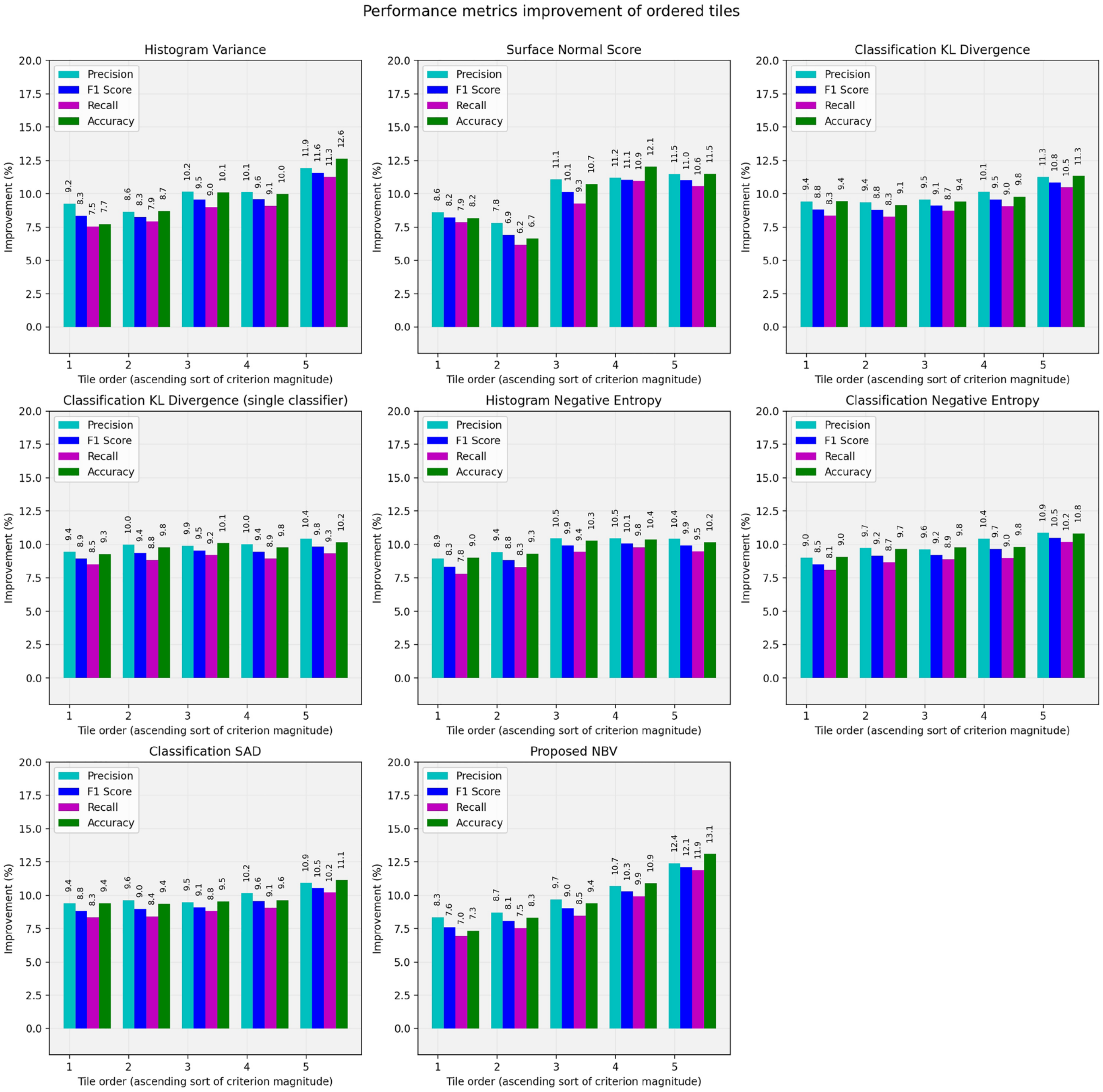


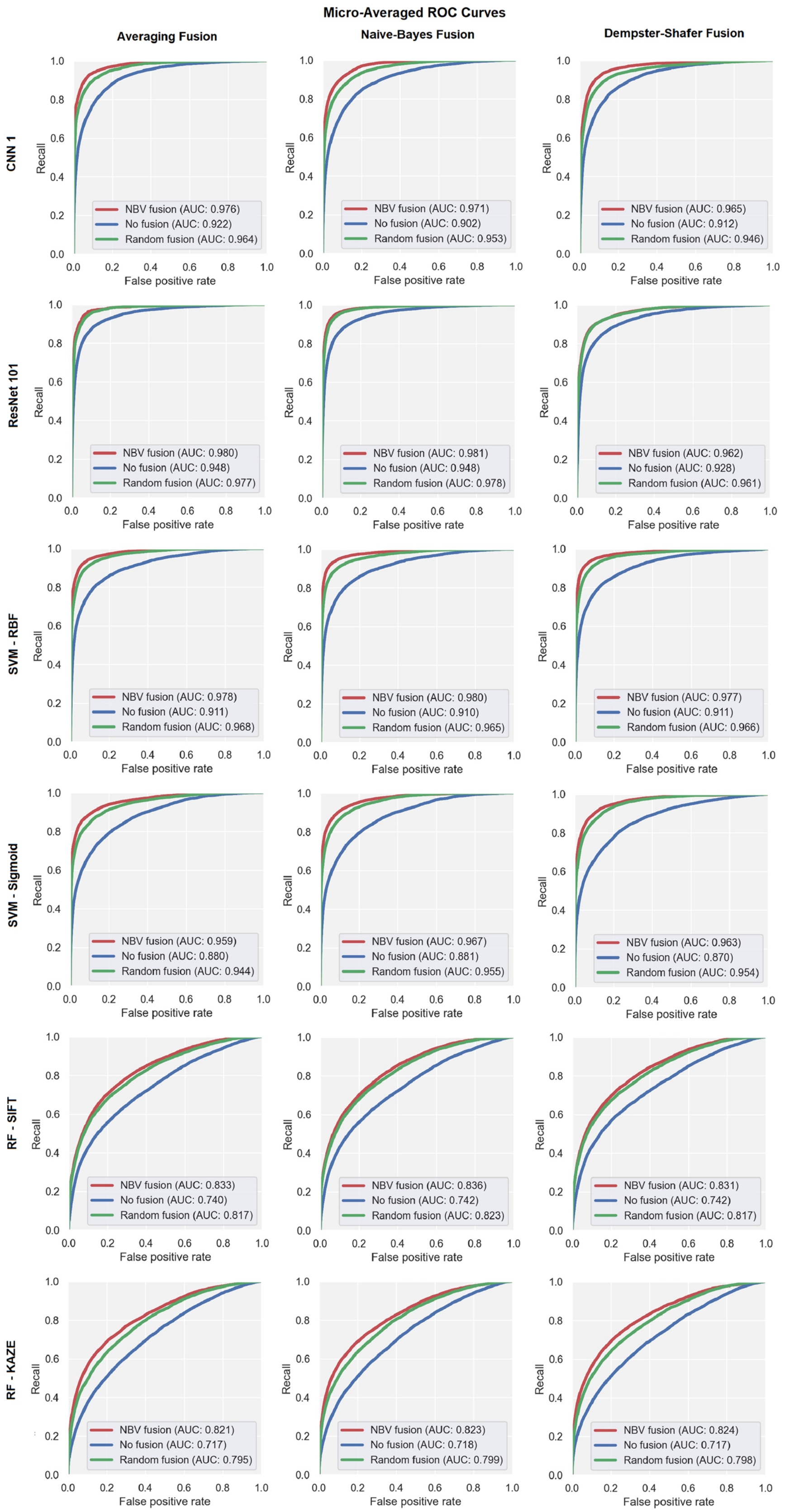
Fig. 10 Performance metrics improvement of tiles in ascending order of scores

The blue curves in the figure, represent the recognition results of the initial view only, while the green curves indicate the effect of fusing with a randomly selected view. The red curves also show the results of utilizing the proposed method. Comparing the three sets of the curves verifies the efficacy of the AOR system in ameliorating the ROC curve and of the NBV method in further enhancing it.

