Sequential Attention Planning for Large-scale Environment Classification

Hosun Lee, Sungmoon Jeong, and Nak Young Chong

Abstract— **In this paper, a sequential attention planning algorithm is proposed and applied to wide-uncertain environment classification with small field-of-view cameras. Attention 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. However, greedy algorithm cannot always guarantee the optimal solution. In order to find the near-optimal solution for attention planning, adaptive submodular optimization is considered under certain assumptions, where the objective function for the internal belief is adaptive submodular and adaptive monotone. First, the amount of information of individual attention area is modeled as dissimilarity variance among the environment data set, respectively. 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 about current environment using a cascade of nearest neighbor classifiers, enabling to maximize the expected information gain. The effectiveness of the proposed algorithm is verified through experiments that can significantly enhance the uncertain 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 [1]. Along these lines, 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 wide-area uncertain environment 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 [2][3]. The path of the robot is optimized to obtain the maximum information which is required to perform a given task. Active vision has been studied to plan proper actions for the optimal visual observation in order to improve the performance of the vision system [4] [5]. However, they assume free access to the whole information without any consideration of the limited sensing coverage. An estimation module is therefore needed to cope with the limited sensing coverage.

In case of human visual system, there are already several researches about the attention planning which show efficient eye-movement for maximizing the local information [6] [7]. This attention planning problem under limited sensing coverage can be described as a sequential feature selection problem which is a fundamental problem in pattern recognition and data mining. Sequential feature selection problem arises when a system is limited to select a feature to determine the unknown target entity at each time [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. The given task is performed with the result of the feature selection. Notably, conventional sequential feature selection is a supervised learning approach which fully trains the classifier with an associated class label to directly achieve a desired classification performance. However, it is difficult to generate the balanced training data with label information in robotic application. Therefore, selecting the informative feature without label information is challenging problem; unsupervised sequential feature selection. Recently, the property of adaptive submodularity is investigated to ensure the bounded performance of the informative path planning optimization and examined on its adaptive features [10] [11]. Based on this property, we attempt to solve the attention planning problem in the unsupervised scenario, where class information is unavailable directly.

In this work, we discuss the attention planning with the small field-of-view camera in large scale data processing. We assume that the system can obtain the information partially at each sampling step by selecting a position of attention called fixation; an attention behavior can be defined by a sequence of fixation. Then, our goal is how to plan the next fixation based on the current information of observed data and unlabeled training data set called prior knowledge to enhance the perception ability to the uncertain environment. There are two main components of attention planning: (1) sequential selection of fixation using prior knowledge to maximize the information gain, and (2) update of prior knowledge by discarding the non-informative subset of prior knowledge according to currently observed data. First, the information is defined as the amount of uncertainty reduction from the information theory, and an information gain function is modeled to measure the expected information

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

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 the 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 dissimilarity to current observation. After then, 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.

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

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

II. Problem Statement

Before describing the proposed algorithm in detail, a general formulation for the attention planning problem is considered. Attention planning is regarded as the general problem of sequential feature selection [8] by our formulation. In sequential feature selection, the most informative feature is gathered from a set of features at each iteration to perform the given task. The feature selection is decided depending on the obtained feature set and the internal belief. For the case of attention planning, a fixation is selected in each time step and used to observe the part of the entire target data. Finally, the class of the target data is classified with observed data from the class set space, $\mathbb{C} = \{c_1, c_2, \cdots\}$.

