Exploring the Differences Between Tourists and Locals in Urban Settings Through Multi-labeled Geotagged Photos: The Case of Tokyo

Ahmed Derdouri [a,*], Toshihiro Osaragi [a]

[a] School of Environment and Society, Tokyo Institute of Technology – derdouri.a.aa@m.titech.ac.jp, osaragi.t.aa@m.titech.ac.jp

* Corresponding author

Abstract: Understanding the behaviors of both locals and tourists is essential for good city planning, especially in tourism-dependent cities. This study aimed to explore the disparities between the two groups on the basis of their geotagged photos taken in Tokyo during the last decade (2009–2019). The photos were collected from the photo-sharing platform Flickr. Locals and tourists were then identified. Next, a transfer-learning-based convolutional neural network model was developed to multi-label photos into eight general categories reflecting the major frequented activities/locations, including nature, amusement, and culture. Additional information was assigned to these records, including distances to various nearest points of interest. Qualitative and quantitative methods were used to investigate the differences between locals and tourists. Results showed that tourists have a strong preference for amusement while locals are attracted to nature. In contrast to tourists who are not followed by job obligations, locals' photos are mostly taken during the weekends. Given their familiarity with the area, locals tend to cover a wider spatial extent compared to tourists who are concentrated near the Yamanote railway loop line connecting most of the tourist attractions.

Keywords: multi-label classification, MobileNetV2, geotagged social data, Flickr, Tokyo

1. Introduction

Understanding how tourists and locals behave in urban settings is crucial to tourism and urban planning. While studies in urban planning focus generally on investigating the behaviors of residents, including travel choices and mobility patterns to name a few, little light has been shed on exploring those concerning tourists even though they form an important population group in tourism-dependent cities (Hasnat and Hasan, 2018). There is no doubt that tourists bring a positive impact into the communities they visit, including economic advantages (e.g., investments, job opportunities, etc.) and social benefits (e.g., cultural exchange). However, in poorly designed cities, such dynamic and heterogeneous groups might contribute to serious consequences, such as traffic jams, congestion while using public facilities, and degradation of the urban environment (Liu et al., 2018). These outcomes might cause overwhelming experiences for arrived visitors and residents alike. Therefore, to prevent such cases, it is vital to explore the behavior of both groups by analyzing their distribution across space and over time together with the nature of frequented places.

Prior to the big data era, studies on urban planning based their analyses essentially on data collected through traditional survey methods and aggregated mobile phone datasets. While the latter can be collected with relative ease compared to the former, which requires lengthy fieldwork, both are budget consuming. Crowdsourced data collected through social media networks offer a new alternative to these methods. Social data, known also as user-generated content, fall into two main types: texts (e.g., Twitter) and photos (e.g., Flickr). Most studies on tourism and urban planning used text-based social data because of their availability, large volume, and easy-to-handle character. However, in recent years, more and more researchers are employing images in their studies, especially in tourism-related ones. This change is attributed to the fact that photos serve as a spatiotemporal reflection of tourist activity given the wealth of information they contain themselves apart from their associated metadata (Zhang et al., 2019). Another reason is the advancement of deep-learning techniques that make it possible to extract vital information from images.

Image labeling is one of the applications of deep learning on images. Many techniques have been proposed and successful models developed to perform image labeling tasks. However, most of these concentrated on single-label image classification and paid little attention to the challenging multi-label classification (Cevikalp et al., 2020). In tourism-related studies, a single-label image classification can hardly describe the content of photos, especially those taken outdoors. Thus, for a better understanding of the content of photos and ultimately get insights about the nature of places where they were taken along with popular times and seasons, opting for a multi-label classification approach is essential. Such insights would provide planners with the mobility behaviors of tourists—which are often overlooked—and subsequently offer a solution to common issues such as seasonal overcrowding.

With these considerations in mind, this study comes to extend the existing literature by focusing on the main objective of comparing spatiotemporally the behavior preferences of tourists and locals via visual content analysis of their geotagged photos. Specifically, the two

research objectives are (1) to multi-label geotagged and time-stamped photos taken by both groups and, eventually, (2) compare the spatial and temporal distributions based on their popular activities and/or most frequented places.

The remainder of this article is structured as follows. Section 2 provides an overview of past studies regarding the exploration of tourists' behaviors by using pictorial content analysis. Section 3 presents the study area and the key steps of the applied methodology. Section 4 presents the obtained results of the analysis. Finally, Section 5 draws the main conclusions, discusses the limitations of the study, and suggests directions for future studies.

2. Past Work

Research on visual content analysis is divided into two categories (Zhang et al., 2019): (1) the traditional method, which involves manually defining the visual content of a photo and supplementing it with attached textual material. and (2) the emerging method, which involves using computer vision technologies to decode the content of a photo. In the tourism literature, several studies have been published that analyze tourist behaviors in urban settings using the traditional method, but those relying on computer vision technology are relatively scarce.

