IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 November 2022, accepted 14 January 2023, date of publication 23 January 2023, date of current version 26 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3239014

RESEARCH ARTICLE

Estimating Bounding Box for Point of Interest Using Social Media Geo-Tagged Photos

THANH-HIEU BUI
College of Technology and Design, University of Economics Ho Chi Minh City (UEH University), Ho Chi Minh City 70000, Vietnam
e-mail: hieubt@ueh.edu.vn

This work was supported by the University of Economics Ho Chi Minh City, Vietnam.

ABSTRACT The accuracy and completeness of information in geographical databases are very important for many location-based applications and services. However, the incompleteness in geographical databases is currently an issue. One consequence of this is that the geographic bounding boxes of many points of interests (POIs) have not been known. This paper studies the problem of estimating geographic bounding boxes for POIs using geo-tagged photos contributed by public users on social media. We present a novel approach using relevant geo-tagged photos of POIs to estimate geographic bounding boxes for the POIs. In the proposed method, we extend to apply survival analysis with random distance variable for our estimation. We demonstrate the superiority and effectiveness of our proposed approach over competing methods.

INDEX TERMS Geographic bounding box, POI, survival analysis, geo-tagged photos, Flickr.

I. INTRODUCTION

Gazetteers have played a vital role in many different domains and applications due to their wide coverage and useful geographical information. In literature, there are many domains where gazetteers are used such as toponym resolution [53], geo-tagging tweets [54] and geo-tagging named entities [55]. To effectively support for such applications, we need to join a gazetteer with other geo-coded data. With the fast growing of GPS technology, many devices such as smart phones, tablets and so on can capture current GPS coordinates. Mapping the current GPS coordinates to a place is useful for many services, for instance, in booking and dispatching services such as GrabTaxi. The accurate information about places and their boundaries is an important factor which makes gazetteers become useful on many applications. However, there is a problem with geographic information in gazetteers. That is, bounding boxes of many POIs are not available (e.g. GeoNames). In this paper, we propose a novel approach to address this challenge by using relevant geo-tagged photos of POIs to estimate geographic bounding boxes for the POIs. Estimating bounding boxes for POIs has several benefits such as efficiently supporting for reverse geo-coding queries and better monitoring spatial relationships between POIs. In this study, we take minimum bounding rectangle (MBR) as the bounding box of POI. The main contributions in this paper can be summarized as follows:

- We provide a systematic study for the problem of estimating MBRs for POIs using social media geo-tagged photos; from the best of our knowledge, this is the first time such a study is presented.
- We propose a novel approach that extends to apply survival analysis with random distance variable on estimating MBRs for POIs.
- We present a novel point of view on survival analysis model when extending survival analysis model with random distance variable.
- We evaluate our proposed method and report the accuracy of estimated MBRs.

The remainder of the paper is organized as following. Section II discusses related works. Section III presents data acquisition and processing. Problem definition and methodology are presented in Section IV and Section V respectively. Performance evaluation is given in Section VI. Section VII concludes the paper and discusses about the future work.

The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Kolomvatsos.

II. RELATED WORK

The studies in literature related to our work can be divided into: (1) spatial extent estimation of geographic entities and POI identification, (2) automatic gazetteer expansion or enrichment.

