Informative Sequential Patch Selection for Image Retrieval

Zhihao Shen, Sungmoon Jeong, Hosun Lee, and Nak Young Chong
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai Nomi Ishikawa, Japan
{*shenzhihao*, *jeongsm*, *hosun_lee*, *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 important and informative parts of area. Especially in the image retrieval tasks, small observations could be sequentially selected to improve the performance of image retrieval with less computational costs than whole observations at once and the enhanced retrieval performance could be used to select the next best-view again in a cyclic process. In this paper, we have investigated the effects of selected image patches, which might be either overlapped with a certain ratio or non-overlapped with previous observations, in this cyclic process. The adaptive patch selection algorithm is also described as follows: (1) A current observation is decided by its own information gain model which is designed by a similarity value between current observed information and training dataset. (2) After then, the system will update the information gain model by discarding the irrelevant training data with the current observation. 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 a better performance than an independent observation method. Experimental results also have shown that the model selects the informative patches around the important contents to retrieve the target images such as the sky, building and so on.**

***Index Terms*— Image retrieval, Small field-of-view, Next best-view, Sequential Selection, Visual attention**

I. Introduction

Intelligent robots make decisions and adjust their actions to achieve various tasks based on an information from different environments. However, they need to preform their tasks with restrictions and limitations such as operating time, battery capacity, and/or limited sensing coverage. There are some researches that has been done for solving robot navigation problems in limited operating time and battery capacity [1] [2]. However, it requires high computational cost because of the large-scale environment. To avoid this huge computational cost, the robots can be equipped with a small field-of-view camera and capture a part of image from large-scale environment sequentially. Similar with this robotics problem, there are various studies to solve large-scale image retrieval problem based on an informative image patch selection method such as a viewpoint planning and saliency-based visual attention [3] [4]. However most of studies assume that the system can fully access to the large-scale environment 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, it is difficult for the robots to understand a large environment with partial information. The robots need to decide where to take the image patch at each time. There are some studies that are inspired by a human eye movement [5]. Assuming that humans can observe a small field-of-view at each time, they decide the best view point based on their knowledge, and they try to keep a target-relevant memory for the next view selection. By mimicking this concept, a robot with limited sensing coverage sequentially selects an observation from large-scale environment and a target-relevant memory is updated by comparing the similarity between the current observations and the previous memory.

Attention-path planning algorithm [6] was proposed to enable the robot to sequentially select a part of environment with limited field of view as shown in Fig. 1). It is similar to the human visual perception [7] to plan an eye movement by combining a current visual stimulus and a prior knowledge. This algorithm is based on two main components: (1) observation selection based on an informativeness of each fixation. (2) a prior-knowledge update by calculating a similarity score between the current observation and the prior-knowledge. The robot can sequentially enhance the retrieval accuracy for the target environment with a cyclic process using the two components. However, each sampled image patch is independent between each other so it causes the sampling artifact to represent the large-scale environments. Moreover,some informative features can be omitted or only observed once with this independently defined patches.

In this research, we propose a method to adaptively select a best observation to retrieve the images. All image patches are partially overlapped by their neighbors similar to a continous sampling from the robot with small field-of-view camera. Each image patch is represented by the set of feature vectors to select crucial position on the target environment depending on the whole prior knowledges. It is an arduous work to manually label a large set of training data. Therefore, an unsupervised approach [8] is considered to solve this problem. An information gain is designed by submodular function and a simple greedy algorithm is used to find a near optimal solution for the most informative position under the current state.

