LANDMARK IMAGE DISCOVERY USING NETWORK CLUSTERING

ALA'A AHMED MOHAMMED AL-ZOU'BI

UNIVERSITI SAINS MALAYSIA

2022

LANDMARK IMAGE DISCOVERY USING NETWORK CLUSTERING

by

ALA'A AHMED MOHAMMED AL-ZOU'BI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosphy

March 2022

ACKNOWLEDGEMENT

First, and foremost, I thank Allah (SWT) for his blessings, generosity and for giving me strength during this long journey of this thesis.

I take this opportunity to express my heartfelt gratitude to my supervisor Dr Keng Hoon Gan, for her kindness, support, encouragement and guidance during this research. I could not have imagined having a better advisor and mentor for my study.

My thanks and appreciation are also extended to my brothers and sisters for their encouragement and support. Very special thanks and gratitude to my dear sister Om Rami and my brother in law Dr Musa . No words can express my appreciation for what they have presented for me.

Also, I would like to express my sincere and heartfelt thanks to my dear friends inside and outside Malaysia who supported me during my study.

Last but not least, this work is dedicated to my late parents.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

PENCARIAN IMEJ MERCU TANDA MENGGUNAKAN PENGGUGUSAN RANGKAIAN

ABSTRAK

Sejumlah besar koleksi gambar disimpan dalam internet secara atas talian dan koleksi ini terus berkembang dengan pesat. Kekayaan dan ketersediaan maklumat visual ini mengembangkan aplikasi visi komputer. Oleh yang demikian, teknik yang cekap diperlukan untuk menyusun dan mengatur sebilangan besar imej-imej ini. Khususnya, gambar-gambar mercu tanda membentuk sebahagian besar daripada koleksi tersebut. Tindakan melombong imej mercu tanda bergantung kepada pengumpulan koleksi imej berskala besar kepada penggugusan objek yang mereka gambarkan. Pengelompokan ini merupakan tugas yang sangat mencabar kerana variasi dalam rupa bentuk imej tersebut yang disebabkan oleh pencahayaan, perbezaan skala dan sudut pandang pengimejan. Tambahan lagi, faktor kerumitan yang tinggi dalam proses pemadanan imej dan dimensi tinggi data imej. Beberapa algoritma pengelompokan telah digunakan dalam sorotan kajian untuk pengelompokan imej mercu tanda seperti pengelompokan spektrum; Peralihan min, pencincangan min dengan pengembangan pertanyaan. Kebanyakan algoritma ini bergantung pada kaedah berasaskan RANSAC untuk operasi pengesahan geometri yang digunakan untuk pemadanan imej. Walau bagaimanapun, kaedah pengesahan ini mempunyai overhed perkomputan yang tinggi, sekaligus menjadikan masa jalanan juga tinggi. Oleh itu, objektif utama kajian ini adalah untuk membangunkan pendekatan pengelompokan mercu tanda yang berkesan yang mampu mengenal pasti imej mercu tanda dalam koleksi imej internet berskala besar dengan kerumitan perkomputan yang dikurangkan sepanjang proses pengelompokan. Bagi mencapai objektif ini, tesis ini mengemukakan tiga sumbangan utama.

Pertama, kami mengemukakan pelaksanaan algoritma penggugusan berstruktur untuk rangkaian KNN-SCAN. Dalam algoritma ini, imej-imej dikumpulkan dalam kelompok berdasarkan imej biasa dalam senarai K-Nearest Neighbours mereka. Seterusnya, kami mengemukakan kaedah pengesanan geometri berasaskan Hough Transform untuk pemadanan imej bagi memperbaiki kualiti senarai KNN dan meningkatkan ketepatan penggugusan. Akhirnya, kami mengemukakan sebuah teknik untuk mengurangkan jumlah pengiraan kesamaan yang digunakan dalam proses penggugusan. Kami menilai prestasi pendekatan pengelompokan yang dicadangkan menggunakan semua kaedah yang dicadangkan pada set data imej penanda aras dan mendapati bahawa ia memberikan peningkatan yang ketara. Khususnya, pengelompokan imej mercu tanda menggunakan kaedah KNN-SCAN mempunyai peningkatan 8.6% berbanding kaedah sebelumnya yang diakui terbaik pada set data yang sama dari segi nilai mean average precision (mAP). Ia juga mempunyai hasil yang setanding dari segi perolehan jika dibandingkan dengan kaedah terkini. Di samping itu, kaedah pengesahan geometri yang dicadangkan mempunyai peningkatan yang ketara dari segi keberkesanan perolehan dengan nilai mAP masing-masing, 0.48 dan 0.43 pada set data Oxford dan Paris. Manakala nilai mAP bagi kaedah berasaskan RANSAC yang berprestasi terbaik mempunyai nilai 0.35 dan 0.27 pada set data yang sama. Kaedah pengesahan geometri mempunyai peningkatan prestasi dari segi masa pengesahan juga. Akhir sekali, pengurangan kesamaan yang dicadangkan berada pada purata 68% daripada pengiraan kesamaan yang diperlukan dalam proses pengelompokan.