A. SPATIAL EXTENT ESTIMATION OF GEOGRAPHIC ENTITIES AND POI IDENTIFICATION

Place is an important concept in geography that has been studied extensively and it strongly relates to social, economics, cultural and political aspects. The spatial extent of a place can be estimated using geo-tagged data [1]. In the study of Chen et al. [2], the authors presented a novel approach to detect the spatial extents of places with vague boundaries. The boundary of place is defined based on the density of geo-tagged photos which are mapped in to a region. Kernel Density Estimation (KDE) is applied for estimating the boundaries of regions where geo-tagged photos are spare. In the approach, the authors firstly identify a set of clean points. After that, they estimate the boundary for the rest points by using KDE. This approach has a limitation that is if a place has disjoint regions, each region will forms its own boundary. In addition, this approach might not define the boundaries for places with less geo-tagged photos. In another study, Parker and Downs [3] proposed a novel approach to generate geometric footprints, which delineate the region occupied by a spatial point pattern, by clustering data points and then creating a minimum convex envelope to enclose each cluster. This study utilizes two density-based clustering techniques for footprint generation. Firstly, DBSCAN algorithm is applied to separate non-core points, core points, or statistical noise. Next, a footprint is generated from the non-core and core points in each cluster based on convex hulls. Secondly, the authors applied Fuzzy-Neighborhood (FN)-DBSCAN algorithm to assign points to clusters depending on membership values. Two methods are introduced for defining footprints with FN-DBSCAN: (1) hull-based methods and (2) contouring techniques. The second method shows more flexible for footprint generation, as it gives a continuous surface of membership values from which accurate contour can be described. A heuristic approach of parameter selection for FN-DBSCAN is also represented. Alani et al. [4] used Voronoi diagrams to approximate the extent of places from their centroids to create polygonal boundaries of places. In this study, the authors presented a Voronoi diagram method for creating approximate regional extents from centroids that are inner and outer to regions. The resulting approximation gives a real extent measurement and it can be applied to support in responding geographical queries based on assessing spatial relationships such as direction, distance and common boundary length. The experimental results of the approach have been analyzed in the context of a semantic modeling system which joins the centroid data with adjacency and hierarchical relations between the related place names. In another work, Somodevilla et al. [5] introduced a notion of fuzzy MBR to model the spatial extent of a geographical location. However, the authors did not provide any evaluation on the quality or the accuracy. In this study, the authors introduced a fuzzy set approach to present the spatial area of a place utilizing thematic, spatial, and temporal reasoning. The authors took point locations and created an inscribed rectangle that is the maximum rectangle inner the location. They also built a fuzzy minimum bounding rectangle which contains all points of place. The area between inscribed rectangle and fuzzy minimum bounding rectangle is regarded as the fuzzy area. The final minimum bounding rectangle is estimated based on membership value of points in fuzzy area. Montello et al. [14] specified the common core region of downtown Santa Barbara by sending invitation to ask participants for drawing the boundaries of the downtown. In their study, the authors discussed the application of vague spatial concepts in particular vague regions. In addition, the authors accepted the premise that the characteristics of geographic information system will be improved if they clarify queries including vague terms. The study concentrates on methods to specify the referents of queries regarding vague regions in geospatial information systems. For instance, the users can query the map with the terms such as "Northern California" or "are around the Eiffel Tower". Understanding vagueness is important and this fact has been perceived in geographic information science for the long period of time. There are many researches on how to represent vagueness mathematically or computationally. This work does not focus on the formal structure of vague spatial concepts, but it is necessary for information systems. This work addresses the problem of using behavioral methods to determine what people mean when using vague terms, especially vague spatial terms. The detailed example of the empirical determination for the downtown area of Santa Barbara is presented. Jones et al. [21] utilized a search engine to collect geographic entities which are related to a vague place name, and used the locations of harvested entities to estimate the vague boundary. In this study, the authors presented and evaluated a method that utilized knowledge collected from the Web to model the extent of boundaries for vague places. The approach is based on the reality that when a vague place is referenced in a text document, it is usually followed by references to other more accurate places that conjugate with the extent of the value place. After that, the method of density surface modeling is applied to specify regions connected with the most often co-occurring places. The promising results of evaluation for the method on both vague and precise are presented. The application using a geographical Web search engine is demonstrated. The study examines the density surface modeling approach in more depth, expresses Web harvesting techniques, and gives evaluation of the approach presented including application to geographical Web search engine. Geo-tagged Flickr photos, which include textual tags and locations, are utilized in many researches on estimating boundaries of vague places [22], [23], [24]. Grothe et al. [22] introduced an automated approach of footprint generation utilizing the statistical evaluation of a set of points, which are supposed to lie in the region. In this work, the authors applied and compared two statistical techniques, Kernel Density Estimation and Support Vector Machines (SVMs). The overall proposed approach is assessed using precise regions. The results for two techniques evaluated by means of statistical classification measures show a slight superiority of SVMs. Lastly, a priori choices for the input parameter are inferred from the results. The footprints of imprecise regions are created in an automatic process. In another study, Intagorn et al. [23] proposed a novel method that specifies noise in social annotations. The authors evaluated the method on some countries and US states. The evaluation results show that the proposed method can learn considerable better boundaries than an alternative method. The authors also demonstrated that the proposed method can learn reasonable boundaries of vague places without ground truth. Li et al. [24] utilized spatial footprints as information of human interaction with the environment. More specific, the authors used Flickr geo-tagged photos to provide views about places. Spatial footprints, i.e. geo-tags, associated with photos can describe place locations and spatial extents as well as the relations of places. This information about place can be used to examine the way that people perceive their landscape. It also can be integrated into existing gazetteers for location-based services and geographic information retrieval.