II. 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,

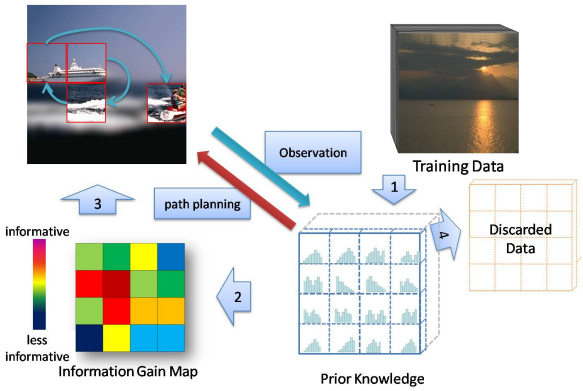


Fig. 1. Concept of attention path planning: 1. The dataset are divided by several small patches and each patch is represented by local feature vector that stores in a memory as a prior knowledge, 2. Informativeness of each patch, 3. Best patch selection, and 4. Prior knowledge update by discarding target-irrelevant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

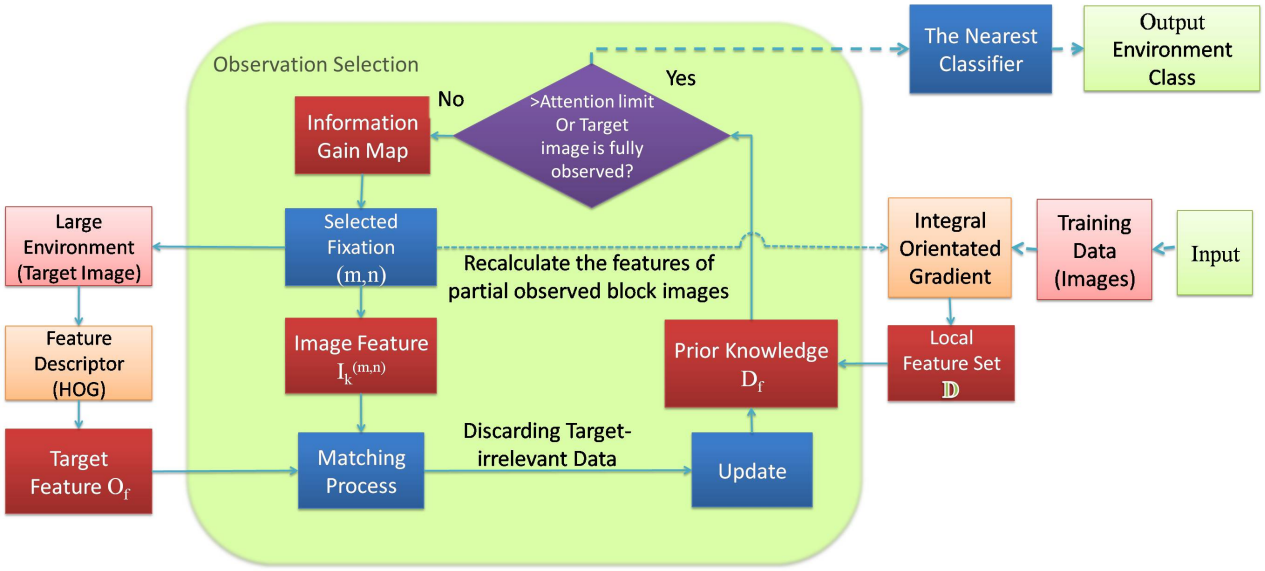


Fig. 2. Overall architecture of sequential patch selection for image retrieval

we propose that the target image is divided into rectangular blocks according to the coverage of the robot's camera and each block is partially overlapped by their neighbors, which can be seen in Fig. 3(b). Hence, there are more observable patches than non-overlapped condition. The framework of the adaptive observation selection is presented in Fig. 2. There are three components of the framework which are presented in the following.