LANDMARK IMAGE DISCOVERY USING NETWORK CLUSTERING

ABSTRACT

Significant amounts of Internet photo collections are stored online and continue to grow rapidly. This wealth and availability of visual information enable the development of several computer vision applications. Therefore, there is a need for efficient techniques for structuring and organizing this large number of images. In particular, landmark images form a large portion of such collections. Mining of landmark images relies on clustering to group large-scale image collections by the object they depict. The grouping process is a very challenging task due to the variations in the object's appearance, which can be caused by illumination conditions, differences in scale and imaging viewpoint. In addition, the high complexity of the image matching processes and the high dimensionality of image data. Several clustering algorithms have been used in the literature for landmark image clustering such as spectral clustering, Mean shift, min-Hashing with query expansion. Most of these algorithms depend on RANSAC-based methods for geometric verification operations which are used for image matching. However, these verification methods have a high computational overhead, and hence, a high run-time. Therefore, the main objective of this study is to develop an efficient landmark clustering approach capable of recognizing landmarks in large-scale Internet image collections with reduced computational complexity throughout the clustering processes. To achieve this objective, this thesis presents three main contributions. First, we present our implementation of the structural clustering algorithm for networks, namely, KNN-SCAN. In this algorithm, images are grouped into clusters based on the common images in their K-Nearest Neighbors lists. Second, we present a Hough Transform-based geometric verification method for image matching

to refine the quality of KNN lists and improve the accuracy of the clustering. Finally, we present a technique to reduce the number of similarity calculations used in the clustering process. We evaluated the performance of the proposed clustering approach using all the proposed methods on benchmark image datasets and found that it gave significant improvements. Specifically, the landmark image clustering using KNN-SCAN has 8.6% improvement over the best previous methods on the same dataset in terms of mean average precision (mAP). It also has comparable results in terms of recall in comparison with state-of-art methods. In addition, the proposed geometric verification method has a significant improvement in terms of retrieval efficiency with mAP of 0.48 and 0.43 on the Oxford and Paris datasets, respectively. While the mAP value of the best performing RANSAC-based method has 0.35 and 0.27 on the same datasets. The geometric verification method has a performance boost of verification time as well. Finally, the proposed similarity reduction average of 68% of the similarity calculations required in the clustering process.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Over recent years, the Internet has become a repository for massive amounts of visual information. The widespread use of digital cameras, as well as smartphones, enable taking photos everywhere and at any time. As the current social network technologies together with the ubiquitous availability of the Internet allow people to capture and share their life and memories with friends, large databases of visual data have been created, most notably, photos of famous buildings and tourists' attractions. Examples of these image collections include Flickr, Instagram, Facebook, etc. In addition to the visual content, users usually tag their images with some textual description and often add the time and approximate location at which the photo was taken. Images of landmark buildings, in particular, contribute to many of these image collections as people often take images of well-known places and historic buildings (see figure 1.1).

Due to their recognition value, these large image collections and the variety of information linked to them have encouraged the Computer Vision community to build interesting applications in many different domains, such as landmark image recognition. In this type of applications, the user issues a query photo of a specific object (e.g., a picture with his/her smartphone) then the recognition system returns more details about the queried object. In addition, proper groupings of these collections can be used in applications for building 3D models of landmark buildings that take as input multiple views of the same object. Other applications also include image auto anno-

Figure 1.1: Examples of well-known landmarks from around the world

tation, image localization, and tour guides and travel recommendation systems that provide a sequence of places to visit with a set of representative images with diverse viewpoints.

However, these applications need to automatically discover objects in large and unstructured collections of photos before their main function. The task of mining and structuring landmark image collections has been addressed as a *clustering* problem in recent research as in (Philbin & Zisserman, 2008; Weyand, 2016; Q. Zhang & Qiu, 2015). The objective of this clustering is to structure landmark image collection into groups in a way that images within the same cluster depict the same scene or the same object. This automatic organization is an important step towards a higher-level understanding of these collections.

In this research, we focus on how to develop a visual landmark clustering approach capable of automatically discovering photographed landmark buildings in large and

unstructured collections of photos. Note that we do not consider classes of objects, such as "buildings" in general, our goal is to discover objects instances, such as "Eiffel Tower" or the "Louvre".

The need for visual mining is increasing. A significant number of images are taken every day using digital cameras, smartphones, tablets and are uploaded to social networks and image sharing websites. Therefore, there is a need for efficient techniques to organize and structure this large number of images so that further vision and multimedia applications can be developed.

1.2 Problem Statement

As mentioned above, the process of managing such large image collections is a clustering problem, which transforms a set of a large unorganized image into coherent and similar groups, that is, clusters, to facilitate other tasks and applications.

