Informative Sequential Selection of Variable-Sized Patches for Image Retrieval

Zhihao Shen†, Hosun Lee, Sungmoon Jeong, and Nak Young Chong,

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai Nomi Ishikawa, Japan
{shenzhihao, hosun_lee, jeongsm, nakyoung}@jaist.ac.jp

Abstract: To quickly and efficiently analyze a large-scale environment by the camera with limited field-of-view, intelligent systems should sequentially select the optimal field-of-view to observe an important and informative patch of area. Especially in the image retrieval task, small observations should be sequentially selected to increase the performance of image retrieval and the updated performance can be used to select the next best view again in a cyclic process. In this paper, we have investigated the role of selected image patches, which can be either overlapped or non-overlapped with previous observations, in this cyclic process. To evaluate the different patch selection strategies, the adaptive observation selection method is also described as follows: (1) robots select adaptive observations sequentially based on its prior knowledge from the training dataset. (2) After each selection, the prior knowledge will be updated by discarding the target-irrelevant data for the next observation selection. During this process, we have shown that an informative patch, even though a part of selected patch is already observed at previous steps, can enhance the retrieval accuracy and it has better performance than an independent observation method.

Keywords: Image retrieval, Small field-of-view, Partial observation, Next best-view

1. INTRODUCTION

Intelligent robots make decisions and adjust their actions on the basis of information from different environments when they are applied to perform variety of tasks. However, the mobile robots need to preform their tasks always with some limitations such as time, battery capacity and/or limited sensing coverage. There are some researches that has been done for solving robot navigation problems in limited time or battery capacity [1][2]. However, it takes high computation cost, if the robots try to analyze a large-scale environment. For avoiding processing high-dimensional data, the robots are equipped with a small field-of-view camera, and capture local images from large environment sequentially. For retrieving informative image patches at every time, the studies, such as viewpoint planning and saliency-based visual attention [3][4], are assuming fully access to the target image at once.

Generally, the entire environment is sequentially accessed with multiple view images depending on the size of field-of-view. On the other hand, with limited information input, it is hard for robots to recognize a large environment. The robots need to decide where to take the image patch at each time. There are some studies that are inspired from human eye movement [5]. Assuming that humans can observe a small field-of-view each time, they decide the view point based on their knowledge, and they keep target-relevant memory for the next observation selection. The concept that a robot with limited sensing coverage sequentially selects observation from large-scale environment is similar with human eye movement when human move their eyes purposefully to gain the sensory information.

† Shen Zhihao is the presenter of this paper.

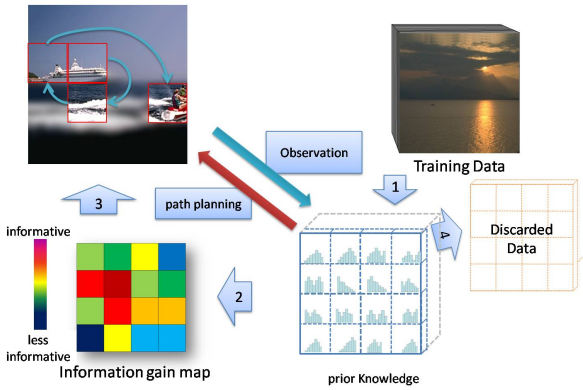


Fig. 1 Concept of attention path planning: 1. The dataset are divided by several small patches and each patch is represented by local feature vector that stores in a memory as a prior knowledge, 2. Informativeness of each patch, 3. Best patch selection, and 4. Prior knowledge update by discarding target-irrelevant training dataset. From steps 2 to 4 are repeated to correctly retrieve the target image.

