Unsupervised Learning Approach to Attention-Path Planning for Large-scale Environment Classification

Hosun Lee, Sungmoon Jeong, and Nak Young Chong

Abstract— **An unsupervised attention-path planning algorithm is proposed and applied to large unknown area classification with small field-of-view cameras. Attention-path planning is formulated as the sequential feature selection problem that greedily finds a sequence of attentions to obtain more informative observations, yielding faster training and higher accuracies. In order to find the near-optimal attention-path, adaptive submodular optimization is employed, where the objective function for the internal belief is adaptive submodular and adaptive monotone. First, the amount of information of attention areas is modeled as the dissimilarity variance among the environment data set. With this model, the information gain function is defined as a function of variance reduction that has been shown to be submodular and monotone in many cases. Furthermore, adapting to increasing numbers of observations, each information gain for attention areas is iteratively updated by discarding the non-informative prior knowledge, enabling to maximize the expected information gain. The effectiveness of the proposed algorithm is verified through experiments that can significantly enhance the environment classification accuracy, with reduced number of limited field of view observations.**

I. Introduction

There has been an increase in the demand for smarter robots that perform a variety of tasks autonomously in different environments. First of all, robots need to obtain enough information about the environment to reach appropriate decisions and actions. However, it is difficult for robots to perform a full-scale measurement of the environment at once, in particular, when they are exploring large areas. Therefore, multiple view images are required to capture data for an entire environment, concatenated into a single high-dimensional data. Over the years, as an effort to avoid dealing with very high dimensional data, many approaches have been tried with low-dimensional data. In this paper, a new learning technique is proposed to sequentially discover the most informative data from large-scale data sets with possible applications to unknown area classification.

Over the years, many mobile robot navigation problems deal with informative path planning for search-and-rescue tasks under limited time or battery capacity [1][2]. The path of robot is optimized to obtain the maximum information required to perform a given task. Likewise, the active vision [3] has been studied to plan proper actions for the optimal observation such as viewpoint planning and saliency map [4][5]. However, they assume full access to the whole information without limited sensing coverage consideration.

H. Lee, S. Jeong, and N. Y. Chong are with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan {Hosun_LEE, jeongsm, nakyoung}@jaist.ac.jp

There have been similar studies on human attention-path planning which show efficient eye movement for maximizing local information [6] [7]. This attention-path planning problem under limited sensing coverage can be described as a sequential feature selection problem [8] [9]. At each time step, the system has to sequentially decide the most informative sequence based on the observed feature set and the internal belief of the target. Notably, conventional sequential feature selection is a supervised learning approach which fully trains the classifier with an associated class label. However, manual labeling of a large set of training data causes a severe bottleneck in robotic applications. In practice, selecting the informative feature without labels is a quite challenging problem. Recently, the property of adaptive submodularity is investigated to ensure the bounded performance of the informative path planning [10] [11]. Based on this property, we attempt to solve the attention-path planning problem in the unsupervised scenario.

In this work, we discuss the attention-path planning with small field-of-view cameras for large area classification. We assume that the robot can obtain the area information partially at each step by selecting proper fixations; attention can be defined by a sequence of fixations. Then, our goal is how to plan the next fixation based on the current observation and unlabeled training datasets called prior knowledge to enhance the robot's uncertain environment perception capability. There are two main components of attention-path planning: (1) top-down approach for sequential selection of fixations using the prior knowledge to maximize the information gain, and (2) bottom-up approach for the prior knowledge update by discarding the non-informative subset of the prior knowledge according to current observation data. First, the information is defined as the amount of uncertainty reduction, and an information gain function is modeled to measure the expected information gain, and maximized using the proposed adaptive submodular optimization framework. Adaptive submodular optimization can be a great help to solve the problem of sequential feature selection without any fully trained classifier. Secondly, a cascade nearest neighbor classifier is implemented to recursively update the prior knowledge depending on the previously selected observation. In each time step, non-informative data of the prior knowledge is discarded by calculating the dissimilarity between the unlabeled dataset and current observation. After then, the sequentially modified prior knowledge is simultaneously used to update the information gain of each attention area for selecting the next fixation and classify the uncertain environment with a nearest neighbor classifier.

