International Journal of
Geo-Information

Article

Exploring the Distribution Patterns of Flickr Photos

Xuan Ding [1] and Hongchao Fan [2,*]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; ding_x@whu.edu.cn

[2] Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, 7491 Trondheim, Norway

* Correspondence: hongchao.fan@ntnu.no; Tel.: +47-73559665

Received: 14 August 2019; Accepted: 15 September 2019; Published: 17 September 2019

Abstract: In recent years, volunteered-geographic-information (VGI) image data have served as a data source for various geographic applications, attracting researchers to assess the quality of these images. However, these applications and quality assessments are generally focused on images associated with geolocation through textual annotations, which is only part of valid images to them. In this paper, we explore the distribution pattern for most relevant VGI images of specific landmarks to extend the current quality analysis, and to provide guidance for improving the data-retrieval process of geographic applications. Distribution is explored in terms of two aspects, namely, semantic distribution and spatial distribution. In this paper, the term semantic distribution is used to describe the matching of building-image tags and content with each other. There are three kinds of images (semantic-relevant and content-relevant, semantic-relevant but content-irrelevant, and semantic-irrelevant but content-relevant). Spatial distribution shows how relevant images are distributed around a landmark. The process of this work can be divided into three parts: data filtering, retrieval of relevant landmark images, and distribution analysis. For semantic distribution, statistical results show that an average of 60% of images tagged with the building's name actually represents the building, while 69% of images depicting the building are not annotated with the building's name. There was also an observation that for most landmarks, 97% of relevant building images were located within 300 m around the building in terms of spatial distribution.

Keywords: VGI; Flickr; GIS; data mining; data distribution

1. Introduction

Volunteered-geographic-information (VGI) images are data associated with a specific geographic location through visual content or metadata such as coordinates and textual descriptions (e.g., location name) [1]. With the continuous contribution of volunteers around the world, these data contain a significant proportion of geographic information about our surroundings. This information has been used in various geographic applications in recent years, such as place-semantic-information extraction [2–4], scene summarization [5–7], and 3D reconstruction [8–14].

However, there are several issues when it comes to acquiring the relevant images for these applications from a large amount of VGI image data. The contamination issue, which was introduced in the work of Raguram et al. [9], means that the image content may not correspond to textual annotations such as tags and titles. Considering that querying with specific keywords (the name of the landmark) is a common way to collect reference data, this issue has two aspects of effects on these studies. First, retrieved images may not correspond to the geographic entity from its content. Second, some content-related images are normally missing due to imprecise descriptions in the annotated tags. Another way to retrieve VGI images is to search images close to a specific geolocation, but how much of the area surrounding the place should be searched remains an open question. Further, there are, most of time, many irrelevant images posted close to a place, and these images cannot be excluded by only setting a region around the place.

In order to solve the abovementioned problem and provide a better understanding of the quality of VGI image data, quality assessment has been conducted in various studies in recent years. Depending on the intended usage of the assessment, studies can be classified into semantic-accuracy analysis and positional-accuracy analysis. While positional-accuracy analysis mainly focuses on the correctness of geographic coordinates, semantic-accuracy evaluates the consistency between visual content and textual annotations. However, images analyzed in these studies are generally semantic images (tagged with specific keywords), while content-related but semantic-unrelated images received little consideration.

In this paper, the existing quality analysis was extended by exploring the distribution patterns of images collected from Flickr. Taking landmarks as an example, images are considered as not only semantic-related but also content-related. In order to retrieve content-related images, in a first step, we used an image-clustering technique to identify representative images of a landmark from semantic-related images. Then, these representative images were treated as visual queries to retrieve content-related images of the landmark. The distribution pattern of these images was studied in terms of two dimensions: (i) semantic distribution and (ii) spatial distribution. More specifically, semantic-distribution patterns show the accuracy of textual annotations (i.e., tags) by providing the proportion of images that were both tagged with the name of the given landmark and that visually represents the landmark. Further, it depicts the proportion of images that only associate with landmarks through visual content. Spatial distribution demonstrates the spatial relationship between content-related images and the actual landmark location.

The main contributions of this paper are twofold. First, we proposed a framework to find images referring to a certain building. Second, we found the empirical searching ranges of images to their corresponding buildings on Flickr based on our experiments in London. This should be an important value to set initial search regions for research communities when searching for images of specific buildings.

The remainder of this paper is structured as follows. We review the work related to this paper in Section 2. In Section 3, we introduce our data-filtering work and the landmark-image retrieval process. Section 4 describes the experiment results, and Section 5 provides a detailed discussion. Section 6 concludes the whole work.

2. Related Work

2.1. Applications Based on VGI Images

In recent years, VGI image data have been widely used in spatial analysis tasks. Georeferenced photos contain not only visual content but also textual metadata (tags, titles, descriptions) added by users to describe the images. This information can be used in many applications. For instance, Popescu et al. [2] and Keßler et al. [3] extracted georeferenced entities from VGI images and utilized them to build a geographic gazetteer. Mackaness et al. [4] used the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm to identify place-related tags with different levels of spatial detail. In addition to studies concerning the extraction of georeferenced information from textual metadata, some researchers are interested in the visual content of VGI images. A scene summarization problem was formulated in selecting a set of images that efficiently represent the visual content of a scene [5]. Kennedy et al. [6] introduced a context- and content-based approach to generate a representative set of images for landmarks. They extended the work of Simon et al. [5] by using location and textual information to automatically identify landmarks on a city scale.

Three-dimensional reconstruction based on VGI images is a complex research topic that benefits from various computer-vision tasks such as image retrieval and scene summarization as cited above. For instance, a comprehensive solution that combined image retrieval, scene summarization, and structure from motion (SFM) techniques to construct 3D landmark models was proposed in [9].

The utilization of scene summarization selects a subset of images that maximize the accuracy and completeness of the reconstruction result while minimizing computation time [15]. Although the image subset have different names in different studies, such as “canonical views” [5,8], “representative images” [6], or “iconic images” [9], their methods are essentially based on image clustering because people prefer to take photos from specific viewpoints, giving rise to many groups of photos with a similar appearance. While the abovementioned works are focused on improving the accuracy or efficiency of reconstruction work, there are also some researchers that focus on improving the image-retrieval process to obtain a better raw dataset before image clustering. Chum et al. [16] proposed visual-query expansion with spatial constraints to obtain more images related to a landmark. Gammeter et al. [12] expanded the small cluster by exploiting cross-media information (Wikipedia, Google), and they proved that the expanded dataset achieved better reconstruction results.