Preprocessing All images which are used as the training data, are categorized into the given training class set, $\mathbb{C} = \{c_1, c_2, \cdots\}$. We extract the local features of each image patch by considering a partial overlap. There is the local feature set which is generated by the training dataset $\mathbb{D} = \{ \mathbb{I}_1, \mathbb{I}_2, \cdots, \mathbb{I}_N\}$ and each image consists of a set of local features like $\mathbb{I}_k = \{f^k_{(1,1)}, \cdots, f^k_{(r,c)}\}$, where a image is divided into r row and c column blocks. $f_{(m,n)}$ represents the local feature of each

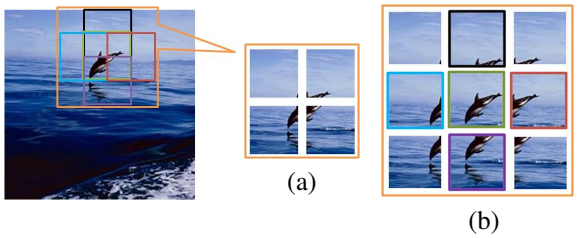


Fig. 3. Image patach selection with different methods. (a) block images without overlap, (b) block images with overlap

block image in the observed position (m,n) and the fixation is pointed with a coordinate (m,n). The prior knowledge will be updated after each selection by discarding 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 selected position (m,n) will be combined into the image feature $I_k = [\cdots, f^k_{(m,n)}]$. Note that I_k is the feature vector of the k-th image in the prior knowledge and I_k is generated by combining the features of the selected block image. $D_f = \{\cdots, I_k\}$ is the prior knowledge that contains all image features. The target feature $O_f = [\cdots, o_{(m,n)}]$ combines all local features that are sequentially observed from target environment. The observed image patches are used to calculate the similarity value with the prior knowledge. The similarity is calculated by using Cosine Similarity (CS) [9] as follows

$$Similarity = CS(O_f, I_k) = \frac{O_f \cdot I_k^T}{||O_f|| \times ||I_k||} \tag{1}$$

Where O_f is the feature of target environment and I_k is the image feature of the k-th image of the prior knowledge. We consider an adaptive observation selection method, therefore, the most dissimilar data will be discarded to update the prior knowledge. Based on the similarity values, a dissimilarity score can be calculated by

$$Dissimilarity = 1 - CS(O_f, D_f) \tag{2}$$

Information Gain In this step, an informative patch includes the best local features to retrieve the target image. The expected information gain in the selected fixation (m,n) can be measured by

$$G_{(m,n)} = E[1 - CS(\omega, D_f^{(m,n)}) \tag{3}$$

$$D_f^{(m,n)} = D_f \cup f_{(m,n)}] \tag{4}$$

where ω is the average feature vector of the prior knowledge $D_f^{(m,n)}$, which can be calculated by

$$\omega = E[D_f^{(m,n)}] \tag{5}$$

At the each iteration, the information gain is represented by averaging the dissimilarity values between ω and the local feature as

$$I_k^{(m,n)} = I_k \cup f_{(m,n)} \tag{6}$$

This information gain will be recalculated after updating the prior knowledge by discarding the irrelevant training dataset with the target images. The entire procedures are presented in Fig. 4.

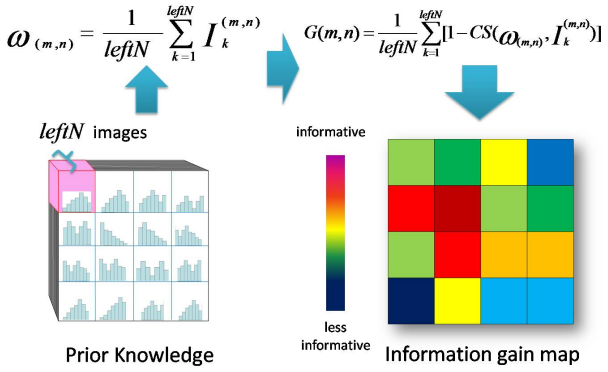
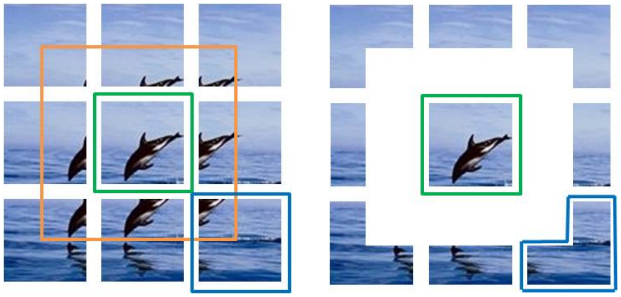


Fig. 4. Information gain map from the prior knowledge with training dataset

III. Environment Classification

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. 5. 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.