Attention and Observation At each iteration, we can select a fixation a from the set of all possible attention $\mathbb{A} = \{a_1, \cdots, a_{n\times m}\}$, where the entire region of the attention on the target data is divided into $n \times m$ rectangular blocks according to the range of the sensing coverage as shown in Fig. 1(a). And 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

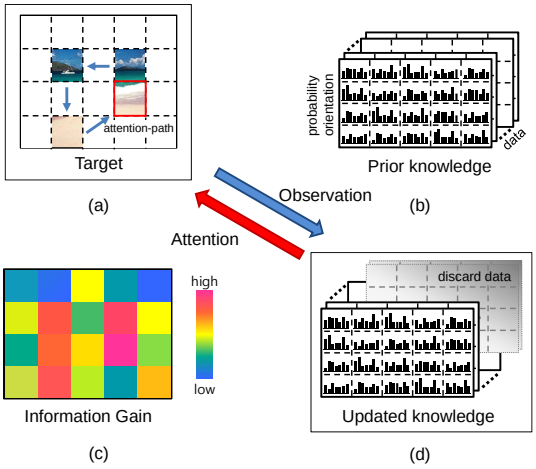


Fig. 1. Concept of attention path planning algorithm: (a) example of sequential attentions, (b) unlabeled training dataset called prior knowledge, (c) information gain of each attention data, (d) update of prior knowledge according to dissimilarity between observed data and prior knowledge. Attention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

the local 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. As results of the attention, a selected fixation set $A \subset \mathbb{A}$ and an observed feature set $O_A \subset \mathbb{O}$ are accumulated sequentially.

Internal prior knowledge In general 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 data are also divided into $n \times m$ rectangular blocks, and regrouped 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 of the attention planning framework. For securing the adaptivity, we update the prior knowledge after making the observations as shown in Fig. 1(d). Finally, the next fixation a is decided sequentially by taking attention A, observed feature set O_A, and updated prior knowledge $\mathbb{D}$.

Information gain The goal of the attention planning is to construct the attention subset that maximizes information gain. It can be rewritten as the following optimization problem formally:

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

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

The optimal attention set A^{opt} can be considered for the above problem, which optimizes the information gain function. Unfortunately, it is 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 properties of adaptive submodularity and monotonicity.