Performance Per Confidence Threshold

Confidence threshold defines the sensitivity of an AOR system to uncertainties in its initial viewpoint recognition. Higher confidence thresholds indicate the tendency of the active vision system to retrieve new viewpoints. It is an influencing parameter of the AOR system, that balances the effort to fetch new viewpoints of objects with the improvement in recognition. Figure 12 illustrates the performance metrics improvement over different confidence

Fig. 11 ROC curves of different test benchmarks



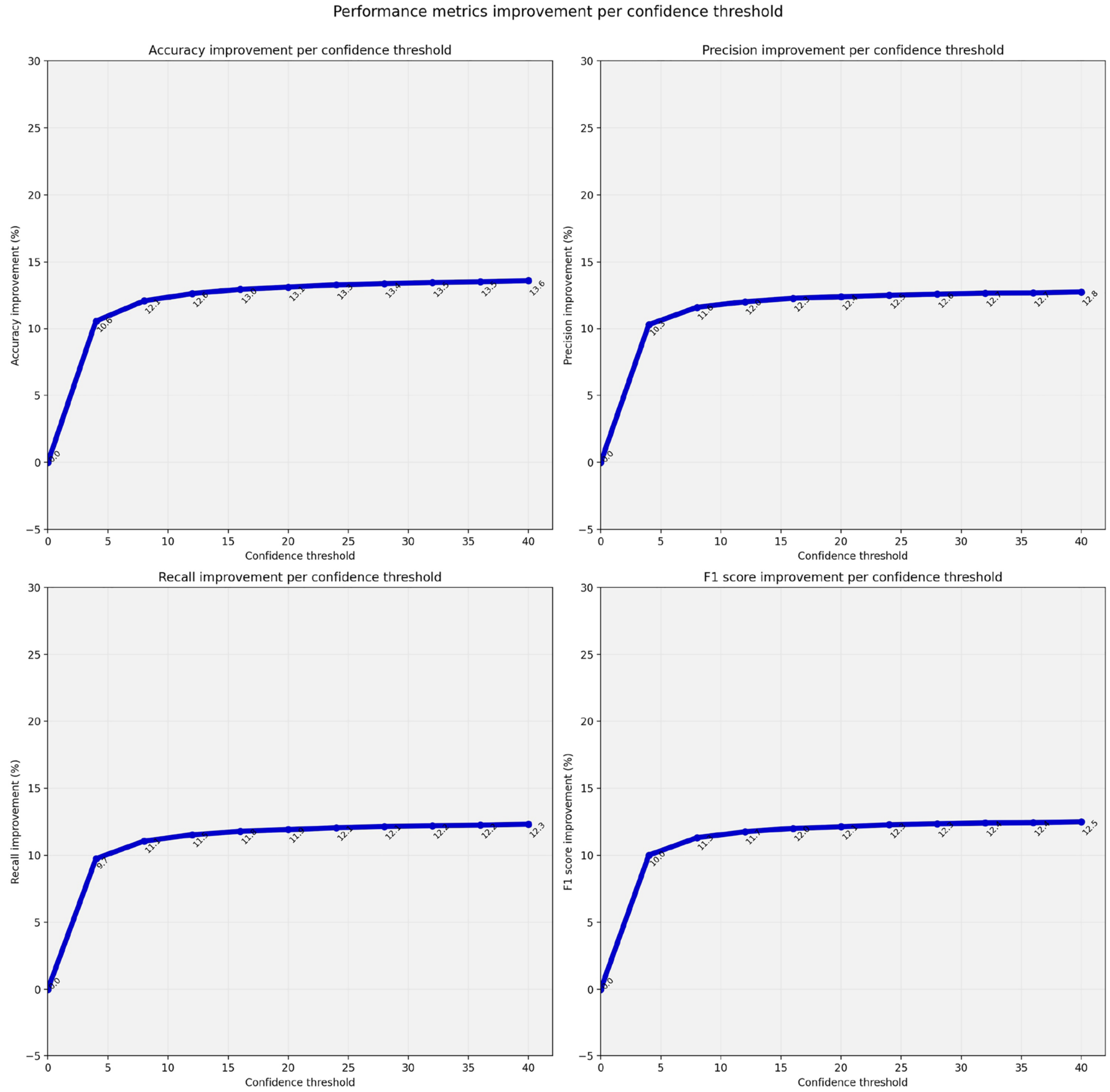


Fig. 12 Performance metrics improvement over different confidence thresholds

Table 1 Summary of differences of the proposed method with random next viewpoint selection and no active vision

Method	Improvement % over no active vision				Average AUC
	Accuracy (%)	Precision (%)	Recall (%)	F_1 Score (%)	
Proposed NBV	20.2	19.7	19.5	19.6	0.923
Random selection	15.1	15.7	14.7	15.2	0.910
No active vision	N/A	N/A	N/A	N/A	0.849

thresholds. Accuracy, precision, recall, and F_1 score were originally 0.55, 0.53, 0.52, and 0.52, respectively at the 0% improvement point in Fig. 12. From the figure, it is obvious 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 relatively high quality initial recognition outputs are not regarded good enough and would require a decision fusion after an active vision procedure. This overreach probably does not contribute to improved classifications as they were performing well from the beginning, but only increasing the energy and time costs of physical camera movements. In contrast, in the lower confidence thresholds, we observe larger enhancements in the performance.