Recently, studies employing computer vision analysis to analyze human behavior within urban areas and beyond have been gradually increasing, and the topic has been getting considerably more attention. Zhang, Chen et al. (2019), for instance, used a deep learning approach to analyze the content of images taken by international tourists visiting Hong Kong. The authors employed the ResNet model to perform single-label image classification. A total of 78 scenes were recognized and then categorized into 12 categories. Another study carried out by Zhang et al. (2019) identified 103 scenes from photos taken by Flickr users visiting Beijing. The authors employed ResNet-101 to single-label the photos. The behaviors and perceptions of tourists from various continents and countries were then compared through comparison via statistical and spatial analyses. Kim et al. (2020) examined the images of tourists arriving in Seoul shared on Flickr. These images were sorted into 14 categories in the already-trained convolutional neural network (CNN) model Inception v3. Furthermore, the authors extracted 11 regions of attractions (RoAs), defined as the most popular areas, using the density-based spatial clustering application with noise algorithm. They found that, among multiple types of attractions, tourists prefer visiting palaces, historical monuments, traditional cuisine, and restaurants the most, although these preferences differ from one RoA to another.

Given the reviewed literature above, the limitations of the published studies are twofold: (1) they considered single-label image classification and neglected the multi-labeling approach, and (2) they did not compare the preferences of tourists and locals. This research is an attempt to fill these gaps.

3. Data and Methods

3.1 Study Area

We selected the special wards of the Tokyo Metropolitan Area as the target area (Figure 1) in this study. In total, there are 23 wards (hereinafter referred to as Tokyo) dotted with tourist attractions that are mainly connected by a complex yet well-organized and efficient railway network. These attractions are among the reasons Tokyo is the most popular destination in Japan among international and domestic visitors. Nearly half the tourists visiting Japan pass by Tokyo according to statistics gathered by the Japan National Tourism Organization. For reference, more than 14 million tourists visited Tokyo in 2018.
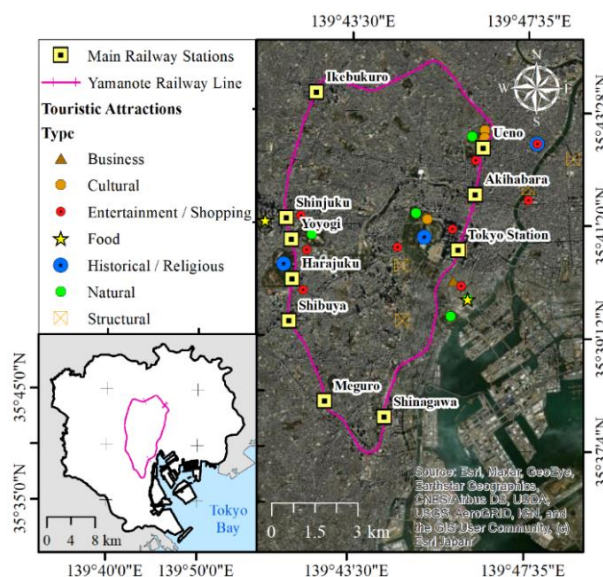

Figure 1. Study area and various tourist attractions by type

3.2 Methodology

The methodology of the analysis was structured in different key steps that are summarized in the following. First, geotagged photos gathered from the photo-sharing platform Flickr were collected and preprocessed. Then, the photo-takers were classified into locals and tourists. Next, the photos were multi-labeled and additional information was assigned to an image's records, such as the minimum distance to a variety of points of interest (POIs), time of the day, working day or holiday, and so on. In the final step, analyses were conducted regarding semantics and spatiotemporal distribution. In the following subsections, a detailed description of each step is provided.

3.2.1 Data Collection and Preprocessing

The primary data source of this study for geotagged photos is Flickr, which is a photo-sharing platform that allows registered users to upload and share their photos online. The platform provides an application programming interface (API) that permits developers to freely access and query its photo database. Therefore, with the use of API, a Python script was developed to

collect records with spatial coordinates located within the boundary box covering the extent of Tokyo and taken from July 1, 2008, to December 31, 2019.

Each collected record consists of a set of descriptive attributes that fall into five categories: (1) photo *ID* and a series of *URLs* linking to various sizes of the photos, (2) temporal attributes (e.g., *date* and *time* information when photos were taken and uploaded), (3) spatial attributes (i.e., *latitude* and *longitude*), (4) textual attributes (e.g., *title*, *tags*), and (5) photo-owner-related attributes (e.g., *user ID*, *country*, and *city*). While attributes (1) and (2) are generated automatically, those of (4) and (5) are optionally filled in by the owner except for the user ID, which is a unique identifier assigned to each user when signing up to the website. With regard to the geographical attributes, these are available when a user allows sharing of their location details. In this study, only records with geographic coordinates were considered.

