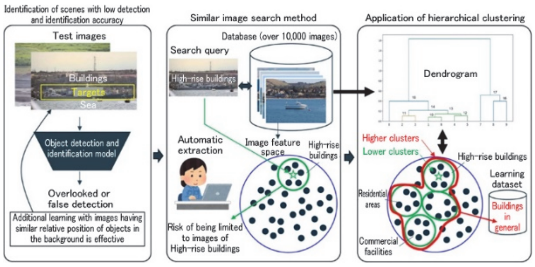
Image Retrieval Techniques for Enhancing the Robustness of Object Detection Models in Autonomous Vehicles


Identification of scenes with low detection and identification accuracy
Test images
Buildings
Targets
Sea
Object detection and identification model
Overlooked or false detection
Additional learning with images having similar relative position of objects in the background is effective
Similar image search method
Database (over 10,000 images)
Search query
High-rise buildings
Automatic extraction
Image feature space
High-rise buildings
Risk of being limited to images of High-rise buildings
Application of hierarchical clustering
Dendrogram
Higher clusters
Lower clusters
High-rise buildings
Residential areas
Commercial facilities
Learning dataset
Buildings in general

AMANE KOBAYASHI[*1] YURIKO YAMANO[*1]

YUSUKE KINOUCHI[*2] KIICHI SUGIMOTO[*3]

For unmanned systems used in offshore surveillance and coastal security, image recognition functions which can detect and identify targets and obstacles from camera images are essential. To achieve high image recognition, learning using images of scenes where accuracy of detection and identification is low is required. However, acquiring images for learning and searching for images with specific scenes requires significant effort. Therefore, a method to automatically search for images of specific scenes from large-scale image databases available online, using images with low detection and identification accuracy as input, has been developed. As a result, the workload required for this task could be reduced. Wide application of this developed technology to products equipped with image recognition functions that use deep learning is planned.

1. Introduction

Image recognition technology using deep learning has rapidly developed, and social implementation of methods to detect and identify specific objects in images, such as people and ships, is actively progressing [(1)]. For example, Mitsubishi Heavy Industries, Ltd. (hereinafter referred to as MHI) has applied an object detection and identification method to the function that detects suspicious vessels using cameras mounted on Uncrewed Aerial Vehicles (hereinafter referred to as UAVs) of unmanned systems used in offshore surveillance and coastal security. In this method, images containing targets which should be detected and identified, such as images of ships, are learnt in advance. A large amount of learning images with backgrounds of various offshore and coastal locations where operations are expected must be prepared. Testing is conducted using a learned object detection and identification model. For images with low accuracy results, such as those where targets are overlooked or non-targets are falsely detected, additional learning with images containing relevant scenes is performed. This additional learning is repeated until scenes with low accuracy are resolved and performance is improved.

Images for additional learning should ideally be images of scenes with objects, such as buildings and the sea in the background, in the same relative positions as those in scenes with low detection and identification accuracy. Acquiring such images involves developing a test plan, application for a filming permit, and procuring equipment. Accordingly, repeating this process for each additional learning session is a costly procedure. Therefore, MHI has developed and applied a method that utilizes large-scale image databases publicly available on the internet [(2)]. However, searching for images with specific scenes that contribute to learning from a large-scale database exceeding 10,000 images also requires significant effort, which is a challenge to improving development efficiency. Furthermore, if variation of the images selected for additional learning is restricted, for example, to images that only show specific buildings such as high-rise buildings in the background, there is the risk of degrading the robustness of the detection and identification model obtained through additional learning.

To address these challenges, MHI has developed a method that uses hierarchical clustering to classify large sets of images by scene and automatically selects images with the specifically desired scene using a similar image search. Hierarchical clustering allows for the adjustment of the

[*1] Control Systems Research Department, Research & Innovation Center

[*2] Research Manager, Control Systems Research Department, Research & Innovation Center

[*3] Principal Manager, Control Systems Research Department, Research & Innovation Center

granularity of scenes to be classified, and allows images to be categorized into, for example, those showing "high-rise buildings", images that show buildings in general, including high-rise buildings, commercial facilities, and residential areas, and images that contain any type of man-made structure. By applying a similar image search to these clusters and using an image with the desired characteristics as a query (image input into the search engine), images of a specific scene with similar characteristics are extracted. With this developed technology, learning with a diverse set of images desired by users is possible by parametrically adjusting granularity of scenes within hierarchical clustering.