2.2. VGI Image Quality Assessment

Most of the researchers introduced above used the textual query method to retrieve images they needed. However, one of the inherent issues of VGI images is that image content may not correspond to the textual description, such as tags and titles. As shown in Figure 1, a photo named “Buckingham Palace” actually represented another landmark, namely, the London Eye. Unmatched tags can cause a certain number of irrelevant images to be retrieved in their work. To investigate how accurate the textual information of VGI images is, Zhong and Zhou [17] used the original keyword-based query (querying an entity with its name) to retrieve Flickr images and checked whether the returned images contained the entity. Their results showed that the mean accuracy of Flickr images was less than 50%. Panteras et al. [18] were interested in using VGI imagery for disaster response. Therefore, the most important issue was whether the images retrieved from the VGI platform reflected the ground truth of the area where the disaster happened. They selected a wildfire disaster as a case study and obtained 255 geotagged images through querying keywords such as fire and wildfire, as well as the toponyms of the affected area. Accuracy assessment showed that 56.2% of user-tagged annotations could be considered reliable. Moxley et al. [19] investigated the correlation between visual features and user-generated tags, with results showing that 57% of the tags could be regarded as visually identifiable tags.

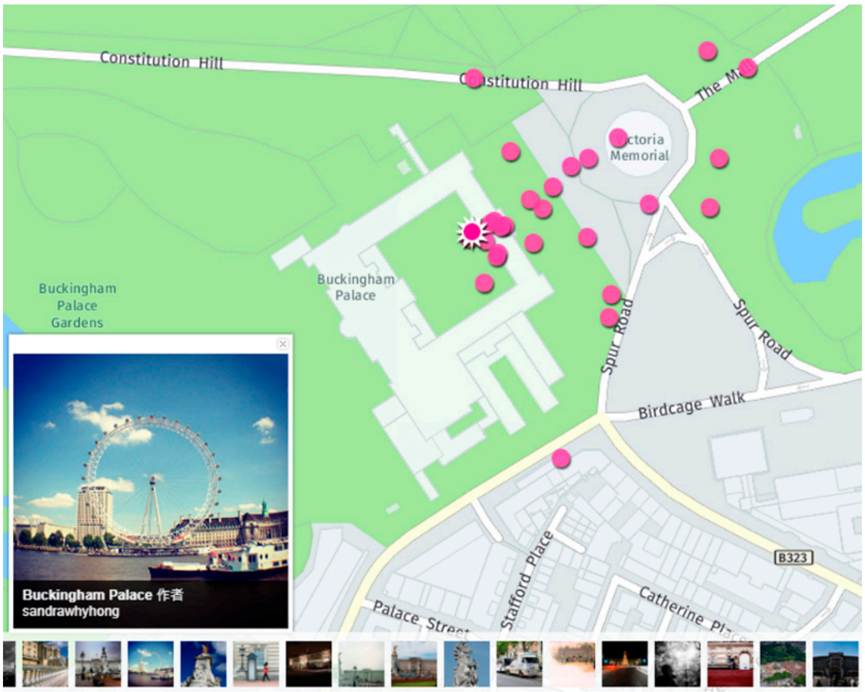


Figure 1. Unmatched-title image.

Since these geotagged images could also be obtained based on their locations, geotag-information accuracy affects the correctness of the retrieved images. For instance, an image in Figure 2, which depicts Buckingham Palace, was located far away from the actual position of the building. To make better use of these VGI images, Hauff [20] evaluated their positional accuracy through measuring the distance between the true location of a venue captured in the photo and the location recorded in its metadata. He selected ten venues and retrieved images that contained the name of the venue in the textual metadata. His experiment results showed that the accuracy of images taken at popular venues was higher than those taken at unpopular venues.

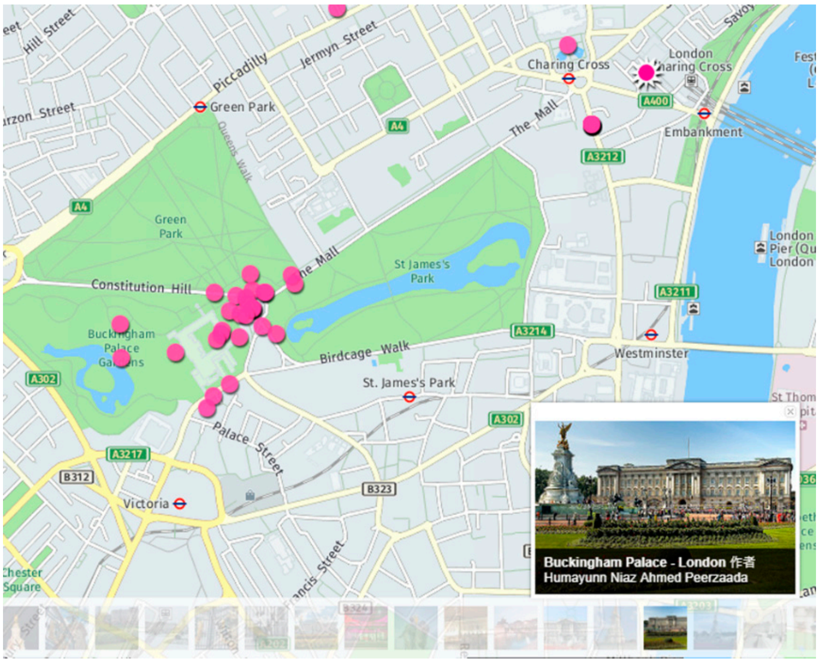


Figure 2. Incorrectly geotagged image.

The work of Zielstra and Hochmair [21] is similar to that of Hauff. They evaluated the positional accuracy of geotagged Flickr and Panoramio images through measuring the distance between image-coordinate pairs and estimated camera position. They formulated a manual estimation of the camera position based on the image content in order to avoid many of the limitations brought about by automatic approaches. Their experiment results showed that street-related features such as buildings had higher accuracy than other features. Senaratne et al. [22] proposed reverse-viewshed analysis to evaluate the location correctness of an image. They selected two points of interests (POIs) and retrieved the photos that tagged the POI names. The location correctness of each photo was evaluated by checking whether the calculated visible area of the image containing the POI position. Their work revealed that, the closer the image to the corresponding POI was, the more accurate the textual and geolocation information is.

However, these works analyzed only a subset of images that related to an entity or a specific geographic location with textual descriptions. Considering that visual content is the key clue to associate it with a location, how many images are content-related and how they are distributed is important for content-based applications such as 3D building reconstruction. We adopted the scene-summarization techniques and the image retrieval introduced above to retrieve most content-related images of a landmark. The detailed process is discussed in the following section.

3. Data and Methods

In this section, the overview of the proposed approach to explore the distribution of VGI images is provided in Section 3.1. Sections 3.2–3.4 then describes the methodology in detail. In Section 3.5, the definition and the distribution analysis aim are stated.

3.1. Overview of Exploring VGI Images Distribution Patterns

The approach to explore the distribution patterns of VGI images entails a series of steps as described below.