Visual content mining is a complex domain with different challenges that are sometimes specific to images. To perform clustering, a measure of similarity is needed between images. These functions are used in deciding to partition the image into an appropriate cluster based on the similarity value, these decisions affect the performance of the clustering algorithm. Accordingly, the effectiveness of the image clustering algorithm relies on the similarity functions of the clustering algorithm.

Over the past years, the most popular method for image matching is performing basic local features matching and a subsequent geometric verification process on the feature correspondences. The number of inliers produced in the verification method is

usually used as the similarity value. This matching process has been used in different stages of the clustering pipeline starting from building matching graphs (e.g. (Philbin, Sivic, & Zisserman, 2011; Philbin & Zisserman, 2008)) and ending with growing and merging the clusters (e.g. (Chum & Matas, 2010; Weyand & Leibe, 2011; Q. Zhang & Qiu, 2015)). However, this querying and verifying is considered to have a very high run-time and high computational overhead and become expensive as the number of images grows very large. Almost all the landmark image clustering approaches depend on RAndom SAmple Consensus or RANSAC (Fischler & Bolles, 1981) to perform the geometric verification process. This algorithm entails high computational overhead since it runs in a hypothesis and test process. This means that the steps of this process are repeated for a number of time to tests several mathematical models in order to specify the best hypothesis. Consequently, this modeling and testing have high run time computations especially when the ratio of the true correspondences low. The need for such high computational functions for image matching emerges from two main factors related to the nature of image representations. These factors are explained as follows:

High dimensionality. High-dimensional data means that the data is represented by a large number of attributes (i.e., data is represented by a sparse vector). As a consequence, the similarity value of the conventional similarity functions loses its significance to differentiate similar and dissimilar objects (Donoho et al., 2000).

In the context of image matching, the basic image representation with local features entails several hundred to few thousands of local features (depending on the keypoints detector) and each keypoint is represented by a high dimensional descriptor. For ex-

ample, the well-known SIFT descriptor (Lowe, 2004) encodes each keypoint in a 128-dimensional vector. Although this high-dimensionality is alleviated using quantized to a vocabulary of visual words (Sivic & Zisserman, 2003), the final image representation still composed of high dimensional vector (depending on the size of the visual vocabulary).

Variations in the appearance. An object mining system has to be able to correctly match images of the queried object despite large changes in the object's appearance. The differences in the object's appearance can be caused by: (1) variations of the *illumination* conditions when the photo is captured such as the time of day that an image was taken, different weather conditions, or different camera settings; (2) Differences in *scale and viewpoint* of the photographed object based on how far and from what angle the image was taken; (3) Partial *occlusions* which are also quite common as people often pose in front of a landmark or scenic landscapes. An example of such variations in depicted in Figure 1.2. These variations might cause problems for the clustering process to distinguish between images since the feature extraction algorithm might represent similar images very differently due to these appearance variations between the images. While most interest point detectors and descriptor offer some level of invariance w.r.t. viewpoints and lighting variations, matching can fail to correctly match the images since the appearance variation would cause their amount of shared content (visual features) to be miscalculated.

Figure 1.2: Appearance changes

1.3 Research Objectives

The overall aim of this study is to develop an efficient visual landmark clustering algorithm capable of grouping landmarks in Internet photo collections. The objective of the algorithm to develop and improve a clustering approach that can overcome the challenges of image clustering and can improve the efficiency of the clustering performance.

This aim can be achieved via the following objectives:

- To propose a fast and efficient geometric verification method for successfully match-

ing images. This method aims to generate accurate image structures to be used in the clustering algorithm. Also, to filter the image dataset from noisy and unrelated images, and hence, to exclude such images from being involved in the image representations or even in the clustering process.

- To propose a landmark image clustering approach that automatically groups images containing the same object (landmark) in Internet landmark photo collections. The purpose of this approach is to group a set of images in clusters where each cluster contain the same object despite the variations in their appearances, differences in the imaging viewpoints, and scales. In addition, this clustering approach aims to be *less dependent* on the high-computational image matching and the subsequent geometric verification processes.

- To propose a technique that reduces the number of structural similarity calculations used in the clustering process. The aim is to alleviate the computational complexity involved in the clustering process.

1.4 Contributions

The major contributions of this research are described as follows:

- We introduce a geometric verification method for image matching. This method uses the Hough Transform voting strategy for identifying and maintaining the correct correspondences between a pair of images. The proposed verification method consists of two stages; First, we detect and remove obvious outliers and increase the inlier ratio of initial candidate matches. Second, we perform a local consistency verification for each correspondence to refine the final matches.

- We present our adaptation for a clustering algorithm for landmark image clustering, namely, KNN-SCAN. This implementation is based on SCAN which is a structural clustering algorithm for networks. We use the k-nearest neighbour (KNN) graph to generate the image structures which are required for the clustering process. This algorithm casts the task of landmark clustering as a community search problem. In this algorithm, vertices (i.e. images) are grouped into clusters by how they share neighbours in their structures. The algorithm finds clusters, hubs, and outliers by using the structure and the connectivity of the vertices as clustering criteria. Accordingly, images sharing common objects in their KNN lists (i.e. image structures) will be then grouped to the same cluster.