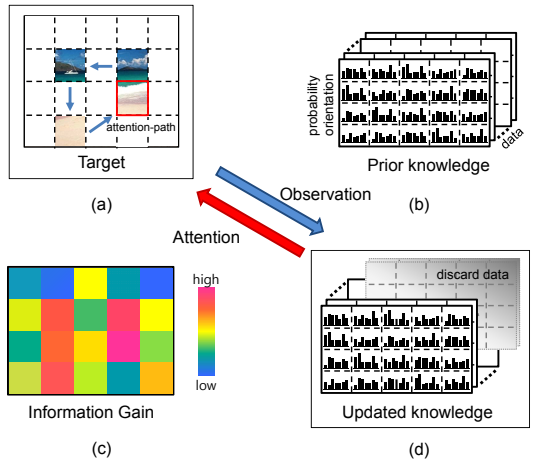


Fig. 1. Concept of attention path planning: (a) example of sequential attention-path, (b) unlabeled training dataset (prior knowledge), (c) information gain of each attention data, (d) update of prior knowledge according to dissimilarity between observed data and prior knowledge. Attention: top-down approach to select a fixation using prior knowledge, Observation: bottom-up approach to update the prior knowledge based on observed data

In summary, the main contributions of this paper are as follows:

- We propose an attention planning framework for large scale area classification with limited field of view.
- We can achieve a near optimal solution not only for the sequence of sensing, but the classification accuracy applying adaptive submodular optimization. In the attention planning process, the most informative fixation is selected by the prior knowledge and then non-informative data is discarded for the next fixation selection, comparing with the current observation. This procedure can guarantee the bounded optimal attention, and the retrieved prior knowledge improves the classification performance.
- The proposed algorithm can be applied to unsupervised sequential feature selection, while conventional algorithms rely on supervised learning approaches.

II. Problem Statement

Attention-path planning can be regarded as the sequential feature selection [8], where the most informative feature is gathered from a set of features at each iteration. The feature selection is decided depending on the obtained feature set and the internal belief. For the attention-path planning problem, a fixation is selected at each time step and used to observe a subset of the target dataset. The observed target data is finally classified with the given training class set, $\mathbb{C} = \{c_1, c_2, \cdots\}$.

Attention and Observation At each iteration, we can select a fixation a from the whole set of possible attention $\mathbb{A} = \{a_1, \cdots, a_{n\times m}\}$, where the target image is divided into $n \times m$ rectangular blocks according to the range of the sensing coverage as shown in Fig. 1(a). Each fixation is represented with a position vector $[i, j]^T$, where i and j are the indices on the column and row directions. By selecting a fixation, a local feature placed on the same position and size as the fixation is observed from the target feature set of the target data $\mathbb{O} = \{o_1, \cdots, o_{n\times m}\}$ which is not available to access before the selection of the fixation. Accordingly, a selected fixation set $A \subset \mathbb{A}$ and an observed feature set $O_A \subset \mathbb{O}$ are accumulated sequentially.

Prior knowledge In sequential feature selection, we are given the prior knowledge for the internal belief in the informativeness of each feature. There is also a local feature set of training data $\mathbb{D} = \{d_1, \cdots, d_N\}$ as the prior knowledge that consists of the unlabeled local feature as shown in Fig. 1(b). Initially, N training dataset are divided into $n \times m$ rectangular blocks, and re-grouped into $n \times m$ cells according to the position of the fixation. Then, all partial data is transformed to the local feature set as the prior knowledge. The prior knowledge is updated after the observations are made as shown in Fig. 1(d). The next fixation is decided sequentially by taking attention A, observed feature set O_A, and updated training data D_A.