As other researchers have reported in previous studies, social data contain erroneous records due to faulty hardware (M. Chen et al., 2019); for instance, a reduced GPS accuracy might produce an average distance error between 7 and 13 meters in some phones (Merry & Bettinger, 2019), or biased data introduced by "active users" who tend to take a large number of photos during a short time could yield findings that are subjugated by the users' behavior (Hollenstein and Purves, 2010). To avoid such cases, we preprocessed the collected records by following these steps:

- Users with less than two records were excluded.
- When photos were taken continuously at the same spot, only one record was kept.
- Similarly, one record was retained when a user took several photos within the same minute.
- In the case of two records of the same user, we filtered out those records that were taken within the same minute in two distinct geographical locations distant from each other by more than 13 meters.

This process resulted in 150,688 remaining photos taken by 3,063 unique users compared to 308,746 photos initially collected and taken by 10,110 users. As a result, all photos linked to these records were downloaded.

3.2.2 *Classification of Tourists and Locals*

The second step of this analysis consists of classifying Flickr users into tourists and locals. The machine learning approach proposed in (Derdouri and Osaragi, 2021) was applied to accomplish this task. The method considers numerous parameters that could explain the variability between the two groups, including those related to weather conditions, human mobility, temporal and spatial entropy, and population density. The method yielded an accuracy of 76% compared to the 71% scored using the typical method based on temporal entropy applied in Chen et al. (2019) and Sun et al. (2015).

3.2.3 *Multi-labeling Photos*

The second step of this analysis consists of multi-labeling the collected photos into eight general labels reflecting the nature of the activities carried out by the photo takers and/or the locations where they were taken.

To achieve this goal, we developed a CNN-based model to multi-label photos as amusement, business, culture, crowd, nature, infrastructure, residence, and other (i.e., objects). Table 1 lists all the considered labels and the possible venues that they are referring to. First, we prepared a training dataset consisting of 1,416 photos. These photos were manually multi-labeled using a C# application developed by the authors to facilitate the process. We then developed a model using the transfer-learning approach based on the MobileNetV2 architecture, which was selected because of its capacity to (i) avoid issues related to overfitting due to the relatively small training dataset as well as (ii) minimize the execution time and memory consumption while minimizing prediction errors.

	Labels	Detailed Subcategories
1	Amusement	Food, shopping, and entertainment
2	Business	Business districts
3	Culture	History-, religion-, and art-related objects and places
4	Crowd	People
5	Nature	Greenspaces, waterbodies, and mountains
6	Infrastructure	Transportation, landmarks, and cityscape
7	Residence	Hotels and homes
8	Other	Objects (books, dishes, etc.)

Table 1. General labels and examples of venues they refer to

We kept the trained parameters unchanged and only fine-tuned the entire network during testing since the MobileNetV2 model had already been trained using the ImageNet dataset. We used the grid search technique to fine-tune and compute the best values for the activation function, optimizer, and the number of epochs hyperparameters. The model was trained and validated on threefold splits of a 70:30 ratio. To ensure well-balanced training and validation samples across the eight groups, we applied the stratification method proposed by Sechidis et al. (2011) and improved by Szymański and Kajdanowicz (2017). In terms of hardware, the model was trained on a 10-Intel-core i9-10900X with 120 GB RAM and a 3.70 GHz CPU.

The grid search technique produced the best-performing model with a mean accuracy of 83.46% on the three splits (standard deviation = 0.00182), a micro-averaged area under the curve of 0.87, and a micro-averaged area under the precision-recall curve of 0.78.

3.2.4 *Assigning Additional Information to Flickr Records*

In this step, multi-labeled Flickr records were merged with other parameters, including (1) distances to the nearest POIs, such as shopping areas, parks, accommodations, which were calculated using ArcMap

	Seasons		Day/Night		Work/Holiday		Steady/Perturbed	
Amusement	164.129	(*****)	423.077	(*****)	89.149	(*****)	8.000	(****)
Business	7.253	(*)	0.239	( )	7.735	(***)	0.015	( )
Crowd	164.756	(*****)	9.937	(****)	98.372	(*****)	9.146	(****)
Culture	40.313	(*****)	0.725	( )	67.516	(*****)	1.419	( )
Infrastructure	5.079	( )	0.351	( )	5.267	(**)	0.300	( )
Nature	67.647	(*****)	7.972	(****)	330.167	(*****)	45.564	(*****)
Residence	15.185	(****)	0.235	( )	4.783	(**)	0.328	( )
Other	351.126	(*****)	178.244	(*****)	171.311	(**)	5.266	(****)

Table 2. Results of the Chi-squared test.

Note: The asterisks in parentheses depict the statistical significance level of the result, in which five levels are considered: (*****) = 0.1%; (****) = 0.5%; (***) = 1%; (**) = 5%; and (*) = 10%. ( ) is used when a p-value is not significant at all considered levels.

based on features extracted from OpenStreetMap data of the study area; (2) information about whether the day is a working day or a holiday (e.g., weekend, Japanese national day); (3) time of the day (i.e., daytime or nighttime); and (4) steady or perturbed states based on the historical records of disasters that impacted the area, such as typhoons and earthquakes.