In the first step, Flickr images are downloaded through the Flickr Application Programming Interface (API). Since there are a significant number of images that do not represent any buildings or that were captured inside a building, which is hard to distinguish, these images are first filtered based on scene-recognition techniques to reduce the computational cost of the following process. The remaining images are deemed as building-related images. In the second step, image clustering is carried out with a subset of the building-related images that are tagged with the name of the landmark, that is, the semantic-related images. This process selects the most representative images of the landmark. Then, the selected representative images are treated as seeds to retrieve part of the content-related images that represent the landmark but are not tagged with its name. Finally, the retrieved images and the semantic-related images are used to analyze their distributions in terms of semantic and spatial distribution. Semantic distribution describes how well the Flickr image tags match the content, while spatial distribution depicts how images are distributed around the corresponding landmarks.

3.2. Flickr Dataset and Irrelevant-Building Images Filtering

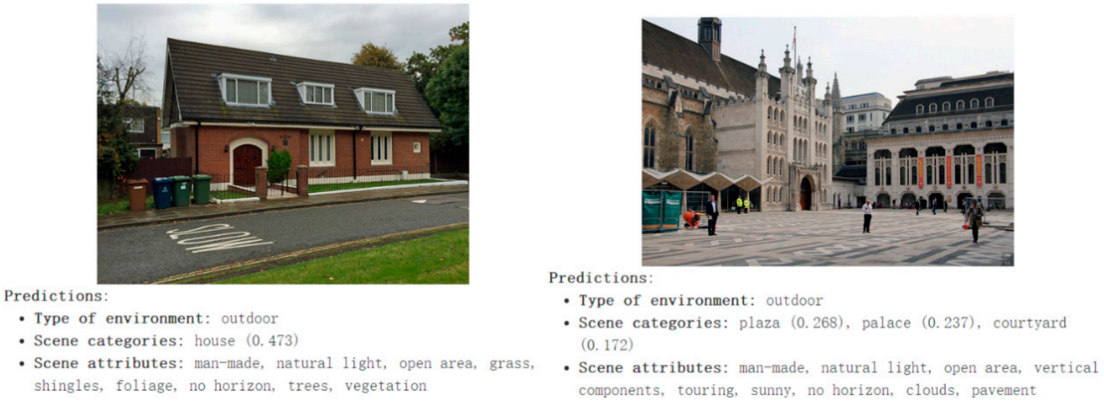
Flickr is one of the most prominent photo-sharing websites, with users uploading as many as 25 million photos on a very-high-traffic day [23]. As per the statistics of Michel [24], from 2004 (the year when Flickr started) until December 2016, a total of 5.87 billion public photos were uploaded. Developers and researchers can retrieve Flickr data through the Flickr API after registration.

The data retrieved from API are in JSON (or XML) format. The extracted information can be parsed into two parts, Flickr images and metadata. Metadata contain different kinds of information about these images, such as photo ID, title, tags, description, coordinates, and even the photo owner's information. Textual information (title, tags, description) is optionally provided by users, while location information is generated in two ways: built-in GPS device receivers and manual geotagging by photographer through web services.

The subjects of Flickr images can be classified into diverse types, such as portrait and landscape images. Most retrieved images without any buildings in their content can hardly provide useful information about a landmark. Therefore, those building-unrelated images were excluded in this work. However, it is infeasible to manually select building-related images from a large number of images. Many applications demand algorithms for the automated selection of building-related images.

Recently, the emergence of large datasets like ImageNet and the rise of Convolutional Neural Networks (CNNs) have significantly improved many computer-vision tasks [25]. Scene recognition, a hallmark task of computer vision, can be applied to automatically filter out useless images by checking image scene context. A scene-recognition image database named Places (http://places.csail.mit.edu/) contains 7 million labeled images of scenes and has greatly improved scene-recognition performance. The latest version of Places is Places2.

Places365-Standard is one of the latest subsets of Places2—365 refers to the 365 labeled scene categories for all 1.8 million images. In this work, we adopted classification network VGG16 that achieved the best performance in top1 accuracy on the Places365-Standard dataset. The trained network was able to automatically classify Flickr images into the 365 specified scene categories. Unfortunately, these scene categories could not directly be used to determine whether an image was building-related due to classification ambiguity. For instance, some building-related images are classified as "house", which is closely related to buildings, while others are classified with ambiguous labels such as "plaza" or "courtyard", as shown in Figure 3.



Predictions:
- **Type of environment:** outdoor
- **Scene categories:** house (0.473)
- **Scene attributes:** man-made, natural light, open area, grass, shingles, foliage, no horizon, trees, vegetation

Predictions:
- **Type of environment:** outdoor
- **Scene categories:** plaza (0.268), palace (0.237), courtyard (0.172)
- **Scene attributes:** man-made, natural light, open area, vertical components, touring, sunny, no horizon, clouds, pavement

Figure 3. Various kinds of scene recognition results of building-related images.

In order to overcome the above problem, a simple workflow has been designed in this work. We randomly selected 400 building-related images and classified them using the trained CNN model. The predicted labels that were obviously unrelated to buildings were filtered out, while the rest of the labels were considered as building-related image labels, as shown in Table 1. Finally, these labels were applied to select the images whose classification result matched the building-related image labels. The workflow is illustrated in Figure 4.

Table 1. Building-related image labels and other labels.

Type	Label Name
Building-related image labels	"apartment_building-outdoor", "building_facade", "church-outdoor", "courthouse", "courtyard", "embassy", "general_store-outdoor", "hospital", "hotel-outdoor", "house", "inn-outdoor", "library-outdoor", "mansion", "museum-outdoor", "oast_house", "office_building", "palace", "parking_garage-outdoor", "synagogue-outdoor"
Other labels	"arcade", "arch", "atrium-public", "barndoor", "bus_station-indoor", "construction_site", "crosswalk", "drugstore", "fire_station", "formal_garden", "harbor", "kennel-outdoor", "loading_dock", "lock_chamber", "museum-indoor", "natural_history_museum", "parking_lot", "pharmacy", "playground", "promenade", "railroad_track", "shopping_mall-indoor", "temple-asia"

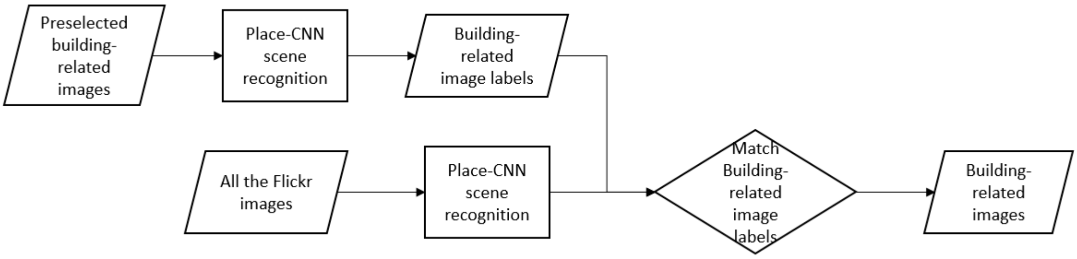

Preselected building-related images
Place-CNN scene recognition
Building-related image labels
All the Flickr images
Place-CNN scene recognition
Match Building-related image labels
Building-related images

Figure 4. Detailed workflow of the data-filtering process.