- We present a technique to reduce the number of structural similarity calculations between the structures (KNN lists) of the images. This method depends on the nature of the KNN lists created from the previous step, where several KNN lists are very similar to each other due to the re-ranking process. Taking advantage of this property, we propose to share the status of the vertex with its highly similar unprocessed neighbourhood vertices. This helps the vertices to avoid unnecessary images comparisons.

- We present a hybridization between the proposed similarity reduction method with a technique that has the most reduction ratio in the literature. This Hybridization is used to ensure the maximum possible reduction of the similarity calculations.

1.5 Structure of the Thesis

The rest of this thesis is organized as follows:

- In Chapter 2, we present a review of the literature that is most relevant to this work. First, we present the related work about the landmark image clustering approaches. Then, we present the geometric verification techniques for image matching. In addition, we review some works abou the reduction of similarity calculations of the clustering algorithm used in this thesis. Finally, we discuss and analyse the gap of these approaches.

- In Chapter 3, we present an overview of the basic concepts and techniques that this work is based on. First, we present an overview of image matching using local features and geometric verification methods using RANSAC and Hough Transform. Second, we review the image retrieval pipeline using the well-known bag of visual words framework with its methods and techniques. Finally, we present an overview of the clustering algorithm used in this thesis, namely, SCAN. We begin with a brief overview of the graph representations and graph clustering methods. Then, we overview the basic definitions and the clustering procedure of SCAN.

- In Chapter 4, we present the proposed methodology for landmark image clustering. First, we present the proposed geometric verification method for image matching. This method is based on the Hough Transform voting technique for refining correspondences between images. We demonstrate the two-phase procedure of this method. In the first phase, we show how to remove apparent outlier correspondences by using their location differences. In the second phase, we explain how we perform a local verification process for more refinement. Second, we present the implementation of KNN-SCAN for landmark image clustering in large unstructured image collections. We describe the method for generating image structures using the image matching and geometric verification in the previous step. Then, we illustrate how to

perform the clustering procedure on the generated image structure. After that, we present our technique for merging the clusters with different viewpoints. Finally, we present the proposed technique for reducing the structural similarity computations in KNN-SCAN. We begin with the basic definitions for applying the reduction technique. Then, we describe the steps to avoid a large amount of these calculations. This chapter also presents the hybrid method that merges the proposed reduction property with a method that has the most reduction ratio to ensure maximum efficiency.

- In Chapter 5, we present the setup details of the experiments and the image datasets. We explain the evaluation measures of the clustering efficiency. We also study the effects of using different parameter values on the clustering performance. In addition, this chapter presents discussions about some implementation details and analysis of some limitations.

- Finally, in Chapter 6 we present the conclusions.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we present a review for the most related methods to our work. Section 2.1 presents the previous work in landmark image clustering. Section 2.2 presents a review of geometric verification methods using Hough transform. Section 2.3 presents the works for reducing the number of similarity calculations. Finally, section 2.4 discusses and analyses the research in these approaches.

2.1 Landmark Image Clustering

Several clustering algorithms and approaches have been implemented for the discovery of landmarks images in large databases. These groupings are used in different tasks and applications in the Computer Vision literature, such as scene summarization, tour guides, locating places of interest and many other applications. In the same context, different types of modalities are used to improve efficiency and reduce the computational overhead in the clustering process.

Some research depends mainly on the visual features to perform the image similarity and the *visual clustering* operation. Some research performs clustering using the GPS data (known as geo-tags) provided with the community contributed image. This approach of clustering is known as geo-clustering. This is done to reduce the computational cost of visual clustering especially when the image dataset is very large. (millions). Some other papers use more modalities, the goal is to find the optimal weight

among groups within different views by maximizing the clustering quality within view and the clustering consistency across views. Next, we briefly overview these methods along with the modalities used in them.

Philbin and Zisserman (2008) generated a sparse matching graph over the entire image dataset then they applied spectral clustering (Ng, Jordan, & Weiss, 2001) to over-segment the connected components in this graph. However, this process has high computation complexity. It also requires to pre-determine the number of clusters. Chum and Matas (2010) and Chum, Perd'och, and Matas (2009) used min-hash and geometric min-hash, respectively, to find pairs of highly similar images in large image collections to be used as seeds for growing the cluster. The seeds are formed via image collisions resulted from similar minimum hashing values. Then, these seeds are used as queries to build clusters by recursive query expansion (Chum, Philbin, Sivic, Isard, & Zisserman, 2007). However, this process makes their clustering prone to contain images of other landmarks. Weyand and Leibe (2011) cast the landmark discovery as a mode search problem. Their Iconoid Shift algorithm performs mode search to find popular views of the object using Medoid Shift (Sheikh, Khan, & Kanade, 2007) which is a variant of Mean Shift (Comaniciu & Meer, 2002) to search for iconic views which they call iconoids. The mode search is performed in local matching graphs and they use homography overlap distance to find images that have locally maximal mutual overlap with their neighbouring images. This algorithm was extended as a hierarchical clustering algorithm called Hierarchical Iconoid Shift (Weyand & Leibe, 2013) to mine and structure the architectural details in landmarks. Papadopoulos, Zigkolis, Kompatsiaris, and Vakali (2011) first create two similarity graphs, one graph is created using the visual features and the second one using textual tags. These graphs are merged into

