論　文

Generating Tourist Attraction Introduction Videos Using Image Search and Short Video Analysis

Takeru Fujiwara[a], Yuanyuan Wang[a] and Wuyi Yue[b*]
[a] *Graduate School of Sciences and Technology for Innovation*
Yamaguchi University, Yamaguchi 755-8611 Japan
[b] *The Kyoto College of Graduate Studies for Informatics*
Kyoto 606-8225 Japan

(Received November 25, 2024)

Abstract

With the resurgence in travel demand post-COVID-19, enhancing access to tourism sites is critical. Web-based information can be complex, making destination selection difficult. This work uses image search to find short videos and related spots to generate tourist attraction introduction videos. It conveys attractions more realistically than text and images. Image search finds tourist spots similar to an input photo, reducing decision time. We collect and label short tourist videos, determine tags for user-input photos, and use cosine similarity to extract relevant spots.

Keywords: Travel demands, tourist attraction introduction, tourist photos, image search, short videos, tourist support.

1 Introduction

Travel demand is recovering as the COVID-19 pandemic subsides. The widespread use of smartphones has made it easier for many people to access tourist information on websites such as Jalan[1] and Rakuten Travel[2]. According to a survey by the JTB Japan Tourism Promotion Association and Values, Jalan.net had an estimated 30.3 million viewers from PCs and 31.9 million from smartphones in 2022, both exceeding 30 million [1]. These numbers are expected to increase as travel restrictions are eased. However, the sheer volume of information on the Web makes it difficult to choose, and potential destinations are often overlooked in the search.

Traditional tourism services have been dominated by text and image recommendations [2], [3]. However, with the increasing number of users of TikTok, YouTube, and Instagram, it is becoming easier to promote tourist attractions using video. There are many advantages to using tourism videos. For example, the information you can get from a single video is diverse, including details about food, events, and accommodations. The motion in videos makes it easier to convey the local

*Professor Emeritus, Konan University
[1] https://www.jalan.net/
[2] https://travel.rakuten.co.jp/

atmosphere and increase the desire to visit. The visual presentation also makes it easy to appeal to foreign tourists who do not understand Japanese.

Therefore, this paper aims to improve the impression and attractiveness of tourist spots by generating videos introducing tourist attractions with a high degree of similarity using user-input photos. By assigning tags and labels to input photos and short videos using image search, extracting tourist short videos with high similarity, and estimating tourist areas based on the location information of the tourist spots targeted by the extracted short videos and the input photos, we propose a method for generating a video introducing tourist attractions. The short videos are arranged according to the order of appearance of tourist attractions on the official regional tourism websites. Short videos, usually 60 seconds or less in length, are mainly uploaded to each video platform. Because they are short, they are less likely to be skipped, which reduces production costs and increases the number of views more easily than regular videos. Furthermore, with the spread of smartphones and SNS, it is easy to browse without time or place restrictions, and the videos are likely to be shared via SNS, which helps to improve the impression and attractiveness of tourist spots.

The paper is organized as follows. Section 2 introduces related studies such as tourism recommendation using Web information, tourist map generation, and their analysis. Section 3 describes the labeling of short videos and the tagging of input photos. Section 4 outlines the procedure for generating a tourist attraction introduction video based on the proposed method. Section 5 describes an evaluation experiment to determine the usefulness of the generated tourist attraction introduction videos. Finally, Section 6 provides a summary and discusses future issues.

2 Related Work

In recent years, there has been significant research and interface development related to the recommendation and mapping of tourist attractions using information on the Web, as well as their analysis [2]-[4]. Some of these studies create pictorial maps from SNS data such as Flickr and Twitter [5], [6], while others construct systems to recommend tourist routes taking into account preferences for tourist spots and travel routes [7]. Furthermore, some research recommends tourist spots based on text information in SNS posts [8], and others propose methods for recommending tourist routes that match travelers' preferences from geotagged tweets posted in tourist spots [9]. Tsuchida [10] also uses Word2Vec, a neural network-based language model, to extract tourist information from SNS data by converting word information into vectors. This method is proposed to find similar places in other regions by word addition from the generated corpus.