3.3. Representative Image Selection

Retrieving all images of a given landmark is a complex task in the field of image retrieval. Normally, image-retrieval tasks retrieve all instances of an object from a large database through a given query image of that object. However, a landmark may have various appearances in different images. Hence, it is impossible to retrieve all relevant images based on a single query image. Further, the issue of the matching process aside, it is hard to manually select images that contain sufficient information about a landmark. Fortunately, according to [9], most landmark images are often taken from certain viewpoints, forming groups of images with a similar appearance. Thus, instead of developing an image-retrieval algorithm that can retrieve most relevant images with a single image, we used an image-clustering technique to discover a set of images that formed a complete visual summary of the landmark. These images were regarded as representative images and then used as seeds to retrieve more relevant images.

In order to calculate the similarity between two images, it is necessary to find a suitable image descriptor. Although global image features, such as gist, were used in recent studies [13,14] to speed up the image clustering process, the cost of it is loss of clustering accuracy. The reason is that global image features describe the overall appearance of an image rather than specific objects in the image. This leads to similar-looking images falling into different clusters due to changes in illumination or scaling. Furthermore, images with similar gist descriptors but irrelevant content are grouped into the same cluster.

To remedy this, local image feature Speeded Up Robust Features (SURF) [26] was used in this paper. This is a scale- and rotation-invariant interest point detector and descriptor that computes and compares much faster than the classical Scale Invariant Feature Transformation (SIFT) descriptor [27]. Compared with global image features, local image features can help to identify actual structural elements of real-world objects and ensure that the intended objects are actually contained in the images. In this work, each image in the raw dataset was represented by a set of 64-dimensional SURF feature vectors. For each pair of images, the features were matched by calculating the nearest Euclidean distance between all feature pairs and verified by applying the distance ratio test [27]. Furthermore, the matched feature pairs were filtered by applying geometric constraints to make sure that image pairs were consistent in appearance as well as geometry. The geometric relationship between image pairs was estimated by a random-sample-consensus (RANSAC) algorithm [28]. During each RANSAC iteration, four matched SURF keypoint pairs were randomly selected to compute a candidate fundamental matrix. The remaining matches that satisfied the fundamental matrix were treated as inliers, otherwise, as outliers. The iteration with the most inliers was retained. If the inliers were less than a threshold (10 in the implementation), the image pairs were regarded as inconsistent. Finally, in order to increase the ratio of correct matches, the image pairs with more inliers than the threshold in both the forward and reverse matching directions were retained.

To identify images with a similar appearance and group them into different clusters, the image-clustering process was performed using density-based spatial clustering for applications with noise (DBSCAN) [29]. This algorithm was configured by two parameters, namely, search radius (Eps) and minimum number of points within the search radius (MinPts). While alternative clustering methods, such as K-Means, were used in similar works [6,9], DBSCAN does not require a predefined cluster number. It is difficult to estimate a proper number of clusters for different landmarks. In addition, DBSCAN is robust to data noises that are common for these VGI images.

In this paper, the distance of two images in terms of their appearance was defined using Formula (1):

$$distance = \begin{cases} \frac{I_{\max}-I_{\min}}{I_{ij}-I_{\min}+1}, I_{ij} > 10 \\ \infty, I_{ij} < 10 \end{cases} \tag{1}$$

where I_{ij} is the number of inliers for image pair (i, j) and I_{max} and I_{min} are the maximal and minimal number of inliers in the whole raw-reference dataset. According to this definition, the more similar two images are, the smaller the distance is. Therefore, the smaller Eps is, images in the same cluster appear more similar, while a larger MinPts can help to ensure higher quality for the clustered images.

However, in the above steps, the value of Eps was set to be relatively small to achieve high quality of clustering results. This caused some relevant images that were less similar to other images to be classified as noise images. In order to retrieve discarded relevant images, the second round of clustering was applied to the noise images. The value of Eps was set higher than during the initial clustering to expose more image clusters. Finally, for images in each cluster generated from both the initial and the second clustering process, the cumulative distance was calculated. Images were ranked according to how well they represented the cluster. Those images with the lowest distance were selected as the representative images of each building. Sometimes, there were images that were irrelevant to the intended landmark that were also grouped into a cluster. This does not happen very often. In some cases, it is difficult to identify whether representative images selected from these groups are captured from less common aspects of the landmark or an unrelated object. Since there is no ideal way to deal with these images yet, a manual check was applied to select representative images that actually depicted the landmark. This step is illustrated in Figure 5.

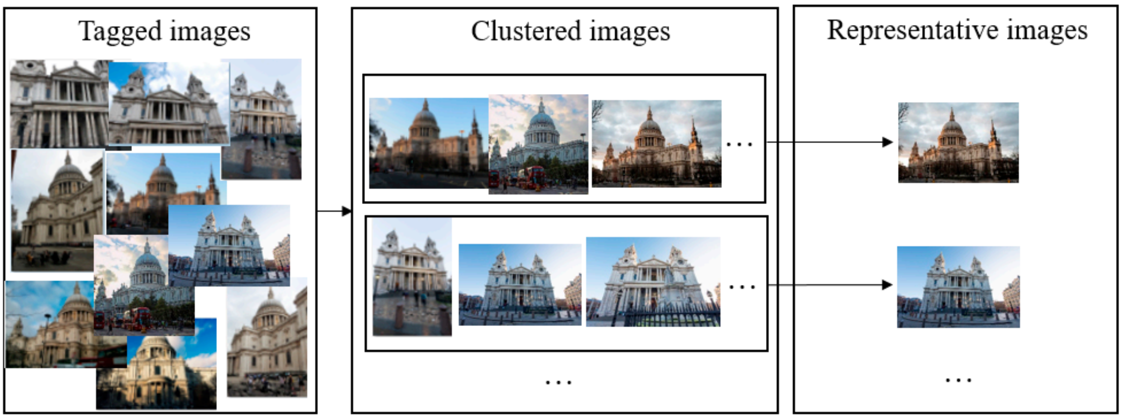


Figure 5. Workflow of representative-building-image selection.

3.4. Relevant Image Retrieval

After the selection of representative landmark images, the next step was to use an image-matching technique to retrieve images that contained the landmark but were not tagged with the name of the landmark in the whole building-related image dataset. As mentioned above, local features perform better than global features in reflecting object information. Therefore, local image features were selected to search for relevant images in this work. While traditional methods of extracting and representing local image features are generally based on carefully hand-crafted methods, deep-learning-based methods have achieved better performance than traditional ones in the past few years [30].

In this work, a deep-learning algorithm named Learned Invariant Feature Transform (LIFT) [30] was applied for local-image-feature extraction and description. The architecture of this deep network-integrated three individual CNNs that corresponded to the three steps in the local-feature-processing chain: feature detection, orientation computation, and feature description. Since the network was trained with image patches that contained feature points for the SFM reconstruction process, the learned features were robust to viewpoint changes and illumination conditions. As shown in Figure 6, there were only nine valid feature matches of the image matching results based on SIFT, while there were 32 valid matched feature pairs that could be found based on LIFT.