a hybrid image graph comprising the union of their nodes and the union of their edges. After that, they perform graph-based image clustering using the SCAN (Xu, Yuruk, Feng, & Schweiger, 2007). Q. Zhang and Qiu (2015) use a modified tree partitioning min-Hash to build a sparse affinity matrix. Then, they generate a set of dense sub-clusters using the neighbours of each image. Image discovery is then accomplished by growing and merging these sub-clusters through exploring the image overlap between these clusters. Geometric verification is applied when the similarity is less than a threshold. Philbin et al. (2011) present the geometric Latent Dirichlet Allocation (gLDA) for landmark image mining. LDA (Blei, Ng, Jordan, & Lafferty, 2003) is a method for semantic clustering in the statistical text community. gLDA augments the position and shape of the visual words and introduces a geometric transformation between topics and images. This approach is applied on a matching graph built by standard image retrieval and geometric verification using LO-RANSAC (Chum, Matas, & Kittler, 2003) is used for re-ranking the top-ranked images. For image scene summarization, Qian et al. (2015) use a graph-growth-based approach to group images with various viewpoints into a cluster (or viewpoint album). Starting from the two most similar images, the graph is grown by adding other images if they satisfy a minimum pre-defined similarity threshold.

The hierarchical agglomerative clustering (Webb, 2003) has been applied in several papers. Quack, Leibe, and Van Gool (2008) organize the geo-tagged images in tile-shaped sub-regions according to their locations, then created a matching graph using feature matching followed by geometric verification using RANSAC. Then, hierarchical agglomerative clustering is applied to each tile. The resulting clusters are analyzed and are classified into objects and events. For scene summarization, Johns and Yang

(2011) create an image similarity matrix is by using pair-wise matching and geometric verification, then they implement hierarchical agglomerative clustering on pairs of matching images to form a set of scenes. Zheng et al. (2009) follow a two-layer clustering scheme that applies both geo-clustering and visual clustering. Geo-clustering is performed using agglomerative hierarchical clustering on the GPS coordinates. Then they create a match graph region by matching images using their local features. This graph is then used for visual clustering is using hierarchical agglomerative clustering. Finally, the textual tags for each visual cluster are processed to generate the appropriate annotation.

Some other methods use a two-level clustering. Bui and Park (2017) use random walk (Harel & Koren, 2001) and constrained clustering. In the random walk clustering step, a large-scale collection of geo-tagged photos is separated into a number of geo-clusters. In the constrained clustering step, they continue to divide the clusters that include many places of interest (POI) into many sub-clusters, where the geo-tagged photos in a sub-cluster associate with a particular POI. Avrithis, Kalantidis, Tolias, and Spyrou (2010) first perform geo-clustering on geo-tagged images using the Kernel Vector Quantization (KVQ) (Tipping & Schölkopf, 2001). They also use KVQ to perform visual clustering in each of the produced geo-clusters. Visual clustering performs several matching and geometric verification to select a minimal subset of images such that each image in the original dataset has at least a certain minimum number of inliers with at least one photo in this subset. However, this method clusters all images including the noise images, and the clusters may represent unuseful objects. The same clustering algorithm is used by Spyrou and Mylonas (2016) to cluster geo-tagged images of a particular geographical area to divide large areas into smaller groups. Then they

further identify meaningful places of interest (POI) by analyzing any textual metadata available to generate appropriate tags.

Some methods depend on using the low-dimensional global features for representing images. Li, Wu, Zach, Lazebnik, and Frahm (2008) use low-dimensional global "gist" descriptors (Oliva & Torralba, 2001) for representing images. K-means clustering is then applied for generating a preliminary grouping of images. In each cluster, geometric verification using SIFT features is performed in subsequent steps to ensure the image similarity and creating the iconic scene graph. For the same task, Kennedy and Naaman (2008) concatenate two global features for representing images, namely, grid colour moment features (Stricker & Orengo, 1995) to represent the spatial colour distributions in the images and Gabor textures (Manjunath & Ma, 1996) to represent the texture. Then, k-means is used for clustering on a single feature vector.