In this paper, the location information obtained from geotags is used to consider the relationships between tourist destinations. There is active research on the extraction and application of information posted on social networking sites such as Flickr and Twitter. Wang and Yue [5] propose a method to calculate the visibility and satisfaction of arbitrary tourist spots using various location information, posting dates, and other metadata attached to photos posted on Flickr. They generate two types of pictorial maps: a map of popular spots and a map of hidden spots that reflect these data. In addition, the construction of tourist recommendation systems based on collected tourist information using Word2Vec has been actively researched. One prototype system discovered latent interests and recommended sights based on those interests [11]. Another system used a pre-trained language model to estimate user preferences and recommend tourist spots during

casual conversations [12].

In this paper, we attempt to address the problem of potential travel destinations often overlooked in search queries by using informative image data instead of textual information. Takahashi [12] addressed the cold-start problem in online store product recommendations and video streaming by using diverse information about chat dialog preferences learned with Word2Vec. A system was also constructed for estimating interlocutor preferences and recommending tourist attractions. In addition, the Azure Video Indexer used in this paper has been used in other research. Wang [13] proposed a video viewing support system for users who skip viewing, a behavior that has increased with the spread of video subscription services. This system uses video scene segmentation.

3 Labeling Short Videos and Tagging Photos

In this section, we describe the methods for labeling short videos and tagging input images. In this paper, Azure AI Video Indexer is used to label short videos, while Azure AI Custom Vision is used to tag input images.

3.1 Short Video Labeling

This paper focuses primarily on videos from the official websites of tourist attractions and YouTube. For each of the 40 selected tourist attractions, we prepared 10- to 20-second videos. Labels are extracted from each video using Microsoft's Azure AI Video Indexer[3]. The labels are indexed in English to facilitate similarity calculations when generating tourist attraction videos. Figure 1 shows an example of labels extracted from a particular video using Azure AI Video Indexer.

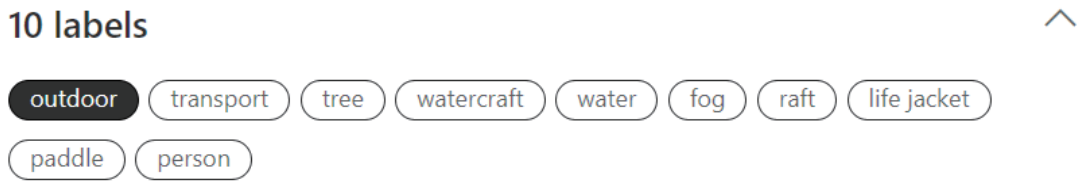

Figure 1: Example of Azure AI Video Indexer label extraction.

3.2 Input Photo Tag Determination

The main labels extracted from the short videos are selected, and the input photos are classified by 25 tags, such as "mountain," "sea," "tower," and "bridge." To tag the input photos, we used Microsoft Azure AI Custom Vision[4] and trained it with about 420 tourist photos. Figure 2 shows an example of tag probabilities and their thresholds determined for a particular photo using Azure AI Custom Vision.

[3] https://azure.microsoft.com/ja-jp/products/ai-video-indexer
[4] https://azure.microsoft.com/ja-jp/products/ai-services/ai-custom-vision