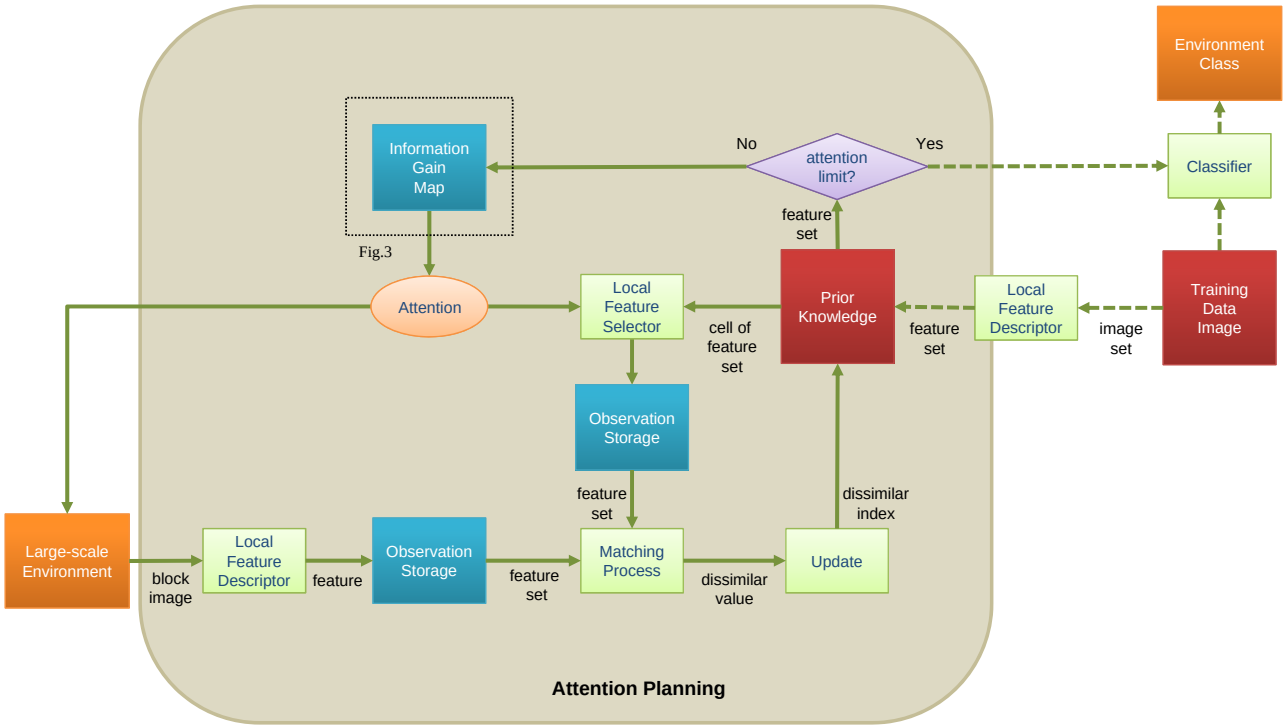


Fig. 2. Unsupervised large-scale environment classification framework

A. Adaptive Submodularity

We consider the adaptive condition 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 performing a fixation is defined as

$$\Delta(a|D_A) = \mathbb{E}\left[f(A \cup \{a\}, \mathbb{D}) - f(A, \mathbb{D})|D_A\right]. \tag{2}$$

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

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

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

Definition 2: (adaptive monotonicity) A function f is called monotone iff, for all $X \subseteq \mathbb{A}$ and all singletons $a \in \mathbb{A}$, we have

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

The expected information gain is nonnegative for all possible attention set.

The information gain of each fixation is sequentially maximized by using the recursive greedy algorithm. Based on the concept of adaptive submodularity, following performance is guaranteed when our information gain function is adaptive submodular and monotone:[11]

$$f(A^{gdy}_{adapt}, \mathbb{D}) \geq (1 - 1/e) f(A^{opt}_{adapt}, \mathbb{D}), \tag{3}$$

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

III. Uncertain Environment Classification

The proposed uncertain environment classification framework in this paper is designed based on adaptive submodular optimization [11] by developing the information gain map and the attention planner. The whole framework of the proposed attention planner and its position in uncertain environment classification are described in Fig. 2.

A. Local feature descriptor and Matching process

Initially, the entire target environment 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 local feature descriptor. In computer vision and image processing, histogram of orientation gradient (HOG) [12] is a well-known feature descriptor used for the purpose of object detection and classification. The technique counts occurrences of gradient orientation in localized portions of an image. There are similar methods such as scale invariant feature transformation (SIFT) [13], GIST [14], and so on, but HOG performs improved accuracy by 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 the label information; 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

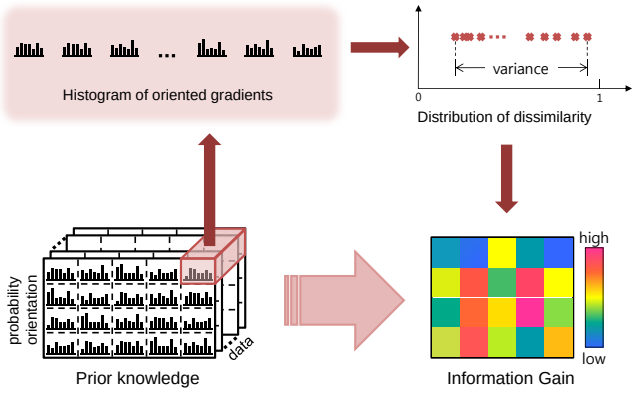


Fig. 3. Information Gain Map

fixation is selected by comparing the expected information gain of each possible fixation that is computed with the local feature set, the prior knowledge. During the test, new observation is obtained from the selected fixation and transformed the partial image to local feature descriptor.