Information gain The attention-path planning is to construct the attention subsets that maximizes the information gain. It can be stated as the following optimization problem:

$$\max f(A, \mathbb{D}) \text{ s.t. } A \subset \mathbb{A}, \tag{1}$$

where the information gain function $f(A, \mathbb{D})$ is modeled to quantify the "informativeness" of an attention.

The attention set A^{opt} can be determined by optimizing the information gain function. It is known to be difficult to obtain even approximate solutions [11]. However, near-optimal performance can be guaranteed with a simple greedy algorithm, if the information gain function satisfies the properties of adaptive submodularity and monotonicity.

A. Submodularity

First, we consider the static case, where the training data $\mathbb{D}$ is not updated.

Definition 1: (submodularity) A function f is called submodular *iff*, for all $X, Y \subseteq \mathbb{A}$ and all singletons $a \in \mathbb{A}$,

$$X \subseteq Y \Rightarrow f(X \cup a, \mathbb{D}) - f(X, \mathbb{D}) \geq f(Y \cup a, \mathbb{D}) - f(Y, \mathbb{D})$$

The information gain of adding a fixation to a smaller attention set is at least as much as adding it to the larger attention set. The information gain is never decreased by adding fixations.

Definition 2: (monotonicity) A function f is called monotone *iff*, for all $X \subseteq \mathbb{A}$,

$$f(X \cup a, \mathbb{D}) - f(X, \mathbb{D}) \geq 0$$

In this case, the information gain of each attention is maximized using a simple greedy algorithm. Let A^{gdy} be the set of attention generated with the greedy algorithm. Then, the following condition is guaranteed, when our information gain function is submodular and monotone:[10]

$$f(A^{gdy}, \mathbb{D}) \geq (1 - 1/e) f(A^{opt}, \mathbb{D}) \tag{2}$$

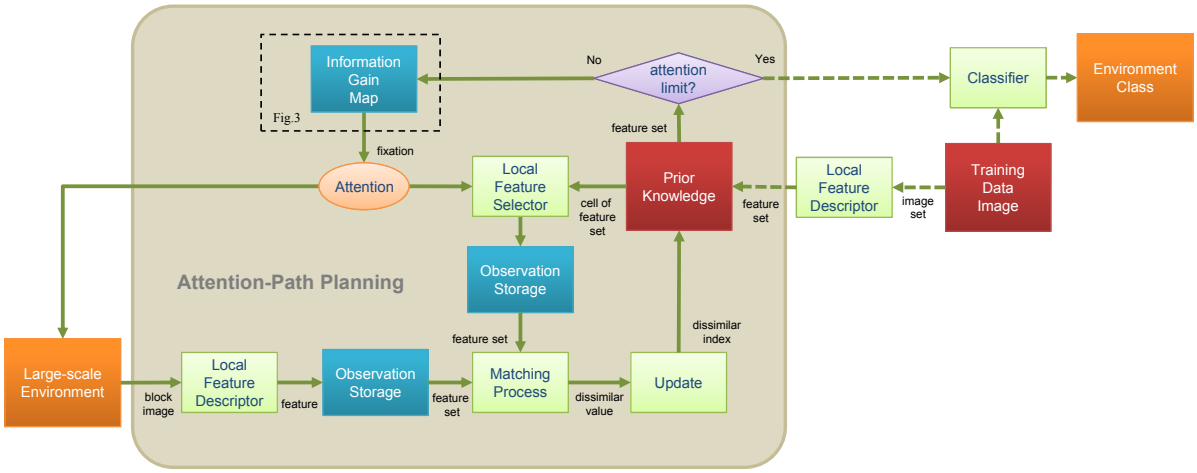


Fig. 2. Unsupervised large-scale area classification framework

B. Adaptive Submodularity

We consider the adaptive case where the training data $\mathbb{D}$ is updated at each iteration. The information is obtained sequentially as the attentions make the observations. Hence, the expected information gain of selecting a fixation is defined as

$$\Delta(a|D_A) = \mathbb{E}\left[f(A\cup\{a\},\mathbb{D}) - f(A,\mathbb{D})|D_A\right]. \tag{3}$$

