Fig. 1. Example of a detected similar geo-region pair. Both are geo-regions where people focus on Chinese-style streetscapes and Chinese foods; Chinatowns in Yokohama, Japan, and Los Angeles, USA.

Here, we focused on the recent trend of the di usion of Social Network Services (SNS). When people travel, many of them take photos depending on their interests and post them to SNS. These photos contain rich information on the objects and activities that take place in the regions that they were taken. In order to extract the targets-of-interest from them, we decided to use Visual Concepts to represent the features of a geo-region. Focusing on the interests of the crowd instead of objective information such as that observed from a map or a satellite image, we expect to obtain more intuitive similarities.

The main contribution of this paper is the proposal of a method for the detection of similar geo-regions based on Visual Concepts in social photos. By applying the method to the YFCC100M dataset [8], we could detect similar geo-regions anywhere around the World (Fig. 1).

The remainder of this paper is organized as follows: Section 2 introduces related work. Section 3 introduces the proposed method for detecting similar geo-regions based on Visual Concepts detected from social photos. Section 4 reports and analyzes the results of the detected similar geo-regions, and nally Section 5 concludes the paper.

2 Related Work

As examples of travel recommendation systems, there are route recommendation systems, and landmark detection and recommendation systems.

As a route recommendation system, Kurashima et al. [3] proposed a method that makes use of the photographers' action history from geo-tagged photos on Flickr. Although they used geo-tagged photos to construct the photographer's behavior model, they focused especially on location information and the movement of each photographer, but not on detailed visual information in the photos. As a landmark detection and recommendation system, Shi et al. [6] proposed a personalized system. They used category information extracted from Wikipedia and recommended landmarks to users. They focused on landmarks and their prede ned categories, whereas our method focuses on geo-regions and do not prede ne categories.

Min et al. [4] proposed a probabilistic topic model called Multimodal Spatio-Temporal Theme Modeling (mmSTTM) and analyzed themes discovered from

twenty popular landmarks around the World. They focused only on popular landmarks and also did not consider similarities between themes, whereas our method focuses on similarities of pairs of arbitrary geo-regions.

As for work which focuses on the similarity between geo-regions, there are two major approaches: Those that refer to texts and those that refer to images.

As work that refers to texts, Shimada et al. [7] proposed a sightseeing spot recommendation system based on text information on the Web. It detects similar geo-regions (locations) by calculating similarities between an user-selected favorite geo-region (location) and each sightseeing spot in their database. In this system, however, similarities cannot be calculated across di erent linguistic areas because they cannot directly compare keywords extracted from Web sites written in di erent languages. Considering globalization in the travel sector, for this reason, we decided not to rely on text, but rather make good use of non-linguistic information that can be extracted from image.

As work that refers to images, Pongpaichet et al. [5] implemented a visual analytics system of social photos and made event models to detect similar events. Each model is described by Bag-of-Visual Concepts that co-occur in a pre-de ned event. They focusd on the similarity between a geo-region and a pre-de ned model in order to detect photos on a certain event, whereas our method focuses on the similarity between geo-regions based on arbitrary events.

Ide et al. [2] proposed a method that automatically detects people's common attention in a geo-region from a large number of geo-tagged photos, and its visualization on a map. By classifying each local area into ve scene categories based on global image features extracted from social photos, they classi ed a geo-region into one of the ve categories. In their method, however, since the category of a geo-region was decided according to the majority of photos that belonged to the scene category, other photos that did not belong to the category were ignored and visual information in them were not used to describe the geo-region. Instead, our method makes use of the information in all the photos available and describes a geo-region with combination (distribution) of objects or activities detected from them.

3 Detection of Similar Geo-Regions based on Visual Concepts in Social Photos

We propose a method for the detection of similar geo-regions based on Visual Concepts in social photos. First, geo-regions are xed by spatially clustering a large number of geo-tagged photos. Next, for each geo-region, their features are described based on Visual Concepts detected from social photos taken there. Finally, the similarity between a pair of geo-regions is calculated and those with high similarity are output as similar geo-regions. Details of the three steps are described below.

3.1 Definition and Decision of Geo-Region

In general, a geo-region is de ned as an area demarcated according to certain criteria: administrative division, climatic zone, and so on. However, since we aim to support travel, such objectively-de ned areas do not necessarily satisfy the users' purposes. Instead, we subjectively de ne geo-regions based on the interests of the crowd.