In the matching process, the local features of the observation and the prior knowledge are compared by using Cosine Similarity (CS) [15]; a similarity between two features of block image can be measured with the cosine of the angle between them. Two vectors with the same orientation have the highest measurement of 1, and two vectors orthogonal orientation have a similarity of 0, independent of their magnitude.

$$CS(A,B) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{4}$$

In addition, we can define the measurement of the dissimilarity as follows,

$$Dissimilarity(A,B) = 1 - CS(A,B). \tag{5}$$

B. Information Modeling

We now propose a method for the information model of the attention. From the fundamental of the information theory, the information is defined as amount of uncertainty reduced by an observation. For the general purpose of the uncertain environment classification with multiple observation using small field-of-view camera, the similarity between local feature of the target environment and the local features of the training environment is measured at each observation. Therefore, each local feature of the training data is an uncertainty until a local feature of target image is observed by the selection of the fixation. In other words, if the images in a cell of the training data is equivalently dissimilar between each training data, it is difficult to predict the target image at the same area before making the observation which is "informative" selection to classify the target image into specific training data. On the other hand, an attention area contains less information if the features of the cell are similar between each feature, and the target feature is easily predicted. We model the information gain function $f(A,\mathbb{D})$ as the dissimilarity variance in a set of prior knowledge $\mathbb{D}$ considering the attention A.

$$f(A,\mathbb{D}) = E\left[Dissimilarity(D_A, \mu_{D_A})\right], \tag{6}$$

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

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

Algorithm 1: *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

We now present semi-Reactive Unsupervised Attention Planner (s-RUAP) which considers both of statistical analysis of unlabeled prior knowledge and relationship between the current observation and the prior knowledge. The prior knowledge is updated by discarding non-informative data which are dissimilar with observation in successive time step. For the purpose of the update, a cascade type of nearest neighbor classifier [17] is implemented to finally discard required number of data for the classifier at each time step. Also, the expected information gain is computed in each time step by using information gain function that takes the attention and the updated prior knowledge. Therefore, attention sequences of each target image can be adaptively generated by maximizing the adaptive submodular function for different target image. The detail computation is described in Algorithm 1. Based on the concept of adaptive submodularity, some performance are guaranteed for the case of variance reduction [10][16]: $f(A^{gdy}_{adapt},\mathbb{D}) \geq (1-1/e) f(A^{opt}_{adapt},\mathbb{D})$. In our adaptive setting, the expected information gain function $\Delta(a|D_A)$ estimates the reduction of dissimilarity variance according to adding fixation and satisfies adaptive submodular and adaptive monotone. The observation set is found by recursively discarding non-informative data and selecting a fixation which maximizes the expected information.

At the final stage of the attention planning, a part of target environment image and successfully retrieved the prior knowledge are obtained.

IV. Experimental Verification

A. LabelMe: urban and natural scene categories

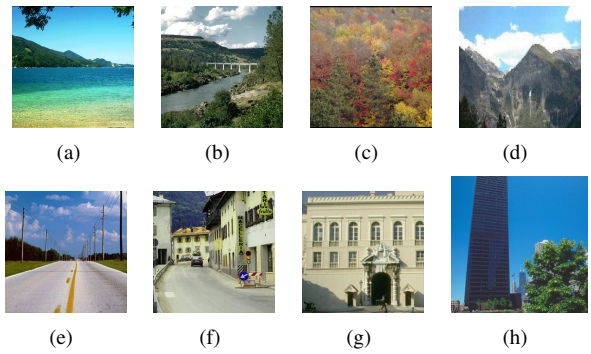


Fig. 4. Samples of LabelMe - urban and natural scene categories: (a) coast & beach, (b) open country, (c) forest, (d) mountain, (e) highway, (f) street, (g) city center, and (h) tall building,

To evaluate the effectiveness of the proposed *s*-RUAP, "LabelMe: urban and natural scene categories" was used, which contains various scenes (size: 256 $\times$ 256 pixel) categorized into 8 classes [14]. Fig. 4 shows the image samples in each class. It is assumed that the whole environment is discovered with small field-of-view (64$\times$64 pixel) camera. Among the entire scene database (total 2080 images: 8 classes $\times$ 260 images), 10% of images in each class are randomly selected for 10-fold cross-validation. By constraining the selectable number of the fixation, the experimental result is evaluated on respectively limited rate of attention. k-nearest neighbor is used to verify the output of the attention planner, the retrieved local feature set, with the performance of the classification.