Mean shift clustering algorithm (Comaniciu & Meer, 2002) is a very popular clustering method in geo-clustering approaches. This algorithm has been used to determine important locations (frequently photographed places) in geographical clusters. These geo-clusters are then further analysed to be used in different tasks. Crandall, Backstrom, Huttenlocher, and Kleinberg (2009) use mean shift clustering on the GPS data to find places of interest. For estimating the location of the image, they trained a classifier on selected images using the visual features and textual tags. They build the full matching graph and segment it using spectral clustering (Ng et al., 2001) to extract the representative images for particular landmarks. Jiang, Qian, Mei, and Fu (2016) presents a personalized travel sequence recommendation from both travelogues and community-contributed photos and the heterogeneous metadata (e.g., tags, geo-

location, and date) associated with these photos. J. Zhang, Wang, and Huang (2017) process the textual data in each geo-cluster for automatic image tagging. While Sang, Fang, and Xu (2017) analyses the textual tags to find appropriate representative places of interest (POI) and themes in a geo-cluster. Recently, Qian et al. (2021) also use the mean shift to discover candidate places of interest. They present a clustering algorithm called LAST that fuses several modalities, namely, location, appearance, semantic, and temporal information. A graph is constructed for each modality, then they apply K-means or spectral clustering for summarizing POIs.

The most recent paper that uses the visual clustering is (Q. Zhang & Qiu, 2015). Several recent landmark clustering papers use geotags meta data and other modalities to reduce the computational effort of building the matching graph and performing the clustering by pre-grouping images based on their location. However, this step limits the applicability of these approaches to images where geotags are available. Our approach continues the work of visual clustering methods with the objective to reduce the computational overhead by creating limited-sized image representations, making geographic pre-clustering unnecessary. Table 2.1 presents a summary of these methods.

Table 2.1: Summary of the landmark image clustering methods related to our work.

Philbin and Zisserman (2008)	generate a matching graph then they apply spectral clustering to over-segment the connected components.
Chum and Matas (2010) and Chum et al. (2009)	use min-hash, to find cluster seeds for growing the cluster by recursive query expansion.
Weyand and Leibe (2011)	Their Iconoid Shift algorithm search in local matching graphs for iconic views of the object using Medoid Shift.
(Weyand & Leibe, 2013)	Hierarchical Iconoid Shift to structure the architectural details in landmarks.
Papadopoulos et al. (2011)	create hybrid graphs, using the visual features and textual tags. they perform graph-based image clustering using the SCAN.
Q. Zhang and Qiu (2016)	generate dense subclusters using tree partitioning min-Hash. grow clusters through connectivity among these clusters.
Philbin et al. (2011)	Create matching graph and apply geometric Latent Dirichlet Allocation (gLDA).
Qian et al. (2015)	use a graph-growth-based approach, images are grouped together if they satisfy a minimum similarity threshold.
Quack, Leibe, and Van Gool (2008)	apply hierarchical agglomerative clustering to matching graphs of tile shaped geographic sub-regions.
Johns and Yang (2011)	they implement hierarchical agglomerative clustering on matching graph of images to form a set of scenes.
Zheng et al. (2009)	generate geo-clusters using agglomerative hierarchical clustering. The same algorithm is used for visual clustering.
Bui and Park (2016)	apply random walk clustering to generate geo-clusters. Then apply constrained clustering to discover places of interest.
Avrithis, et al. (2010)	first perform geo-clustering on geo-tagged images using (KVQ). They also use KVQ to perform visual clustering.
Spyrou and Mylonas (2016)	cluster geo- images of a geographical area Then they identify meaningful places of interest by analyzing textual metadata.
Li et al. (2008)	Apply K-means on global descriptors followed by geometric verification on SIFT features to generate iconic scenes.
Kennedy and Naaman (2008)	Apply k-means for image clustering using grid colour moment features and Gabor textures.
Crandall et al. (2009)	use mean shift on the GPS data to find places of interest. Then apply spectral clustering on each geo-cluster.
Jiang et al. (2016)	Geo-clustering using mean shift then analyze metadata associated to generate travel recommendations.
J. Zhang et al. (2017)	use mean shift process the textual data in each geo-cluster for automatic image tagging.
Sang et al. (2017)	use mean shift analyses the textual tags to find appropriate representative places of interest (POI) and themes.
Qian et al. (2021)	They present (LAST) that fuses several modalities, namely, location, appearance, semantic, and temporal information.

2.2 Geometric Verification

Geometric verification has been used in several image retrieval research. In this section, we present a brief review of some of the related research.

