Utilizing Reverse Viewshed Analysis in Image Geo-Localization

Yuhao Kang
Department of Geography
University of Wisconsin, Madison
Madison, WI, United States
yuhao.kang@wisc.edu

Song Gao
Department of Geography
University of Wisconsin, Madison
Madison, WI, United States
song.gao@wisc.edu

Yunlei Liang
Department of Geography
University of Wisconsin, Madison
Madison, WI, United States
yunlei.liang@wisc.edu

ABSTRACT

When users browse beautiful scenery photos uploaded on a social media website, they may have a passion to know about where those photos are taken so that they could view the similar sceneries when they go to the same spot. Advancement in computer vision technology enables the extraction of visual features from those images and the widespread of location-awareness devices makes image positioning possible with GPS coordinates or geo-tags (e.g., landmarks, place names). In this paper, we propose a novel method for image positioning by utilizing spatial analysis and computer vision techniques. A prototype system is implemented based on large-scale Flickr photos and a case-study of the Eiffel Tower is demonstrated. Both global and local visual features as well as the spatial context are utilized aiming at building a more accurate and efficient framework. The result illustrates that our approach can achieve a better accuracy compared with the baseline approach. To our knowledge, it is among the first researches that combine not only the visual features of photos, but also take the spatial context into consideration for the image geo-localization using high-density social media photos at the spatial scale of a landmark.

CCS CONCEPTS

• Information systems → Geographic information systems; • Computing methodologies → Computer vision

KEYWORDS

Geo-localization, image retrieval, viewshed analysis, spatial heterogeneity, image matching

1 Introduction

With the development of the economy and technologies, tourism has become an important part of human life. Thanks to the widespread of Internet and mobile devices, people from all over the world could browse photos that other tourists took and shared on the websites, such as Flickr, Instagram, etc. When users plan their tour, the beautiful scenery photos would inspire the curiosity of users to find out where those images are taken. A user may visit the same scenery spot during his/her own trip. Therefore, the identification of the photo location from its visual content and some other attributes, what is called *geo-localization* [3], is quite necessary and useful. In the past decade, there has been an intensive research interest in image geo-localization [2,3,6,12]. Some traditional methods combine global and local visual feature detection algorithms to represent images as multi-dimensional vectors and the similarities among them could be computed. After matching the key points of the target image and comparing the similarity of all images in a database, the candidate images that represent similar scenes can be retrieved and the location coordinates are ranked as the position of the input image [2, 6, 8]. Recently, deep neural networks have also been implemented for solving the geo-localization problem [11]. However, challenges remain for geo-localizing high-density of heterogenous images in a fine spatial resolution. To this end, we propose a novel method for image positioning by utilizing spatial analysis and computer vision techniques. Our contribution of this work is threefold:

(1) We preprocessed the dataset with a binary classification: "with" or "without" geographic information in order to improve the efficiency of the training process.
(2) We utilized the reverse viewshed analysis in GIS to reduce the searching area for the potential locations of images. To our knowledge, it is among the first image geo-localization research that takes such spatial context into consideration.
(3) Using large number of images taken around the Eiffel Tower as an example, we validated the accuracy and the effectiveness of our image geo-localization method that outperformed the baseline approach without the reverse-viewshed analysis.

The following of this paper is organized as follows: We present the framework and elaborate the three phases of our methodology in Section 2. Then we take the attraction region around the Eiffel Tower as our experiment area and examine the accuracy and the reliability of the proposed workflow in Section 3. We conclude our work and discuss some future research directions in Section 4.

LocalRec'18, November 6, 2018, Seattle, WA, USA

ACM ISBN 978-1-4503-6040-1/18/11…$15.00
https://doi.org/10.1145/3282825.3282828.