The researches on POI identification or POI discovery, in a broad sense, can be classified into two problems: identifying single POI boundary, and identifying multiple POIs where POIs are regarded as POI clusters, i.e. the clusters of geo-tagged photos associated with POIs. POI clusters can be used to form POI areas. The approaches for discovering a single POI boundary were studied in [10], [11], and [13]. Bui et al. [10] introduced a novel method, namely Boundary-dependent Explicit Semantic Analysis (BESA), to identify the boundary of POI. In this study, POI boundary is represented as a circle with the POI location as the center and the radius of circle is unknown. When the radius is determined as an assured distance from the POI location, textual data of geo-tagged Flick photos inside the circular boundary are presented to a topic vector that each element of the vector is a Wikipedia concept. To detect the appropriate boundary, the authors considered a POI as a circle with increasing radii. By examining the cosine similarities between the vector of a POI center and those of distant positions, the POI boundary is specified. The POI boundary is determined with the radius at which the cosine similarity decreased significantly. The experiment is done on five POIs. The experimental results showed that top 20 highly weighted topics inside the boundaries specified by the proposed approach are more relevant to the POIs than inside other boundaries. These results confirm the effectiveness of BESA for detecting boundaries of POIs. Vu et al. [13] introduced a novel approach for detecting social POI boundary based on geo-tagged tweets. In this study, the authors defined social POI boundary as a cluster including POI center and a convex polygon which forms a geographical region of POI. The authors also formulated a problem of constrained optimization and presented an effective optimal estimation algorithm to solve the problem. The performance of GeoSocialBound algorithm is evaluated on various environments. The analysis results show that the proposed approach can obtain high degree of accuracy. Tran et al. [11] presented a novel approach, namely iterative SoBEst (I-SoBEst), for detecting the social boundary of POI which is represented as a convex polygon. The analysis results show that the complexity of I-SoBEst is linear with the number of records. The experimental results also show the superiority of the algorithm over SoBEst (which was originally referred to as GeoSocialBound in [13]) and competing clustering methods. Previous studies related to identifying multiple POIs usually result in a list of POI clusters [31, 32, 33, 36, 37, 38, 39,40, 41, 42, 43, 44, 45]. Lee et al. [31] applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [30] to discover POIs from the collection of geo-tagged photos and then mining association rules for associative POIs. Similarly, Sun et al. [32] employed DBSCAN clustering to identify POIs from Flickr geo-tagged images. Subsequently, the authors built a recommendation system that provides users with the most popular POIs as well as the best travel routings between the POIs. Höpken et al. [33] used both K-mean and DBSCAN clustering to identify POIs from collection of geo-tagged photos in the city of Munich. The authors then discovered tourists' behavioral patterns in the form of POIs often visited together and frequent visitation sequences. Crandall et al. [36] applied mean shift clustering [35] to discover POIs from a collection of Flickr geo-tagged photos. Zhang el al. [37] discovered POIs by mean shift approach to obtain the geographical clustering result on online sharing Websites image collections. The authors then proposed a POI-based tag matrix completion framework which processes the images within each POI in parallel. Kisilevich et al. [41] proposed P-DBSCAN which is a new and more advanced version of DBSCAN. P-DBSCAN was specifically developed for clustering geo-tagged photos and discovering POIs by taking into account information about the photo owners. Majid et al. [40] applied P-DBSCAN to discovery POIs from a collection of Flickr geo-tagged photos. Next, the authors presented a novel method for POI recommendation that is compatible with users (i.e., personalization) in the given context (i.e., context awareness). Bui [43] used P-DBSCAN to discover POIs as shopping locations from a collection of geo-tagged photos related to shopping in Los Angeles City, California, USA. Subsequently, the author uncovered the spatial-temporal behaviors of shopping users based on their visited POIs. Lyu et al. [44] employed P-DBSCAN algorithm to identify POIs from geo-tagged photos, i.e., obtaining a set of travel POIs. Thereafter, the authors introduced Weighted Multi-Information Constrained Matrix Factorization for personalized travel location recommendation. Similarly, Ameen et al. [45] also utilized P-DBSCAN to discover POIs as travel locations from community-contributed geotagged photos. Next, the authors presented a convolutional neural network and matrix factorization-based travel location recommendation method to address the the travel location cold start problem. There are also many studies using other clustering methods to discover POIs from geo-tagged photos. Kuo et al. [42] presented an end-to-end framework for discovering POIs/AOIs from the spatial and temporal properties and attributes of Flickr geo-tagged photos. Yang et al. [51] introduced a robust noise-resistant approach based on Laplacian for POI identification using geo- and textual-tagged social photos data. Pla-Sacristán et al. [52] proposed two density-based clustering algorithms, namely K-DBSCAN and V-DBSCAN that have a direct applicability on the task of automatic POI identification.

