A METHOD FOR REGIONAL ANALYSIS USING DEEP LEARNING BASED ON BIG DATA OF OMNIDIRECTIONAL IMAGES OF STREETS

T. Oki[1, *], Y. Ogawa[2]

[1] Tokyo Institute of Technology, 2-12-1-M1-27 Ookayama, Meguro-ku, Tokyo 152-8550, Japan - oki.t.ab@m.titech.ac.jp
[2] The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan

Commission IV

KEY WORDS: Omnidirectional image, Deep learning, Big data, Semantic segmentation, Clustering, Computer vision.

ABSTRACT:

In this paper, we propose a method for regional analysis using image recognition technology based on deep learning and big data of street images captured by omnidirectional cameras on vehicles. Specifically, we first construct a classification method of regions based on street images using a pretrained deep learning model (VGG16) for image recognition as a feature extractor. Next, we develop a method to evaluate the landscape and safety of streets based on the ratio of street components (such as buildings, roads, fences, vegetations, sky, street lights) at each shooting point, which is calculated by semantic segmentation.

1. INTRODUCTION

1.1 Background and Research Objective

Various methods have been proposed for characterizing and classifying regions until today. However, it has been difficult to analyse for a wide area with high spatial resolution.

Recently, with the advancement of measuring instruments and IoT technology, a wide variety of big data on buildings and cities is becoming readily available. The development of artificial intelligence (AI) technology to utilize the obtained data is also remarkable. Therefore, the study of new methods of regional analysis that utilize such data and technologies is of significant importance.

Therefore, the purpose of this paper is to construct a novel method for regional analysis using image recognition technology based on deep learning and big data of streetscape images captured by omnidirectional cameras installed in vehicles.

The overview of research process is shown in **Figure 1**. Specifically, we first construct a region classification method based on street images using a trained deep learning model for image recognition as a feature extractor. Next, we propose a method for calculating the compositional ratio of street components (such as buildings, roads, fences, plants, sky, streetlights) at each shooting point using automatic semantic segmentation of each pixel. Using the proposed method, we attempt to evaluate the landscape and safety of streets.

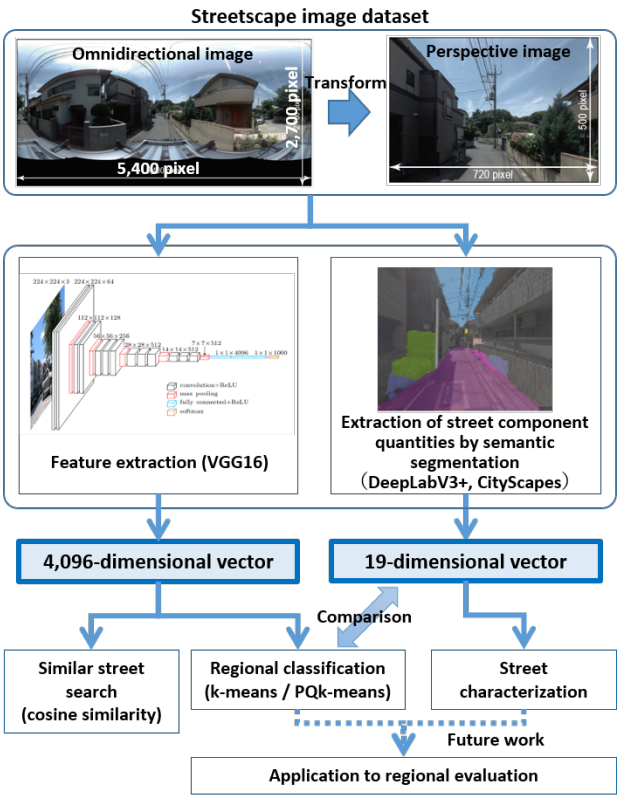

Figure 1. Overview of research process in this paper.

1.2 Related Works

The number of studies that have attempted to evaluate streets using Street View images (provided by Google, Baidu, or Mapillary) has increased in recent years.

For example, Naik et al. (2014) conducted a questionnaire on more than 7,000 subjects using about 1 million images obtained from Google Street View and built a machine learning model to estimate the safety of streets from street view images. The Place Pluse 2.0 dataset, constructed by the MIT Media Lab, is based on approximately 1.11 million Google Street View images collected from 56 cities worldwide. This dataset is based on a pairwise comparative evaluation of six perspectives (safe, lively, boring, wealthy, depressing, and beautiful) on over 80,000

* Corresponding author

subjects. Some studies have attempted to use this dataset as training data to calculate the rank and score of street images by deep learning (Abhimanyu et al. (2016), Min et al. (2020)). In addition, there are some studies that attempt to classify the use of buildings along the street from street images (Liu et al., 2017).

Many other studies have applied Street View images for landscape evaluation. For example, there are studies attempting to quantify the continuity of streetscapes (Ki and Lee, 2020), green view index (Ma et al, 2021), the gap between the street semantic metric and urban renewal (Wang et al., 2019), the relationship between the perception of neighbourhood safety and mental health (Zhou et al., 2019), visual walkability (Laupheimer et al., 2018).