3.2.5 *Statistical, Temporal, Semantics, and Spatial Analyses*

The last step of the analysis is to investigate the differences between tourists and locals through statistical and semantic analyses, in addition to temporal and spatial visualizations of the distribution in terms of their labeled photos. The statistical analysis was performed with the Chi-squared test and ordinary least squares (OLS) while spatial visualization is done by applying density analysis using a hexagonal grid. As for the semantic analysis, we counted the number of the most frequent label combinations in photos taken by locals and tourists.

4. Results

4.1 Statistical Analysis of the Differences Between Locals and Tourists

To investigate the differences between locals and tourists in the perception of the eight labels, we ran the Chi-square test to show how significant these differences are between the two groups during different seasons, day (daytime/nighttime), working days/holidays, and steady/perturbed states. The results are listed in Table 2. Significant seasonal differences between tourists and locals are visible in the perception of amusement, crowd, culture, nature, residence, and objects. On the other hand, business- and infrastructure-related photos do not have much of a difference between the two groups during the four seasons. With regard to the differences concerning daytime/nighttime and steady/perturbed states, fewer differences are observed except for the perception of amusement, crowd, nature, and objects. The most obvious differences can be spotted during working days and holidays because the perception of all labels can refer to different groups.

4.2 Semantic Analysis

The differences between locals and tourists in terms of the top combinations of labels found in single photos taken during the whole study period and during each season (Figure 2) were examined further. Locals show a strong preference for nature during the spring season (26%), for crowded venues during autumn (15%), and for amusement all year round (9%–12%). Object-focused (labeled as other) photos were taken mostly during winter (16%) and autumn (15%), which might suggest that they were taken indoors given the bad/cold weather (e.g., a dish inside a restaurant, brochure inside a train station, an object inside a museum, etc.). The preferences of tourists, on the other hand, are slightly different. They tend to take object-focused photos (16%–22%), which might suggest that they are more curious about the things they are witnessing for the first time. Other than these photos, the combination of labels suggests that tourists take a good amount of amusement-related pictures. Moreover, in contrast to locals, it appears that tourists are less concerned about nature.
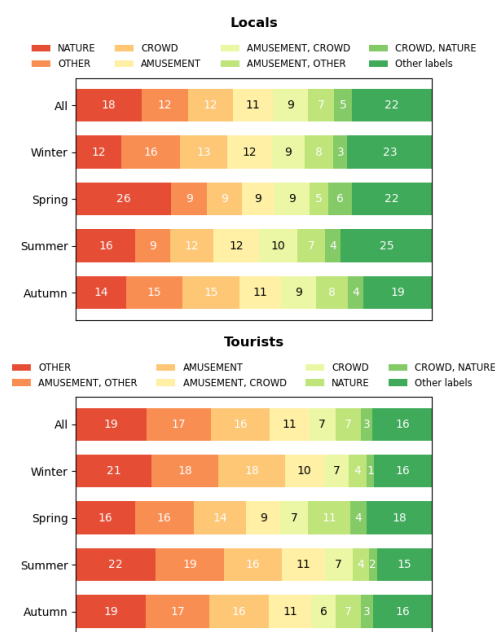

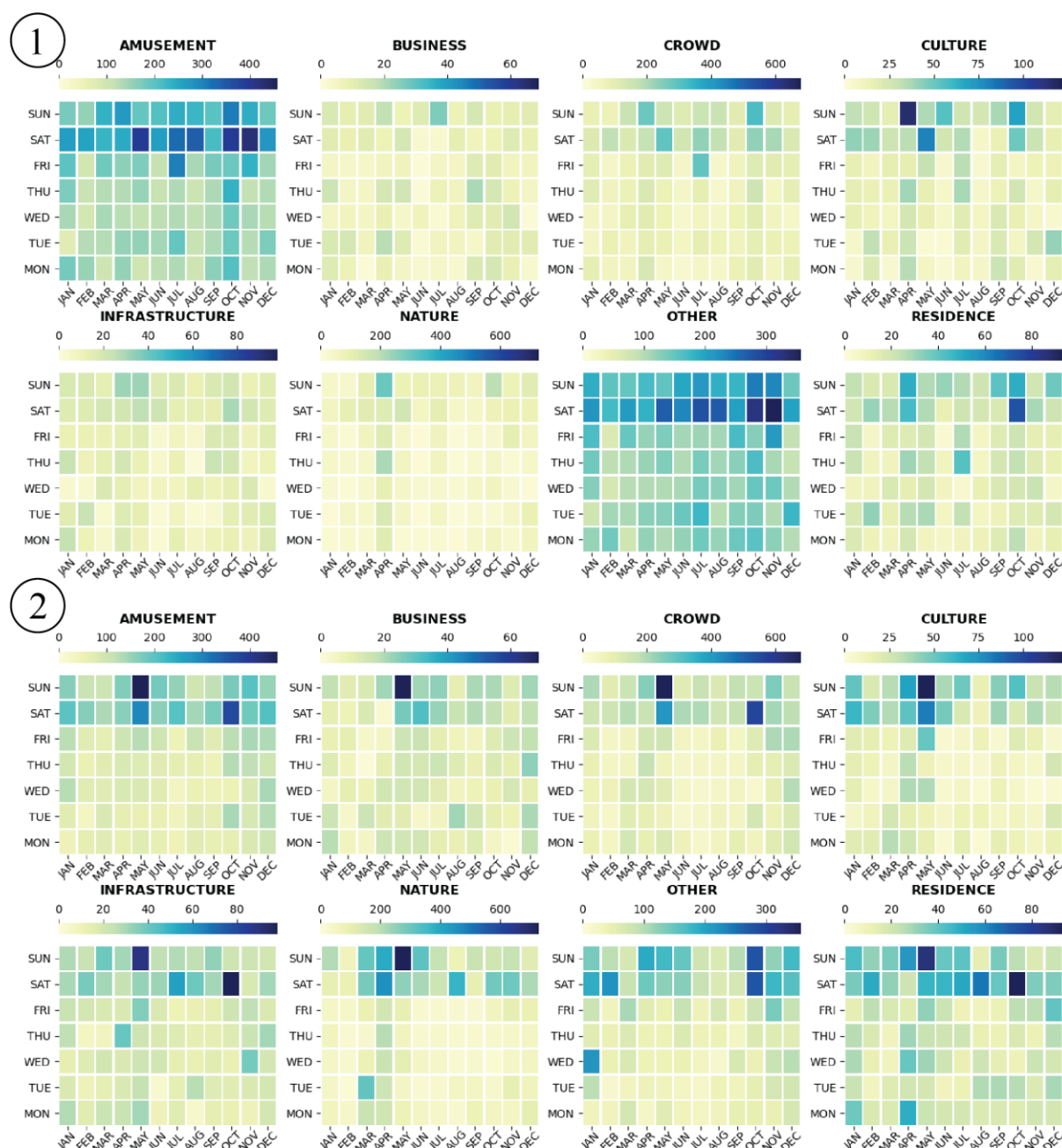
Figure 2. Seasonal statistics of label combinations