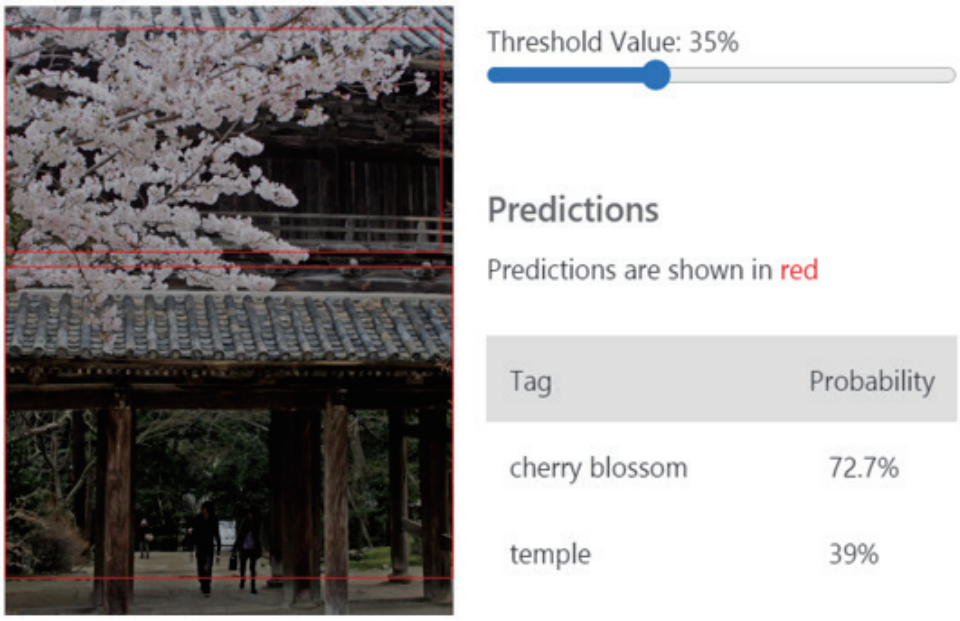

Figure 2: Example of Azure AI Custom Vision's tag probability and threshold determination.

4 Generation of Tourist Attraction Introduction Videos

The generated tourist attraction introduction video consists of three or four short videos of about 10 to 20 seconds that are played in succession. This section describes the procedure for generating a tourist attraction introduction video based on the proposed method.

4.1 Extracting Sights from Short Videos

First, the cosine similarity between the labels extracted from the short video and the tags of the input photos is computed using Word2Vec. Word2Vec is a natural language processing technology that converts words in a sentence or text into numerical vectors. By using Word2Vec, it is possible to more accurately represent word-to-word relationships, meanings, and nuances, enabling highly accurate natural language processing. In this paper, the tourist spots targeted by these short videos with high similarity are extracted as tourist spots with high relevance to the input photos.

4.2 Estimation of Tourist Areas Based on Location Information

Photos taken with GPS-enabled smartphones and cameras record location information; it is very likely that these photos show tourist spots that match the user-input photo. Nearby attractions at that location are geographically close, so they are unlikely to be significantly different from the attractions in the user-input photo. Therefore, if location information is present in the user-input photo this information can be used to include tourist spots that better match the user-input photo.

In addition to image information, Exif data records the date and time the image was taken, the camera model name, and settings such as aperture, ISO sensitivity, and shutter speed. This data also includes location information, called geotags, which record the location at the time the photo was taken. The location information from the user-input photo is compared to the location information of the landmarks in the short video. If the distance between the two points is within the default value, the short video is considered to be in the same region and is given priority. In this paper, the distance between the two points is set to 2 km. Furthermore, the area where the short video is located is estimated based on the location of the target attraction.

4.3 Generating Tourist Attraction Introduction Video

To refine the tourist regions estimated in the previous step, feature words are extracted from the overviews of official tourist websites for these regions. The tourist areas are identified by calculating the cosine similarity between these feature words and the tags in the input photos. The tourist attraction introduction video is generated by arranging short videos based on the order of appearance of tourist spots in the overview from the official tourism site. The name of each tourist attraction and the source of the short video should be displayed at the beginning of each tourist attraction scene in the introductory video.

5 Evaluation Experiment

In this section, we evaluate the usefulness of the generated tourist attraction introduction videos. In this paper, tourist attraction introduction videos were generated for 40 tourist spots in Yamaguchi Prefecture using labels assigned from short videos of these spots and tags determined from input photos. The target tourist spots are listed in Table 1.

Table 1: Target Tourist Attractions Used to Create the Tourist Attraction Introduction Video.

Order	Tourist Spots	Order	Tourist Spots
1	Kintai Bridge	21	Hagi Castle Town
2	Kouyoudani Park	22	Ohira Mountaintop Park
3	Yoshika Park	23	Tokiwa Park
4	Uno Chiyo House	24	Nagato Yumoto Onsen
5	Iwakuni Castle	25	Yuda Onsen
6	Nishiki River	26	Shunan Chemical Complex
7	Tsunoshima Island	27	Irori Sanzoku
8	Motonosumi Inari Shrine	28	Akiyoshidai Safari Land
9	Shimonoseki Aquarium Kaikyokan	29	Omi Island
10	Akiyoshidai	30	Chomonkyo
11	Syuhoudou	31	Chofu Tourism Association
12	Houhu Tenmangu Shrime	32	Kaneko Misuzu Memorial Museum
13	Beppu Benten Pond	33	Kikugahama Beach
14	Matsushita Village School	34	Suooshima
15	Ruriko-ji Five Story Pagoda	35	St. Francis Xavier Memorial Church
16	Senjoujiki	36	Susa Hornfels
17	Kaikyo Yume Tower	37	Kanmon Tunnel
18	Sesshu Garde	38	Kiwara Beach
19	Jakuchi Gorge	39	Funakata Farm
20	Higashi Gohata Terraced Fields	40	Ichinosaka River