Besides, Verma et al. (2019) are unique in that they used auditory as well as visual information in their study of urban perception. Law and Neira (2019) applied an unsupervised learning model based on CNN and principal component analysis to street-level geographic knowledge discovery. In another unique study, Oki and Kizawa (2021) compared the results of impression evaluation in densely built-up wooden residential areas with the subjects' eye movements at that time. They also constructed an impression evaluation model using deep learning and attempted to compare the differences in evaluation structures between humans and AI.

However, all of the above research examples are limited to the evaluation of streets from a specific perspective(s), and few studies consider diverse perspectives. In addition, there are few cases in which the features obtained from street images are applied to the analysis, such as regional classification. Furthermore, few studies consider technical issues such as image pre-processing methods.

2. METHODOLOGY

2.1 Image Feature Extraction

In this study, we use the VGG16 model (Simonyan and Zisserman, 2014), whose parameters have been trained on the ImageNet dataset, as the feature extractor. The VGG16 model is a kind of convolutional neural network originally designed for image classification of 1,000 classes, so it has an excellent performance in image feature extraction. Therefore, we use the vectors (4,096 dimensions) obtained from the first fully connected (FC1) layer of the VGG16 model as the feature vectors of the images.

2.2 Regional Classification Using Feature Vector

The feature vectors enable us: to calculate the similarity between street images based on cosine similarity; to plot the positional relationship of each image on a two-dimensional plane after applying any dimension reduction method (such as principal component analysis, multidimensional scaling, t-SNE); and to perform regional classification using non-hierarchical cluster analysis.

However, when the number or size of images is large, it becomes difficult to apply clustering methods such as the general k-means method from the viewpoint of computational load. Therefore, we show that the PQk-means method by Matsui et al. (2017) can be applied to Dataset 1, which has about 1.8 million images, to classify regions appropriately.

2.3 Street Evaluation

A feature vector captures the features of an image by a machine learning model, but it is difficult for humans to interpret this multi-dimensional vector. Therefore, we use semantic segmentation, an automatic pixel-by-pixel semantic mapping task, to mechanically calculate what and how much is in each image and perform a street evaluation. Specifically, the DeepLab V3+ model (Chen et al., 2018) trained on the CityScapes dataset (Cordts et al., 2016) is used to calculate the composition ratio of elements in each street image.

3. STREET IMAGE DATASET

3.1 Omnidirectional Image of Streets

In this paper, we use the omnidirectional image big data provided by Zenrin Corporation for analysis. A special vehicle equipped with a 360-degree camera on its roof patrolled the areas and took omnidirectional images of the streets at 2.5-meter intervals. Each image is accompanied by basic information such as the location of the shooting point (latitude and longitude) obtained by GPS, the direction of the vehicle (north, east, and south are indicated 0, 90, and 180 degrees, respectively), the date and time of the shooting, and the vehicle ID. As shown in **Figure 1**, the original image is 5,400 pixels wide by 2,700 pixels high and is based on the equirectangular method, which means that the distortion increases toward the top or bottom of the image. In this case, about 35% or 45% of the image (depending on the vehicle) shows the vehicle's roof. In addition, the omnidirectional image can be freely transformed into a cube map, a circumferential fisheye image, or a perspective image. This study converts an omnidirectional image into a perspective image based on the same parameters as Google Street View (120 degrees field of view, 0 degrees elevation angle). Then, we discuss the difference in analysis results between the omnidirectional image and the perspective image (**Figure 1**).

3.2 Dataset

Dataset 1 consists of 1,789,821 omnidirectional images taken in the entire Setagaya Ward (including some outside the ward). Some images from Dataset 1 are sampled for detailed analysis. Specifically, 4,689 consecutive images (Dataset 2), 1,000 consecutive images (Dataset 3), and 1,000 images randomly selected from the entire Setagaya Ward (Dataset 4) taken on the same day by a single vehicle near Sangenjaya Station will be used for analysis.

4. SIMILAR STREET SEARCH AND REGION CLASSIFICATION RESULTS

4.1 Visualization of the Location of Each Image

In this section, after obtaining the feature vectors for 1,000 consecutive omnidirectional images (Dataset 3), we attempted to place the positional relationships of the images on a two-dimensional plane using the multidimensional scaling method (**Figure 2**). Here, the images are divided into two large clusters, and we can see the differences in the amount and distribution of trees and plants between clusters. Even in two consecutive images, the distance between the images is long in some cases due to the proximity of the trees.

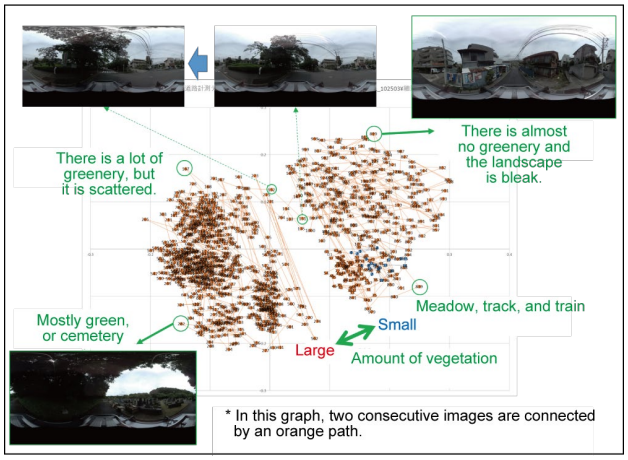

Figure 2. Example of a two-dimensional planar plot of 1,000 consecutive image feature vectors.