Figure 3. Monthly variations of photos taken by tourists (1) and locals (2)

4.3 Temporal Variations of Locals and Tourists

4.3.1 *Monthly Variations*

Figure 3 shows the heatmaps of the number of labels detected during the days of the week in each month in photos taken by tourists (top) and locals (bottom). The activities carried out by locals are concentrated during the days of the weekend (except for business). Tourists, on the other hand, take different-label photos almost every day of each week during all months. This is because tourists do not have job-related obligations. The months of spring and winter seasons are the months when locals take the most photos when going to cultural events and amusement, nature-related, and crowded venues. With respect to tourists, the labeled photos do not follow a fixed pattern.

4.3.2 *Hourly Variations*

Figure 4 illustrates the heatmaps of the number of labels detected during the hours of the week in photos taken by tourists (top) and locals (bottom). For locals, for all labels, it can be observed that activities are concentrated during the weekend from 9 AM to 9 PM. Business-labeled photos are taken almost every day of the week. Except for the amusement-, crowd-, and object-related themes, the observed patterns in the numbers of photos taken by tourists do not follow a fixed pattern in contrast to those observed in locals. The temporal span of activities carried out by tourists is wider, starting from 8 AM to midnight all day of the week.
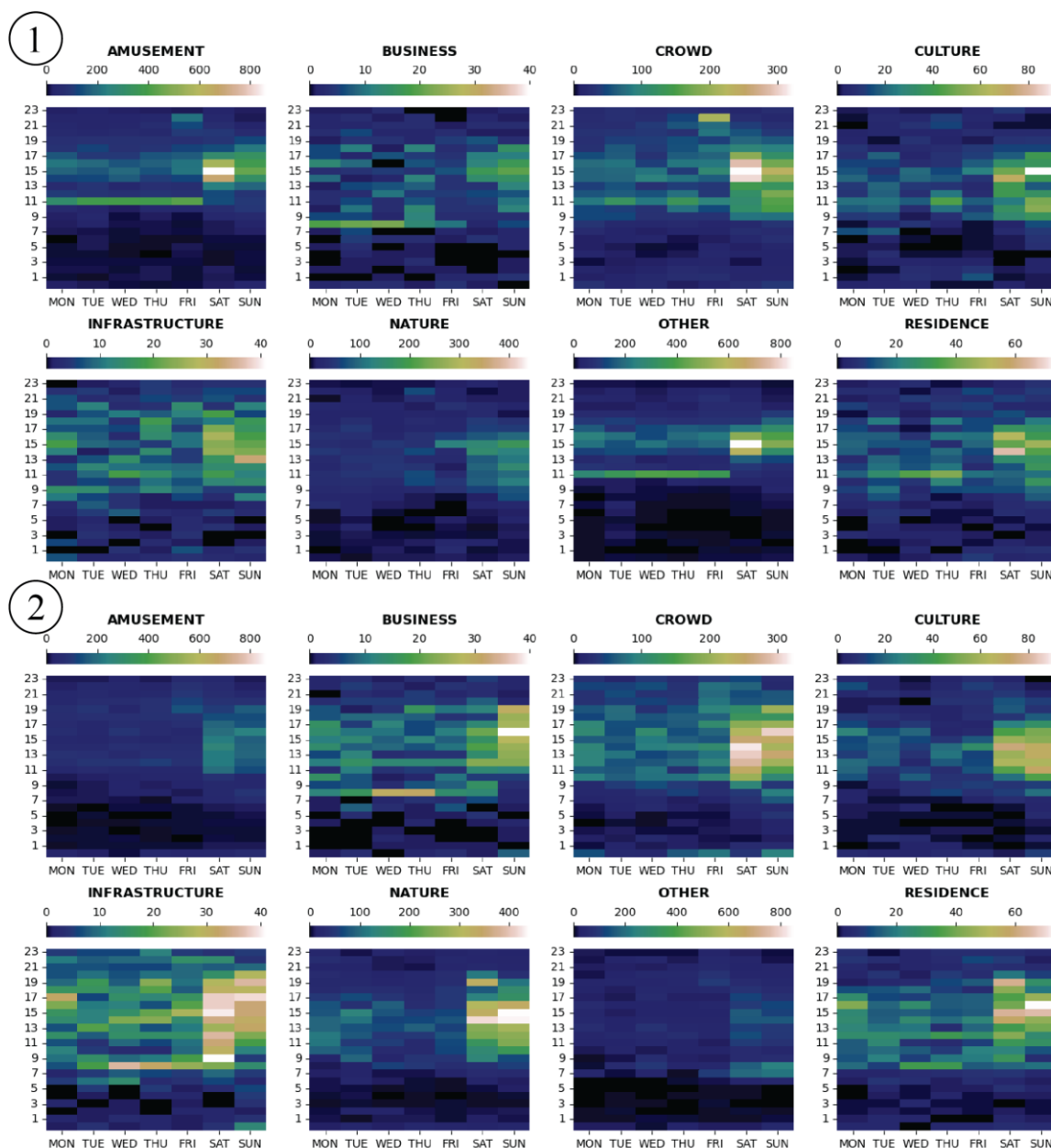
Figure 4. Daily variations of taken photos of tourists (1) and locals (2)