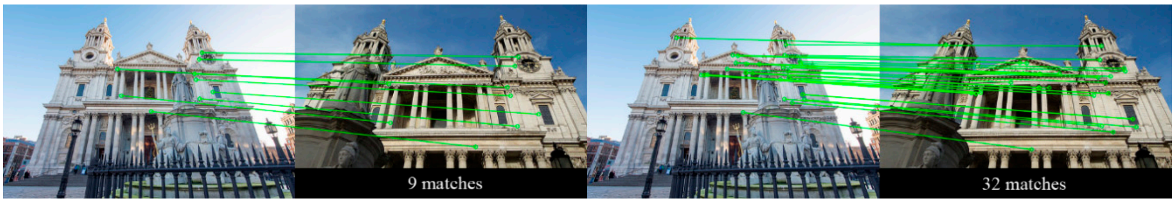


Figure 6. Local-image-feature matching examples of (left) Scale Invariant Feature Transformation (SIFT) and (right) Learned Invariant Feature Transform (LIFT).

In order to ensure the geometric consistency between seed images and matched images, the RANSAC algorithm was used to estimate their geometric relationship. The matched image pairs with inconsistent geometry were filtered. After RANSAC filtering, putative matches with fewer than 10 inliers were discarded, according to our experiences in this study. A problem that remained with the local-feature-matching algorithm was that the matching sequence of the images affected the results in most cases. Figure 7 shows an example of this. There were 11 matching pairs (mostly incorrectly matched) that could be detected when using the target image (selected image for specific building) to match the image to be tested. After transferring the matching order, only four matched pairs remained. Then, a bidirectional matching check was applied to match images from both directions. As a result, an image pair could only be considered a valid match if the number of inliers generated from both matching directions was higher than the threshold (10 were selected from the experiment results). Note that there are some improved algorithms that can achieve better retrieval performance. However, it should also be noted that image retrieval itself is an ongoing research topic in the field of computer vision. Designing a more sophisticated and efficient image-retrieval algorithm is beyond the scope of this research.

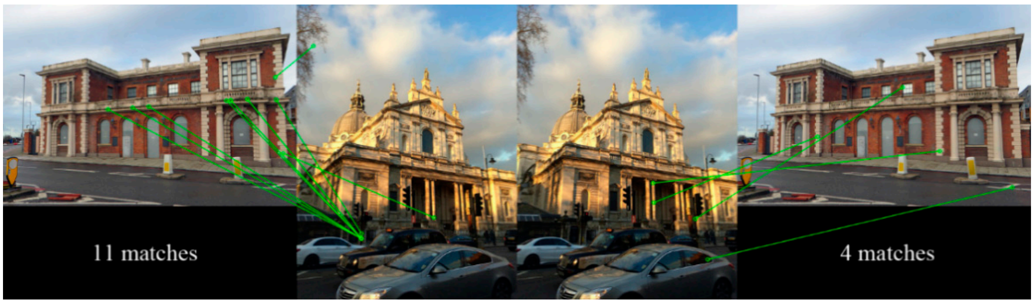


Figure 7. Different matching results of two matching directions.

3.5. Distribution Analysis

As stated previously, Flickr data distribution was analyzed based on two aspects, namely, semantic and spatial distribution.

Since Flickr images can be freely tagged using any words or phrases that users choose, textual metadata are less constrained than authoritative data, such as a geographical gazetteer. As shown in Figure 1, textual metadata are not always associated with the content of the images. In this paper, the aim of semantic-distribution analysis was to explore how many images were annotated with the building's name (semantic-relevant) while the landmark was contained/not contained in its content (content-relevant/-irrelevant). How many images actually represent the landmark but are not labeled with its name (semantic-irrelevant) is another important issue to be addressed. For this purpose, these images were classified into three categories: (a) semantic-relevant and content-relevant images, (b) semantic-relevant but content-irrelevant images, (c) semantic-irrelevant but content-relevant images.

With regard to spatial distribution, we were interested in how these relevant images are distributed around a landmark. The spherical distance between images' recorded position and the actual position of the buildings (extracted from OpenStreetMap building footprints) was calculated for spatial distribution analysis.

4. Experiment Results

4.1. Image-Retrieval Results

The datasets used in the present study were downloaded from the Flickr API (https://www.flickr.com/services/api/) based on their locations. We retrieved all publicly available and geotagged Flickr images of London from 2016, and acquired more than 489,000 images in total. After the data-filtering process outlined in Section 3.1, 41,463 building-related images were retained. We selected six landmarks with sufficient images to test our retrieval approach and analyze their distribution patterns.

The name of each landmark was used as a query condition to retrieve tagged images for the image-clustering process. Since tags were lowercase and their blanks were removed before they were stored in the database, we queried, for instance, "buckinghampalace" for Buckingham Palace. These tagged images were labeled as positive or negative based on whether the landmark was present in the image or not. These corresponded to the images of categories *a* and *b* that were defined in Section 3.3, respectively. Table 2 gives the count of positive and negative images of each landmark.

Table 2. Statistics of each kind of image for each selected building. Tagged images: tagged with building name, retrieved images: retrieved based on representative images.

Building Name	Tagged Images		Representative Images	Retrieved Images
	Positives	Negatives		
St Paul's Cathedral	143	67	18	329
Buckingham Palace	154	139	17	357
Westminster Abbey	97	64	18	402
National Gallery	44	35	7	114
British Museum	37	87	3	89
Imperial War Museum	50	5	10	52

Figure 8 presents the performance of the image-clustering results based on different Eps combinations in the two-step clustering process. The performance of our retrieval approach was measured in terms of recall. The recall used here was the proportion of positive images retrieved based on the representative images to the total positive images of the landmark. To ensure that most relevant images were classified into the correct groups, MinPts was set as 2 in the implementation. That means that at least two similar images were needed to form a cluster. Representative images that achieved the best recall were used as seeds to search for relevant images. For instance, we selected representative images obtained from an Eps combination of 20 and 25 as the queries for these landmarks, except for the National Gallery, because the highest recall was achieved based on this combination, whereas images of the National Gallery tend to be captured from a limited area at the front of the building. Most of these images also appear similar and need to be distinguished with a low Eps (5 and 10 in the implementation). Table 2 gives the number of representative images and the retrieved images of each building.