4.2 Results of Similar Street Search

Figure 3 shows the results of similarity retrieval based on the cosine similarity between image feature vectors for the perspective images of Datasets 2 and 4.

Because humans make judgments based on the overall impression of the entire image, the similarity between two consecutive points (2.5 m apart) seems to be high, but the cosine similarity is only about 80% (**Figure 3(a)**). On the other hand, there are cases where the cosine similarity exceeds 80%, even for images of discontinuous points (**Figures 3(b) and 3(c)**). A closer look at the extracted images reveals similarities in the individual components even though the overall impression is different, such as the presence of trees, the composition of the sky and road, power lines, and road lines. This is thought to be due to the mechanism of convolutional neural networks.

In interpreting the results of this analysis, it should be noted that there is a difference in perception between humans and AI.

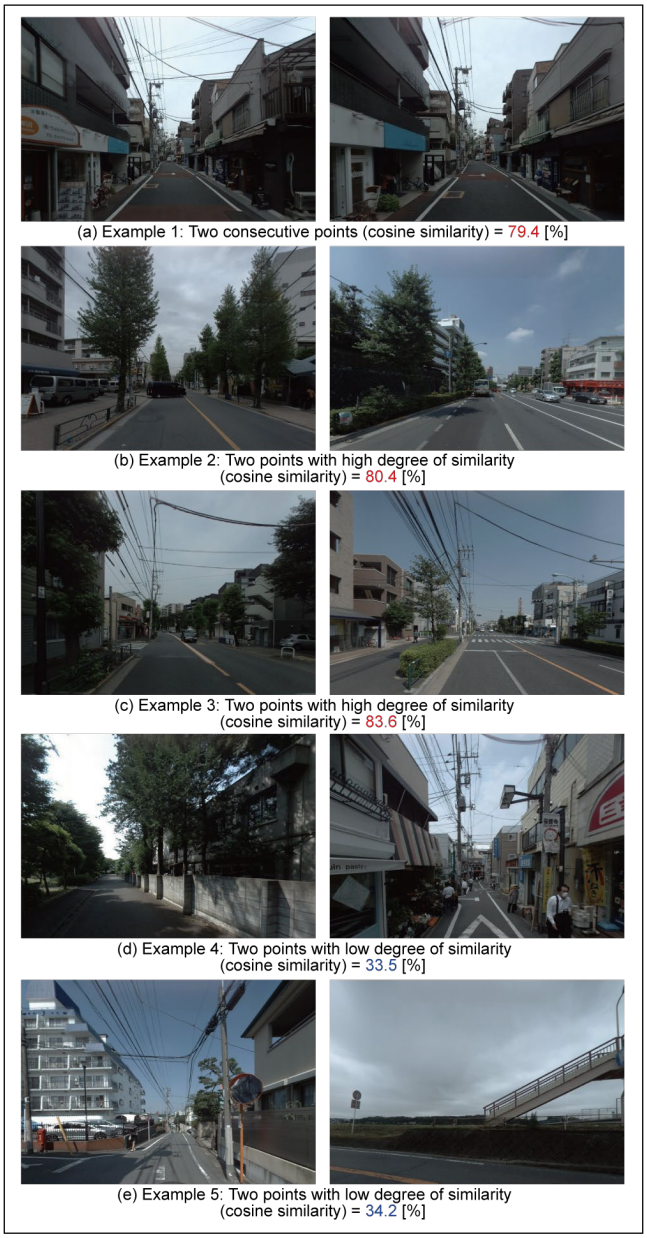

Figure 3. Examples of similar street search based on cosine similarity.

4.3 Results of Regional Classification (Dataset 2)

Through the analysis in Sections 3.1 and 3.2, we confirmed the usefulness of the image feature vectors extracted by the VGG16 model for both omnidirectional and perspective images. In this section, we attempt to classify regions based on image feature vectors using the general k-means method, using a series of images (Dataset 2) as an example. Here, we examine how the classification results change when using omnidirectional images and when using perspective images.

Figure 4 shows the results of the regional classification when the number of clusters is set to 10. There is no significant difference between the case using the omnidirectional image (**Figure 4(a)**) and that using the perspective image (**Figure 4(b)**), and it can be seen that the same streets are generally classified into the same clusters. Note that there are some cases where the cluster-ID frequently