Definition 3: (adaptive submodularity) A function f is called submodular *iff*, for all $X,Y \subseteq \mathbb{A}$ and all singletons $a \in \mathbb{A}$,

$$X \subseteq Y \Rightarrow \Delta(a|D_X) \geq \Delta(a|D_Y)$$

The expected information gain of adding an a fixation to a smaller attention set is at least as much as adding it to the larger attention set.

Definition 4: (adaptive monotonicity) A function f is called monotone *iff*, for all $X \subseteq \mathbb{A}$ and all singletons $a \in \mathbb{A}$,

$$\Delta(a|D_X) \geq 0$$

The expected information gain is nonnegative for all attentions.

In this adaptive case, the information gain of each attention is sequentially maximized using the recursive greedy algorithm. Based on the concept of adaptive submodularity, the following condition is guaranteed, when our information gain function is adaptive submodular and monotone:[11]

$$f(A^{gdy}_{adapt},\mathbb{D}) \geq (1-1/e) f(A^{opt}_{adapt},\mathbb{D}), \tag{4}$$

where A^{opt}_{adapt} and A^{gdy}_{adapt} are the optimal attention in the adaptive case and the set of attention generated with the recursive greedy algorithm, respectively.

III. Unknown Environment Classification

The proposed unknown environment classification framework is designed based on adaptive submodular optimization [11] by developing the information gain map and the attention-path planner: proactive and semi-reactive versions. The whole framework of the proposed attention-path planner and unknown area classification are described in Fig. 2.

A. Local Feature Descriptor and Matching process

The entire target image is partially accessed by dividing it into $m \times n$ fixation areas depending on the size of field-of-view, and the edge information of each fixation area is represented by a local feature descriptor. In computer vision and image processing, histogram of orientation gradient (HOG) [12] is a well-known feature descriptor used for object detection and classification. The technique counts occurrences of gradient orientation in localized portions of an image. Compared to other methods such as scale invariant feature transformation (SIFT) [13] and GIST [14], HOG performs improved accuracy using a dense grid of uniformly spaced cells and overlapping local contrast normalization. Initially, all the block images in the training data set are transformed to the local feature set respectively without labels; the training data contains the label information only for the classification. The block images of the target scene are transformed separately at each observation. In the attention planning process, the next fixation is selected by comparing the expected information gain of each possible fixation that is computed with the local feature set and the prior knowledge. During the test, new observations is made with the selected fixation and the images are transformed to the local feature descriptor.

In the matching process, the local features of the observation and the prior knowledge are compared using Cosine Similarity (CS) [15]; a similarity between two features of block image can be measured with the cosine of the angle between them given by