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

Fig. 5. Concept of Observation Selection

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. 5(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.

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. 5(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 this environment classification method with variable field of view is described in Fig. 2.

Integral Oriented Gradient There are many features detection methods base on visual features such as edges, contours, corners and regions for solving the problem in computer vision and image processing [10]. A high-performance feature detector should show the robustness to solve a computer vision task with various image conditions. In this research, one of the well known primitive feature such as Histogram of orientation gradient (HOG) [11] is used to represent the local block images.

It is necessary to recalculate the local features when the target image is observed sequentially with variable field of view. On the one hand, the boundary information of the small image patches will be neglected, it takes a lot of time in the local feature recalculating process, especially, when the training data is a large quantity and the overlap rate is high. There are some researches have been done for decreasing the computational complexity, like integral histogram [12] [13]. The local features can be easily calculated by applying the integral histogram. However, if the partial view is not a regular rectangular, it is also complex to calculate the feature vectors.

Every dataset are represented by two matrices, an orientation matrix and a gradient matrix. And two matrices have the same size of the image to store the orientation and gradient values at the same position. When a partial image patch is added into the current observation, the new feature vector is generated by extracting information of same partial area in the orientation matrix and gradient matrix. An image pixel mask is used to indicate either pixel has been observed or not. With this method, partial image patches are used to easily calculate the feature vectors even if the observation is irregular.

IV. Experimental Verification

A. LabelMe: urban and natural scene categories

The number of training dataset [14] is 2080 (size: 256 $\times$ 256 pixel) and the dataset is categorized by 8 classes (8 classes $\times$ 260 images). We assume that the field-of-view of the camera is limited to 64 $\times$ 64 pixels. Hence, the target environment is divided into 4 $\times$ 4, 5 $\times$ 5, and 7 $\times$ 7 blocks respectively under the 0% (non-overlap), 25% and 50% overlap conditions. The 10-fold cross-validation is applied to verify the general performance of the image retrieval task. The test is performed with several limited the number of selected fixations, and the retrieval accuracy and its final observation rate are stored.

Fig. 6. 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.

B. Experiment results

Fig. 7 shows the experiment results, the fixed field-of-view (green solid line) and the variable field-of-view (green dashed line) in the 25% overlap condition, and the average classification accuracy is 71.47% and 71.37% respectively with 100% attention rate. Fig. 7 also shows the experiment results of the fixed field-of-view in 50% overlap condition (red solid line) and an average of 73.07% classification accuracy is achieved with 100% attention rate. Non-overlap condition (blue) is shown as a contrast experiment, the classification accuracy is 72.64%. The classification decision is evaluated with the simple nearest neighbor algorithm [15]. 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 is generated by limiting the number of observations. Note that the classification processing does not stop until the selection times equals to observation limits or the target image is fully observed.

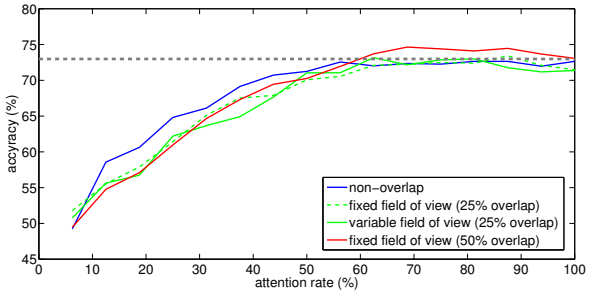


Fig. 7. Comparison of image retrieval accruacy between different selection approaches. Balck dashed-line presents the baseline accuracy using HOG feature with fully observed image.