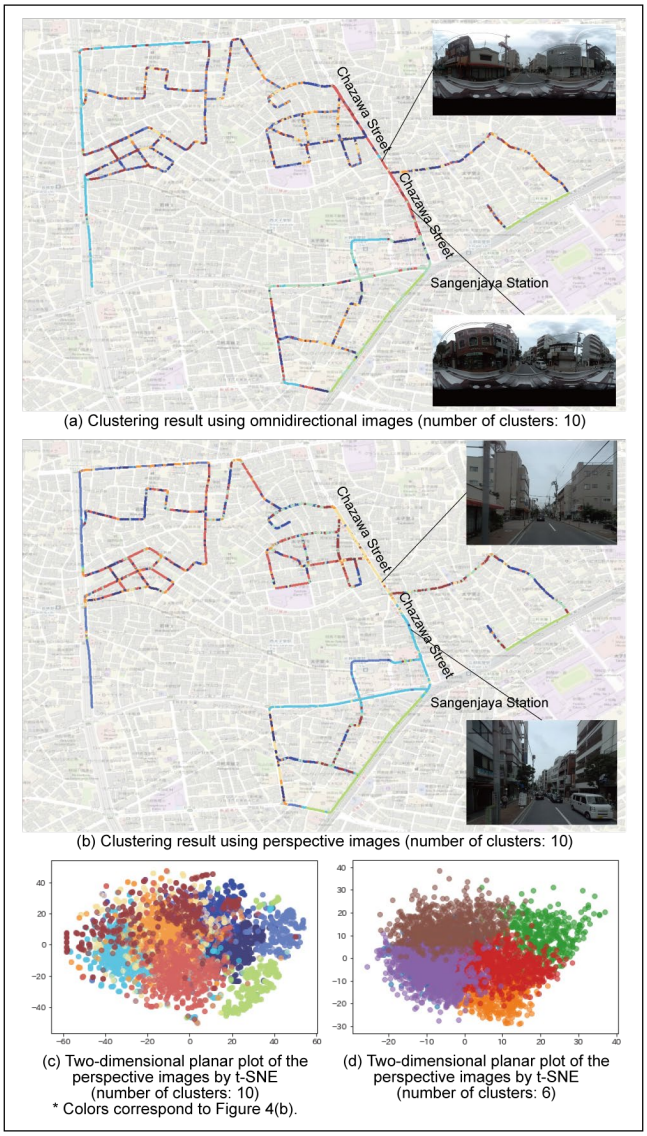

Figure 4. Comparison of clustering results using different type of images in the same region.

changes within the same street. For instance, Chazawa Street, which runs north-south on the east side of Sangenjaya Station, is generally classified into one cluster in the omnidirectional image. However, in the perspective image, the clusters change in the middle of the image. The north cluster has a relatively wide frontage of buildings. By contrast, the south cluster has a narrow frontage of buildings, and the building facades tend to be complicated.

In omnidirectional images, the area occupied by the buildings in the image is relatively small, so these differences in building characteristics tend to be overlooked. On the other hand, in perspective images, only one feature in one direction can be considered from a single image. It is necessary to use them differently according to the purpose of the analysis.

4.4 Regional Classification Results (Dataset 1)

As the number of images to be classified increases, it becomes difficult to apply the k-means method used in the previous section. However, anticipating the future era of street image big data, it is important to consider regional classification

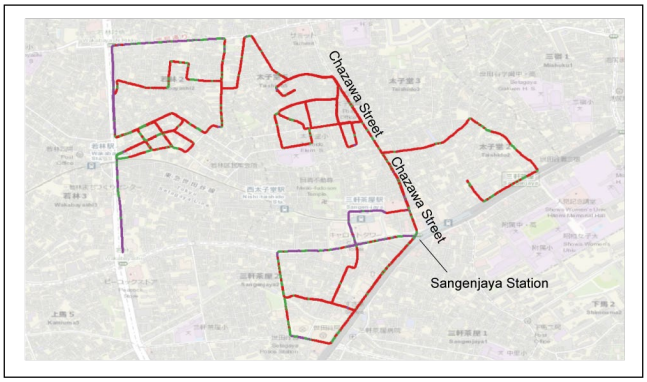

Figure 5. Result of regional classification using PQk-means.

methods for large datasets. Here, we attempted to classify approximately 1.8 million images in Dataset 1 into 10 clusters using PQk-means (Matsui et al., 2017), which can compress the 4,096-dimensional feature vectors extracted by the VGG16 model to a lower dimension such as four dimensions.

Focusing on the same images as in Dataset 2 (**Figure 5**), PQk-means generally classified them into three clusters; the percentage of images belonging to one cluster (red) is high, but this trend may change by increasing the number of dimensions after compression (reducing the loss of features).

Next, **Figure 6** shows the spatial distribution of clusters #0 to #9 in the whole Setagaya Ward. Clusters #3, #8, and #9 are unevenly distributed on the east side of Setagaya Ward because the degree of reflection of vehicles in the omnidirectional images is higher in these clusters comparing with other clusters. We randomly extracted 100 images from each cluster to validate the regional classification and applied the semantic segmentation described in Section 2.5 to each image. In other words, we attempted to see how the distribution of street components differs among clusters (**Figure 7**).

First, the interpretation of the seven clusters (except clusters #3, #8, and #9) is described below:

[Cluster #0] The large proportion of roads, sidewalks, and sky and the relatively small proportion of buildings indicate that the areas are likely to be along main roads or wide roads.

[Cluster #1] The large proportion of vegetation and terrain indicates areas with relatively lots of greenery.

[Cluster #2] The small proportion of roads, fences, walls, vegetation, and sky and the small proportion of buildings indicate dense residential areas.