4.4 Spatial Distribution of Tourists and Locals

The collected 11 years of geotagged photos of tourists and locals were mapped using photos associated with spatial coordinates. These photos were assigned to a 1 km hexagon grid covering the study area. Figure 5 illustrates the spatial distribution of photos taken by tourists (top) and locals (bottom) based on the main themes of the photos. Additionally, we ran an OLS to analyze the relationships between photos taken by both groups and the distances to the nearest POIs.

Both groups took photos mostly in the Yamanote Line's vicinity: Ueno, Akihabara, Tokyo station, Asakusa, Ikebukuro, Shibuya, and Shinjuku. These stations are gateways to the most popular tourist attractions of Tokyo. Areas in the southwest were not visited by any group of Flickr users owing to a shortage of attractions.

Photos taken by tourists were mainly concentrated in areas 2–6 km from the centroid of the study area. These clusters were not evenly distributed across this zone, as they were mainly located near the Yamanote Line stations close to tourist attractions. In terms of the content of the taken photos, it appears that tourists prefer amusement mainly in areas located near the Tokyo station and Akihabara. The OLS results indicate that these amusement areas are near shopping ($\beta = -8.6 \times 10^{-4}$, p=0.000) and commercial ($\beta = -1.1 \times 10^{-4}$, p=0.000) areas; indoor accommodations ($\beta = -9.9 \times 10^{-5}$, p=0.000) such as hotels, hostels, and guesthouses; and waterways ($\beta = -1.2 \times 10^{-4}$, p=0.000). The spatial distribution of crowd-related photos almost follows that of amusement-related ones. However, the OLS results show that these crowded venues are near outdoor venues near trees ($\beta = -2.0 \times 10^{-4}$, p=0.000) and waterways ($\beta = -6.0 \times 10^{-5}$, p=0.000) and slightly far from shopping areas ($\beta = 2.0 \times$