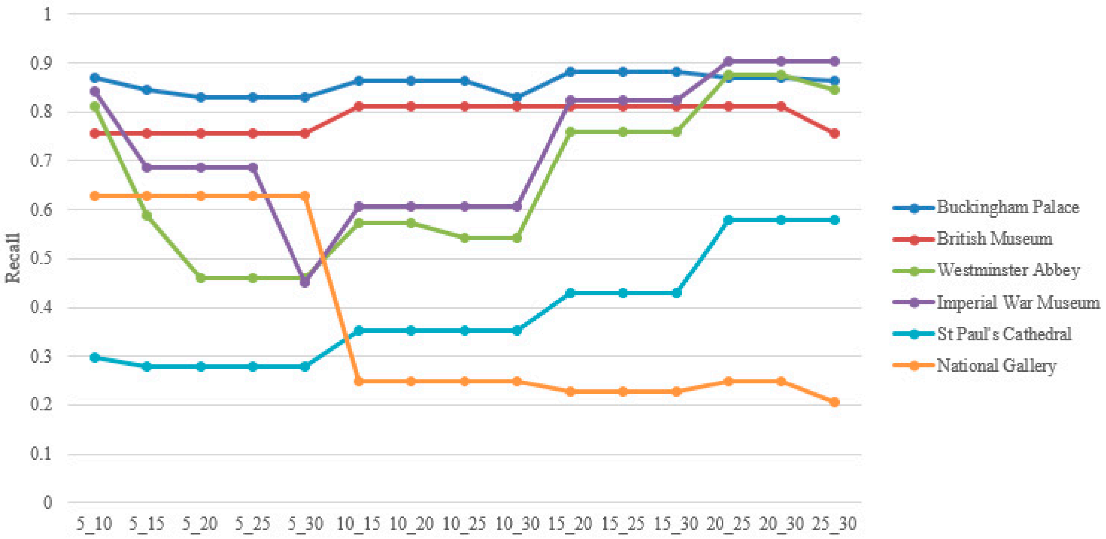


Figure 8. Recall of different Eps combinations for each landmark.

4.2. Distribution Results

As stated in Section 4.1, landmark images were classified as *a* (semantic-relevant and content-relevant), *b* (semantic-relevant but content-irrelevant), or *c* (semantic-irrelevant but content-relevant). The images of categories *a* and *b* were labeled manually, while images of category *c* were obtained through our retrieval approach. Figure 9 reveals that most landmark images were semantic-irrelevant but content-relevant. In other words, they belonged to category c. This reflects the fact that the majority of Flickr contributors did not make an effort to tag images besides uploading them in a relatively correct position. The proportion of category *b* images of the British Museum was higher than that of other landmarks. This is due to the fact that many images taken inside the museum failed to be filtered out, such as images of a building model, and were classified into category *b*.

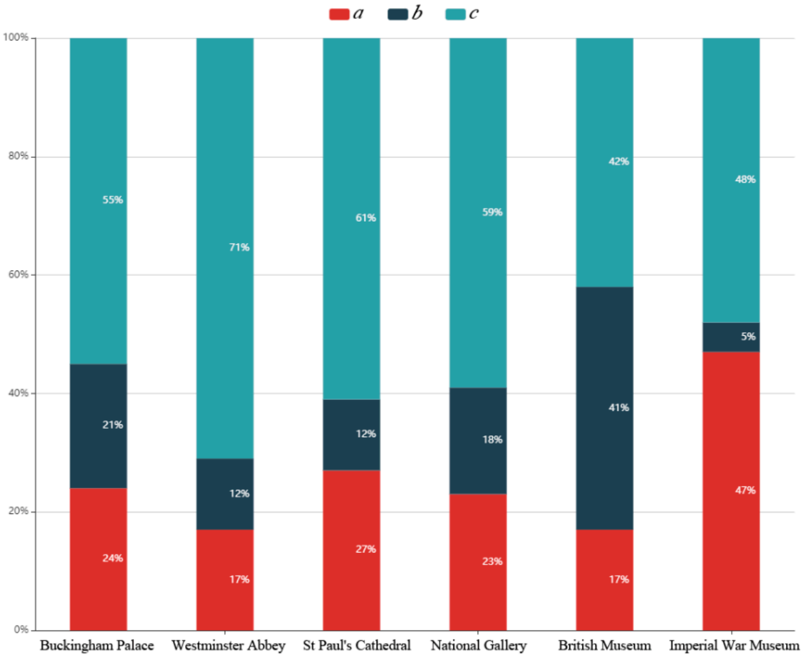


Figure 9. Proportions of category a (semantic-relevant and content-relevant), b (semantic-relevant but content-irrelevant), and c (semantic-irrelevant but content-relevant) images of each landmark.

Drawing upon the definition of precision and recall that was used in the information-retrieval scenario, the two indicators were used in this work to evaluate the performance of retrieving relevant images of a specific landmark through querying its name. While precision can be calculated by the proportion of category *a* images from all category *a* and *b* images, recall was evaluated through the proportion of category *a* images from all category *a* and *c* images. Therefore, the mean precision of the image-retrieval approach based on keyword searching was 60%, and the mean recall was 69%.

The spatial distributions of landmark images are shown in Figure 10. Content-relevant images of each landmark are visually depicted with red dots. It is obvious that most of them were located around the building, while only few images had relatively long distance from their corresponding buildings. For instance, images of Buckingham Palace and the National Gallery are mainly located in front of the building. This is due to there being an open square in front of them, and the front side of the building is more representative than the other sides. In the case of Westminster Abbey and St Paul's Cathedral, the facades of the different sides are quite distinctive. Hence, their images were mainly distributed around of these sides.

Figure 11 presents a more intuitive view of the count distribution of the distance between images and their corresponding buildings. These images were divided into different layers based on the distance between the recorded position and the corresponding landmark. For instance, "0–50" stands for images within 50 m of the building, and "50–100" represents images in the 50 to 100 m range. It is clear that images were mainly located in the 50–150 m range, as shown in Figure 12, where the light- or dark-green layers are always wider than the other layers and lie at the bottom of the graphs.

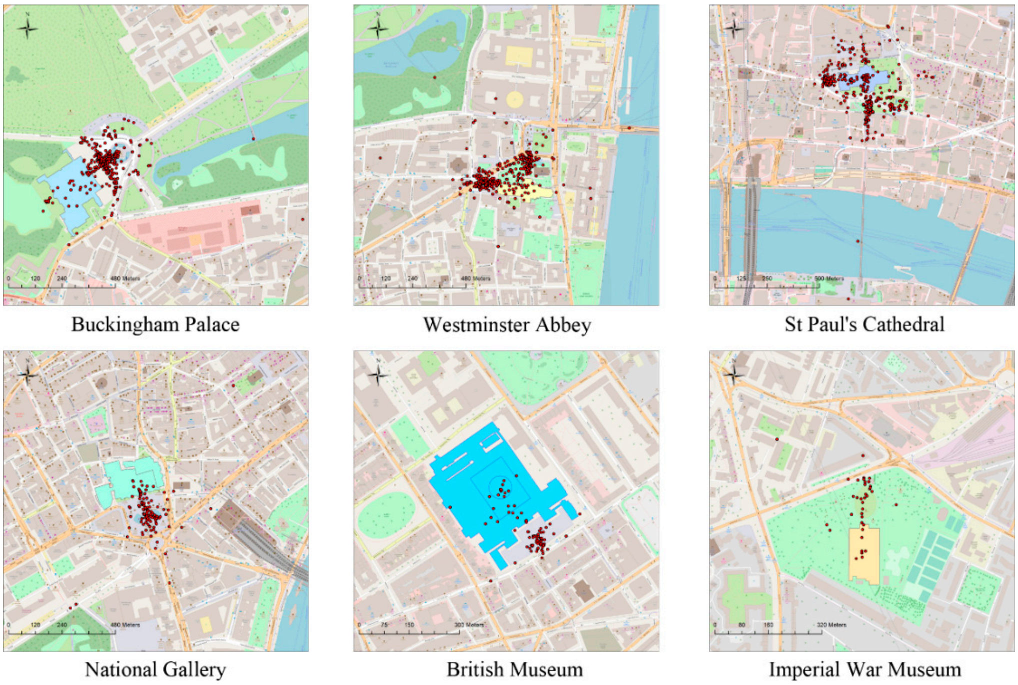


Figure 10. Spatial distribution of relevant images (red dots) of each landmark.