[Cluster #4] The proportion of roads and sidewalks is small, but that of fences and walls is large. The areas have few characteristics.

[Cluster #5] The proportion of roads, sidewalks, buildings, and fences is large, and the proportion of the sky is small. There is a possibility that the areas are well-developed residential areas or areas dedicated to medium and high-rise residential buildings. In elevated roads, the elevations above the roads tend to be misidentified as buildings and classified into this cluster.

[Cluster #6] The proportion of poles is large, and the proportion of the sky is also somewhat large. These areas have few features.

[Cluster #7] The proportion of sidewalks and sky is slightly large. These are comparatively standard areas.

Next, the interpretation of the three clusters (#3, #8, and #9) with a large reflection of a vehicle is as follows:

[Cluster #3] The proportion of roads, vegetation, and cars is small, and the proportion of buildings is large. There is a possibility that these are dense residential areas where the car

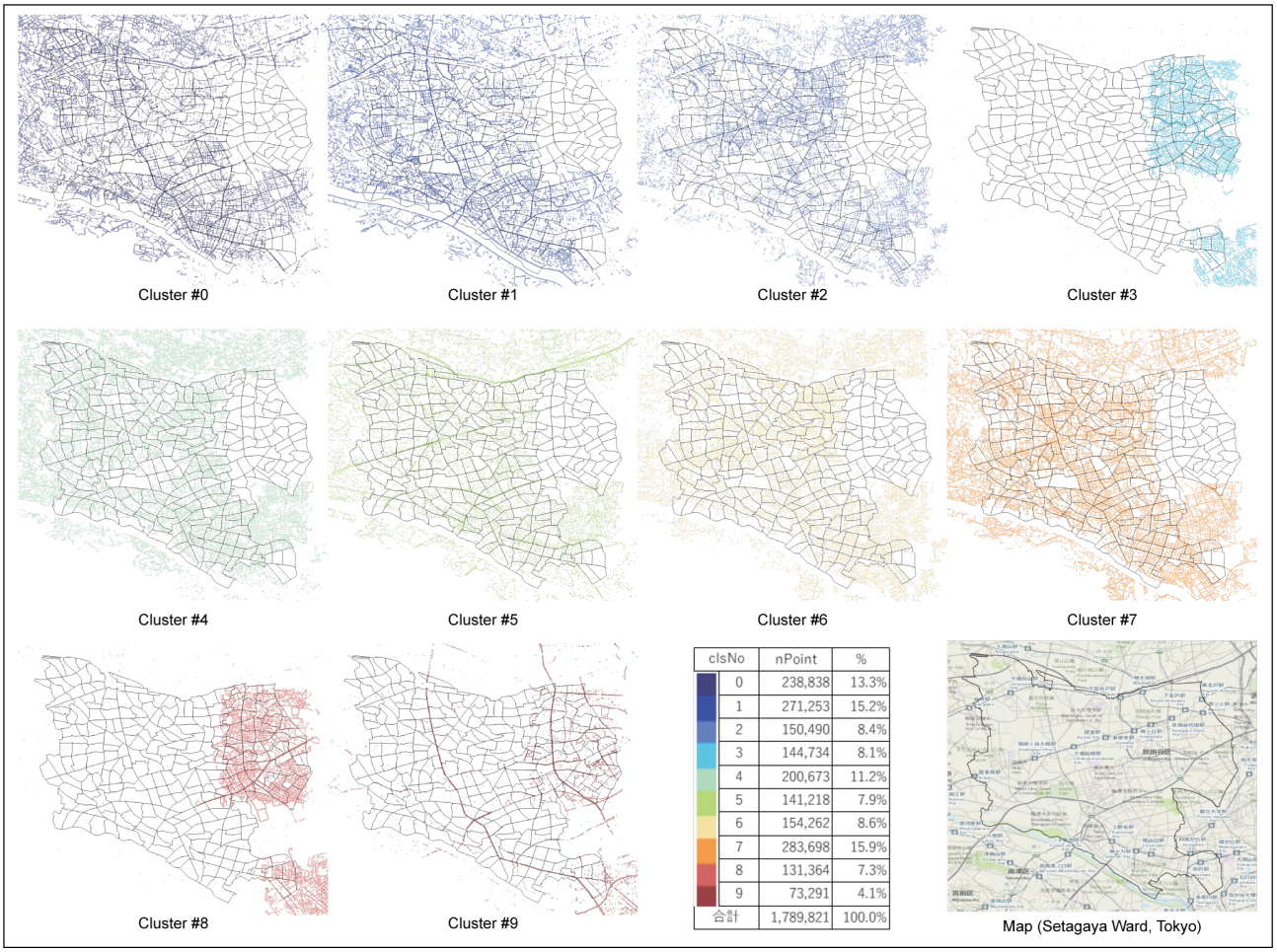

Figure 6. Clustering results of about 1.8 million images (database 1) the whole Setagaya Ward using PQk-means (number of clusters = 10).