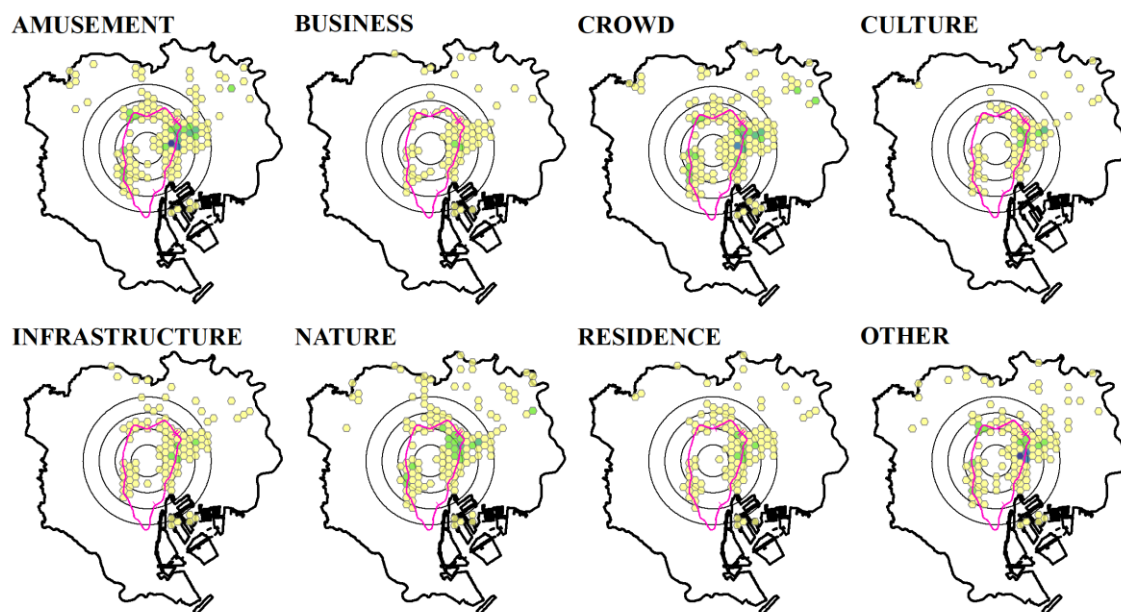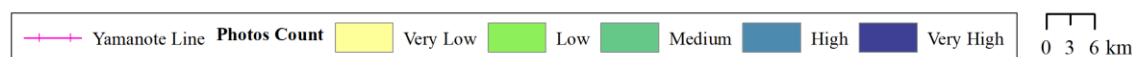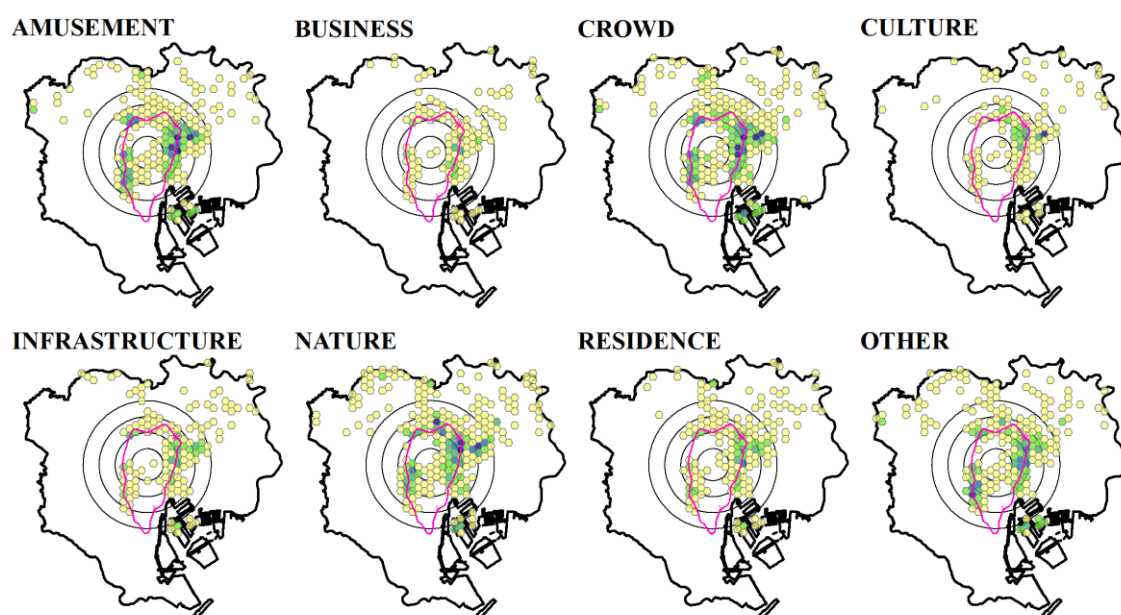
TOURISTS

LOCALS

Figure 5. Spatial distribution of tourists (top) and locals (bottom)

10^{-4}, p=0.000). For the other-label-related photos, OLS results show that they are concentrated near shopping areas (β = -7.5 × 10^{-4}, p=0.000), parks (β = -2.1 × 10^{-4}, p=0.000), public facilities (β = -1.9 × 10^{-4}, p=0.000), and recreation grounds (β = -1.7 × 10^{-4}, p=0.000).

Locals visited a wider area across Tokyo, with "high" and "very high" concentrations in a buffer zone 8 km from the centroid of the study area. These highly dense clusters diminished beyond the 8 km buffer zone. They are on the north side of the study area and far away from the Yamanote Line. Labels with high concentrations are amusement, crowd, culture, object-focused, and

nature. The OLS results suggest that the amusement areas are in the proximity of shopping areas (β = -9.6 × 10^{-4}, p=0.000), indoor accommodations (β = -1.7 × 10^{-4}, p=0.000), and recreation grounds (β = -1.6 × 10^{-4}, p=0.000). Crowd-related photos are spatially distributed the same way as those representing amusement, and the OLS results show that these crowded venues are near shopping areas (β = -4.7 × 10^{-4}, p=0.000) and outdoor venues, such as green spaces including trees (β = -2.2 × 10^{-4}, p=0.000) and parks (β = -1.4 × 10^{-4}, p=0.000).