When the attention rate is less than 60%, non-overlap case achieves the best performance. While the attention rate is higher than 60%, the performance of 25% overlap cases or

non-overlap case are almost same. However, it is obverse that the 50% overlap case shows the best performance in Fig. 7. Based on the experiment result, the gray dashed line is a base line of combining all feature vectors in 50% overlap case. It is not necessary to observe the entire environment, by sequentially selecting observations, the system can achieve same performance in the 50% overlap case just with 60% attention rate.

According to the experiment results and the notion of observation selection algorithm, because one image is divided into more blocks in the overlap condition, the system can select a better fixation to make the first view than non-overlap case. The observation is selected based on the concept that the informative position best preserves the dissimilarity across the whole training data. With out enough information input, namely, if the attention rate is less than 60%, the system needs to make its decision on the basis of variant local information that observed from the large environment. The information may be used many times in the overlap case. However, after discarding enough redundant training data with sufficient information input, the remaining training data are so similar. It proves that though some information is used many times, this information is still important for classifying the most similar data from the similar training data.

Fig. 8 (a), (b), (c) and (d) show the observation selection sequence of non-overlap, 25% overlap in fixed field of view, 25% overlap in variable field of view and 50% overlap case respectively. The 50% attention rate is achieved at step 8 of non-overlap case (shows in Fig. 8 (a) step 8), step 11 of 25% overlap case (shows in Fig. 8 (b) step 11) and step 14 of 50% overlap case (shows in Fig. 8 (d) step 14). It achieves 48.44% attention rate in step 11 of the variable field of view with 25% overlap (shows in Fig. 8 (c) step 11). While the attention rate is less than 50%, the system try to discarding dissimilar images from training dataset quickly, it takes 11 steps in 25% overlap and 14 steps in 50% overlap. The red dot is the center position of attentions, it shows that the differences among non-overlap case and 25% overlap cases are not so obvious, expect more observations are extract from the target image. In the 50% overlap case, as the red dot shows, more detailed information is accessed by the system. After discarding plant of images, the images of training dataset are so similar to each other that it is difficult to classify which class the target image belongs to. So, when the attention rate is higher than 50%, it can be seen that the system takes more views around the building's boundary in overlap case. The boundary information is used repeatedly in several observations, which promotes the system can carefully classify the images of target class from the rest of training data.

V. Conclusion And Future Works

In this study, we proposed a partially overlapped partition method for adaptive observation selection in large scale environment retrieval system. The new partition method provides a flexible observation selection and the system attains an important image patch to correctly retrieve the target image. Experimental results have shown that a partial observation could have enough information to efficiently retrieve the target image without the entire observations. It means that the proposed model can efficiently and effectively extract an informative area only and discard a meaningless region to enhance the image retrieval performance. For the improvements, an adaptive method will be investigated to automatically adjust overlapping ratio during the image retrieval tasks. Moreover, the proposed model will be applied to solve different tasks such as image classification with limited sensing coverage and next best view selection for the mobile robot by considering the three dimensional action spaces.

VI. 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] S. Jeong, S.-W. Ban and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment", Neural Networks, 21(10):1420-1430, 2008

[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] Y. Yang, H.T. Shen, Z. Ma, Z.Huang, X. Zhou, "L21-Norm regularized discriminative feature selection for unsupervised learning", in: Proceedings of the International Joint Coferences on Artificial Intelligence (IJCAI2011), 2011, pp. 1589-1594

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

[10] 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

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

[12] Viola,P., Michael Jones, J.: "Robust real-time face detection". International Journal of Computer Vision 57(2), 137-154(2004)

[13] Porikli, F.M.: Integral histogram: A fast way to extract histograms in cartesian spaces. Proc. of the IEEE CVPR, vol. I, pp. 829836 (2005)