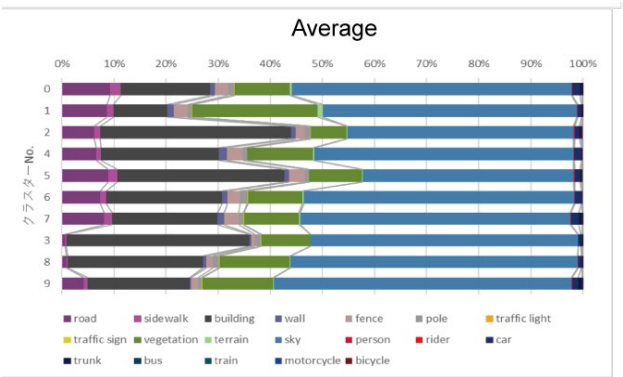

Figure 7. Comparison of street component distributions among clusters based on semantic segmentation.

ownership rate is low. It is similar to Cluster #2.

[Cluster #8] Comparatively standard areas.

[Cluster #9] The proportion of roads, sidewalks, vegetation, sky, and cars is large, and the proportion of buildings is small. It is considered to be well-developed residential areas, similar to Cluster #1. Many of the images taken on Kannana-dori and Kanpachi-dori also belong to this cluster.

5. APPLICABILITY TO STREETSCAPE AND SAFETY EVALUATION

5.1 Purpose of the Analysis

The previous section showed the results of a kind of AI model that mechanically extracted image features and applied them to similarity search and regional classification. However, to judge the validity of the classification results, it was necessary to calculate the percentage of street components in each image using semantic segmentation and discuss the results based on the calculation. In this way, if the images were given meaning, it would be easier for humans to understand the images and compare them with what they would sense.

In this section, we discuss the possibility of simple landscape evaluation and safety evaluation using street image big data and semantic segmentation. Here, we will use the perspective images of Dataset 2 as an example, which are continuous images but contain streets with various characteristics.

5.2 Results of Street Characterization

Figure 8 shows the visualization results of the percentage of street components based on semantic segmentation. Because the perspective images are close to the human field of view, for example, when looking at the percentage of vegetation, it is not necessarily the case that the value is higher for streets with a