Attention-path planning algorithm [6] was proposed to enable the robot to sequentially access to a part of the target environment with limited field of view (the concept of attention path planning as shown in Fig. 1). It is similar to the human visual perception [7] that visual stimulus information combines with prior knowledge and task goals to plan an eye movement. The Attention-path planning algorithm mainly contains two components: (1) observation selection based on the informativeness of each fixation. (2) at each step, the robot selects an informative patch from the target environment based on its prior knowledge and updates its prior-knowledge based on cur-

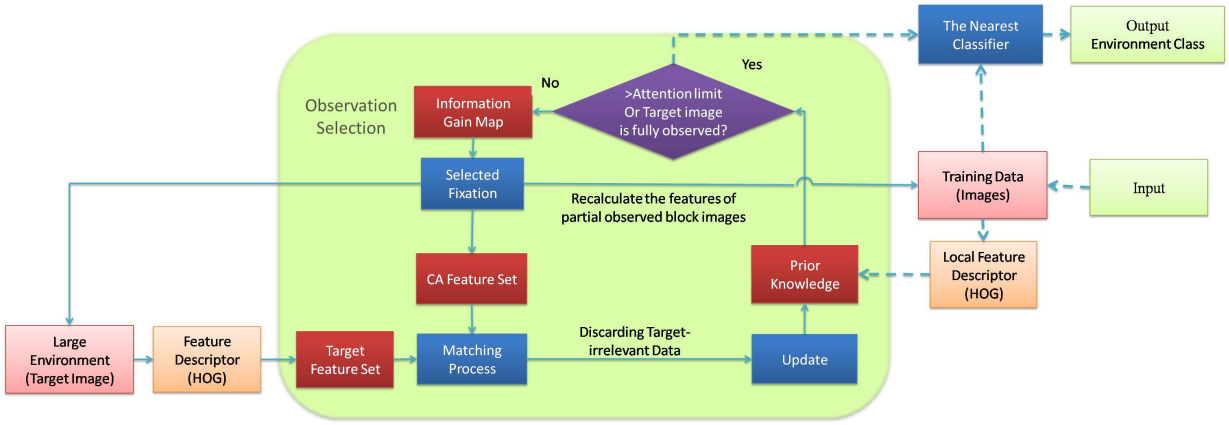


Fig. 2 Overall architecture of sequential patch selection for image retrieval

rently selected observations. With this cyclic process, robot can classify the target environment without accessing whole environment. However, each image patch that the robot observed from the target environment is fixed in specific position and fully independent of each other.

In practice, some informative features can be omitted or only observed once with this independently defined patches. It is necessary to place the patches adaptively to enhance the retrieval performance. In this research, we proposed a method by considering the partially overlapped patches to select observation more adaptively. All blocks are partially overlapped by their neighbors. And the size of the block images is the same case of the small field-of-view camera on robot. All image data are represented by feature vectors. The fixation selection means to select crucial position on the target environment depending on the whole prior knowledge. It is an arduous work that manually labelling a large set of training data. We try to solving this problem in the unsupervised scenarios. The fixation is selected based on the concept that best preserve the informativeness of the fixation derived from the whole prior knowledge. A simple greedy algorithm is used to guarantee that the selected fixation is the most informative position under the current state.

2. PROBLEM STATEMENT

In the attention-path planning algorithm, all images are divided into non-overlapped blocks. It becomes less informative when a valuable feature is divided into different patches, which can be seen in Fig. 3(a). Therefore, in this research, we propose that the target image is divided into rectangular blocks according to the coverage of the robot's camera and each block is 50% overlapped by their neighbors, which can be seen in Fig. 3(b). Hence, there are more observation selection options than non-overlapped condition. The framework of the adaptive observation selection is presented in Fig. 2. There are three components of the framework are illuminated in the following.

Preprocessing All images which are used as prior knowledge are categorized into the given training class set, $\mathbb{C} = \{c_1, c_2, \cdots\}$. We extract the local features of

Fig. 3 Image patach selection with different methods. (a) block images without overlap, (b) block images with overlap

the block images by dividing images with overlap. There is a training data set of all images $\mathbb{D} = \{ \mathbb{I}_1, \mathbb{I}_2, \cdots, \mathbb{I}_N\}$ and every image contains local features such as $\mathbb{I}_k = \{f^k_{(1,1)}, \cdots, f^k_{(r,c)}\}$, where all images are divided into r row and c column blocks. $f_{(m,n)}$ represents the feature of the block image in the position (m, n) and the fixation is represented with a coordinate (m, n). The prior knowledge will be updated after each selection by discarding a number of target-irrelevant images for the next selection.