5.1 Experiment Summary

A questionnaire survey was conducted to evaluate the usefulness of the generated tourist attraction introduction videos. In the survey, subjects watched three demonstration videos that were actually generated and answered the questionnaire. A total of 11 subjects participated in this experiment (9 males and 2 females). The demonstration first played a tourist attraction introduction video generated from the input photo shown in Figure 3. Next, a tourist attraction introduction video generated from the input photo shown in Figure 4 was played. Finally, a tourist attraction introduction video generated from the input photo shown in Figure 5 was played.

Figure 3: Input photo for the 1st tourist attraction introduction video.

The questionnaire is as follows:

Q1: The tourist attraction introduction video was appropriate for the input photos.

Q2: The video helped me understand the attractions of the tourist spots.

Q3: I wanted to visit the tourist spot after watching the video.

Q4: The order of the tourist spots presented in the video was appropriate.

Q5: The length of the tourist attraction introduction video was appropriate.

Q6: Please describe any positive aspects of this tourist attraction introduction video.

Figure 4: Input photo for the 2nd tourist attraction introduction video.

Figure 5: Input photo for the 3rd tourist attraction introduction video.

Q7: Please describe any improvements you would like to see in this tourist attraction introduction video.

The questionnaire was rated on a 5 point likert scale (1: strongly disagree, 2: somewhat agree, 3: neither agree nor disagree, 4: somewhat agree, and 5: strongly agree) for Q1 to Q5. Q1-Q3 focus on the content of the tourist attraction introduction videos, while Q4 and Q5 focus on their format. Q6 and Q7 are open-ended questions with a comment box for detailed feedback. Based on these survey results, we evaluated the usefulness of the tourist attraction introduction videos we generated.

5.2 Experimental Results

The results show that the average of Q1 was about 3.9, Q2 was about 4.2, and Q3 was about 3.6 (see Figure 6). These results suggest positive feedback regarding the appropriateness of the generated tourist attraction introduction video for the input photos, the ability of the video to convey the attractions, and the viewers' desire to visit the featured locations.

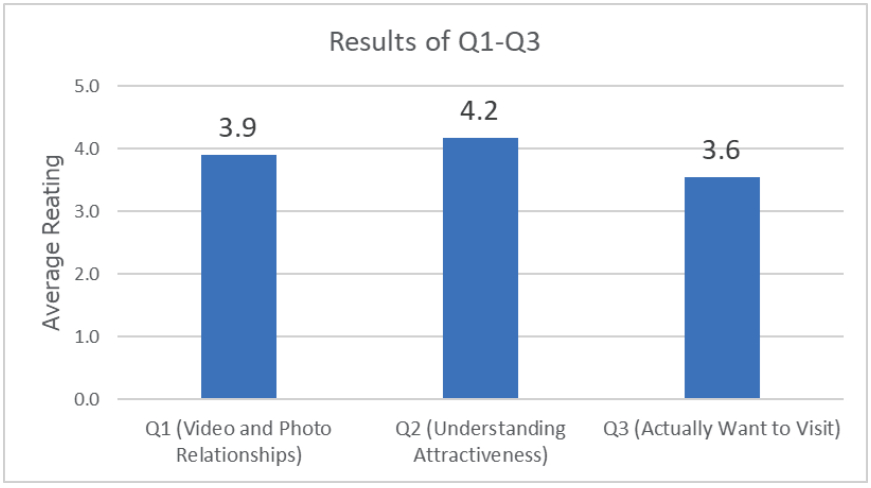

Figure 6: Results for the content of generated videos in Q1-Q3.

The results show that the average ratings for Q4 (the appropriateness of the order of the sights) and Q5 (the appropriateness of the video length) are very positive (see Figure 7). Specifically, Q4 has an average rating of approximately 4.0, and Q5 has an average rating of approximately 4.3, indicating that the order and length of the generated tourist attraction introduction videos were well received.

In Q6, respondents were asked to describe the positive aspects of the tourist attraction introduction video. The following are some of the responses:

- I think it would be very helpful when selecting tourist attractions to be able to suggest tourist attractions as videos from the input images.
- The atmosphere of the photo was reproduced in a video introducing the tourist attractions.
- The video was short and easy to watch.
- Eye-catching landscapes and landmarks were attractively displayed.