2 Methodology

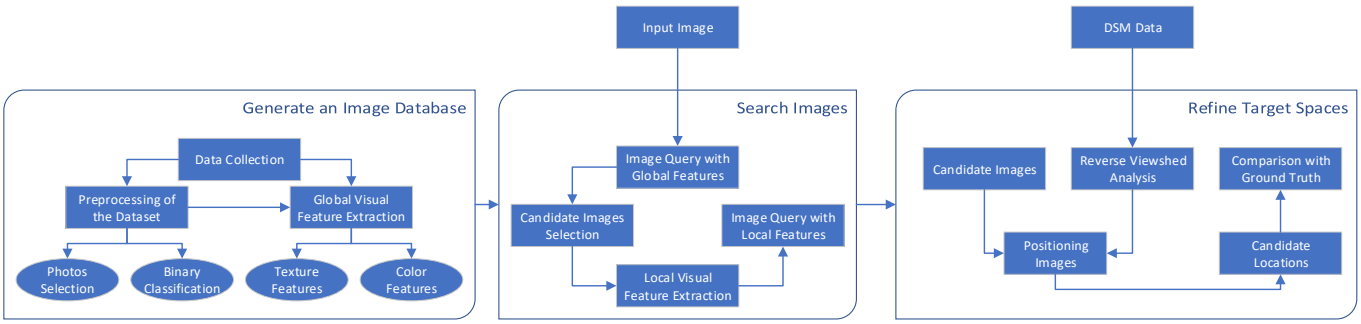

Figure 1. Framework of the Image Geo-Localization System

Under ideal conditions, positioning an image is the process of identifying images that have the same view stored in a large database. However, matching the visual features of images is quite a computationally challenging process. Therefore, considering the tradeoff between the accuracy and the efficiency, the process of image geo-localization could be simplified as "select a small set of candidate images similar to the target image and with geo-location from an image database". There are three phases in our research: (1) preprocessing of the original image data, (2) creation of an image database, and (3) matching the target images and location refinement. Figure 1 shows the framework of our approach.

2.1 Generating an Image Database

2.1.1 Preprocessing of the Dataset. Before the construction of an image database, the preprocessing of image data is employed with regard to the data bias, error of metadata, and so on. The user-contribution threshold strategy was employed in the selection of geotagged photo given a region [4]. Another important question is that, not all photos contain geo-location information. For example, images with only certain objects (like human faces) might not be geo-indictive. This kind of photos might be less valuable for image geo-localization and would cause the redundancy of the image database. A binary classification of "with geo-information" and "without geo-information" is quite important for the image training process. In this research, we utilized the Supported Vector Machine (SVM) for a supervised binary classification to find out photos with and without geo-location information. And a subset of training photos is generated randomly and each photo is manually labeled "with geo-information" or "without geo-information". In order to build this model, key points of each image are extracted using a descriptor of the scale invariant feature transform (SIFT). Using a bag of visual words (BOVW) structure, each image is represented as a 128-dimensional vector. Then a binary classification model is trained using SVM such that all images are labelled with or without geo-information. Before constructing the image database, photos collected from Flickr are tested in this model and those without location information are removed. By utilizing this preprocessing procedure, the efficiency of the proposed methodology could be improved.

2.1.2 Global Visual Feature Extraction. The state-of-art computer vision technology makes it possible to extract visual features and to quantify image contents. In this research, both global and local features are utilized, as the mix usage of those two representations have been proved with a great performance in several image recognition works [6, 8]. The global features including color and texture features that could describe the spatial layout of an image. In comparison, the local features could identify the real objects' structure elements and match them accurately, but with high computational cost. Therefore, mixing both of them could help balance the accuracy and efficiency. In this research, the global-feature based extraction is utilized for indexing and matching for candidate images while the local-feature based extraction will take more details into consideration. Combing these two types of features will ensure the selected photos have the similar photographic composition and contain the same objects in the target photo. As previous studies suggest that color features and texture features are among the most important factors in georeferencing images [5]. We chose the color and edge directivity descriptor (CEDD) that was proposed by [1] and combined all 144 features in total. Specifically, the image is separated into a number of small blocks. Then a 10-bin quantized histogram that represents different colors, is generated following a set of fuzzy rules from the HSV color space. Afterward a 24-bin quantized histogram is represented based on the subdivision of the previous colors. As for the texture feature extraction, several digital filters proposed by MPEG-7 Edge Histogram Descriptors are utilized to describe the edges of various types [12], including vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. Finally, the CEDD will be represented as a total of 144-dimensional vector for an image.

