

A Review on Image-based Geolocation

Jaydeep Ranpariya¹, Roshan Saravanan², Alvin James³, Pranav Abraham⁴ and Dr Mohammed Zabeeulla A N⁵
¹⁻⁵Dept. of Computer Science & Engineering, Jain (Deemed-to-be) University, Ramanagara District, Karnataka - 562112
Email: 19BTRCR010@jainuniversity.ac.in, 19BTRCR0{10, 12, 52, 54}@jainuniversity.ac.in,
z.mohammed@jainuniversity.ac.in

Abstract— This work provides a comprehensive review of image-based geolocation techniques using deep neural networks. The rapid growth of digital images has made image-based geolocation a popular research area in computer vision. The review introduces the concept of image-based geolocation, followed by an overview of deep neural networks and their applications in geolocation. The work then provides a detailed analysis of the existing literature on image-based geolocation techniques, including their advantages, limitations, and challenges. This review also highlights recent advancements in deep neural network-based geolocation techniques and discusses their potential applications in various fields such as urban planning, transportation, and tourism. Finally, the work outlines future research directions that could enhance the accuracy and effectiveness of image-based geolocation using deep neural networks.

Index Terms— Geolocation, Google Street View Images, Image Processing, Deep Learning, EfficientNet, CNN-LSTM.

I. INTRODUCTION

Geolocation, or the process of determining the location of an object or a person, has become a crucial research topic in recent years due to its numerous practical applications, such as navigation, augmented reality, and location-based services. The most popular and widely used method for geolocation depends on image analysis, where the location of the image is estimated depending on its visual features. Geolocation is a predominant and challenging problem in machine/computer vision and geo-informatics due to the inherent complexity of identifying a location from diverse visual features in the image. This makes it difficult for most models to adapt and generalise feature representations that can identify the characteristics of an image that defines a location.

With the increasing availability of geo-tagged images and the swift development of deep learning techniques that have shown great potential for image-based geolocation due to its ability to learn discriminative features from large amounts of data—A growing number of people have been expressing interest in using deep neural networks for image-based geolocation. This has resulted in a wealth of valuable literature in this problem domain. Therefore, this review aims to provide an exhaustive bird's eye view of recent advances in image-based geolocation. We will review several influential research works that have contributed to developing deep learning-based methods and algorithmic approaches for image-based geolocation, ranging from feature matching to 3D modelling and CNN (Convolutional Neural Network)-based methods. We will discuss the key ideas and technical approaches proposed in the literature and their strengths and limitations, highlighting some of the main challenges and future research directions in the problem domain.

In conclusion, this review aims to provide an exhaustive and up-to-date overview of state-of-the-art image-based geolocation. We hope this review will be useful for researchers and practitioners interested in this area and

stimulate further research and development. The insights gained from this work will provide valuable insight for researchers and practitioners interested in developing image-based geolocation methods and outline future directions to pursue in the domain.

II. LITERATURE REVIEW

A. Using Convolutional Neural Networks (CNNs)

The work by Suresh et al. [1] presents a geolocation approach using computer vision and deep learning. The task is framed as a classification problem, where the most probable state among 50 classes is predicted. Previous work used models extensively trained on large, online datasets tailored to specific locations. To address this, the authors created the 50States10K dataset, which contains 0.5 million Google Street View images of the country. The dataset is sampled using uniform sampling and population density-aware sampling. It contains 125000 samples from 50 states, with 2500 samples (10000 unique images) in each state. Each sample contains four rectified, 256x256 Street View images from the sample location, facing the cardinal directions. A deep NN based on the ImageNet architecture [7] is trained, and 4 unique evaluation metrics are used to evaluate the effectiveness of the model.

Standard CNN classification was used as a baseline method to classify each image as a separate training example. To increase the model's accuracy, three variants of the 50-layer ResNet architecture [6] were tested to determine the best way of integrating the information from multiple views of the same location. The Early Integration scheme concatenates all four views to form a twelve-channel input image, sharing information between images in all network layers. The Medium Integration scheme uses a CNN to get a low dimensional summary of each image, concatenates the features of each, and runs a single layer perceptron on the new feature for classification. The Late Integration scheme runs a shared CNN over each input image and then combines the results by taking the maximum over each output class.