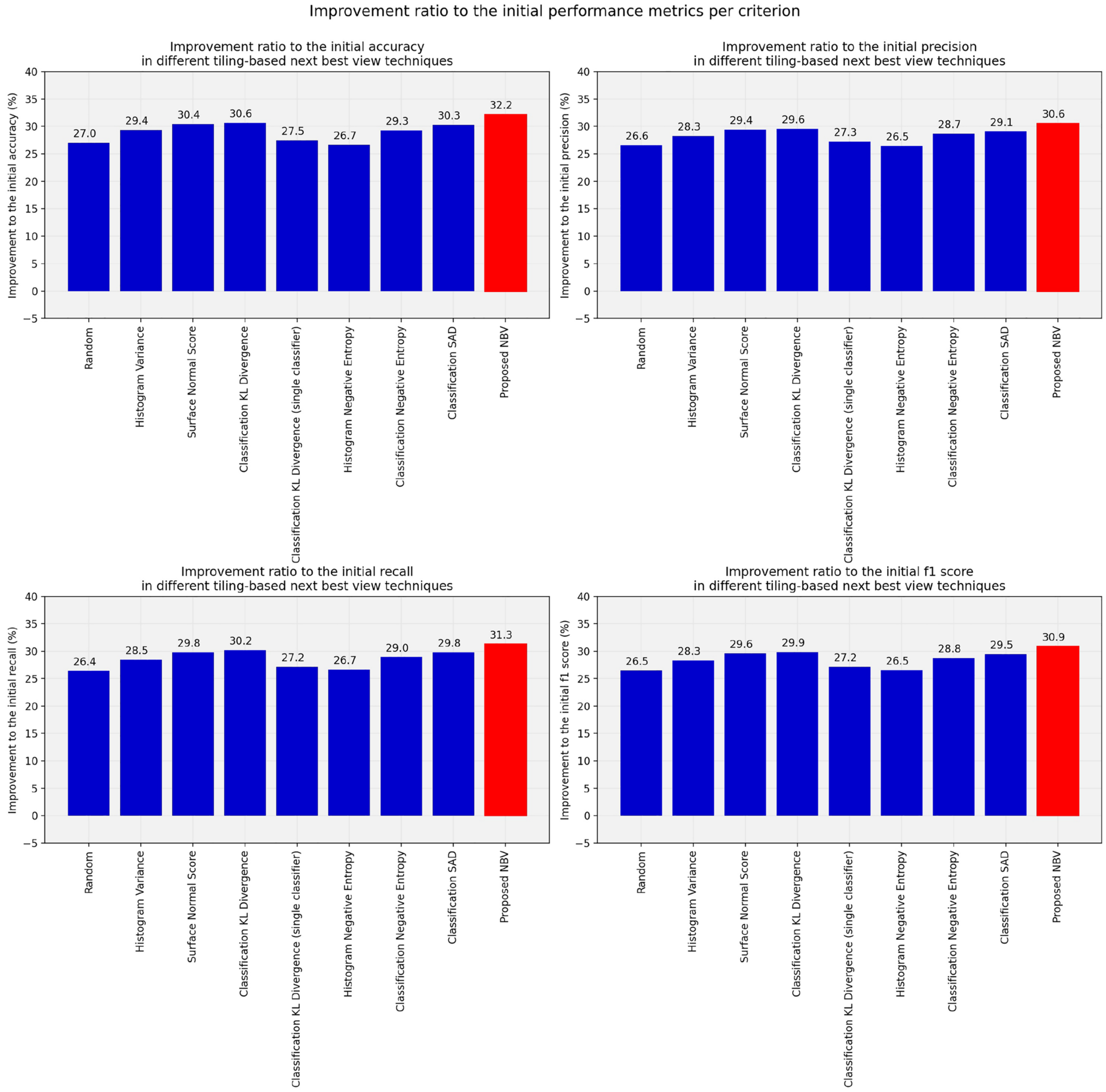


Fig. 13 Improvement ratio to the performance metrics of initial viewpoints with severe adverse artifacts per measure

This underpins the importance of active object recognition and next best view as crucial improvement steps for the smaller group of classifications with high uncertainty.

Discussion

Table 1 summarizes the improvements in accuracy, recall, precision, and F_1 score compared to no active vision. It also shows the average AUC difference of the proposed NBV with random selection of next viewpoint and no active vision. By observing the experimental results it is clear that the proposed NBV is applicable for improving accuracy, recall, precision, and thus F_1 score of the active object recognition systems. In the results, we observe that the active object recognition systems with a random selection of next viewpoint attain 15.1% and 15.2% accuracy and F_1 score improvement on average compared to their initial performance values. It denotes the effectiveness of AOR in general. With the proposed next best view method, the same AOR systems fulfill 20.2% and 19.6% relative accuracy and F_1 score enhancements, which amounts for 5.1% and 4.4% further improvement over a random AOR. It is also worth mentioning that in the case of severe occlusions and other adverse artifacts in the initial view, which are usually the reason an AOR system seeks a new viewpoint, the proposed NBV method improves the performance metrics even more. While Fig. 9 illustrates the improvement ratios per each criterion, averaged over non-altered clear initial views, along with mildly or severely altered ones, Fig. 13 shows improvement ratios only in the case of initial views with severe adverse artifacts (i.e. stronger alterations). From the figure, we see 32.2% and 30.9% relative improvement in accuracy and F_1 score.

An interesting aspect of the tile ranking using the foreshortening score in Fig. 7 is that sometimes the tiles with the penultimate score reach better ranks than the highest scoring ones. Those cases occur perhaps when the higher scoring tile has a very steep object surface with respect to the image plane of the camera in the initial view. Compared to less steep surfaces, a very steep one may impede the proper view of the respective object side from the perspective of the initial view.

It should be also 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. Moreover, from the two figures, it can be inferred that the proposed KL divergence method works better compared to the ones based on negative entropy or sum of squared differences of classifications. 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. Although the proposed NBV technique is fairly simple and lightweight, for speedier NBV deductions, one can drop the classification dissimilarity, which is comprised of classifying the peripheral tiles and computing KL divergence for them, 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.