B. Experimental results

Fig. 5 shows the result of the information gain map and corresponding attention subregion selected at each iteration. *s*-RUAP generates the information gain map, Fig. 5(a), at every time step, yielding a different attention path as shown in Fig. 5(b).

The effectiveness of the proposed planners is evaluated with the simple nearest neighbor algorithm between the numbers of the prior knowledge remained in each class. Specifically, the class information is used only for the verification. Note that the class information is not used by *s*-RUAP. Fig. 6 is the result of environment classification with 16 fixations. It should be noted that only with approximately 20% of attention region, the planner reach the performance obtained with the whole target image. A full (100%) attention region gives an average of 67.36% classification accuracy seen in Table. I. However, the random path selection approach can yield the similar performance with 90% of attention region. For the verification of the clustering performance, the maintenance rate is defined as the rate of remained number of data in the final result of the prior knowledge. According

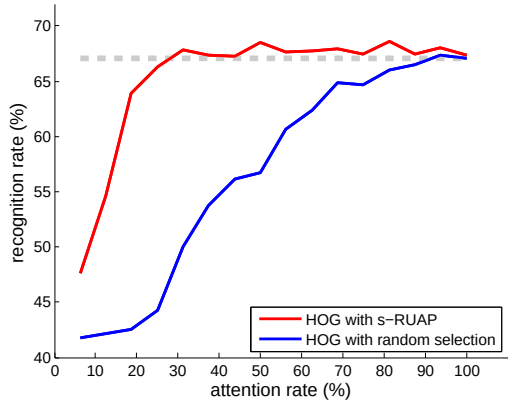


Fig. 6. Performance of environment classification rate with 8 classes(4×4 blocks): HOG with all (100%) of attention (dashed gray), HOG with Sequential attention selection by *s*-RUAP (red) , and random selection (blue).

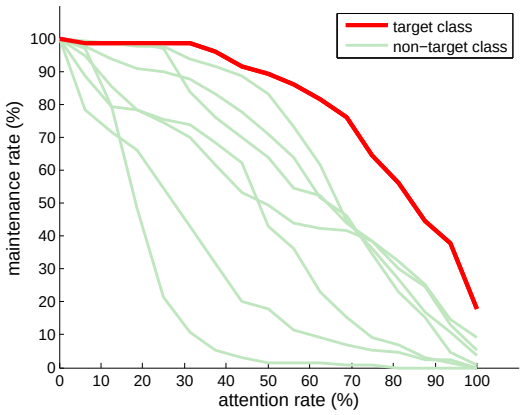
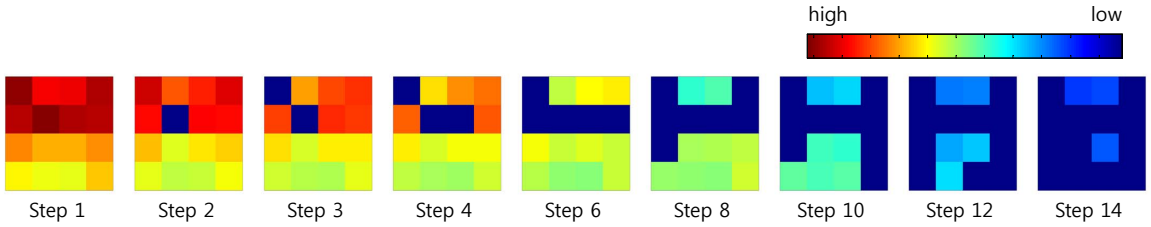


Fig. 7. Example of maintenance rate in training dataset at each attention: target environment class (red), non-target 7 classes (green).

to our test result as shown in Fig.7, the maintenance rate in non-target class rapidly decreases at each attention step by discarding unnecessary data, but the rate in the target class remains higher.