2. Search technology for images with a specific scene using hierarchical clustering

As shown in **Figure 1**, this chapter describes the developed image search technology that incorporates both a similar image search and hierarchical clustering. This technology is useful when selecting images with a specific scene from a large-scale database, starting with the identification of scenes which had low detection and identification accuracy (such as scenes where the target is overlooked or non-targets are detected) during learned model testing. First, the similar image search method is described as existing technology, then preprocessing is performed to ensure that the similar image search focuses on the same relative position of objects in the background. Finally, the specific-scene image search technology that combines this with hierarchical clustering is explained.

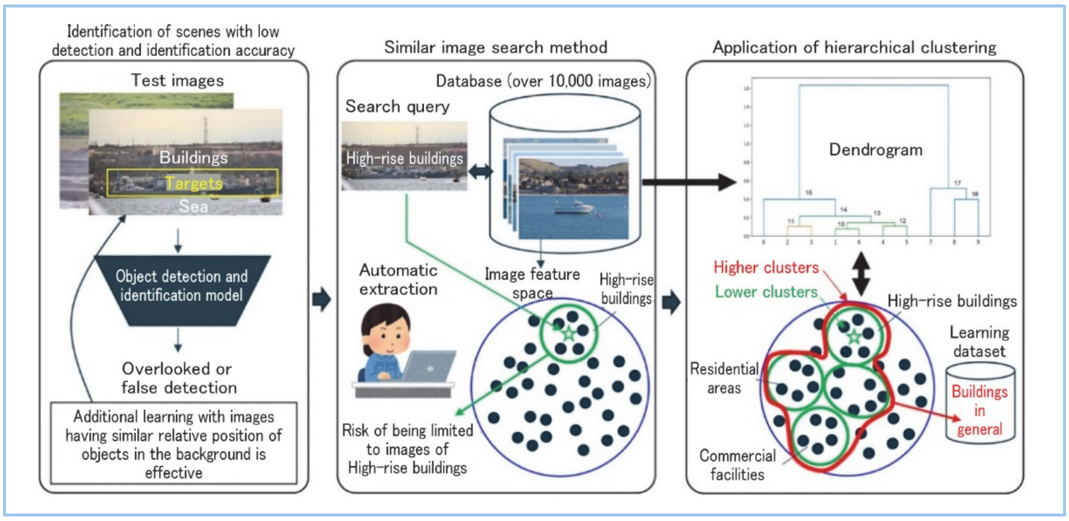

Identification of scenes with low detection and identification accuracy
Test images
Buildings
Targets
Sea
Object detection and identification model
Overlooked or false detection
Additional learning with images having similar relative position of objects in the background is effective
Similar image search method
Database (over 10,000 images)
Search query
High-rise buildings
Automatic extraction
Image feature space
High-rise buildings
Risk of being limited to images of High-rise buildings
Application of hierarchical clustering
Dendrogram
Higher clusters
Lower clusters
High-rise buildings
Learning dataset
Residential areas
Buildings in general
Commercial facilities

Figure 1 Outline of the developed method

(1) Similar image search method

First, all images to be compared are input into a feature extractor equipped with a neural network and feature vectors are output. Similarity between features of two images is calculated using cosine similarity between feature vectors using the following equation.

$$\cos(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j / \|\boldsymbol{x}_i\| \|\boldsymbol{x}_j\|$$

Here, $\boldsymbol{x}_i$ is the feature vector of the query image, and $\boldsymbol{x}_j$ is the feature vector of an image included in the search target database. Cosine similarity is calculated for all images in the database, and images with high scores are identified as images with features similar to those of the query image. However, simply picking up images with high similarity as learning candidates carries a risk of overfitting. Therefore, an approach to select images with a certain degree of variation, even among those similar to the query image, is required.

(2) Preprocessing that focuses on the background

In the similar image search, the feature extractor is applied to the entire area of an image. However, when objects targeted for detection and identification such as ships are present, the influence of their type, position, and direction is significantly large, and performing a similar image search that focuses on the background is not possible. As such, preprocessing to trim the area surrounding the target and mask the target area in black is performed as shown in

Figure 2. In this way, an image focusing solely on the relative position of objects in the background, for example, with a building at the top and the sea at the bottom, can be created. Then, feature vectors of the preprocessed image are output using the feature extractor. The feature extractor is a publicly available neural network from various IT companies that has learned to capture a wide range of image features. In this report, since the objective is to determine the relative position of objects in the background as a feature, Vision Transformer[3], which can divide relative positions within an image into small patches and vectorize their sequence as features, was adopted as the feature extractor. This trimming and feature extraction process is applied not only to the query image but also to all images included in the search target database.