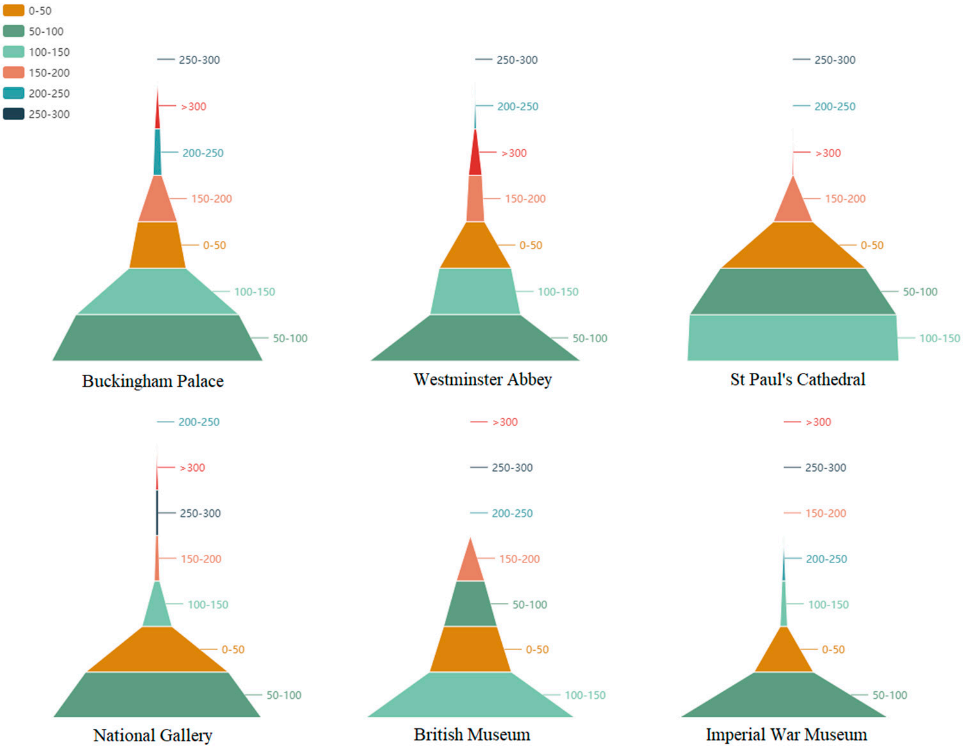


Figure 11. Count-distribution visualization of each distance layer. The wider the layer is, represents the more images were distributed within the corresponding range.

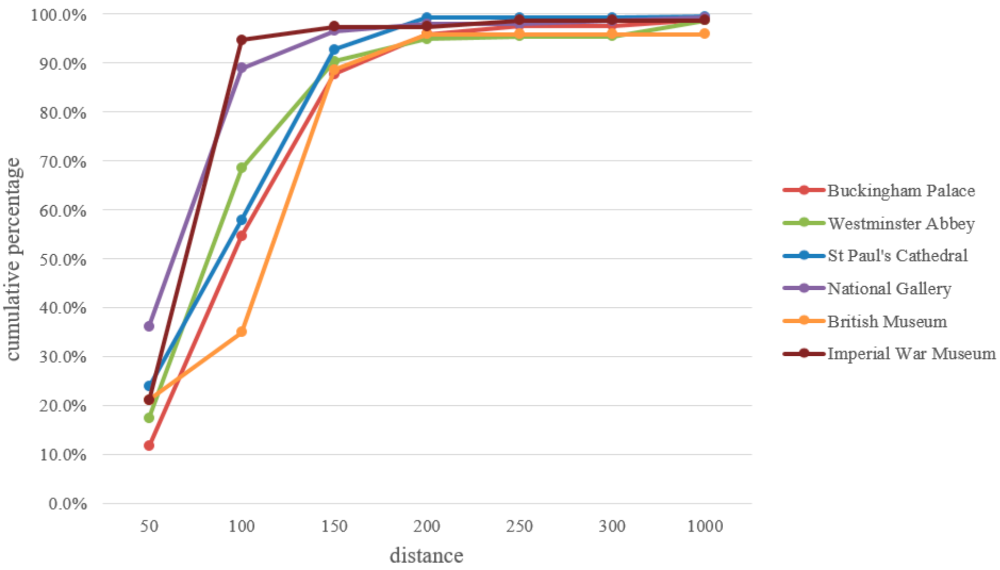


Figure 12. Cumulative-count-distribution graph of distance between images and landmarks.

Figure 12 shows the cumulative count distribution of the distance from images to corresponding buildings, with "50" meaning the proportion of images within 50 m of the building. It is clear that more than 97% of the images were distributed, on average, in the range of 300 m from the buildings. A possible explanation is that it was difficult to capture these landmarks at a distance of over 300 m.

5. Discussion

This study utilized an approach combining image-clustering and -matching techniques to retrieve landmark images, and explored the semantic and positional distribution of these images. The image-retrieval results demonstrated that our approach retrieved a relatively large number of images that visually represented the landmark, but were not annotated with the name of the landmark. With respect to the distribution pattern of landmark images, semantic distribution showed that the accuracy of tags that annotated the image was 60%, on average. This means that 60% of the images were not only tagged with the name of the landmark but also represented it. This result is similar to the observation of [18,19]. The mean recall of the textual query of six landmarks, meanwhile, was only 31%. According to the result of [12], the larger the collection of images that correspond to a scene is, the better the performance of its 3D reconstruction is. It can thus be concluded that image-retrieval work based on the keyword-searching approach misses more than 69% of relevant images. That is a large amount of valuable data for 3D building-reconstruction analysis.

Spatial distribution showed that most relevant images were located within 300 m of the building. Therefore, compared to collecting images by using keyword searches, query images based on its geographic coordinates obtain more relevant images. Further, this finding could greatly reduce the searching range (i.e., 300 m around a given location) of the image-collection process.