One of the first approaches is the fast spatial matching (FSM) method presented by Philbin, Chum, Isard, Sivic, and Zisserman (2007). They generate hypotheses from single correspondences exploiting local feature shape. Then, they generate and evaluate all possible transformations and apply local optimization whenever a new best model is found. The main drawback of this approach is the high computational complexity due to evaluating all hypotheses on all the correspondences. Jegou, Douze, and Schmid (2008) present weak geometric consistency (WGC) method that uses the geometry of the local features. The Hough voting space is built using the scale and rotation of the SIFT feature matches. Each match votes independently for scale and rotation using two histograms. Assuming that truly matched features will share similar orientation and scale difference, peak matches in the histograms are selected as inliers. Li, Larson, and Hanjalic (2015) present the pairwise geometric matching (PGM). They begin with a pruning step that enforces 1-vs-1 correspondences. Then they perform a two-stage procedure to handle noise in the voting process caused by inaccurate feature matches. First, a voting process similar to WGC mentioned above to estimate the dominant ranges of orientation and scale in the putative matches. PGM uses pairwise geometric relations by generating vectors between pair of correspondences to measure length ratio and the angular differences of the two vectors to vote for rotation and scale differences, respectively. Finally, a score is given for correspondences that incorporate the geometric estimation.

Tolias and Avrithis (2011) propose the Hough Pyramid Matching (HPM) for retrieval re-ranking. HPM uses a hierarchical voting space using 4 transformation parameters (rotation, scale, and translation which includes x-axis and y-axis). Every match votes for a single transformation on each level. A score value is calculated for each correspondence by counting the number of other correspondences in the same bin. The similarity between the two images is evaluated by aggregating the score values for all the correspondences. The retrieval efficiency of HPM is extended in Avrithis and Tolias (2014) to include the soft assignment (Philbin, Chum, Isard, Sivic, & Zisserman, 2008) of visual words on the query image.

Schönberger, Price, Sattler, Frahm, and Pollefeys (2017) present a vote-and-verify method. This method uses a similar hierarchical voting space to (HPM). The voting scheme is used to generate a set of hypotheses to be verified in a subsequent step. To select the best hypothesis, each bin in the fine level is given a score, the bins with the highest scores are selected. Then, the hypotheses are verified in decreasing order of their score using the same procedure of FSM (Philbin et al., 2007). Similarly, Yuan, Li, Wan, and Yau (2018) uses 4D hierarchical voting space with a consequent verification step. However, multiple hypotheses are used for verification. First, they employ a hypothesize-and-verify strategy for all similarity transformations, then they apply the pair-wise geometric verification of (Li et al., 2015) for hypothesis testing. These hypotheses are used to identify the best group of correspondences while remaining other groups of correspondences.

Shen, Lin, Brandt, Avidan, and Wu (2012) propose to apply various predefined scale and rotation transformations to the features of a pre-defined query region and

produce a 2D translation voting map for the query region. Specifically, 64 queries when using 8 quantization levels for scale and rotation each are issued for every query which makes this method computationally expensive. Zhong, Zhu, and Hoi (2015) propose a fast solution to this problem with direct spatial matching (DSM) approach. Instead of performing multiple queries, they directly calculate the ratio of the difference in scale for each matched feature through the parameters of feature shape.

Closely related to our approach is the work of Zhang, Jia, and Chen (2011). They use a 2D Hough voting space based on the relative translation differences of feature correspondences. They generate the geometric-preserving visual phrases (GVP) which are groups of visual words of specific length falling in the same voting cells. However, in our proposed method we exclude correspondences with random voting cells. In addition, we perform and local consistency check for each matching feature for further verification.

Lu, Zhang, and Tao (2016) use multiple feature matches where each query feature descriptor is matched with k-nearest features in the second image. Similar to the PGM method mentioned above, they generate vectors between pairs of features to vote for scale and rotation differences. For each matching feature, the voting is performed in a local accumulation array and another general voting array. The final inliers are determined when the correspondences of the peak parameters in the local voting space match the peak voting parameters in the general array.

Another work that has a close verification procedure to our method is presented by Jiang and Jiang (2018). They propose a Hierarchical Motion Consistency Constraint

Table 2.2: Summary of the geometric verification methods related to our work.

Philbin et al. (2007)	FSM generate and evaluate all possible transformations exploiting local features shapes.
Jegou et al. (2008)	WGC votes for scale and rotation of the SIFT feature matches then fides peak values.
Li et al. (2015)	PGM votes similar to WGC on 1-to-1 matches then uses pairwise geometric relationships for voting.
Tolias and Avrithis (2011)	HPM uses a hierarchical voting space using 4 parameters the score is calculated using bins in each level.
Avrithis and Tolias (2014)	extends (HPM) to include the soft assignment.
Schönberger et al. (2017)	vote-and-verify uses the voting space of (HPM) then verifies inliers using the same procedure of FSM.
Li et al.	uses the voting space of (HPM) with multiple hypotheses for verification. Then apply (PGM) for verification.
Shen et al. (2012)	apply various scale and rotation transformations to the features of a pre-defined query region.
Zhong, Zhu, and Hoi (2015)	(DSM) directly calculate the ratio of the difference in scale for each matched feature using the feature shape.
Zhang et al. (2011)	(GVP) generate "visual phrases" using a 2D Hough voting space based on the relative translation differences.
Lu, Zhang, and Tao (2016)	the voting is performed in a local and in a general array. inliers have local and general peak voting parameters.
Jiang and Jiang (2018)	Generate motions. motions with large changes are removed. local consistency verification with neighbours is performed.

for the unmanned aerial vehicle (UAV) (e.g. drones). They first establish a 2D motion translation for corresponding features. Then, motions with large direction and length changes are removed. In the second step, a local consistency constraint is performed using the neighbouring motions. The final matches are then refined using RANSAC. This method depends on the data retrieved from the UAV to project the movement of the correspondences as "motions". In our method we depend only on the locations of the features to measure the consistency. Table 2.2 presents a summary of the these methods.