$$CS(A,B) = \frac{A\cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{5}$$

We can now define the degree of dissimilarity as follows,

$$Dissimilarity(A,B) = 1 - CS(A,B). \tag{6}$$

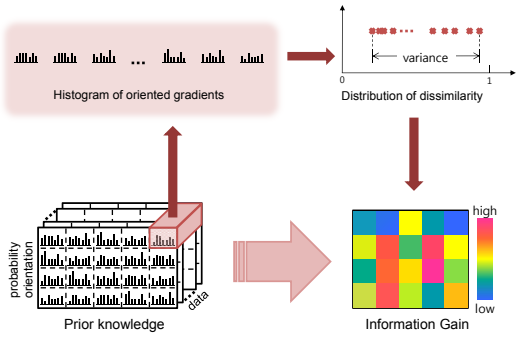


Fig. 3. Information Gain Map

B. Information Modeling

We propose a quantitative model of attention defined as the amount of uncertainty reduced by an observation. For the purpose of uncertain environment classification using a small field-of-view camera, the similarity between the local features of the target image and training images needs to be assessed at each observation. Initially, the local features of the training dataset remain uncertain until the fixation of target image is selected. It is advantageous to select the most informative cells in the training dataset for the target data assessment. Note that a cell is "informative" if the features of that cell are significantly dissimilar from each other across the training dataset. As it is difficult to make a prediction about the target image based on prior knowledge, any new observation of the target image is combined that should correspond to the informative cell. In contrast, a cell is "less informative", if its features are similar to each other across the training dataset. We model the information gain function $f(A,\mathbb{D})$ as the dissimilarity variance in a set of prior knowledge $\mathbb{D}$ considering attention A.

$$f(A,\mathbb{D}) = E\left[1 - CS(D_A, \mu_{D_A})\right], \tag{7}$$

where μ_{D_A} is the mean of the feature vectors on the attention A. Using this function, we can compute an expectation over observation for possible fixations. Finally, the expected information gain over all attentions can be represented as the information gain map (Fig. 3).

C. Proactive Unsupervised Attention-path Planner (PUAP)

In Proactive Unsupervised Attention-path Planner (PUAP), the attention-path is selected using the statistical analysis of the corresponding features of prior knowledge to each fixation and submodular optimization technique, as described in Algorithm 1. At the beginning of the algorithm, the information gain map is created initially from the prior knowledge. By following this map, a new fixation which has the maximum information gain is selected sequentially. The proposed information gain function can be used as the objective, since the observation reduces the dissimilarity variance existing on the training dataset. There are many studies that indicate that variance reduction is monotone and submodular [16]. Based on the submodularity, a simple greedy algorithm can guarantee some bounded performance for the attention selection: $f(A^{gdy},\mathbb{D}) \geq (1-1/e) f(A^{opt},\mathbb{D})$. With this offline learning technique, an informative observation sequence for environment classification can be obtained in an efficient way, without considering the given target dataset.

Algorithm 1: PUAP

input : dataset $\mathbb{D}$, attention limit l
output: attention $A \subseteq \mathbb{A}$, dataset $\mathbb{D}$

begin
  $A \leftarrow \emptyset$, $D_A \leftarrow \emptyset$, $O_A \leftarrow \emptyset$;
  foreach $a \in \mathbb{A}$ **do**
    compute $f(a,\mathbb{D})$;
  end
  for $i = 1$ **to** l **do**
    Select $a^* \in \arg\max_{a \notin A} f(A \cup a, \mathbb{D})$;
    Set $A \leftarrow A \cup a^*$;
    Observe $\mathbb{O}(a^*)$;
  end
  Select $D_{dis} :=$ `DiscardData(`O_A,D_A`)`;
  Set $\mathbb{D} \leftarrow \mathbb{D} \setminus D_{dis}$;
end

D. semi-Reactive Unsupervised Attention Planner (s-RUAP)

We now present semi-Reactive Unsupervised Attention Planner (s-RUAP) which considers both the statistical analysis of the prior knowledge and relationship between the current observation and the prior knowledge. The prior knowledge is updated by discarding non-informative dataset which are dissimilar with observations in successive time steps. For the purpose of the update, a cascade type of nearest neighbor classifier [17] is implemented to finally discard required numbers of dataset at each time step. Also, the expected information gain is computed at each time step using the information gain function that takes the attention and the updated prior knowledge. Therefore, the attention sequence of each target image is adaptively generated by maximizing