The Medium Integration scheme achieves the highest accuracy on the testing data split with a 38.32% accuracy for the top-1 metric. There is a significant improvement in accuracy when considering the best in top-N results, reaching 71.87%. The network performed poorly in regions of low population density, likely due to the distribution of Street View images. The network's performance [1] was also evaluated in head-to-head rounds of GeoGuessr against human subjects, with the medium integration network beating humans in 4 out of 5 rounds.

The future work scope of this paper is to extend the model to perform regression on latitude and longitude, evaluate results with different network architectures, and generalise the network to in-the-wild photographs.

B. Using a Geospatial Coordinate System

The authors introduce a method that tackles the problem of ascertaining a photo's location estimating the camera's position relative to a sizable, geo-registered 3D point cloud [2]. This work proposes two new techniques to tackle the problem of determining the location of a query image in a geo-tagged coordinate system: Improved RANSAC (Random Sample Consensus) is achieved through the utilisation of statistical data available regarding the conjunction of points in three dimensions in images, as well as through establishing bi-directionality of image vectors with 3D points. An evaluation of the method on several large datasets shows competent landmark recognition and the ability to locate the camera pose within metres. These techniques are evaluated on various datasets and cross-checked with recent work on image localisation.

The dataset [2] used for 3D point cloud-based pose estimation includes the Landmarks dataset [10], which consists of images from the original reconstructions. In addition, the San Francisco dataset is utilised, which comprises perspective images obtained by cropping 640x480 resolution images from omnidirectional landscapes. The third dataset is the Quad dataset, which offers improved accuracy for location ground truth in the form of identifiable landmarks or buildings.

The RANSAC scheme chooses points from a subspace of paired 3D points that have a high occurrence probability jointly in new images. The likelihood of a subset is proportional to the empirical concurrence frequency of the 3-dimensional points in the images in the database. The summation over future selections is approximated using the concurrence frequency of the 1st points to avoid an expensive forward search. The proposed [2] bidirectional matching scheme two-step process to estimate the pose of a camera. Firstly, a conventional forward matching scheme is applied to identify a set of primary matches. Then, a sampled subsection of these matches with low distance ratios is selected as preferred matches. Next, an inverse matching procedure is performed, with the above-sampled matches serving as new geo-points to search for in the images. The primary matches are then augmented with the results of this inverse matching, and the final pose estimation is conducted based on the augmented set of matches.

The work also [2] assesses the effectiveness of exploiting concurrent point statistics in a RANSAC scheme. The corresponding results showed that using concurrence to help sampling was beneficial, particularly when the number of iterations was little. Despite the additional computational burden of concurrence/co-occurrence sampling, the total amount of time taken by RANSAC was only marginally affected. The performance was further boosted by bidirectional matching, with the highest performance achieved when both co-occurrence and bidirectional matching were used. The registration precision was estimated by considering the problem where it requires categorising into one of 1000 different landmarks. When using average point descriptors, the classification accuracy was 98.1%, whereas the use of all descriptors resulted in a slightly lower accuracy of 97.9%. The false registration rate was estimated to be 0.68%. The accuracy of estimated camera poses was then evaluated using the Quad dataset, with a registration rate of approx. 68% using average descriptors and around 73% using all available descriptors, with a mean localisation error of 5.5 metres and a median of around 1.6 metres. The method was also tested on the SF (San Francisco) dataset of Chen and checked with their state-of-the-art system for large-scale location identification, with a false registration rate between 4.1% and 5.3%. Finally, the method was evaluated on the Dubrovnik and Rome datasets, achieving a registration rate of perfect 100% in Dubrovnik and around 99% in Rome.

For a medium-resolution image query, the system takes a few seconds, and further untrue registrations can be highly decreased by identifying objects or considering contextual information.

C. Using CNN-LSTM

The work done by Weyand, Tobias, Kostrikov, Ilya, Philbin, and James [3] introduces a CNN designed to ascertain the location of a picture from its pixel data. The literature elaborates that pictures contain cues such as patterns, vegetation, weather, landmarks, or architectural details that can be used to guess the picture's location. The proposed approach to the problem is to subdivide the surface by covering the vast expanse of numerous multi-scale geographic cells and utilising millions of geotagged images to train a deep neural network, the resultant model surpasses prior methodologies, achieving exceptional levels of precision.