B. AUTOMATIC GAZETTEER EXPANSION OR ENRICHMENT

Gazetteer expansion or enrichment is a process of adding new records as new places or the missing attributes of an existing place to a gazetteer. Popescu et al. [6] proposed a novel approach for creating and enriching a geographical gazetteer, called Gazetiki. The authors extracted geographic entities from Wikipedia, Panoramio, and web search engines. The entities are then categorized, coordinated and ranked. The Geonames database can be enhanced and complemented from information in Gazetiki. The experimental results show that the proposed method can give a richer structure and an improved coverage in comparison with another known study of automatically building a geographic database. Oliveira et al. [7] proposed a novel approach for enriching the GeoSEn gazetteer [17] by using the geographical information that is gathered through crowd sourcing. The authors expanded the spatial hierarchy of the gazetteer by adding places at what appears to be district and street granularity. The authors presented a case study for evaluating the extended version of the GeoSEn system. The analysis results show an acceptable accuracy and precision regarding the known VGI quality issues. Geo-tagging Flickr photos and videos can be applied to enrich places in gazetteers with location specific photos as showed in the study of Serdyukov et al. [8]. In this work, the authors presented a language model for mapping Flick geo-tagged photos to places on the earth utilizing user tags to photos. The authors leveraged a grid based method to separate the earth surface into cells of equal sizes then predicting a specific cell for each photo. The authors used smoothing techniques for refining the cell prediction. In another study, Kordopatis -Zilos et al. [9] introduced a bag of tags method that the probability of a tag being used by users is determined for describing a region. The tags in each cell are weighted based on spatial entropy that the tags which are user specific or general are assigned less weight. The authors offered several refinements over a language model-based approach [18] which has been showed to have competing performance. The authors showed that the proposed refinements result with excellent improvement regarding the geo-tagging precision and the accuracy of the geo-tagging output. Additionally, the authors introduced an in-depth analysis of the performance for the proposed method, as well as the contribution of each refinement and the effects when increasing size of the training dataset. Moura et al. [12] introduced a method that use linked data sources to put gazetteer data together. The linking data sources bring enriching gazetteer with a set of semantic and geographic relationships. Therefore, this activity helps to solve the typical GIR problem such as altering and disambiguation. The study presents the result of efforts to aggregate two linked data sources of gazetteer data, namely DBPedia and GeoNames. Hu et al. [15] introduced a novel method to collect local location names from geo-tagged housing advertisements. The authors utilized the posts on advertising websites such as Craigslist, which usually include local place names. The proposed method has two steps: natural language processing (NLP) and geospatial clustering. In NLP step, place name candidates are extracted from the textual content of advertising posts. The geospatial step concentrates on performing multi-scale geospatial clustering on the coordinates associated with extracted candidate location names. The performance of proposed method is compared with six baselines. The result of proposed approach is also compared with four existing gazetteers to present the not-yet-recorded local location names uncovered by our approach. Smart et al. [16] proposed a mediation framework to retrieve and merge different gazetteer resources to create a meta-gazetteer that produces enhanced versions of place name information. The proposed method joins different information of place name from many gazetteer sources that relates to the same geographic location. The approach also uses many similarity metrics to specify equivalent toponyms. Oliveira et al. [18] introduced a novel approach for gazetteer enrichment based on VGI data sources. In reality, VGI environments are not built to work as gazetteers. But, they usually include more up-to-date and detailed information than gazetteers. The proposed approach is used in geo-parser environments by utilizing its heuristics set besides enriching the gazetteer. The authors presented a case study with geo-parsing Twitter messages which focuses on the micro-texts to exanimate the performance of enriched procedure. Gelernter et al. [19] introduced a novel approach to identify sources of new local gazetteer entries in crow-sourced Wikimapia and OpenStreetMap geo-tags which contain geo-coordinates. The authors built a fuzzy matching algorithm based on Support Vector Machine algorithm that examines both approximate geo-coding and approximate spelling to detect duplicates between the gazetteer and crowd-sourced tags to absorb novel tags. The proposed algorithm creates candidate matches from the gazetteer and next ranks those candidates by word form or geographical relations between gazetteer candidate and each tag. The proposed approach compares a baseline of edit distance for candidate ranking. Keßler et al. [20] introduced a novel bottom-up method for gazetteer generation using geo-tagged photos. This work mentions about the building blocks of geo-tags and