2.2 Searching Images

After developing the image database, two steps are required to select photos that have the most similar scenes compared to a specific input image. First, we extract global features for the input image and search for the most similar k images as candidates. However, images that have similar global visual features (i.e., color and texture features) might be located into different places.

Therefore, we further extract local features of candidate images and that of the input image using the SIFT descriptor and quantize them into the BOVW structure for indexing and searching. A set of candidate images will be selected after scene matching for the target space refinement.

2.2.1 Local Visual Feature Extraction. In this paper, we utilized one of the most popular and mature approaches, David Lowe's scale invariant feature transform (SIFT) [9] for local feature extraction. As mentioned above, the global-feature-based extraction has less computational cost and is good for a sparse resolution, while the local-feature-based extraction is much more suitable for detailed retrieving with a small-scale dataset. Local feature matching could ensure whether or not an input image contains the same object or scene in the database. The SIFT approach describes the edge strength and orientation of local characteristics. A high density of points of interest from an image will be extracted and represented by a 128-dimension vectors derived from 8-orientation histograms within a 4×4 spatial grid. Therefore, each image could be represented by a set of 128 dimensional key points in the local-feature matching process.

2.2.2 Image Matching. High-density points of local features represent the visual content robustness of an image whereas may cause a representation challenge because each point has 128 dimensions and different images have varying count of key points. Here, we use the BOVW approach to summarize and index visual vocabularies of images. This method follows an analogy to represent the image document as word frequencies. Key points are randomly sampled and grouped into C clusters using K-means clustering approach, where C is the number of visual groups or codebooks. Then, each descriptor extracted from the input image could be grouped into the nearest cluster. Finally, we compute the histogram of the codebooks in an image as follows:

$$h = [t_0, t_1, t_2, \dots, t_{C-1}] \quad (1)$$

where t_i is the frequency of the visual word i appears. In order to find out the nearest cluster centroid of the descriptor, the L2 distance between two features is computed:

$$d_{feature}(f_1,f_2)=\sqrt{\sum_{i=1}^{n}(h1_i-h2_i)^2} \quad (2)$$

Following the steps mentioned above, each image in the candidate image dataset could be represented as a C dimensional vectors. After scene matching, each candidate image will be sorted according to the similarity of the local features. In this paper, the top-n nearest candidate images are selected after local-feature-based scene matching. Location information of these images could also be retrieved from the georeferenced image database in order to infer the location of an input image.

2.3 Refine Target Spaces

The above-mentioned steps are purely based on the visual content of images. By comparison, spatial attributes of the metadata, which also has valuable information that can help derive potential locations of the input photo are ignored. A novel method based on reverse viewshed analysis is proposed to refine the target space. We extract candidate locations based on global and local visual features at first. Then, reverse viewshed analysis for a point of interest (POI) (e.g., a landmark) is utilized to determine whether the POI is visible from an area in which the image locates.

Viewshed analysis is a mature method that have been implemented in geographic information systems [7]. A viewshed is an area that is visible from a specific location. It is created from a digital elevation model (DEM) based on an algorithm which could evaluate the elevation between the viewpoint and the target place. Line of sight is created between the two places and only when some locations exist across the two with higher elevation, the line will be blocked and the target place will be determined to be out of the viewshed. Otherwise, it would be included. Therefore, by utilizing the method, if the POI lies outside the visibility of the location of the image, the image will be removed from the candidate georeferenced image set. In this way, the potential space of images will be refined and more accurately represented. In order to utilize reverse viewshed analysis in our research, the Digital Surface Model (DSM), which represents the surface of the earth and the elevation of buildings should be generated. The Shuttle Rader Topography Mission (SRTM) digital elevation model is utilized. And it has a 30-meter spatial resolution and could cover most of the world with an absolute vertical height error less than 16m [10].