The CNN is trained depending on Inception architecture with batch normalisation. The softmax output layer has 1 output for every target cell in the data dividing. The dataset [3] comprises 126 million pictures with geolocations mined from the web, with minimal filtering. The dataset is split into 90+ million training images and 33+ million validation images, evaluable classification into one of 26,263 cells. The model comprises 97,321,048 parameters and is trained until the validation set attains optimal accuracy. Furthermore, measures are taken to eliminate any (near-) duplicate images in the training set from the test sets by filtering out all images from the discarded cells.

The model [3] gives a discrete probability distribution all over the earth, tagging each geographical cell with a probability that the input picture was taken inside it. The network can localise many photos, including landmark buildings, street scenes, and nature scenes like mountains, waterfalls or beaches. It can also allow us to see its uncertainty about a photo, generating a probability distribution that presents several modes representing feasible positions. The features learned by the model are also shown to be capable of being used for getting an image and achieving state-of-the-art (SOTA) yielded on the dataset of INRIA Holidays. Integration of an LSTM architecture [8] with this model facilitates incorporating contextual information to anticipate the positions of uncertain images, enhancing its precision on photo collections by half the margin.

The work [3] was tested on a dataset of more than 2.3 million geotagged Flickr pictures worldwide. The model could localise three per cent of the images at street level and ten per cent at the level of the city. 28% of the pictures were properly localised at the country level and 48% at the continental level. The model could output confidences for possible locations even if an image could not be confidently localised.

The geographic cell discretisation and model size effect on PlaNet's performance is also discussed [3]. Results show that a model with 74 million parameters performs similarly to a full model with 97 million parameters. Nonetheless, the model's efficiency begins to decline when the parameter count is further reduced. Additionally, the study compares the model to two image retrieval baselines that allocate each query photo's location to the nearest matching image from a pool of more than 6 million training images. The outcomes demonstrate that the model surpasses the retrieval approaches at the regional level and beyond; however, it falls short at the street level.

Examinations reveal that the classifier's certainty in accurately identifying the location experiences the greatest decline when the region featuring the distinctive mountain formation, the snow-covered mountain range, or the foreground palm trees is obstructed.

The authors also discuss using LSTMs (long short-term memory) [3] models for sequence geolocation, which can localise images with ambiguous or insufficient information. Outcomes of the LSTM addition on Google Plus

photo albums show that the model can localise images within a country level (750 km) with 68.6% accuracy and within a region level (200 km) with 75.5% accuracy. The LSTM model is trained on a dataset with millions of public photo albums with geotags from Google Plus, split into around 23+ million training albums and 6 million testing albums. The model comprises an Inception part, an LSTM unit [8], and a softmax layer. The model results are compared to three baselines: the single-image model, and the average model. The LSTM model [8] outperforms the other models, as it can capture long-term relationships and assign nearby locations unless there is very strong evidence of a location change. Nevertheless, it is important to acknowledge that the straightforward LSTM model has a drawback in that a considerable number of albums begin with multiple images that offer no valuable visual clues.

D. Using Dominance Set Clustering

Another study discusses a method to determine dominant clusters [4] containing user-defined elements. The authors explain how a parameterised variant of a program can be used to maximise a function and how a weight can be assigned to global attributes based on the area under the normalised score value between the matched reference images and query. Examples of the advantages of having a post-processing step are given, such as the disadvantage of heuristic approaches and the potential for wrong localisations when using KNN-based matching. Lastly, an assessment is conducted on the comparison of global features such as GIST, HSV histogram, and CNN, which are employed in computing the similarity between the query image and the identified corresponding reference images.

The two datasets used to evaluate the proposed algorithm in this work are the WorldCities and the 102 thousand Google Street View Images dataset [4]. The Google Street View Images dataset covers 200+ km of urban streets in Orlando, Pittsburgh, and Manhattan. The WorldCities dataset is a high-resolution dataset of 300 thousand street view images that covers fourteen different cities from all parts of the world. The experiment uses all the images sourced from various locations across the globe to find the best available match. The query images are mined from Flickr, while the reference images are mined from Google Street View.