The proposed NBV relies on splitting bounding boxes in the initial viewpoint into several tiles. The tiling scheme controls the granularity of the next viewpoint options. The tiling strategy is also founded on the supposition that any object surface is visible in the tiles around the object bounding boxes. Even though, it is a fairly reasonable assumption, there is no guarantee for that in every tile.

In comparison to a single recognition stage, the recognition performance improvements of the proposed NBV method come with some computational overhead. This is common trait among all active object recognition systems, since an active vision system conducts more observations than a traditional single-frame vision system. Nevertheless, the proposed NBV method is computationally light compared to many other active vision systems [6, 10, 12, 13, 15, 18]. It performs merely one more move to complete the task of object recognition, while to decide for that single next best move it does not demand a series of images taken beforehand; it uses just the current view. The three criteria of the ensemble method to analyze the current view are also not computation intensive relatively. Histogram variance (second moment), which processes the histogram of a tile instead of its whole image, is intrinsically faster than local image processing methods. Foreshortening 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 two criteria. It consists of a combination of classifications that are performed very fast in modern computers. It should be noted that, depending on the robotic system being employed and the application, the most time-consuming part of an active vision system is probably the physical camera movement.

Conclusion

In this paper a next best view approach for active object recognition systems was presented. It divides an initial image of an object into several areas to analyze each one for clues of better next views. An ensemble of three different techniques is used to examine each area: foreshortening, histogram variance, and classification dissimilarity. The proposed method suggests the next viewpoint after taking the appearance and 3D shape information of merely a single initial view. It also does no need a training set of specific views of objects or their 3D models.

In order to evaluate the proposed approach in standard and reproducible way, a dataset was gathered, which can be used by other researchers in the future to test their active object recognition systems. The experimental results verified efficacy of the presented next best view approach in improving accuracy, recall, precision, and F_1 score. On average, 20.2% and 19.6% improvement in accuracy and F_1 score compared to classifications of the initial view was achieved. As a part of an active vision system, the proposed next best view technique becomes more powerful in improving the object recognition performance in the presence of heavy occlusions and other unfavorable conditions in the initial view. In comparable initial view conditions, we report 32.2% and 30.9% average accuracy and F_1 score improvements compared to the initial performance values.

Future efforts can be directed toward investigating alternative tiling modes of the initial view. In addition, to make sure that every tile in the ensemble contains the object's surface, it is possible to detect presence of the object's surface in a tile and dismiss those candidate tiles without any object surface from the computations of the proposed ensemble. Another promising area of work in this direction can be exploring other ensemble methods in lieu of the current weighted voting strategy. A potentially interesting way to combine the tile scores would be a meta-learning approach to automatically digest the scores from the proposed criteria and weight them on the basis of different environmental factors.

Acknowledgements 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.

Author Contributions Conceptualization: PH, MN, MN; methodology: PH, MN, SKP; formal analysis and investigation: PH, MN, MN; writing—original draft preparation: PH; writing—review and editing: PH, SKP, MN; funding acquisition: MN, MN; resources: MN, MN, PH, SKP; supervision: MN, MN.

Funding Information 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.

Data Availability Yes. Dataset available at https://github.com/pouryahoseini/Next-Best-View-Dataset.