V. Conclusion and Future Works

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



(a)

(b)

Fig. 5. Result of attention planning with *s*-RUAP: (a) Information Gain Map and (b) attention-path of *s*-RUAP

TABLE I

CONFUSION MATRIX (IN PERCENT) BETWEEN THE CLASSES.

		identified class							
		coast	open country	forest	mountain	highway	street	city center	tall building
test class	coast	74.57	11.59	0.89	0.77	11.39	0.22	0.48	0.10
	open country	10.46	70.41	11.18	1.68	5.72	0.19	0.22	0.14
	forest	0.19	2.91	94.74	1.44	0.14	0.29	0.10	0.19
	mountain	3.37	27.33	23.22	40.31	2.64	1.59	1.42	0.12
	highway	12.96	11.49	1.59	0.50	69.59	1.37	2.19	0.31
	street	1.01	4.64	7.40	0.82	5.50	73.82	2.67	4.13
	city center	10.29	3.34	7.74	0.46	4.50	6.42	57.52	9.74
	tall building	3.39	4.71	12.69	2.81	6.27	5.07	7.16	57.88

As future works, various experiments and analyses will be examined with various training data. In detail, we are considering the tradeoff between proactive and reactive approach. This algorithm can be improved by considering various pattern recognition techniques such as distance measurement between target object and training data and robust classifier instead of simple nearest neighbor algorithm. Then, it can be also applied on wide field of robotics, for instance, if the small agent systems such as wheeled-robot and UAV perform a localization task in an uncertain environment with a small field-of-view camera.

REFERENCES

[1] H. Liu and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*, Springer,1998

[2] A. Singh, A. Krause and W.J. Kaiser, "Nonmyopic Adaptive Information Path Planning for Multiple Robots", *International Joint Conference on Artificial Intelligence*, Pasadena, California, USA, pp.1843-1850, 2009

[3] G.A. Hollinger, B. Englot, F.S. Hover, U. Mitra and G.S. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity", *The International Journal of Robotics Research*, vol 32 no. 1, pp.3-18, 2013

[4] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active Vision", *International Journal of Computer Vision*, vol 1 no. 4, pp.333-356, 1988

[5] Y. Su, S. Shan, X. Chen and W. Gao, "Hierarchical ensemble of global and local classifiers for face recognition", *IEEE Transactions on Image Processing*, vol 18 no. 8, pp.1885-1896, 2009

[6] J. Najemnik and W.S. Geisler, "Optimal eye movement strategies in visual search", *Nature*, vol 434 no. 7031, pp.387-391, 2005

[7] L.W. Renninger, P. Verghese and J. Coughlan, "Where to look next? Eye movements reduce local uncertainty", *Journal of Vision*, vol 7 no. 3, pp.1-17, 2007

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

[9] L. Avdiyenko, N. Bertschinger and J. Jost, "Adaptive Sequential Feature Selection for Pattern Classification", *International Joint Conference on Computational Intelligence*, Barcelona, Spain, pp.474-482, 2012

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

[11] D. Golovin and A. Krause, "Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization", *Journal of Artificial Intelligence Research*, vol 42 no. 1, pp.427-486, 2011

[12] D. Navneet and B. Triggs, "Histograms of Oriented Gradients for Human Detection", *IEEE Computer Society Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp.886-893, 2005

[13] D.G. Lowe, "Object Recognition from Local Scale-Invariant Features", *The Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, Kerkyra, Greece, pp.1150-1157, 1999

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope", *International Journal of Computer Vision*, vol 42 no. 3, pp.145-175, 2001

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

[16] A. Das and D. Kempe, "Algorithms for Subset Selection in Linear Regression", *Proceedings of the 40th annual ACM symposium on Theory of computing*, Victoria, BC, Canada, pp.45-54, 2008

[17] V. Athitsos, J. Alon, and S. Sclaroff, "Efficient Nearest Neighbor Classification Using a Cascade of Approximate Similarity Measures", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, pp.486-493, 2005