The authors present a different approach to the demanding problem of geo-localization using image matching in a structured database of images with known GPS [9] coordinates throughout the city [4]. The approach casts the geo-localization as a clustering problem of local image features. The framework builds-level features which enable local matching between images. Local features are extracted from instances of reference images, organised in a k-means tree, and then dynamically collected nearest neighbours depending on how unique the two nearest neighbours are relative to their corresponding query feature. Dominant Set clustering [4] is then used to cluster the featured images, which permits a variable number of nodes in the cluster and exploits the distinct power of each nearest neighbour (NN) feature. This allows for the dynamic selection of nearest neighbours for every query attribute based on its discrimination value. Furthermore, a post-processing step based on a restricted dominant set is implemented to select the most suitable reference image and utilise the location of the strongest match to estimate the position of the query image.

The approach proposed is multiple times faster than existing approaches [4], hence allowing for multiple different clusters (at different locations) to be obtained. A quantitative collation of the proposed approach with other novel methods shows that the suggested method can make a 10% improvement over the other methods without post-processing.

E. Using Complex 3D Representations and Bayesian Classification

Other work done in the problem domain discusses the challenges of auto geo-localization of the provided videos and images and how the current state of computer vision technology is such that it is only effective in areas with high visitor footfall and a substantial volume of geo-tagged images [5]. The literature discusses the challenges and opportunities of data-driven geo-localization methods, like ambiguity (uncertainty reduction) and similarity of visual attributes, undesirable photography effects, and lack of unified reference data. The primary challenges and opportunities discussed are broadening the scope of localisation by leveraging cross-view imagery and large-scale collections of user-generated images, while also focusing on identifying geographically significant features.

The three main approaches proposed by Amir R. et al. [5] are data-driven, semantic, and geometric matching-based. Data-driven approaches rely on extracting features from large-scale data of geo-tagged images and learning them to be invariant concerning unwanted effects, such as viewpoint or lighting changes. Semantic geo-localization refers to the process of determining the location of an image based on semantic and high-level visual cues. Such as human-related characteristics, natural properties, and textual tags. Geometric matching-based methods compare the query image with a highly-detailed scene model, which leads to improved geolocation

accuracy. Challenges of each approach are discussed, such as constructing high-quality 3D textured coloured scene models and performing 2D to 3D matching as opposed to them and geometric alignment of images to untextured (non-coloured) models.

The availability of unprecedented data resources, new applications, and novel techniques largely drives the upcoming trends in the geospatial localisation field. It outlines [5] new geo-referenced data resources, such as web-scale databases, which can be used for geo-localization. The authors state that the geo-location of images can be used to understand their content and how human-in-the-loop localisation frameworks can effectively use multimodal cues to achieve practical and accurate geo-localization.

The work [5] also considers the upliftments in the Spatial and temporal precision of satellite imagery, which have reduced the Ground Sampling Distance to less than 0.35 metres and temporal resolution of less than one day. This shift has opened up new opportunities for temporal geo-localization, largely ignored in existing frameworks. The authors also contemplate the potential of deep learning-based methods to improve the geo-localisation field substantially.

According to work done [5], there are techniques to effectively use large-scale datasets on the web for geo-localization and extracting geographic features. It introduces a visual data mining method that uses weak supervision to identify connections between frequently occurring visual elements in historical images that are both spatially and temporally related, to capture their underlying visual style. Moreover, a cross-view feature translation technique is presented, which significantly expands the scope of methods for geo-locating to cover significant land areas of the world where ground-level reference photos are unavailable. Furthermore, another technique described goes well with street images and top-down air-view images. The final proposal involves a method enabling semantic labelling of datasets and query views through a Bayesian classification of the street view queries and the descriptors applicable to the airborne database. The authors propose a method [5] that combines 2D-3D alignment with a novel descriptor for non-photographic depictions. The descriptor is based on the local curvature of the image and is used to exactly match the queried image to a 3D model of the view. The effectiveness of the proposed technique is demonstrated through its evaluation of multiple datasets, highlighting its ability to precisely geo-locate non-photographic representations with an accuracy as fine as 100 m².