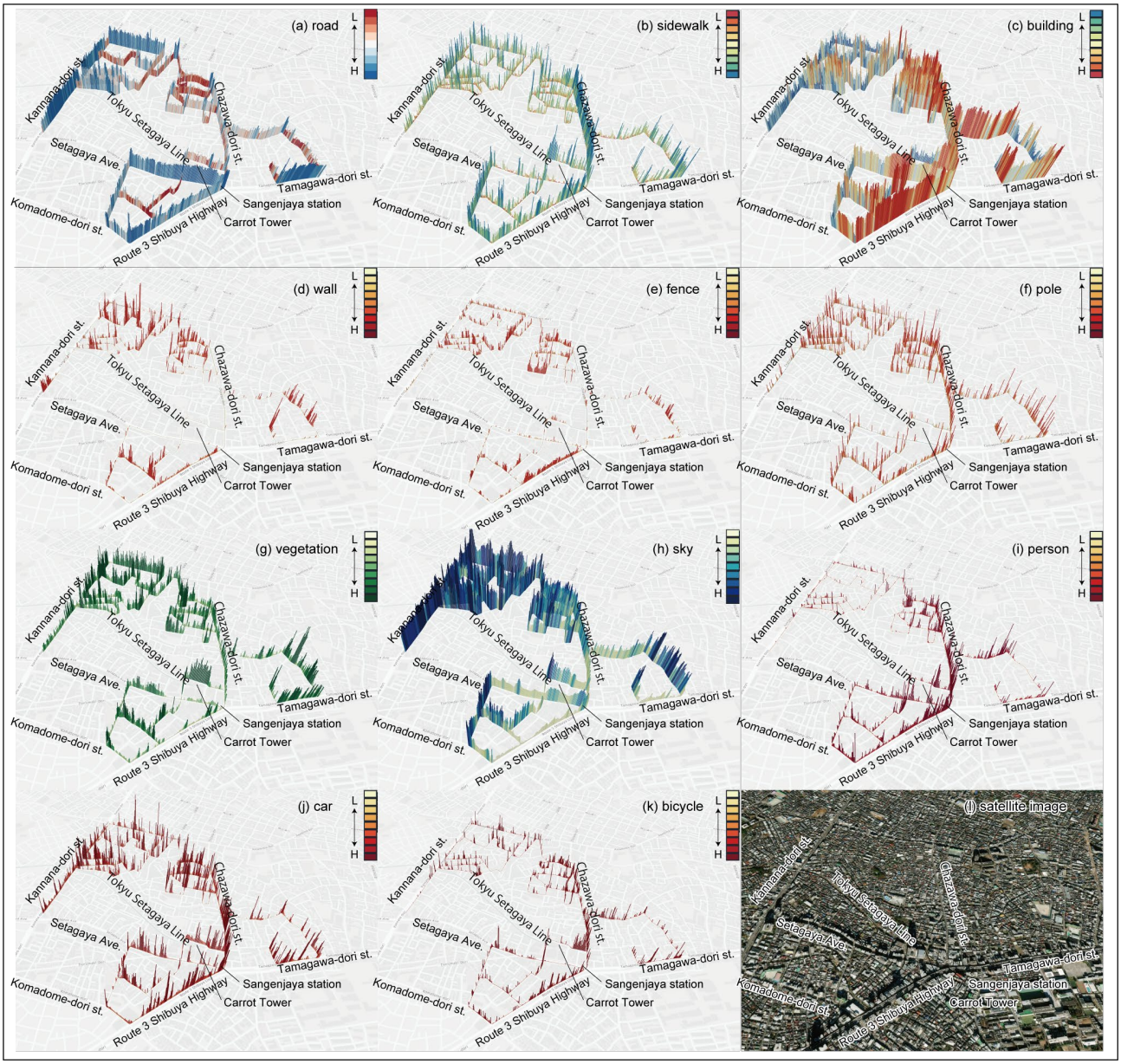

Figure 8. Visualization of street features by semantic segmentation. The length and colour of the bars represent the percentage of each component.

large absolute number of trees and plantings. In other words, it is possible to extract the amount of greenery perceived at eye level while passing through the streets. Other examples are listed below:

- The degree of sidewalk maintenance ((b) sidewalk)
- The degree of oppression from buildings (equivalent to D/H in the existing index; (c) building)
- The number of crime-resistant or collapsible block walls ((d) wall and (e) fence)
- The number of poles and the degree of street lighting maintenance ((f) pole)
- Sky rate ((h) sky) (Nishio and Ito, 2020)
- Street liveliness and crowd density ((i) person)
- Car ownership rate or amount of on-street parking ((j) car)
- The number of bicyclists and the number of abandoned bicycles ((k) bicycle)

These data can be applied to various types of regional analysis (e.g., walkability evaluation, crime prevention, and disaster prevention performance evaluation). Compared with the spatial distribution of crimes published in the crime information map (Metropolitan Police Department, 2016) (**Figure 9**), there appears to be a relationship between the spatial distribution of the occurrence of crimes related to children and women and the spatial distribution of the amount of vegetation and sky. In addition, the range of applications is expected to expand further by combining with a deep learning model for estimating human impressions from street image big data (Kizawa and Oki, 2021).

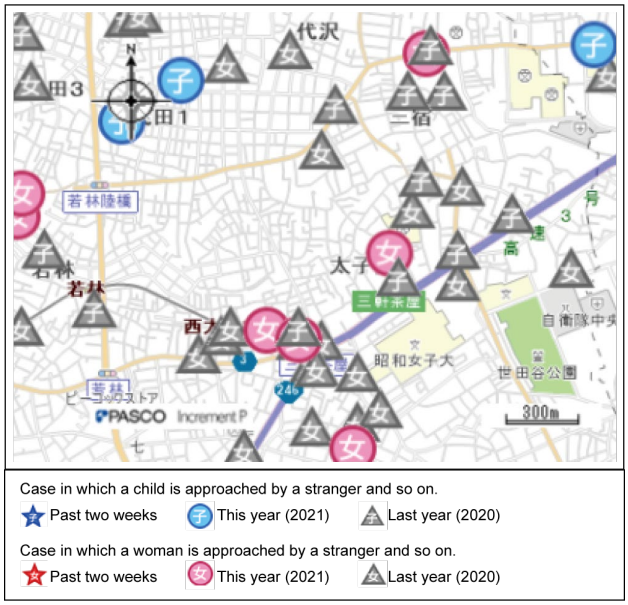

Figure 9. Spatial distribution of crime in the neighbourhood published in the Crime Information Map (Metropolitan Police Department, 2016).

6. DISCUSSION

We directly adopted pre-trained models without fine-tuning them for the test datasets, i.e., omnidirectional street-view images. For example, the VGG16 model used for feature embedding is trained on the ImageNet dataset that differs hugely from the test dataset. Due to the significant difference between the training and test datasets, the generated feature vectors may not adequately describe the omnidirectional street-view images. Such differences may affect the accuracy of similar street searching and regional classification (Section 4). In order to maximize the generalization ability of the pre-trained models, we will use transfer learning to customize them. Another option is to use a model trained on a similar dataset (street imagery) instead of VGG16.

The accuracy of semantic segmentation in this paper is not always high. In order to improve the accuracy of the semantic segmentation, it is desirable to use images with as high a resolution as possible, but this increases the processing time. In some cases, the CityScapes dataset used for training does not correspond to the unique landscape components of Japan, especially in residential areas, and the accuracy of segmentation is compromised. We are waiting for the development of a dataset using Japanese street spatial images for training. It is also necessary to consider how to efficiently collect street images in narrow streets where cars cannot pass, such as in dense wooden housing areas.

Further study is also needed on the appropriate number of clusters for regional classification and of dimensions when applying PQk-means clustering.

7. SUMMARY AND CONCLUSIONS

In this paper, we have presented a method for regional analysis using big data of omnidirectional street images taken by vehicles and clarified its usefulness and issues.

Using the method described in this paper, which converts omnidirectional images taken by a vehicle into perspective images and calculates the percentage of street components using semantic segmentation, it is possible to quickly and efficiently evaluate a wide range of street scenery and safety. Quantitative verification of the degree to which the task can be accelerated and made more efficient is a future issue.