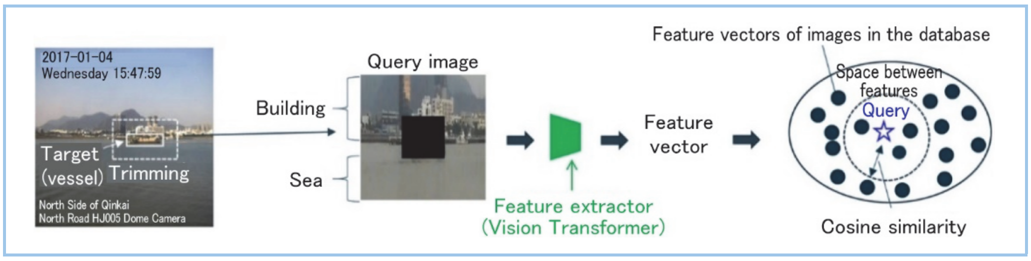


Figure 2 Preprocessing focusing on the background

(3) Specific-scene image search technology

Before implementing the similar image search process, hierarchical clustering is applied to images in the database to form a hierarchical structure represented by a dendrogram. To create the dendrogram, the following steps (i) to (iv) are performed.

(i) Distance between the feature vectors of all images in the database is measured using all combinations.

(ii) Pairs of images with close inter-vector distances are grouped together to form small-scale clusters.

(iii) Distance between clusters is evaluated and clusters are combined.

(iv) Step (iii) is repeated to form large-scale clusters.

To measure the distance between feature vectors $\boldsymbol{u}, \boldsymbol{v}$ and the distance between clusters in steps (ii) and (iii) above, various algorithms can be applied using Python libraries[4]. In this report, average linkage, which evaluates the following Euclidean distance r_k and distance between clusters R, was adopted.

$$r_k = \|\boldsymbol{u} - \boldsymbol{v}\|_2, R = \frac{1}{N}\sum_{k=1}^{N} r_k$$

Here, N represents the total number of combinations of image pairs contained within two clusters. An example of a dendrogram obtained by applying hierarchical clustering to a small-scale database is shown in **Figure 3**. The vertical axis represents the distance r_k between images #0-9, as well as the distance R between clusters #10-17. As the transition from a lower level to a higher level increases, cluster scale also increases, and the variation of the images included expands. The method to adjust the variation of a desired specific scene is described as follows.

The determination flow for clusters containing images of a specific scene is shown in **Figure 4**. First, a similar image search is executed as in (1), where the cosine similarity between the query image and all images in the database is evaluated to determine the most similar image. Next, standard deviation σ of the distribution of cosine similarity values is calculated for images contained in the lowest-level cluster of the dendrogram consisting of two or more clusters containing the query image. With the mode of the distribution defined as the peak position, peak distance d to the next cluster at the same level to be combined is measured. When peak distance d satisfies the following conditions, the clusters are combined, and the same process is implemented for a higher level cluster.

$$d \leq n\sigma, \quad n\text{: integer}$$

By repeating the above process, a hierarchical level where the conditions in the equation are no longer satisfied and clusters can no longer be combined will eventually be achieved. The cluster that is created as a result of the combinations up to this level is adopted as the cluster to represent the specific scene. By adjusting the value of parameter n, the variation of images within the specific-scene cluster can be adjusted.