2.3 Reducing the Number of Structure Similarity Calculations

The procedure of SCAN computes the structural similarity of all vertices in the network, including those that are not assigned to any clusters. This entails a large number of similarity calculations, and consequently increase the computational overhead. To solve this problem, a number of approaches have been proposed to reduce the number of similarity calculation. A brief overview of these approaches is given below.

Shiokawa, Fujiwara, and Onizuka (2015) introduce a method called SCAN++. They propose a new data structure containing vertices that are two hops away from a given vertex. It is designed based on the property that a vertex and its two-hop-away vertices are expected to share large parts of their neighbourhoods due to the large values of clustering coefficients in real-world graphs. Thus, SCAN++ avoids computing structural similarity between vertices that are shared between such data structures. Chang, Li, Qin, Zhang, and Yang (2017) propose the pruned SCAN (pSCAN) approach for structural graph clustering. Identification of all cores is the key to structural graph clustering. pSCAN maintains an upper bound counter and a lower bound counter for the number of similar neighbours of each vertex. The idea is to perform structure similarities as long as the status of the vertex is not determined yet. Once the vertex reaches the core or not a core status (using the counters), the algorithm stops and moves to the next un-processed vertex. pSCAN avoids a large number of similarity computations and is faster than other methods.

Mai et al. (2017) propose a parallel algorithm, called anySCAN, which produces an approximate result in the beginning and then progressively refines it during the execution. This method first generates a set of preliminary groups by randomly examining

un-processed vertices to find cores and connect them with their neighbourhood into "super-nodes". The counters of pSCAN are also used here to limit the calculations. Super-nodes are eventually merged based on their connectivity. Zhao, Chen, and Xu (2017) propose another approach for anytime SCAN. They applied anytime theory and an active learning strategy to find the same clustering result on large-scale networks as the original SCAN. LinkScan* method (Lim, Ryu, Kwon, Jung, & Lee, 2014) use an edge sampling technique for reducing the number of structural similarity evaluations. This is achieved by reducing the links considered and obtained an approximate result of the SCAN. Some other methods present parallel or distributed solutions. Shiokawa, Takahashi, and Kitagawa (2018) propose ScaleSCAN, which is the multi-core implementation for SCAN. Che, Sun, and Luo (2018) present the ppSCAN to parallelize pruning-based SCAN algorithms on multi-core CPUs.

Recently, Inoubli, Aridhi, Mezni, Maddouri, and Mephu Nguifo (2020) implemented a novel SCAN-based distributed graph clustering algorithm on BLADYG framework. Wen et al. (2017) present an index-based SCAN to answer the query for any given parameters.

2.4 Discussion and Gap Analysis

In this section, we discuss some of the drawbacks of the previous works related to this research.

Landmark image clustering. Several image clustering approaches discussed above build a matching graph (dissimilarity matrix) for the entire dataset and then perform clustering as in (Johns & Yang, 2011; Philbin et al., 2011; Philbin & Zisserman, 2008;

Quack et al., 2008). The graph is then used to analyze the connectivity and the structure of images to apply the graph clustering approaches. To build this matching graph, each image is used as a query and matched with all other images in the dataset. The matching process is performed using measuring the local features where each feature from the first image corresponds to its nearest feature in the second image. Then a subsequent RANSAC-based geometric verification is applied to verify the correspondences. Finally, the query image is connected to all the matching images in the dataset that have a minimum inliers threshold. However, this querying and verifying the match for the whole dataset is considered to have a very high runtime and high computational overhead. Building a matching graph for several thousand images can take days and even months to complete using a cluster of computers as reported in the experiments of (Weyand, Hosang, & Leibe, 2012).

The methods that use min-hash as in (Chum & Matas, 2010; Chum et al., 2009) perform only local exploration of the matching graph. These methods produce seed images discovered by collisions of min-hashing to start building the clusters by using recursive query expansion and geometric verification. It is therefore expected to be faster than the methods of the full matching graphs. However, although these methods may discover and cluster hard images taken from an extreme viewpoint, the recursive querying and matching with geometric verification also require high runtime and computation overhead. Besides, this technique can lead to an "object drift" if there is limited control in the cluster growing task. This means that unrelated can be added to the clusters. For example, the "All souls" and "Radcliffe camera" of the Oxford dataset which are two different but adjacent in the location are grouped into one cluster as reported in their papers.