ACKNOWLEDGEMENTS

This research is part of a research project related to the 2020 Tokyo Tech Challenging Research Award. Zenrin Co., Ltd. provided the omnidirectional image data essential for this study. Associate Professor Toshihiko Yamazaki and Lecturer Yusuke Matsui of the University of Tokyo provided technical advice in implementing PQk-means. Ms. Risa Yamanaka (Oki Lab., Tokyo Institute of Technology) assisted in some image processing tasks. In addition, many valuable comments from anonymous reviewers led to the improvement of this paper. We would like to take this opportunity to express our gratitude to everyone involved.

REFERENCES

Abhimanyu, D., Naik, N., Parikh, D., Raskar, R., Hidalgo, C. A., 2016: Deep learning the city: Quantifying urban perception at a global scale, arXiv:1608.01769.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Hartwig, A., 2018: Encoder-decoder with Atrous separable convolution for semantic image segmentation. arXiv:1802.02611.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016: The cityscapes dataset for semantic urban scene understanding. arXiv:1604.01685. https://www.cityscapes-dataset.com, (Accessed August 31, 2020)

Ki, D., Lee, S., 2020: Analyzing the effects of green view index of neighborhood streets on walking time using Google Street View and deep learning, *Landscape and Urban Planning*, 205. DOI: https://doi.org/10.1016/j.landurbplan.2020.103920.

Kizawa, S. Oki, T., 2021: Impression evaluation analysis of streets using big data of street images and crowdsourced questionnaire. *Proc. of the Conference on Geographic Information Systems*, 30.

Laupheimer, D., Tutzauer, P., Haala, N., Spicker, M., 2018: Neural networks for the classification of building use from street-view imagery, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2, 177-184. DOI: https://doi.org/10.5194/isprs-annals-IV-2-177-2018.

Law, S., Neira, M., 2019: An unsupervised approach to geographical knowledge discovery using street level and street network images, *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI 2019)*, 56–65. DOI: https://doi.org/10.1145/3356471.3365238.

Liu, L., Silva, E. A., Wu, C., Wang, H., 2017: A machine learning-based method for the large-scale evaluation of the qualities of the urban environment, *Computers, Environment and Urban Systems*, 65, 113-125, DOI: https://doi.org/10.1016/j.compenvurbsys.2017.06.003.

Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., Zhang, C., Ren, F., 2021: Measuring human perceptions of streetscapes to

better inform urban renewal: A perspective of scene semantic parsing, *Cities*, 110.
DOI: https://doi.org/10.1016/j.cities.2020.103086.

Matsui, Y., Ogaki, K., Yamasaki, T., Aizawa, K., 2017: PQk-means: Billion-scale clustering for product-quantized codes. *Proceedings of the 25th ACM international conference on Multimedia*, 1725–1733.
DOI: https://doi.org/10.1145/3123266.3123430.

Metropolitan Police Department, 2016: Crime information map. http://www2.wagmap.jp/jouhomap/Map?mid=2&mpx=139.648190513433&mpy=35.634099649473576&bsw=1519&bsh=664. (Accessed August 31, 2021)

MIT Media Lab: Place pulse dataset 2.0, https://www.media.mit.edu/projects/place-pulse-new/overview/, (Accessed September 7, 2021).

Min, W., Mei, S., Liu, L., Wang, Y., Jiang, S., 2020: Multi-task deep relative attribute learning for visual urban perception, *IEEE Transactions on Image Processing*, 29, 657-669.
DOI: https://doi.org/10.1109/TIP.2019.2932502.

Naik, N., Philipoom, J., Raskar, R., Hidalgo, C., 2014: Streetscore: Predicting the perceived safety of one million streetscapes, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 793-799.
DOI: https://doi.org/10.1109/CVPRW.2014.121.

Nishio, S., Ito, F., 2020: Relationship between sky ratio and spatial components and determinants of sky visibility by sky map shape, Poster presentation at the conference on Geographic Information Systems (P-25).

Oki, T. Kizawa, S., 2021: Evaluating visual impressions based on gaze analysis and deep learning: A case study of attractiveness evaluation of streets in densely built-up wooden residential area, *The XXIV ISPRS CONGRESS 2021 Digital Edition*, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 887–894.
DOI: https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-887-2021.

Simonyan, K., Zisserman, A., 2014: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.

Verma, D., Jana, A., Ramamritham, K., 2019: Machine-based understanding of manually collected visual and auditory datasets for urban perception studies, *Landscape and Urban Planning*, 190.
DOI: https://doi.org/10.1016/j.landurbplan.2019.103604.

Wang, R., Yuan, Y., Liu, Y., Zhang, J., Liu, P., Lu, Y. Yao, Y., 2019: Using street view data and machine learning to assess how perception of neighborhood safety influences urban residents' mental health, *Health & Place*, 59.
DOI: https://doi.org/10.1016/j.healthplace.2019.102186.

Zhou, H., He, S., Cai, Y., Wang, M., Su, S., 2019: Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning, *Sustainable Cities and Society*, 50.
DOI: https://doi.org/10.1016/j.scs.2019.101605.