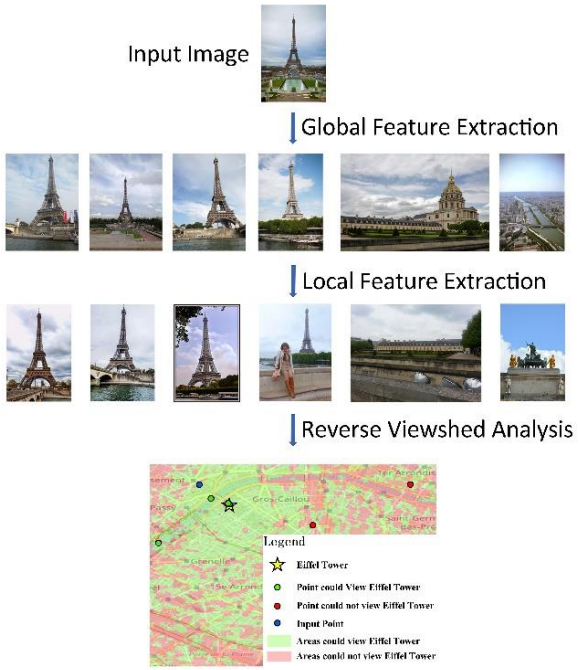

Figure 2. An Example of Workflow

The spatial refinement process consists of three steps. First, a target POI should be selected. By utilizing the reverse viewshed analysis, the visible areas of a target POI would be generated first. And the space would be classified as "visible area" (Figure 2 map, green polygons) and "no visible area" (pink polygons). After that, we analyzed all locations of images to figure out whether they are contained by the visible area. And the image is considered to be remained in the candidate image dataset if the target POI is visible

from this image. Otherwise, the image will be removed. Besides, some content might reflect the target POI accurately, but the coordinates of the image might be geotagged incorrectly. These images will also be considered as invalid images and removed. After removing the images that are not visible to the target POI, the potential locations of the input image will be spatially refined. Finally, we computed the mean/median center of the top *n* candidate images according to their similarity of the local visual features and set it as the predicted location for the input image.

3 Experiments

We took the attraction regions of the Eiffel Tower as an example to evaluate the performance of our framework. Considering the Eiffel Tower is one of the most famous tourist attractions around the world, numerous tourists visit there each year and provide sufficient image data, it is an ideal landmark location for testing. We collected 418,031 Flickr photos (the total size of images is over 100 GB) that are located within 2-km radius of the Eiffel Tower and uploaded by 20,358 users from Jan. 1st, 2012 to Dec. 31st 2017. A training dataset is generated and the distance error (DE) of each image is calculated in order to find out the suitable size of candidate images. And we also make a comparison between the result of our framework and that without reverse viewshed analysis as a baseline approach to test the performance of our framework.

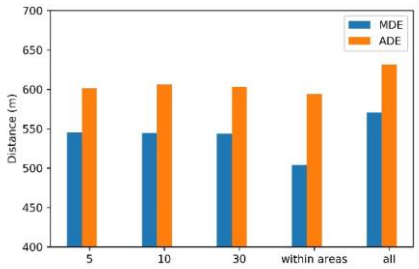

Figure 3. Distance Errors of Different Datasets

As abovementioned, an SVM filter for detecting whether photos have geo-information is utilized. Even though only 145,923 images with geo-information from the original dataset are selected in the further analysis, the study region still has a very high density of images (more than 10, 000 photos per km^2). Many complex objects and detailed sceneries are captured in the image database.

Figure 2 shows an example of the whole workflow for our experiment. First, we extract the global visual features of the input image for searching. To ensure that photos taken in the same place and with similar visual features are enough, we retrieved the top 500 images from the image database as the candidate dataset. It shows that some images have similar visual angles compared with the photo input, while others don't. Then, those candidate images will be utilized for local visual feature matching and top 100 images that have the highest similarity will be returned as the candidate dataset. Later, each photo will be represented on the map and the reverse viewshed analysis is implemented. For example, four left candidate images after local visual feature search locate in different areas that could view the Eiffel Tower, while the right-two false images locate outside the observer area. The reverse viewshed analysis help remove those false images effectively. Afterwards, we defined the following indices called *Average Distance Error (ADE)* and *Median Distance Error (MDE)* for assessing the distance between the predicted locations and the actual location of testing images according to the average and median of the Distance Error respectively:

$$ADE = \frac{1}{n}\sum_{i=1}^{n} de_i \quad (3)$$

We select the top *n* images from the candidate image dataset after the scene matching of local visual feature retrieving. And de_i is the distance between the i_{th} candidate and the ground-truth. The ADE calculates the average distances among all candidate images and the testing photo's location, while MDE calculates the median value of all distances.