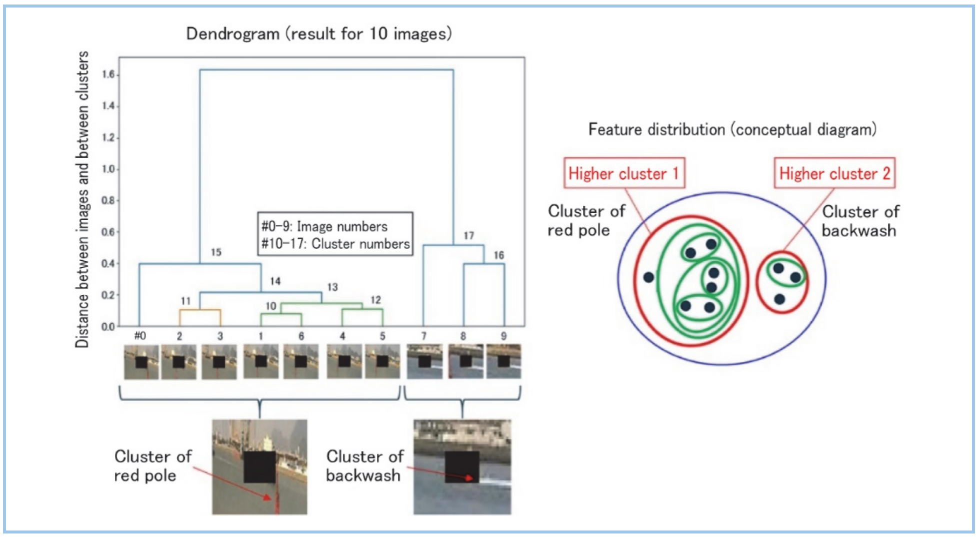


Figure 3 Example of hierarchical clustering application

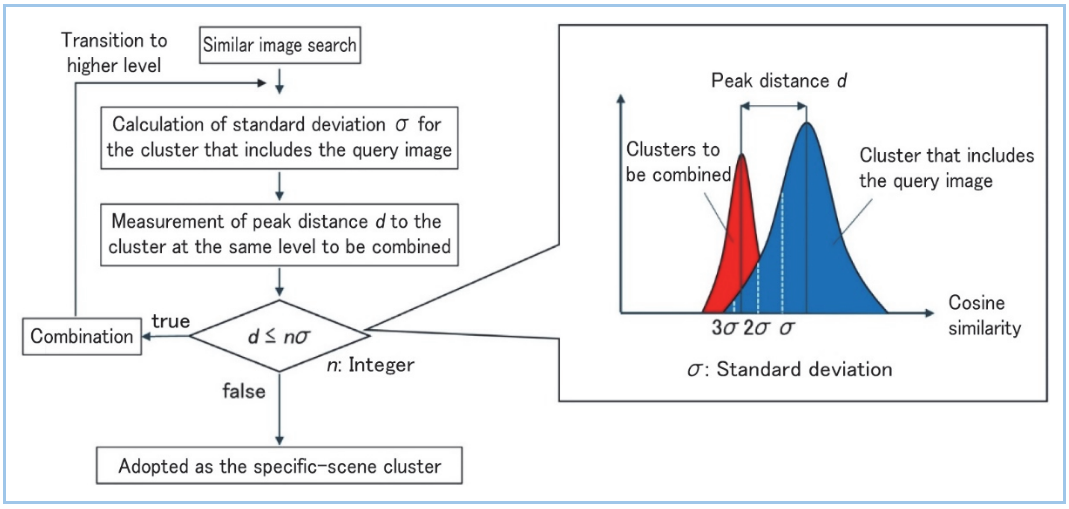

Transition to higher level
Similar image search
Calculation of standard deviation σ for the cluster that includes the query image
Measurement of peak distance d to the cluster at the same level to be combined
Combination
true
$d \leq n\sigma$
n: Integer
false
Adopted as the specific-scene cluster
Peak distance d
Clusters to be combined
Cluster that includes the query image
Cosine similarity
3σ 2σ σ
σ: Standard deviation

Figure 4 Determination flow for specific-scene image clusters

3. Evaluation using public databases

In this chapter, the search technology for specific-scene images as described in Chapter 2 was evaluated using 10,080 images (7,080 from SeaShips [(5)] and 3,000 from Mcships [(6)]) from public databases, as the search targets. As a result of applying hierarchical clustering, the dendrogram shown in **Figure 5** was obtained. Here, clusters could be organized based on each image feature. However, visually analyzing was determined to be impractical. For this database, identifying image clusters with a specific scene, buildings at the top and the sea at the bottom, was set as the problem, and three query images as shown in **Figure 6** were also set. Although these are all categorized as having buildings, they represent scenes with high-rise buildings, residential areas, and commercial facilities at the top, respectively.

Similar image searches were executed independently for each of the three images; higher-level clusters were repeatedly combined starting from the cluster which contained the most similar image. Distribution width 3σ of the cosine similarity and peak distance were calculated at each combination step, and the results are shown in Figure 6. Distribution width of the cosine similarity tends to increase with each repeated combination, indicating that the variation of images within the cluster increases. Since the peak position changes with each combination, no fixed trend is observed in peak distance during the transition from a lower level to a higher level. However, when the combination process for cluster #1964 was implemented, a large difference exceeding 3σ occurred, and images far from the query image were found to be mixed into the cluster after the combination. The above analysis is consistent with the logic described in Section (3) of Chapter 2, and cluster #1964 could be identified as a specific-scene image cluster.