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

[15] 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

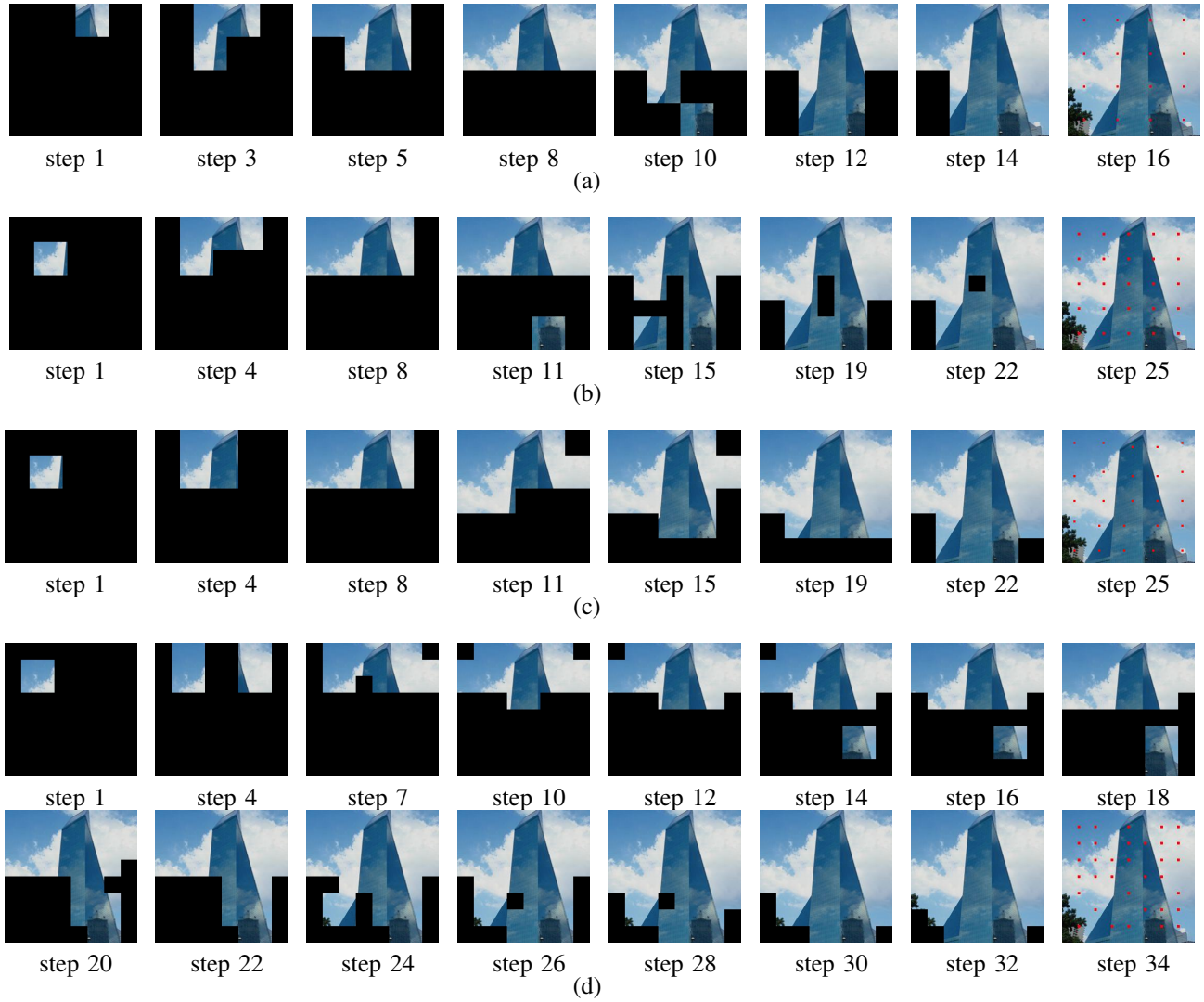


Fig. 8. Examples of sequential patch selection with different approaches. (a) attention sequences with non-overlap case, (b) attention sequences with 25% overlap case (fixed field of view), (c) attention sequences with 25% overlap case (variable field of view), (d) attention sequences with 50% overlap case(fixed field of view) .