TABLE 1. POI names and POI types according to Geonames.

POI Name	Geonames Category	POI Location
California Academy of Sciences (CAS)	S.SCH	37.76993, -122.4658
De Young Museum (DYM)	S.MUS	37.77139, -122.46861
Mission Dolores Park (MDP)	L.PRK	37.75965, -122.42608
Museum of Modern Art (MMA)	S.MUS	37.7852, -122.401
Japanese Tea Garden (JTG)	L.PRK	37.76965, -122.46969
Alamo Square (AS)	L.PRK	37.77629, -122.43467
Ferry Building (FB)	S.BLDG	37.79543, -122.39356
Asian Art Museum (AAM)	S.MUS	37.7808, -122.419
Pier 39 (P39)	S.MALL	37.80965, -122.41025
Contemporary Jewish Museum (CJM)	S.MUS	37.78577, -122.40394

relationships between them to officially define the notion of geo-tag. Based on their discussion, the authors demonstrated an extraction process for gazetteer entries which takes in account the emergent semantics of geo-tagged photos collections and give a group-cognitive perspective on named places. The authors set up an experiment for specifying place names and assigning adequate geographic footprints by using clustering and filtering algorithms. The experimental results of three place names with different geographic feature types including Soho, Camino de Santiago and Kilimanjaro are evaluated. The authors showed how the proposed approach can be combined with other approaches. The discussion about complementing existing gazetteers is also presented.

III. DATA ACQUISITION AND PROCESSING

In this section, we describe how to collect POIs and how to gather geo-tagged photos from a social media photo-sharing platform as well as to extract relevant geo-tagged photos associated with the POIs. We present details of each task as follows.

A. COLLECTING POI

To collect a set of POIs and their locations, we use an open geographic database namely Geonames which contains a great amount of geographical names [29]. There are many feature classes of geographic concepts in Geonames such as parks, area, spot, building, farm and so on. But the following five representative categorized types of POIs are taken into account in the study: park (L.PRK), building (S.BLDG), mall (S.MALL), museum (S.MUS), and school (S.SCH). The POI names and POI types are summarized in Table 1. All POIs in our study are located in San Francisco City, California, USA.

B. COLLECTING GEO-TAGGED PHOTOS

Geo-tagged photos are collected from Flickr, a well-know social media photo sharing platform. We use Application Programming Interface (API)[1] provided by Flickr to collect geo-tagged photos. More specifically, we use flickr.photo.search function to retrieve geo-tagged photos. The geo-tagged photos are collected within San Francisco City, California, USA and having taken date before March 2022. Our dataset consists of 1,159,472 Flickr geo-tagged photos. For data processing on the next step, we adopt the following three essential attributes from the geo-tagged photos:

- tags: actual UTF-8 tags of a geo-tagged photo
- latitude: latitude of a geo-tagged photo's location
- longitude: longitude of a geo-tagged photo's location

TABLE 2. Relevant geo-tagged photos for POIs.