The authors also present a GPS-aided image-based quick identification method [5] of urban landmarks on a GP S [9] enabled mobile device. Using a compact and compressed representation of complex three-dimensional scenes and demonstrating that the suggested method can automatically recognize the accurate architectural location and restore a rough perspective of historical paintings and photographs about the 3D layout of the site. This is accomplished by finding correlations between the information in the photos and the information scraped from the Internet. The scattered sources of information are combined, allowing for efficient and accurate queries with any information the user recalls regarding a picture.

The main challenges and opportunities that are discussed regarding data-driven geo-localization methods include ambiguity (reducing uncertainty) and similarity of visual attributes, undesirable photography effects, and lack of unified reference data [5]. Solutions and directions for future research are provided for each challenge. The primary challenges and opportunities discussed are broadening the scope of localisation by leveraging cross-view imagery and large collections of user-generated images and identifying geographically informative features.

III. CONCLUSION

To sum up, the reviewed papers put forth a range of strategies for tackling the complex task of geo-localization through image matching. The discussed techniques employ various approaches such as CNNs [1, 3], RANSAC [2], bidirectional matching [2], and dominant set clustering [4], among others, to improve the accuracy of landmark recognition and camera poses estimation. The results demonstrate the effectiveness of these methods in achieving competent landmark recognition. Camera poses estimation and location recognition with high accuracy. These methods outperform previous approaches and attain state-of-the-art results in image retrieval, even when visual information is ambiguous or insufficient. We observe that there is no optimal way to solve the geo-localization problem, as each approach has advantages and limitations. However, some of the most promising approaches presented in the review paper include

- The use of large geo-registered three-dimensional point clouds. And, bidirectional matching of image vectors with three-dimensional points to improve landmark recognition and camera pose estimation [2].
- Partitioning the earth's surface into geographic subdivisions, followed by the training of a CNN-LSTM model using millions of geotagged images. This method has demonstrated superior performance in image retrieval and location prediction [3].

- Building-level feature extraction, k-means tree, dynamic nearest neighbours collection, and dominant set clustering to achieve accurate geo-localization in an organised database of reference images throughout the city [4].

Future research should focus on combining these approaches and exploring new techniques to improve the accuracy and effectiveness of image-based geolocation. These papers contribute to developing new and effective geo-localization techniques with important applications in various fields, including navigation, tourism, and environmental monitoring.

REFERENCES

- [1] Suresh, Sudharshan & Chodosh, Nathaniel & Abello, Montiel. "DeepGeo: Photo Localization with Deep Neural Network", 2018
- [2] Yunpeng Li, Noah Snavely, Daniel P. Huttenlocher, Pascal Fua, "Worldwide Pose Estimation Using 3D Point Clouds", Large-Scale Visual Geo-Localization (pp.147-163), 2016
- [3] Weyand, Tobias & Kostrikov, Ilya & Philbin, James, "PlaNet - Photo Geolocation with Convolutional Neural Networks", 2016
- [4] Eyasu Zemene, Yonatan Tariku, Haroon Idrees, Andrea Prati, Marcello Pelillo and Mubarak Shah, "Large-scale Image Geo-Localization Using Dominant Sets", 2017
- [5] Amir R. Zamir, Asaad Hakeem, Luc Van Gool, Mubarak Shah & Richard Szeliski, "Large-Scale Visual Geo-Localization", 2016
- [6] He, K., Zhang, X., Ren, S., & Sun, J., "Deep Residual Learning for Image Recognition", 2015
- [7] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", 2010
- [8] Sepp Hochreiter, Jurgen Schmidhuber "LONG SHORT-TERM MEMORY", 1997
- [9] Abulude, Francis & Akinnusotu, Akinyinka & Adeyemi, Adewale. "GLOBAL POSITIONING SYSTEM AND IT'S WIDE APPLICATIONS", 2015
- [10] Weyand, T., Araujo, A., Cao, B., & Sim, J. "Google Landmarks Dataset v2", 2020