For the decision of geo-regions, we focus on geo-tagged photos. With the di usion of mobile devices equipped with a GPS function, more and more geo-tagged photos are posted to SNS. Since these photos directly re ect people's interests, we apply mean-shift clustering to locations where the photos were taken to decide geo-regions. Since some clusters (geo-regions) contain few photos, only those that contain su cient numbers of photos ($_p$) to describe the geo-regions, are used.

3.2 Definition of Similarity

When a pair of geo-regions are similar, they should share something in common. Since our purpose is travel support, it should be based on subjective information in view of travelers. Therefore, as a criterion for measuring the similarity, we focus on the existence of similar objects or activities.

3.3 Description of Geo-Region's Feature

Since Visual Concepts detected from social photos indicate objects and activities that appear in photos taken by people at geo-tagged locations, they are used as a feature to describe the interests of the crowd in a geo-region. Visual Concepts are extracted and used as features to describe geo-regions as follows.

Extraction of Visual Concepts in Social Photos Recently, more and more people use SNS. They often take photos of objects and activities which aroused their interests and post to SNS. Di erent people have di erent interests: One may be interested in the landscape, but others may be interested in a landmark or food. We considered that Visual Concepts could be used to describe such diverse interests. Recent advance in deep learning methods has allowed the robust and accurate detection of Visual Concepts. Therefore, in the proposed method, we describe image contents of social photos by combinations (distributions) of Visual Concepts detected in photos.

Let the number of categories of Visual Concepts that can be detected by a detector be N, for an input image, the probability of each Visual Concept is calculated and described as an N-dimensional vector $\mathbf{v}_l$.

Description of Geo-Region's Features A geo-region's feature is described as follows. Let P_l be a set of photos taken in a geo-region c_l. For each photo

$p_{l;i} \in P_l$, the feature vector of $p_{l;i}$ is described as $v_{l;i}$. The feature vector of a geo-region c_l is described as $V_l = \sum_i v_{l;i}$.

We should notice that some visual concepts are general and detected from many photos. Thus, we apply a TF-IDF-like approach to each component of v_l. TF and IDF are calculated as follows. Let the n-th component of V_l and $v_{l;i}$ be $V_l(n)$ and $v_{l;i}(n)$, respectively ($n = 1; 2; \ldots; N$), then,

$$TF_{l;i}(n) = \frac{v_{l;i}(n)}{\sum_{m=1}^{N} v_{l;i}(m)}. \tag{1}$$

Let P be all the photos in the dataset, then,

$$IDF(n) = \log \frac{|P|}{|\{p_j \mid v_j(n) > \theta_{vc}; p_j \in P\}|}. \tag{2}$$

Here, θ_{vc} is a threshold for judging whether we should consider a Visual Concept exists in a photo or not. Finally, by calculating the following equation, we obtain the geo-region's feature vector $V_l^0 = (V_l^0(1); V_l^0(2); \ldots; V_l^0(N))$, where

$$V_l^0(n) = \sum_{i=1}^{|P_l|} TF_{l;i}(n) \cdot IDF(n). \tag{3}$$

3.4 Similarity Calculation

Once the geo-regions are described by feature vectors, we calculate the similarity between a pair of them. For a given pair of geo-regions $c_i; c_j$, their similarity is calculated by Zero-mean Normalized Cross Correlation (ZNCC) as,

$$S_{i;j} = \frac{\sum_{n=1}^{N} (V_i^0(n) - \overline{V_i^0})(V_j^0(n) - \overline{V_j^0})}{\sqrt{\sum_{n=1}^{N} (V_i^0(n) - \overline{V_i^0})^2 \sum_{n=1}^{N} (V_j^0(n) - \overline{V_j^0})^2}}. \tag{4}$$

$$\overline{V_l^0} = \frac{1}{N} \sum_{n=1}^{N} V_l^0(n). \tag{5}$$

Finally, pairs with similarities higher than a threshold θ_s are detected as similar geo-regions.

4 Experiment

In this section, we report and analyze experimental results by showing examples of the pairs of similar geo-regions detected by actual social photos.

4.1 Dataset

We used YFCC100M (Yahoo! Flickr Creative Commons 100 Million) dataset [8], which contains 100 million photos and videos from Flickr taken all over the World. Each photo in the dataset is accompanied with metadata (tags, timespan, location and so on). Since the proposed method needs location information of photos, we only used geo-tagged photos in the dataset. In addition to metadata, each photo is annotated with visual information detected by $N = 1;570$ Visual Concept classi ers using CNN. We used them to describe the Visual Concepts in each photo.