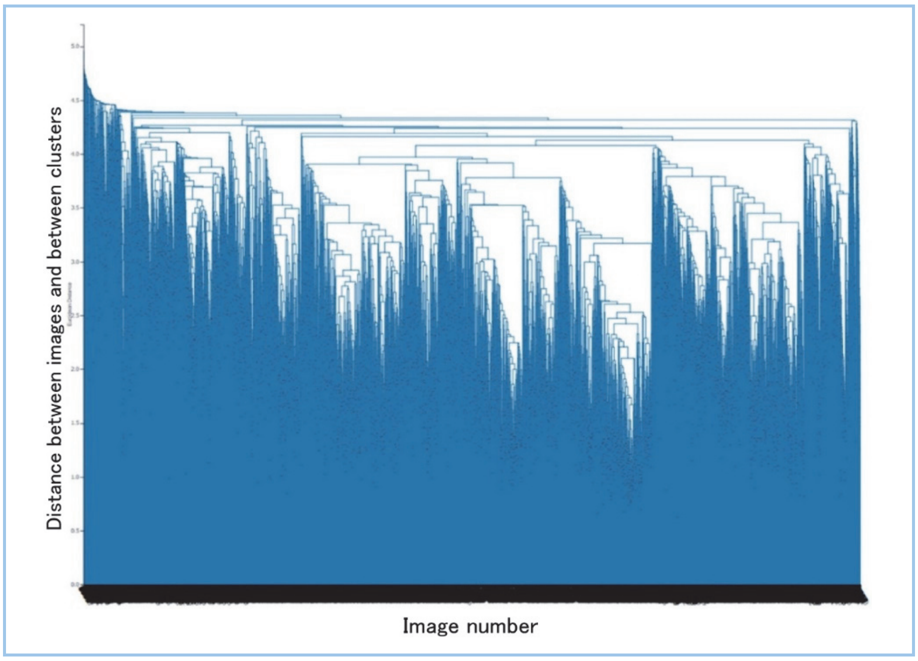


Figure 5 Dendrogram of a public database visualized using hierarchical clustering

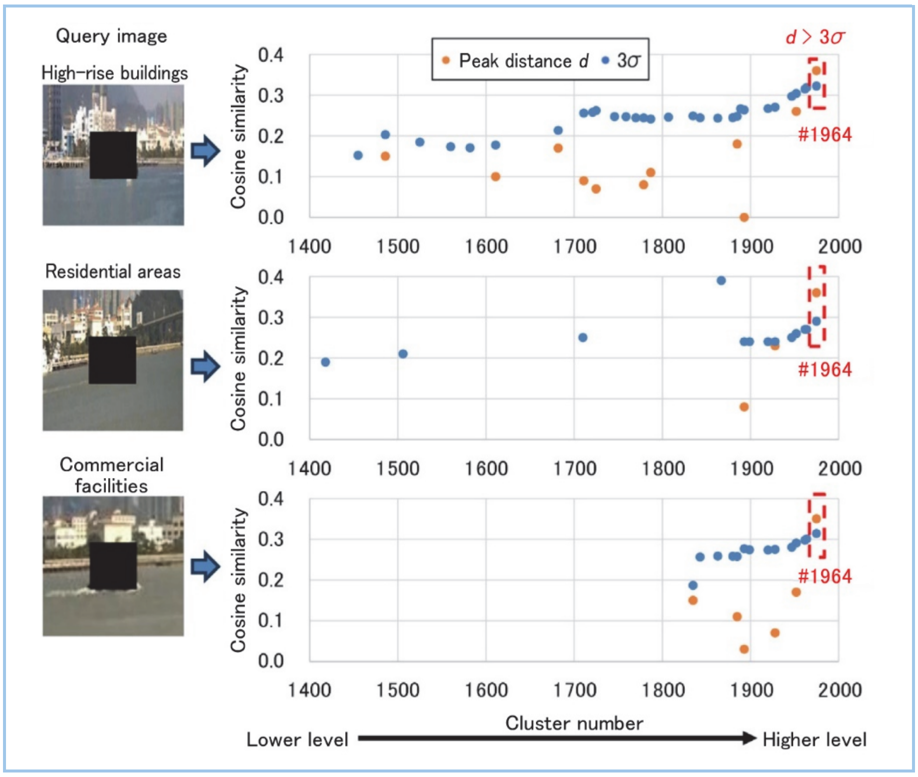


Figure 6 Query images (left) and determination results for specific-scene image cluster (right)