Observations and Prior Knowledge If the fixation is decided, then, the local features $f_{(m,n)}$ of the block image in the position (m, n) will be included in the classification-aim (CA) feature $I_k = [\cdots, f^k_{(m,n)}]$. Note that I_k is the feature vector of the k-th image in the training dataset and I_k is generated by combining the features of the block images in the position that has been selected already. $D_f = \{\cdots, I_k\}$ is the dataset of classification-aim feature sets of all images. The target feature set $O_f = [\cdots, o_{(m,n)}]$ combines all local features that are sequentially observed from target environment.

Information Gain The observed image patches of the target environment are used to compare the similarity with the prior knowledge and the most dissimilar data will be discarded for updating the prior knowledge. The similarity between two images is measured by using Cosine Similarity (CS) [8].

$$Similarity = CS(O_f, I_k) = \frac{O_f \cdot I_k^T}{||O_f|| \times ||I_k||}$$

Where O_f is the feature of target environment and CA-

feature I_k is the feature of the k-th image in prior knowledge.

Therefore, the dissimilarity can be defined as:

$$Dissimilarity = 1 - CS(O_f, I_k)$$

In the observation selection processing, a fixation is informative means that the fixation best preserve the dissimilarity across the whole training data, the information gain of each fixation can be calculated as:

$$I_k^{(m,n)} = [I_k^{(m,n)}, f_{(m,n)}]$$

$$G_{(m,n)} = E[1 - CS(\omega, D_{f_{(m,n)}})]$$

at each iteration, the information gain is the average of the dissimilarity between ω and the CA-features of each image in position (m,n). ω is the mean feature of the remaining images in fixation (m,n), which can be calculated as:

$$\omega = E[D_{f_{(m,n)}}]$$

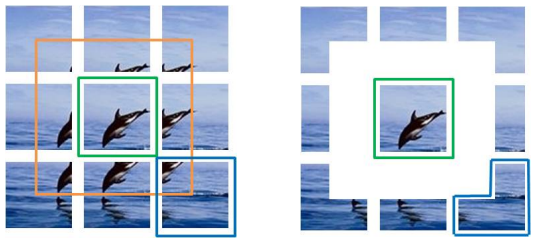
Information gain will be recalculated after updating the prior knowledge. The information gain over all fixations is shown in a gain map which shows the most informative position.

3. ENVIRONMENT CLASSIFICATION

Feature Descriptor It is important to find a good feature detection method to represent the partial image that observed from the target environment and all block images in the training dataset [9]. A high-performance feature detector should show robustness to changing image conditions. Histogram of orientation gradient (HOG) [10] is a well-know feature descriptor which can be used to extract local features form block images. Initially, each block image will be calculated HOG feature vector. Then, the feature vector is saved in the prior knowledge. Based on the current prior knowledge, The most informative position that best presents the data diversity across the whole prior knowledge is selected as the observation fixation.

There are two types of adaptive observation selection methods: fixed field of view and variable field of view. These two methods are presented in Fig. 4. One image is divided by several parts of patches from left most top position to right most bottom position with the flexible size of window (field of view of robot's camera) $a \times b$ pixels. Then, the window (view point) is sliding with a certain interval $a \times (1 - overlap\_rate)$ or $b \times (1 - overlap\_rate)$ along the x and y image axes.

Fixed Field of View In this work, we assume that the robot can observe a partial area in fixed size. The target image is partially accessed at every time steps by observing the most informative fixation. If the overlapped area contains an important image content, usually we need to keep such important features within an current observation. Therefore, some areas of the target image can be observed several times. As show in Fig. 4(a), supposing that the block image in the green box is the selected as a best view point, and the areas within the orange box are partially overlapped with its neighbors. If the next observation is selected among these eight block images, some partial information will be used again in the following observations.



(a) fixed field of view (b) variable field of view

Fig. 4 Concept of Observation Selection

Variable Field of View To reduce the feature vectors from an previously observed area, an previously observed area could be discarded and the rest of area is only used to calculate the feature vectors. As shown in Fig. 4(b), the green block is selected as informative image patch and it is used to extract the feature vectors. After then, the partial area of the block images that was not observed, like the block image in the blue concave polygon, need to be recalculated its feature vector. So the training data is updated in two aspects: discarding dissimilar images from training data and recalculating the feature vectors of the block images that are partially observed in the previous step. There is non-overlapped area among the observations that are selected in each iteration. The observations that are selected in each iteration are independent of each other. The whole framework of the environment classification with variable field of view is described in Fig. 2.