Algorithm 2: *s*-RUAP

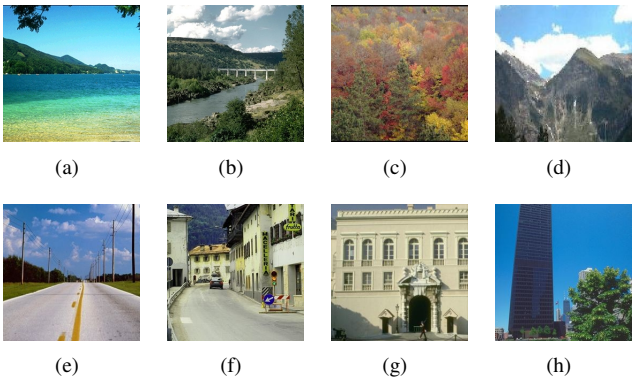
input : dataset $\mathbb{D}$, attention limit l
output: attention $A \subseteq \mathbb{A}$, dataset $\mathbb{D}$

begin
  $A \leftarrow \emptyset$, $D_A \leftarrow \emptyset$, $O_A \leftarrow \emptyset$;
  for $i = 1$ **to** l **do**
    foreach $a \in \mathbb{A} \setminus A$ **do**
      compute $\Delta(a|D_A)$;
    end
    Select $a^* \in \arg\max_a \Delta(a|D_A)$;
    Set $A \leftarrow A \cup a^*$;
    Observe $\mathbb{O}(a^*)$;
    Select $D_{dis} :=$ `DiscardData(`O_A,D_A`)`;
    Set $\mathbb{D} \leftarrow \mathbb{D} \setminus D_{dis}$;
  end
end

the adaptive submodular function for the target image, as detailed in Algorithm 2. Based on the concept of adaptive submodularity, variance reduction guarantees certain performance [10][16]: $f(A^{gdy}_{adapt}, \mathbb{D}) \geq (1-1/e) f(A^{opt}_{adapt}, \mathbb{D})$. The expected information gain function $\Delta(a|D_A)$ estimates the reduction of dissimilarity variance according to added fixations and satisfies adaptive submodularity and adaptive monotonicity. The observation set is found by recursively discarding non-informative datasets and selecting a fixation which maximizes the expected information. Finally, a part of target image and retrieved prior knowledge are obtained.

IV. Experimental verification

A. LabelMe: urban and natural scene categories



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. LabelMe (urban and natural scene categories): (a) coast/beach, (b) open country, (c) forest, (d) mountain, (e) highway, (f) street, (g) city center, and (h) tall building,

To evaluate the effectiveness of the proposed PUAP and *s*-RUAP, "LabelMe: urban and natural scene categories" was used by randomly selecting 2080 various scene images (size: 256×256 pixel) categorized into 8 classes (8 classes $\times$ 260 images) among the entire scene database [14], as seen in Fig. 4. It is assumed that the whole area is discovered with a small field-of-view (64×64 pixel) camera. 10-fold cross-validation is used to assess the performance of the data retrieval of the attention-path planners and the classification accuracy. By limiting the number of fixations, the experimental result is evaluated on limited rate of attentions, respectively.

B. Experimental results

Fig. 5 shows the information gain map and corresponding attention subregions selected at each iteration. Initially, PUAP and *s*-RUAP generate the same information gain map as shown in Step 1 in Fig. 5(a). PUAP employs only this initial information gain to select the attention subregions. Thus, the entire selection path conforms exactly to the initial map, detailed in Fig. 5(b). In contrast, *s*-RUAP updates the information gain map at every time step, yielding a different attention path as shown in Fig. 5(c) that conforms to the information gain map seen in 5(a).

The effectiveness of the proposed planners is evaluated with the simple nearest neighbor algorithm between the numbers of the training data remained in each class. Note that the class information is not used in PUAP and *s*-RUAP. Fig. 6 is the result of environment classification by limiting the number of the fixation from 1 to 16. It should be noted that only with approximately 20% of attention regions, both planners reach the performance obtained with the whole target image. An average of 67.36% classification accuracy is achieved with 100% attention regions as seen in Fig. 6. However, the random path selection approach can yield the similar performance with 90% of attention regions. For the verification of the retrieval performance, the maintenance rate is defined as the rate of remained number of data in the final result of the prior knowledge. Fig.7 shows that the maintenance rate in non-target classes rapidly decreases at each attention step by discarding unnecessary data, while the rate in the target class remains higher.