Since even state-of-the-art image-retrieval approaches cannot retrieve all images that actually depict a landmark, some relevant landmark images were also omitted in our approach, especially for St. Paul's Cathedral and the National Gallery, which could be photographed from quite a wide range. Images that were taken from some specific viewpoints cannot be grouped into any clusters. Moreover, there were many images that only captured part of the buildings, such as a window of St. Paul's Cathedral that is hard to match. Nevertheless, this indicated that the proportion of content-related images missed by keyword searching was more than we calculated.

In this work, the spatial-distribution pattern was only explored for a number of landmarks in London. According to [21], positional accuracy may vary in different regions due to the varying percentage of images that were taken with GPS-equipped units. Nevertheless, we may argue that the distribution pattern explored in this work should be similar to those in other cities because the quality characteristics of Flickr as a sort of VGI data should be more or less similar due to the mechanism of data contribution.

6. Conclusions

This paper gave insight into the distribution pattern of VGI images associated with landmarks through text annotations as well as visual content. The semantic-distribution results showed that retrieving images of a specific building based on keyword searching misses a significant number of content-related images. These missed images may provide additional information about landmarks that could be used to improve image-based applications, such as 3D reconstruction. With respect to spatial distribution, the statistical results showed that most relevant images were located within 300 m of the landmarks, which would greatly reduce the data-collection range for tasks that retrieve VGI images based on geolocation information.

In the future, the framework for retrieving Flickr images will be optimized to be more efficient. The spatial distribution of Flickr images in other big cities will also be analyzed in order to check/verify the research findings in this paper.

Author Contributions: Hongchao Fan contributed toward creating the original ideas of the paper and designed the experiments. Xuan Ding prepared the original data, performed the experiments and analyzed the experimental data under the supervision of Hongchao Fan. Xuan Ding wrote the first draft of the manuscript, while Hongchao Fan revised and edited it.

Funding: This work is supported by the NSFC (National Natural Science Foundation of China) project No. 41771484.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. [CrossRef]
2. Popescu, A.; Grefenstette, G.; Moëllic, P.A. Gazetiki: Automatic creation of a geographical gazetteer. In Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, Pittsburgh, PA, USA, 16–20 June 2008; pp. 85–93.
3. Keßler, C.; Maué, P.; Heuer, J.T.; Bartoschek, T. Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In *Proceedings of the 3rd International Conference on GeoSpatial Semantics*; Springer Science and Business Media LLC: Berlin, Heidelberg, 2009; Volume 5892, pp. 83–102.
4. Mackaness, W.A.; Chaudhry, O. Assessing the Veracity of Methods for Extracting Place Semantics from Flickr Tags. *Trans. GIS* **2013**, *17*, 544–562. [CrossRef]
5. Simon, I.; Snavely, N.; Seitz, S.M. Scene Summarization for Online Image Collections. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8.
6. Kennedy, L.S.; Naaman, M. Generating diverse and representative image search results for landmarks. In Proceedings of the 17th International Conference, Beijing, China, 21–25 April 2008; pp. 297–306.
7. Avrithis, Y.; Kalantidis, Y.; Tolias, G.; Spyrou, E. Retrieving landmark and non-landmark images from community photo collections. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 153–162.
8. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo tourism: Exploring photo collections in 3D. In *ACM Transactions on Graphics (TOG)*; ACM: New York, NY, USA, 2006; Volume 25, pp. 835–846.
9. Snavely, N.; Seitz, S.M.; Szeliski, R. Modeling the world from internet photo collections. *Int. J. Comput. Vis.* **2008**, *80*, 189–210. [CrossRef]
10. Zheng, Y.-T.; Zhao, M.; Song, Y.; Adam, H.; Buddemeier, U.; Bissacco, A.; Brucher, F.; Chua, T.-S.; Neven, H. Tour the world: Building a web-scale landmark recognition engine. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1085–1092.

11. Agarwal, S.; Snavely, N.; Simon, I.; Seitz, S.M.; Szeliski, R. Building rome in a day. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 72–79.
12. Gammeter, S.; Quack, T.; Tingdahl, D.; Van Gool, L. Size does matter: Improving object recognition and 3D reconstruction with cross-media analysis of image clusters. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 734–747.
13. Raguram, R.; Wu, C.; Frahm, J.-M.; Lazebnik, S. Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. *Int. J. Comput. Vis.* **2011**, *95*, 213–239. [CrossRef]
14. Frahm, J.-M.; Heinly, J.; Zheng, E.; Dunn, E.; Fite-Georgel, P.; Pollefeys, M. Geo-registered 3D models from crowdsourced image collections. *Geo-Spat. Inf. Sci.* **2013**, *16*, 55–60. [CrossRef]
15. Snavely, N.; Seitz, S.M.; Szeliski, R. Skeletal graphs for efficient structure from motion. In Proceedings of the 2008 IEEE Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; Volume 1, p. 2.
16. Chum, O.; Philbin, J.; Sivic, J.; Isard, M.; Zisserman, A. Total recall: Automatic query expansion with a generative feature model for object retrieval. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
17. Zhong, H.; Zhou, Y. Combining content and quality indicators in ranking ambiguous query results on Flickr. In Proceedings of the Twenty-Third Australasian Database Conference-Volume 124, Melbourne, Australia, 31 January–3 February 2012; Australian Computer Society, Inc.: Darlinghurst, Australia, 2012; pp. 109–116.
18. Panteras, G.; Xu, L.; Croitoru, A.; Crooks, A.; Stefanidis, A. Accuracy Of User-Contributed Image Tagging In Flickr:A Natural Disaster Case Study. In Proceedings of the International Conference on Social Media & Society, London, UK, 11–13 July 2016; pp. 1–6.
19. Moxley, E.; Kleban, J.; Xu, J.; Manjunath, B.S. Not all tags are created equal: Learning flickr tag semantics for global annotation. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, New York, NY, USA, 28 June–3 July 2009.
20. Hauff, C. A study on the accuracy of Flickr's geotag data. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013; pp. 1037–1040.
21. Zielstra, D.; Hochmair, H.H. Positional accuracy analysis of Flickr and Panoramio images for selected world regions. *J. Spat. Sci.* **2013**, *58*, 251–273. [CrossRef]
22. Senaratne, H.; Bröring, A.; Schreck, T. Using Reverse Viewshed Analysis to Assess the Location Correctness of Visually Generated VGI. *Trans. GIS* **2013**, *17*, 369–386. [CrossRef]
23. A Year Without A Byte. Available online: http://code.Flickr.net/2017/01/05/a-year-without-a-byte/ (accessed on 23 June 2019).
24. How Many Public Photos are Uploaded to Flickr Every Day, Month, Year? Available online: https://www.flickr.com/photos/franckmichel/6855169886 (accessed on 23 June 2019).
25. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*; Massachusetts Institute of Technology Press: Cambridge, MA, USA, 2014; pp. 487–495.
26. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
27. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
28. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
29. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 1996; Volume 96, pp. 226–231.
30. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*; Springer: Cham, Germany, 2016; pp. 467–483.