As shown in **Figure 7**, the images included in cluster #1964 contained incorrectly extracted images in addition to those showing the desired relative position of objects. Furthermore, some desired images were present among those not included in cluster #1964 and were overlooked. However, no critical search errors occurred, such as false extractions where the image contain a man-made structure other than a building, or overlooking a building that was small in the image. To evaluate the accuracy of the search results, the rate of overlooking an image (number of images overlooked and not included in the cluster, divided by the total number of desired images in the database) and the rate of false extraction (number of undesired images divided by the total number of images in the cluster) were calculated, resulting in a rate of overlooking an image of 5.7% and a rate of false extraction of 2.3%. These results demonstrate that the vast majority of desired images were successfully extracted from the large-scale database with an extremely low proportion of false extractions.

Based on the above results, the characteristics of the search technology presented in this report were analyzed. First, as shown by the dendrogram in Figure 5, in order to perform a visual analysis of a dendrogram from a large-scale database exceeding 10,000 images, automatic specific-scene cluster determination logic is essential. Additionally, since the same cluster was identified as the specific scene for all three query images, extracting images of buildings in general, regardless of the specific type such as a high-rise building or commercial facility, is considered possible with this technology. With conventional similar image search methods, a query of a high-rise building would only retrieve images showing high-rise buildings. Consequently, all images of the desired building type needed to be prepared as queries. With this developed technology, however, just one image of the building is needed for extraction. Furthermore, regarding the image variations included in the determined cluster, selection can be controlled by adjusting the threshold for the distance between peaks; setting a wider threshold would result in identifying all types of man-made structures, while setting a narrower threshold would result in identifying only high-rise buildings. That is, the peak distance threshold can be utilized as a parameter to adjust the variation of images to be used as candidates for additional learning.

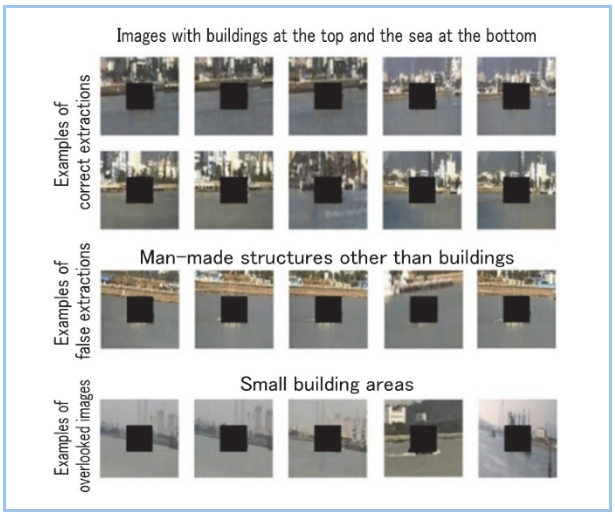

Images with buildings at the top and the sea at the bottom
Examples of correct extractions
Man-made structures other than buildings
Examples of false extractions
Small building areas
Examples of overlooked images

Figure 7 Images determined as the specific scene, false extractions, and overlooked images

4. Conclusion

In this report, a method to automatically search for images of scenes where accuracy of detection and identification is low from a large-scale database and extract them as additional learning data. This method contributes to a reduction in the workload required for image acquisition and searching for relevant scenes, as well as an improvement in development efficiency. Furthermore, even when a high-rise building image is used as a search query, extracting image clusters with a wide variation of other buildings can be allowed. This is expected to prevent overfitting in additional learning and facilitate the construction of high-quality object detection and identification models that function robustly under various conditions. In the future, MHI aims to further advance image recognition technology by utilizing this method in various applications and identifying any challenges.

References

(1) L. Jiao et al., A Survey of Deep Learning-Based Object Detection, IEEE Access, vol. 7, pp. 128837-128868, 2019.

(2) A. Kobayashi et al., Object Detection Developed with Small Number of Acquired Images by Multi-step Training Method in Image Recognition Using Deep Neural Network, Mitsubishi Heavy Industries Technical Review, Vol. 61, No. 1, 2024

(3) Dosovitskly et al., AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE, arXiv:2010.11929v2, 2021.

(4) The SciPy community, Hierarchical clustering, https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

(5) Shao et al., SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection, Transactions on Multimedia, vol. 20, no. 10, pp. 2593-2604, 2018

(6) Zheng and Zhang, Mcships: A Large-Scale Ship Dataset For Detection And Fine-Grained Categorization In The Wild, ICME, pp. 1-6, 2020.