V. Conclusion and Future works

In this paper, an attention control algorithm is proposed for the environment classification system. The goal of the proposed algorithm is to enable the system to decide the next fixation area using previous observations and updating prior knowledge to perform the given task. Specifically, the information gain function is modeled as dissimilarity variance of the training data to select the most informative fixation where the information gain is maximized. The information gain function can guarantee a near optimal solution to maximize classification performance because of adaptive submodularity. Experimental results on uncertain environment classification show encouraging performance of PUAP and *s*-RUAP. Also, the proposed information model can represent the informativeness of each fixation.

As future works, extensive experiments and analyses will be performed considering the trade-off between proactive and reactive approaches. This algorithm can be further improved by employing various pattern recognition techniques such as distance measurement between target object and training data and robust classifier. It can be also applied to a wide range of robotics, particularly, when wheeled and aerial vehicles perform uncertain environment exploration and map building with a small field-of-view camera.

References

[1] A. Singh, A. Krause and W.J. Kaiser, "Nonmyopic Adaptive Information Path Planning for Multiple Robots", *Proc. Intl. Joint Conf. on Artificial Intelligence*, 1843-1850, 2009

[2] G.A. Hollinger, B. Englot, F.S. Hover, U. Mitra and G.S. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity", *Intl. Jour. of Robotics Research*, 32(1):3-18, 2013

[3] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active Vision", *Intl. Jour. of Computer Vision*, 1(4):333-356, 1988

[4] Y. Su, S. Shan, X. Chen and W. Gao, "Hierarchical ensemble of global and local classifiers for face recognition", *IEEE Trans. on Image Processing*, 20(11):1885-1896, 2009

[5] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):1254-1259, 1998

[6] J. Najemnik and W.S. Geisler, "Optimal eye movement strategies in visual search", *Nature*, 434(7031):387-391, 2005

[7] L.W. Renninger, P. Verghese and J. Coughlan, "Where to look next? Eye movements reduce local uncertainty", *Jour. of Vision*, 7(3):1-17, 2007