Code Availability Yes. Code available at https://github.com/pouryahoseini/Next-Best-View.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Bajcsy R, Aloimonos Y, Tsotsos JK. Revisiting active perception. Autonom Robots. 2018;42(2):177–96.
2. Hoseini P, Blankenburg J, Nicolescu M, Nicolescu M, Feil-Seifer D. An active robotic vision system with a pair of moving and stationary cameras. In: International symposium on visual computing. Springer; 2019. p. 184–195.
3. Hoseini P, Blankenburg J, Nicolescu M, Nicolescu M, Feil-Seifer D. Active eye-in-hand data management to improve the robotic object detection performance. Computers. 2019;8(4):71.
4. Hoseini P, Paul S.K, Nicolescu M, Nicolescu M.N. A surface and appearance-based next best view system for active object recognition. In: VISIGRAPP (5: VISAPP). 2021. p. 841–851.
5. Hoseini P, Paul SK, Nicolescu M, Nicolescu M. A one-shot next best view system for active object recognition. Appl Intell. 2021;1–20.
6. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J. 3d shapenets: a deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1912–1920.
7. Zeng R, Zhao W, Liu Y-J. 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; 2020. p. 7050–7057
8. Wei W, Yu H, Zhang H, Xu W, Wu Y. Metaview: few-shot active object recognition. 2021. arXiv preprint arXiv:2103.04242.
9. Barzilay O, Zelnik-Manor L, Gutfreund Y, Wagner H, Wolf A. From biokinematics to a robotic active vision system. Bioinspir Biomimet. 2017;12(5):056004.
10. Atanasov N, Sankaran B, Le Ny J, Pappas GJ, Daniilidis K. Nonmyopic view planning for active object classification and pose estimation. IEEE Trans Robot. 2014;30(5):1078–90.
11. Potthast C, Sukhatme GS. Next best view estimation with eye in hand camera. In: IEEE/RSJ intl. conf. on intelligent robots and systems (IROS). Citeseer; 2011.
12. Potthast C, Sukhatme GS. A probabilistic framework for next best view estimation in a cluttered environment. J Vis Commun Image Rep. 2014;25(1):148–64.
13. Doumanoglou A, Kouskouridas R, Malassiotis S, Kim T-K. Recovering 6d object pose and predicting next-best-view in the crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 3583–3592.
14. Rebull Mestres J. Implementation of an automated eye-in hand scanning system using best-path planning. Master's thesis, Universitat Politècnica de Catalunya. 2017.
15. Bircher A, Kamel M, Alexis K, Oleynikova H, Siegwart R. Receding horizon " next-best-view" planner for 3d exploration. In: 2016 IEEE International conference on robotics and automation (ICRA). IEEE; 2016. p. 1462–1468.
16. Lehnert C, Tsai D, Eriksson A, McCool C. 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; 2019. p. 3890–3897.
17. Cui J, Wen JT, Trinkle J. A multi-sensor next-best-view framework for geometric model-based robotics applications. In: 2019 International conference on robotics and automation (ICRA). IEEE; 2019. p. 8769–8775.
18. Jia Z, Chang Y-J, Chen T. A general boosting-based framework for active object recognition. In: British machine vision conference (BMVC). Citeseer; 2010. p. 1–11.
19. Edmonds M, Yigit T, Yi J. 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). IEEE; 2020. p. 863–868.
20. Xu Y, Hu J, Wattanachote K, Zeng K, Gong Y. Sketch-based shape retrieval via best view selection and a cross-domain similarity measure. IEEE Trans Multimedia. 2020;22(11):2950–62.
21. Lauri M, Pajarinen J, Peters J, Frintrop S. Multi-sensor next-best-view planning as matroid-constrained submodular maximization. IEEE Robot Autom Lett. 2020;5(4):5323–30.

22. Morrison D, Corke P, Leitner J. Multi-view picking: next-best-view reaching for improved grasping in clutter. In: 2019 International conference on robotics and automation (ICRA). IEEE; 2019. p. 8762–8768.
23. Almadhoun R, Abduldayem A, Taha T, Seneviratne L, Zweiri Y. Guided next best view for 3d reconstruction of large complex structures. Remote Sens. 2019;11(20):2440.
24. Palomeras N, Hurtós N, Vidal E, Carreras M. Autonomous exploration of complex underwater environments using a probabilistic next-best-view planner. IEEE Robot Automat Lett. 2019;4(2):1619–25.
25. Gonzalez RC, Woods RE. Digital image processing. 4th ed. London: Pearson; 2018.
26. Ammirato P, Poirson P, Park E, Košecká J, Berg AC. A dataset for developing and benchmarking active vision. In: 2017 IEEE international conference on robotics and automation (ICRA). IEEE; 2017. p. 1378–1385.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.