As shown in Figure 3, we tested the sensitivity for the choice of top *n* photos. Besides, we also take a comparison between our approaches and the methodology without reverse viewshed analysis. We selected 1, 5, 10, 30, and all candidate images located within the visible area, and all candidate images even outside the visible area respectively. And we calculate the ADE and MDE between the mean/median centers and the ground-truth. It is found using MDE performs better than ADE. When returning top 5, 10 and 30 images, the corresponding ADM are within about 600m while the MDE are slightly less than 550m. Most importantly, the error of all images within the visible area is always smaller than that of without the refinement of reverse viewshed. It proves that our methodology is more effective to certain degree for improving the accuracy of image geo-localization.

4 Conclusions and Discussions

In this paper, we proposed an image geo-localization framework, which utilized a binary SVM classifier, the reverse viewshed analysis, and computer vision techniques. Taking the Eiffel Tower area as an example, our approach works well for estimating the actual coordinates of the input scenery images within about 600m in average. The performance of our approach that takes the spatial context into consideration outperformed the baseline approach without the viewshed information. In future we will further investigate what kind of images could be linked to geo-information from the perspective of cognitive behaviors. Another direction is that how to select the representative location. In this research, we simply use the mean/median center of candidate images. Some other researches once utilized clustering algorithms based on density, or comprehensive ranking system for referring the location. Besides, we only took the Eiffel Tower as an example and the 30-meter DSM for viewshed analysis. We plan to extend the scale of our research workflow and utilize higher-resolution of LiDAR data for improving image geo-localization accuracy.

ACKNOWLEDGMENTS

This work was partially supported by the National Student Innovation and Entrepreneurship Training Project (No. 201810486033). Partial support was also provided by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

REFERENCES

[1] Chatzichristofis, S.A. and Boutalis, Y.S. 2008. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 5008 LNCS, (2008), 312–322.

[2] Crandall, D.J. et al. 2009. Mapping the world's photos. Proceedings of the 18th international conference on World wide web - WWW '09 (2009), 761.

[3] Divecha, M. and Newsam, S. 2016. Large-scale geolocalization of overhead imagery. Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '16. (2016), 1–9.

[4] Gao, S. et al. 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science. 31, 6 (2017), 1245–1271.

[5] Hays, J. and Efros, A.A. 2008. IM2GPS: Estimating geographic information from a single image. 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR. 05, (2008)..

[6] Kennedy, L.S. and Naaman, M. 2008. Generating diverse and representative image search results for landmarks. Proceeding of the 17th international conference on World Wide Web - WWW '08. (2008), 297.

[7] Kim, Y.H. et al. 2004. Exploring multiple viewshed analysis using terrain features and optimisation techniques. Computers and Geosciences. 30, 9–10 (2004), 1019–1032.

[8] Li, J. et al. 2013. GPS estimation for places of interest from social users' uploaded photos. IEEE Transactions on Multimedia. 15, 8 (2013), 2058–2071.

[9] Lowe, D.G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 60, 2 (Nov. 2004), 91–110.

[10] USGS Earth Explorer: Download Free Landsat Imagery: 2015. https://gisgeography.com/usgs-earth-explorer-download-free-landsat-imagery/.

[11] Weyand, T. et al. 2016. Planet - photo geolocation with convolutional neural networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2016), 37–55.

[12] Won, C.S. et al. 2002. Efficient use of MPEG-7 edge histogram descriptor. ETRI Journal. 24, 1 (2002), 23–30.