POI Name	No. Geo-tagged Photos
California Academy of Sciences	9,067
De Young Museum	12,316
Mission Dolores Park	5,058
Museum of Modern Art	11,766
Japanese Tea Garden	4,460
Alamo Square	2,428
Ferry Building	7,485
Asian Art Museum	3,153
Pier 39	9,814
Contemporary Jewish Museum	1,329

C. EXTRACTING RELEVANT GEO-TAGGED PHOTOS FOR POI

Social media users are able to tag a POI name (e.g., De Young Museum) on their photos taken at the POI to describe their interest about the POI. Therefore, geo-tagged photos which are tagged with a POI name are considered relevant to the POI. We can easily gather relevant geo-tagged photos of a particular POI by searching geo-tagged photos with the POI name in their tags. Since relevant geo-tagged photos of a POI are usually taken around the POI location, we limit our searches with geo-tagged photos within 5 km from the POI location. In reality, a POI name may appear in many forms such as alternative names, abbreviation. Thus, we perform keyword-based searches by querying different words for a POI name. Relevant variations for a POI name would include its abbreviated name and its alternative names. By using query processing, for each POI we can obtain the set of relevant geo-tagged photos which are tagged with the POI name or the relevant variations of the POI name. Table 2 presents the number of relevant geo-tagged photos for each POI that we extract from the dataset.

IV. PROBLEM FORMULATION

Before we formally define the problem, we give definitions of some basic concepts and terms.

Definition 1 (Geo-Tagged Photo): A geo-tagged photo ρ can be defined as:

$$\rho = (\delta, l, \tau, T, d) \tag{1}$$

[1] www.flickr.com/services/api

where δ is the photo ID; l is geo-tags or a pair latitude–longitude coordinates where the photo ρ was taken; τ is the title of the photo; the photo ρ is annotated with a set of textual tags denoted as T; d is the textual description of the photo.

Definition 2 (POI): A POI P is defined as a unique specific site (e.g., a museum or a park). In our model, a POI has two attributes: POI name and POI location. We use N_P to represent the POI name and $L_P = (lat_P, lng_P)$ to denote its location or corresponding geographical attribute in terms of longitude and latitude coordinates.

Definition 3 (Collection of Geo-Tagged Photos Relevant to a POI): The collection of geo-tagged photos relevant to a POI P denoted as C_P is a set of geo-tagged photos around the POI location which include POI name N_P or its relevant variations in their tags. We can specify C_p as flowing:

$$C_P \triangleq \{\rho\} \tag{2}$$

where $\rho.T$ includes N_P or its relevant variations.

Our goal in this work is to estimate the geographical bounding box, i.e. MBR, for an arbitrary POI based on the distribution of relevant geo-tagged photos associated with the POI.

V. METHODOLOGY

A. SOLUTION OVERVIEW

The schematic overview of the proposed approach including data acquisition and processing stage is illustrated in Figure 1. The detailed description for each block is explained as follows. The first block presents for data acquisition and preprocessing stage as described in Section III. In the second block, for each POI, we estimate the survival probability of relevant geo-tagged photos by distance to the POI location on each geographic axis of North, South, East, and West axes. In the third block, on each axis we specify the first distance where the survival probability is less than a threshold. In the last block, based on the specified distances, the estimated MBR of POI is created as illustrated in Figure 3.

B. SURVIVAL ANALYSIS

Survival analysis is considered as a field of statistics that studies about survival time until an event of death or failure [25], [26], [28]. It is applied broadly in many areas such as economics, biology, sociology and engineering [27]. We denote T as a random time variable which presents for the time until an event happens. In traditional survival analysis model, the event is a ''death'' or ''failure''. There are four main functions used on survival analysis, which are: the failure function, the probability density function, the survival function and the hazard function.

1) FAILURE FUNCTION

the failure function or cumulative distribution function (CDF) of a time random variable T, denoted as $F(t)$, is specified as the probability to die or fail before a certain time t. The failure function is expressed as follows:

$$F(t) = Pr\{T \leq t\} \tag{3}$$

2) PROBABILITY DENSITY FUNCTION

The probability density function (PDF), denoted as $f(t)$, is defined as the derivate of the failure function:

$$f(t) = \frac{\partial F(t)}{\partial t} \tag{4}$$

3) SURVIVAL FUNCTION

The survival function, denoted as $S(t)$, is the survival probability up to a certain time t. It is also the complementary cumulative distribution function (CCDF) of the lifetime. The survival function is defined as:

$$S(t) = 1 - F(t) = Pr\{T > t\} \tag{5}$$

4) HAZARD FUNCTION

The hazard function denoted as $h(t)$ provides the failure rate at time t conditioned on the instance being still survival or alive at time t, i.e. the expected number of failures happening at or close to time t. The hazard function is defined as:

$$h(t) = f(t)/S(t) = -S'(t)/S(t) \tag{6}$$

C. ESTIMATING MBR FOR POI BASED ON SURVIVAL ANALYSIS

The approach of survival analysis is generic and this can be extended for applying to any random variable. Since we are applying survival analysis technique to another domain as in the context of our problem, we need to define our terms. In the forthcoming of this paper, we use a random distance variable R which presents for the distance until an event happens. As mentioned previously, the geo-tagged photos around a POI location which are tagged with the POI name or its relevant variations are considered relevant to the POI. We measure the survival probability of the relevant geo-tagged photos based on their distance to the POI location. We extend to apply survival analysis for this measurement. To this end, we consider the distance variable as same as the time variable in traditional survival analysis model [25], [26], [28]. In practice, traditional survival analysis model is used for modeling the survival probability of an event that depends on the time with some explanation factors.

When we consider the random distance variable as same as the random time variable, we can model the survival probability of the relevant geo-tagged photos to a POI based on the distances of the geo-tagged photos to the POI location. We denote $r_0 = 0$ as the initial distance and Δr as increasing interval for the initial distance. We represent $r_1, r_2 \ldots r_k$ as respective distances when increasing r_0 with $\Delta r, 2\Delta r, 3\Delta r, \ldots, k\Delta r$. If we regard the distance r_i as the survival distance for the event: a relevant geo-tagged photo ρ is still survival if the distance from ρ to the POI location is greater than r_i, we can model the survival probability of

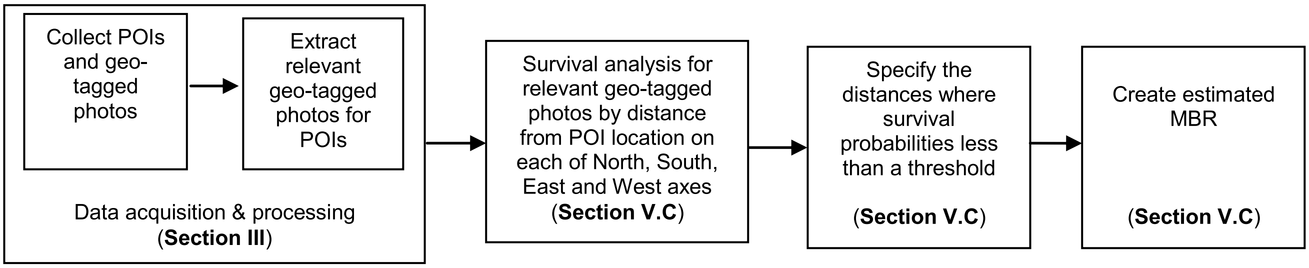

Collect POIs and geo-tagged photos
Extract relevant geo-tagged photos for POIs
Data acquisition & processing (Section III)
Survival analysis for relevant geo-tagged photos by distance from POI location on each of North, South, East and West axes (Section V.C)
Specify the distances where survival probabilities less than a threshold (Section V.C)
Create estimated MBR (Section V.C)

FIGURE 1. The schematic overview of the proposed approach including data acquisition and processing.

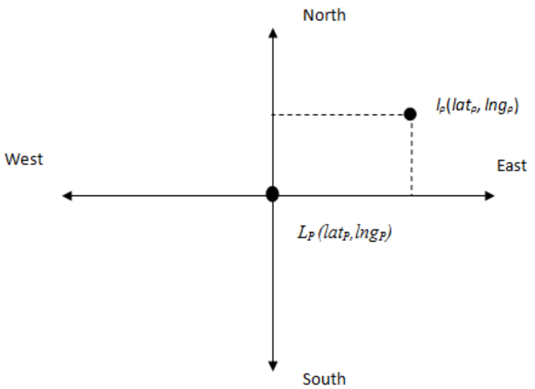


FIGURE 2. The POI location and four axes.

the event during each increasing interval Δr. We express main functions in our model which are used to describe the survival probability based on distances from the positions of the relevant geo-tagged photos to the POI location including the failure function, the survival function, the probability density function and the hazard function as following.