Finally, in Q7, respondents were asked to describe areas for improvement in the tourist attraction introduction video. The following are some of the responses:

- Small geographic area of tourist attractions.
- Viewers are more likely to be interested if detailed information about the sights can be presented in the form of maps and introductions.

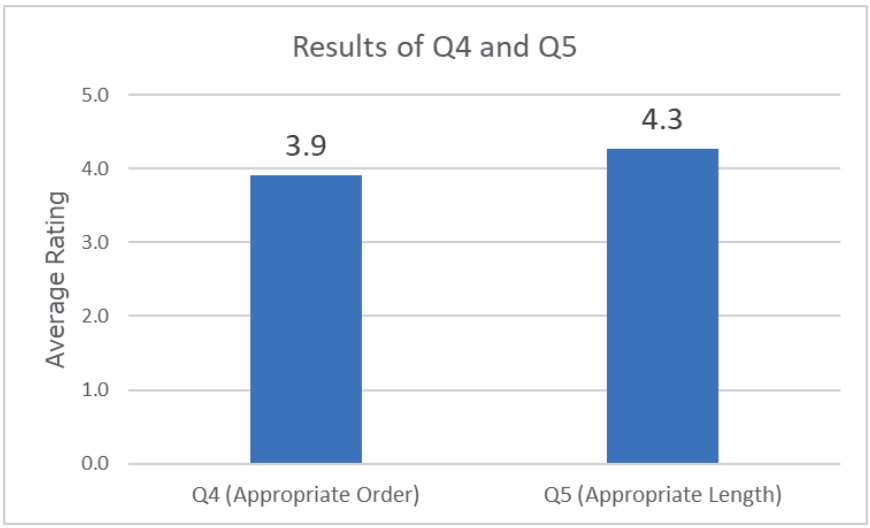

Figure 7: Results for the format of generated videos in Q4 and Q5.

- It would be better to mention the specific city.
- I thought it would be helpful to have more details about each attraction, such as what kind of place it is.
- Expanding the area beyond the prefecture could increase the diversity of the video.
- An introduction to cherry blossoms would be beneficial.

5.3 Discussion

The results of the evaluation of the content of the tourist attraction introduction video from Q1 to Q3 were all above 3.5, which is a good result. However, Q3 was slightly lower than Q1 and Q2. This may be due to the fact that detailed information about tourist attractions, as indicated by the opinions expressed in the descriptive responses to Q7, has not reached a sufficient level. The opinions expressed as requests included the desire for detailed information on the location of tourist spots, geographical relationships, and maps. Although Q1 had a good result with an average rating of 3.9, the tourist attraction introduction video used for the initial evaluation did not include a short video labeled "cherry blossoms," and the theme of the entire video was temples and shrines. Considering the length of the video, the number of short videos used in the tourist attraction introduction video was limited to three or four, which did not allow for the inclusion of short videos labeled with cherry blossoms. Since the opinion "An introduction to cherry blossoms would be beneficial" was received in Q7, further improvement is needed in the calculation of similarity. On the other hand, Q2 had a very good average score of 4.2, and the descriptions in Q6 suggest that the respondents were able to better understand the attractions of the tourist spots.

The evaluation results for the format of tourist attraction introduction videos in Q4 and Q5 were very good. However, although the average rating of Q4 was about 3.9, the geographical relationship

of tourist spots and the storyline in the tourist attraction introduction video were pointed out in the comments on the points for improvement in Q7, indicating that improvements are needed to attract users. Meanwhile, Q5, which evaluates the length of the video, had a very good result with an average rating of 4.3. This suggests that the use of short videos, which was the subject of this paper, produced excellent results.

In this paper, the target spots were biased toward natural tourist attractions because we selected tourist spots within Yamaguchi Prefecture. Therefore, when the labels extracted from the short videos were more outdoor-oriented and the input photos showed more urban structures, the similarity was reduced. As mentioned in the comments of Q7, the small geographical area itself is considered a disadvantage in terms of attracting users' interest. Therefore, it is necessary to expand the geographical scope of the target.

6 Conclusion

In this paper, we collected tourist short videos and photos, assigned labels to the short videos, and determined tags for user input photos by learning with the tourist photos. We calculated the cosine similarity between video labels and photo tags and extracted tourist spots from short videos with high similarity. Based on the location information of the extracted tourist spots and the location information of the input photos, we estimated the tourist area. Finally, the short videos were arranged according to the order of appearance of the tourist attractions on the official tourism websites in the estimated region to generate a tourist attraction introduction video.