Note that the obtained values of R^2 and adjusted R^2 of locals are better than those of tourists for all labels.

While the observations of locals are thrice as many as those of tourists, this result could suggest that locals form a homogenous group with quasi-similar characteristics versus tourists who are from different backgrounds and traits. Besides, foreign travelers are not familiar with the area and tend to sometimes travel randomly in a city. Except for the main tourist spots, they are likely to move independently from the locations of POIs (small spatial correlation) in contrast to locals who consider familiar POIs when moving.

5. Discussion and Conclusions

The findings of this study suggest that various disparities exist between tourists and locals on several levels. First, major seasonal distinctions between the two groups are visible in the perception of amusement, crowd, culture, nature, residence, and objects. Conversely, business- and infrastructure-related photos do not make much of a difference. Second, temporal variations show activities carried out by locals are mostly concentrated during the days of the weekend. By contrast, tourists take different-label photos almost daily during all months because they do not have job-related obligations. Third, locals have a strong preference for nature especially during the spring season, for crowded venues during autumn, and for amusement all year. Tourists, however, have different tastes. They tend to take object-focused photos often, which might suggest that they are more curious about the things they are witnessing for the first time. Aside from these shots, the combination of labels shows that visitors take numerous entertainment-related pictures. Furthermore, visitors tend to be less attracted to nature. Finally, in terms of spatial distribution, tourist images are mostly clustered in areas 2–6 km from the centroid of the study area. These clusters are not widely spread across the study area, as they are mostly concentrated near Yamanote Line stations. Locals visit a wider area.

While the results of this study contribute to the understanding of how locals and tourists behave in Tokyo based on their geotagged photos, the research has many shortcomings that could be addressed in future studies. The first is the limited number of categories considered to reflect the nature of activities or locations both group populations tend to carry out or go to. Specifically, the "other" label is general and might refer to different things. Second, Flickr does not account for the actions of the general public, given that not all Internet users take pictures, much less post them on social media. Thus, integrating data from other sources may be beneficial.

This research may be improved in several ways. To begin with, geotagged records from other media networks may be used, either to supplement Flickr data or compare and analyze the differences of outcomes based on various data sources. Considering more specific categories derived from the ones suggested in this study is another promising path for further detailed analysis and insights to detect more differences between locals and tourists. Likewise, other architectures could be compared for better multi-label image classification accuracy.

Additionally, another CNN-based model could be developed to determine the nature of the environment—either indoors or outdoors—where the photos are taken.

6. References

Cevikalp H, Benligiray B and Gerek ON (2020) Semi-supervised robust deep neural networks for multi-label image classification. *Pattern Recognition* 100: 107164. DOI: 10.1016/j.patcog.2019.107164.

Chen M, Arribas-Bel D and Singleton A (2019) Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems* 21(1): 89–109. DOI: 10.1007/s10109-018-0284-3.

Derdouri A and Osaragi T (2021) A machine learning-based approach for classifying tourists and locals using geotagged photos: the case of Tokyo. *Information Technology & Tourism*, 1-35. DOI: 10.1007/s40558-021-00208-3.

Hasnat MM and Hasan S (2018) Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies* 96: 38–54. DOI: 10.1016/j.trc.2018.09.006.

Hollenstein L and Purves R (2010) Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 2010(1): 21–48. DOI: 10.5311/JOSIS.2010.1.13.

Kim D, Kang Y, Park Y, et al. (2020) Understanding tourists' urban images with geotagged photos using convolutional neural networks. *Spatial Information Research* 28(2): 241–255. DOI: 10.1007/s41324-019-00285-x.

Liu Q, Wang Z and Ye X (2018) Comparing mobility patterns between residents and visitors using geo-tagged social media data. *Transactions in GIS* 22(6): 1372–1389. DOI: 10.1111/tgis.12478.

Merry K and Bettinger P (2019) Smartphone GPS accuracy study in an urban environment. *PLOS ONE* 14(7). Public Library of Science: e0219890. DOI: 10.1371/journal.pone.0219890.

Sechidis K, Tsoumakas G and Vlahavas I (2011) On the Stratification of Multi-label Data. In: *Machine Learning and Knowledge Discovery in Databases* (eds D Gunopulos, T Hofmann, D Malerba, et al.), Berlin, Heidelberg, 2011, pp. 145–158. Lecture Notes in Computer Science. Springer. DOI: 10.1007/978-3-642-23808-6_10.

Szymański P and Kajdanowicz T (2017) A Network Perspective on Stratification of Multi-Label Data. *arXiv:1704.08756 [cs, stat]*. Available at: http://arxiv.org/abs/1704.08756 (accessed 15 June 2020).

Zhang K, Chen Y and Li C (2019) Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing. *Tourism Management* 75: 595–608. DOI: 10.1016/j.tourman.2019.07.002.