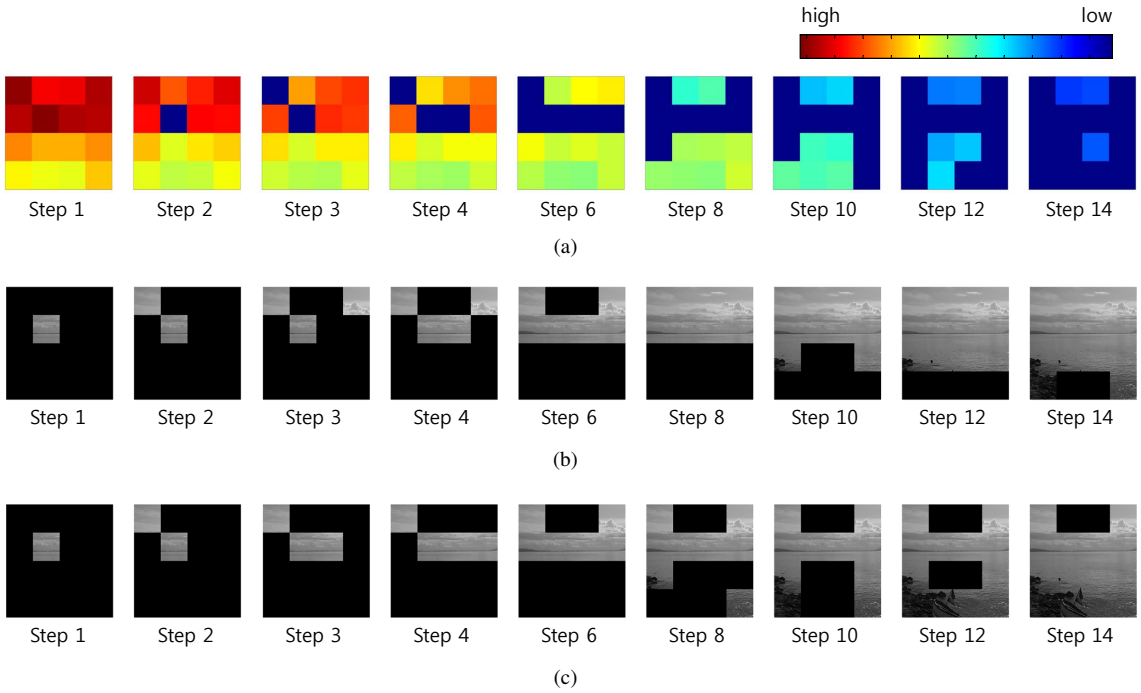


Fig. 5. Result of attention-path planning with PUAP and *s*-RUAP: (a) information gain map, (b) attention-path of PUAP and (c) attention-path of *s*-RUAP

[8] T. Rückstieß, C. Osendorfer and P. van der Smagt, "Sequential Feature Selection for Classification", *AI 2011: Advances in Artificial Intelligence*, Springer, 132-141, 2011

[9] L. Avdiyenko, N. Bertschinger and J. Jost, "Adaptive Sequential Feature Selection for Pattern Classification", *Proc. Intl. Joint Conf. on Computational Intelligence*, 474-482, 2012

[10] A. Krause and C. Guestrin, "Near-optimal Nonmyopic Value of Information in Graphical Models", *Proc. Uncertainty in Artificial Intelligence*, 324-331, 2005

[11] D. Golovin and A. Krause, "Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization", *Jour. of Artificial Intelligence Research*, 42(1):427-486, 2011

[12] D. Navneet and B. Triggs, "Histograms of Oriented Gradients for Human Detection", *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 886-893, 2005

[13] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features", *Proc. IEEE Intl. Conf. on Computer Vision*, 1150-1157, 1999

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", *Intl. Jour. of Computer Vision*, 42(3):145-175, 2001

[15] H.V. Nguyen and L. Bai, "Cosine Similarity Metric Learning for Face Verification", *Computer Vision-ACCV 2010*, Springer, 709-720, 2011

[16] A. Das and D. Kempe, "Algorithms for Subset Selection in Linear Regression", *Proc. Annual ACM Symposium on Theory of Computing*, 45-54, 2008

[17] V. Athitsos, J. Alon, and S. Sclaroff, "Efficient Nearest Neighbor Classification Using a Cascade of Approximate Similarity Measures", *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 486-493, 2005

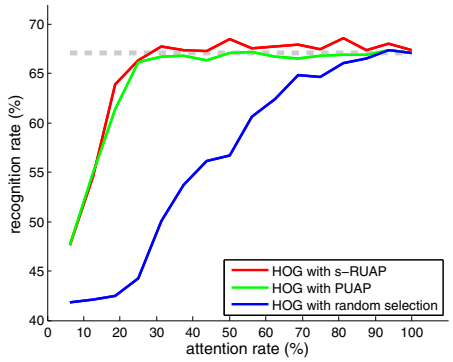


Fig. 6. Area classification with 8 classes (4×4 blocks): HOG with whole images (dashed), *s*-RUAP (red), PUAP (green), and random selection (blue).

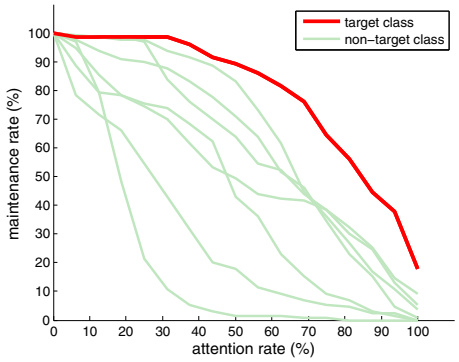


Fig. 7. Maintenance rate in training dataset at each attention: target environment class (red), non-target classes (green).