As future work, we plan to evaluate the usefulness of the tourist attraction introduction videos through participant feedback and to explore methods for calculating the similarity between video scenes and photographs using image features.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Number JP21K17862 and by MEXT, Japan.

References

[1] Kankokeizai News: [Data] 1st place: Jalan.net, 2nd place: Rakuten Travel, 8th place: JTB Japan in "Ranking of estimated number of visitors to tourism-related sites in 2022," Tourism Association, 2023 (in Japanese). https://www.kankokeizai.com/%e3%80%90%e3%83%87%e3%83%bc%e3%82%bf%e3%80%911%e4%bd%8d%e3%81%98%e3%82%83%e3%82%89%e3%82%93net%e3%80%812%e4%bd%8d%e6%a5%bd%e5%a4%a9%e3%83%88%e3%83%a9%e3%83%99%e3%83%ab%e3%80%818%e4%bd%8d%ef%bd%8a/ (Accessed: December 23, 2023).

[2] H. Kawamura, K. Suzuki, M. Yamamoto and H. Matsubara, "Tourism Informatics," *Journal of Information Processing Society of Japan*, vol. 51, no. 6, pp. 642-648, 2010 (in Japanese).

[3] I. Saito, "Analysis of Tourism Informatics on Web," Special Issue: Tourism Informatics and Artificial Intelligence, *Journal of the Japanese Society for Artificial Intelligence*, vol. 26, no. 3, pp. 234-239, 2011 (in Japanese).

[4] M. Hirota, M. Endo, D. Kato and H. Ishikawa, "Discovering Hotspots Using Photographic Orientation and Angle of View from Social Media Site," *International Journal of Informatics Society*, vol. 10, no. 3, pp. 109-117, 2019.

[5] Y. Wang and W. Yue, "Proposal and Evaluation of a Pictorial Map Generation Method based on Degrees of Popularity and Satisfaction for Tourist Spots using SNS Photos," *Journal of the Japan Personal Computer Application Technology Society*, vol. 16, no. 1, pp. 1-10, 2021 (in Japanese).

[6] S. Gao and T. Ushiama, "Automatic Generation of Pictorial Maps using SNS," in *Proc. 9th Forum on Data Engineering and Information Management*, no. D7-4, pp. 1-8, 2017.

[7] Y. Takenobu and T. Okuno, "Construction of a Tourist Route Recommendation System considering Preferences for Tourist Spots and Travel Routes," in *Proc. 83rd National Conference*, vol. 1, pp. 335-336, 2021 (in Japanese).

[8] M. Obara, K. Morita, M. Fuketa and J. Aoe, "Extraction of Tourist Information from Contents of Tweets and Building an Analysis System," in *Proc. 29th Annual Conference of the Japanese Society for Artificial Intelligence*, vol. 29, pp. 1-3, 2015 (in Japanese).

[9] H. Nakajima, H. Niizuma and M. Ota, "Travel Route Recommendation using Tweets with Location Information," *IPSJ SIG Technical Report*, vol. 2013-DBS-158, no. 28, pp. 1-6, 2013 (in Japanese).

[10] T. Tsuchida, D. Kato, M. Endo, M. Hirota, T. Araki and H. Ishikawa, "Analyzing Relationship of Words Using Biased LexRank from Geotagged Tweets," in *Proc. 9th International Conference on Management of Digital EcoSystems*, pp. 42-49, 2017.

[11] R. Kaiji and Y. Higaki, "Recommendation System of Tourist Spots using Large Amounts of Tourism Information," *IEICE Technical Report*, vol. LOIS2015-14, pp. 29-34, 2015 (in Japanese).

[12] M. Takahashi, R. Asahara and M. Inaba, "Tourist Spot Recommendation from Open-Domain Dialogue using Pre-trained Language Models," *JSAI Technical Report, SIG-SLUD*, no. 96, pp. 27-32, 2022 (in Japanese).

[13] Y. Wang, I. Hashimoto, Y. Kawai and K. Sumiya, "Proposal of a Video Viewing Support System based on User Viewing Operations and Geographical Relationships," *DBSJ Japanese Journal*, vol. 20-J, no. 4, pp. 1-6, 2022 (in Japanese).