4. EXPERIMENTAL VERIFICATION

4.1. LabelMe: urban and natural scene categories

Fig. 5 LabelMe (urban and natural scene categories): (Class1) coast/beach, (Class2) open country, (Class3) forest, (Class4) mountain, (Class5) highway, (Class6) street, (Class7) city center, and (Class8) tall building.

The training dataset [11] contains 2080 images (size: 256 $\times$ 256 pixel) is categorized into 8 classes (8 classes $\times$ 260 images). We assume that the size of the small

field-of-view camera is 64 × 64 pixel. In the non-overlap condition, the entire image is divided into 4 × 4 blocks, the images will be divided into 7 × 7 blocks in the overlap condition. The test dataset is the first folder that contains 208 images when applying 10-fold cross-validation to assess the performance of image retrieval accuracy. We limit the selection number of fixation and evaluate the experiment result on limited rate of attentions. The observation rate is the observed proportion of the target image.

4.2. Experiment results

Fig. 6 shows the experiment results of the fixed field-of-view and the variable field-of-view, the test dataset contains 1456 images. Non-overlap condition (blue) is shown as a contrast experiment.

The retrieval decision is evaluated with the simple nearest neighbor algorithm [12]. The total number of the remained training data is 30 images, which means that the decision is made from 30 images by counting the number of the training data remained in each class. The result in Fig. 6 is generated by limiting the number of observation limits from 1 to 49.The average performance of image retrieval with overlap is better than the case without overlap. The best performance of non-overlap case, fixed field-of-view and variable field-of-view are 69.3%, 71.18% and 70.93% respectively. An average of 68.41%, 68.34% and 68.37% accuracy is achieved respectively, while fully accessed the target image.

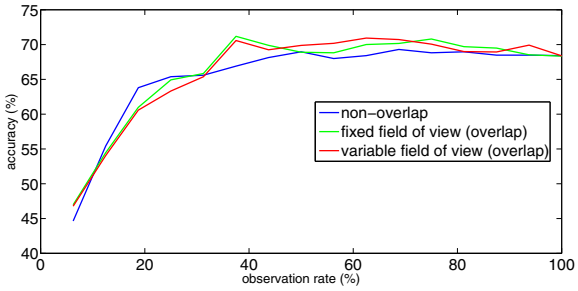


Fig. 6 Comparison of image retrieval accruacy between different selection approaches.

5. CONCLUSION AND FUTURE WORKS

In this study, we proposed a partition method for adaptive observation selection for image retrieval system. The new partition method provides a flexible observation selection way for image retrieval problem. The performance of the proposed method fluctuates with observation rate, and it shows better result than the non-overlap experiment in a certain range. According to the experiment result, this algorithm can achieve a good performance even when the target image is not fully accessed.

In the future works, we need to figure out how large area is overlapped, and whether the overlapped area is important or not. This algorithm can be used to detect the crucial area from the target image. The crucial area can be employed to pattern recognition techniques. The crucial area selection pattern can make an efficient selection of observation and improve the image retrieval performance.

6. ACKNOWLEDGEMENT

This project was supported by the EU-Japan coordinated R&D project on "Culture Aware Robots and Environmental Sensor Systems for Elderly Support" commissioned by the Ministry of Internal Affairs and Communications of Japan and EC Horizon 2020.

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

[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] 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

[4] 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

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

[6] Hosun Lee; Sungmoon Jeong; Nak Young Chong, "Unsupervised Learning Approach to Attention-path Planning for Large-scale Environment Classification", Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1447 1452, 2014

[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

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

[9] Yali Li, Shengjin Wang, Qi Tian, Xiaoqing Ding, "A survey of recent advances in visual feature detection", Neurocomputing, Pages 736751, Volume 149, Part B, 3 February 2015

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

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

[12] 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