1) FAILURE FUNCTION

We denote $F(r)$ as the failure function or the cumulative distribution function of random distance variable R. $F(r)$ is the probability that the relevant geo-tagged photos of a POI exist before a certain distance r to the POI location. The failure function is expressed as follows:

$$F(r) = P(relevant\ geo\text{-}tagged\ photos\ exist\ inside\ distance\ r) \\ = P\{R \leq r\} \tag{7}$$

2) SURVIVAL FUNCTION

The survival probability in our model is considered as the probability that relevant geo-tagged photos still exist outside a given distance r to the POI location. The survival function at a given distance r, denoted as $S(r)$, is defined as following:

$$S(r) \\ = P(relevant\ geo - tagged\ photos\ exist\ outside\ distance\ r) \\ = 1 - P(relevant\ geo\text{-}tagged\ photos\ exist\ inside\ distance\ r) \tag{8}$$

Therefore, we have:

$$S\,(r) = P\,\{R > r\} = 1 - F\,(r) \tag{9}$$

3) PROBABILITY DENSITY FUNCTION

The probability density function of the survival distance R is defined as the probability that the POI has relevant geo-tagged photos in the short interval per unit distance. It can be expressed as following:

$$f\,(r) = \lim_{\Delta r \to 0} \frac{P[r \leq R < r + \Delta r]}{\Delta r} \tag{10}$$

4) HAZARD FUNCTION

The hazard function $h(r)$ is defined as the event rate at distance r conditional on survival until distance r or later (that is, $R \geq r$). It can be formulated as follows:

$$h\,(r) = \lim_{\Delta r \to 0} \frac{P[(r \leq R < r + \Delta r)|R \geq r]}{\Delta r} \tag{11}$$

$$h\,(r) = \frac{f\,(r)}{1 - F(r)} = \frac{f\,(r)}{S(r)} \tag{12}$$

We use life table method [26] to estimate the survival function:

$$\hat{S}\,(r_i) = \prod\nolimits_{j=1}^{i-1} (1 - \hat{q}_j) \tag{13}$$

where:

$\hat{q}_j = d_j/n_j$ is the conditional probability of failure in the interval; d_j is the number of relevant geo-tagged photos failure, i.e. existing in the interval; n_j is the number of relevant geo-tagged photos exposed in the interval.

For each POI P with the POI location L_P=(lat_P, lng_P), we consider the axes of four directions including North, South, East, West as illustrated in Figure 2. For each geo-tagged photo ρ in C_P presented for the collection of relevant geo-tagged photos of the POI P, the location or geo-tags of the geo-tagged photo ρ is denoted as $l = (lat_\rho, lng_\rho)$. The latitude distance from the geo-tagged photo to the POI location is specified by:

$$D_{lat}(\rho, P) = |lat_\rho - lat_P| \tag{14}$$

Similarly, the longitude distance from the geo-tagged photo to the POI location is:

$$D_{lng}\,(\rho, P) = |lng_\rho - lng_P| \tag{15}$$

The latitude distances and longitude distances are then converted into meter unit. We apply survival analysis for the relevant geo-tagged photos of the POI on four axes including West, East, South and North axis. Depending on the latitude

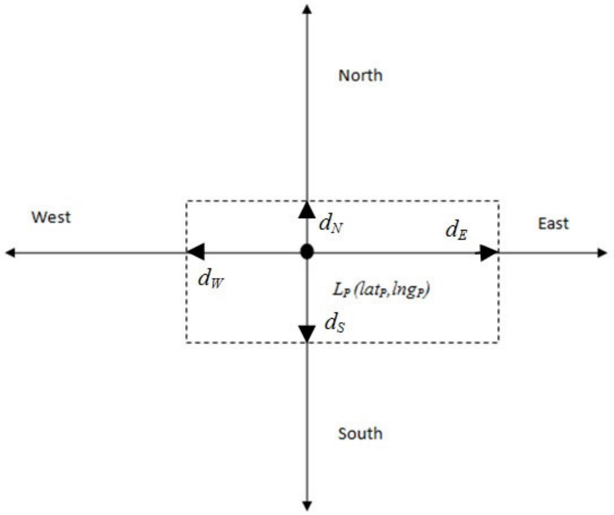


FIGURE 3. The illustration of estimated MBR.

and longitude coordinates of the relevant geo-tagged photos of the POI, we can infer the set of relevant geo-tagged photos on each of West, East, South and North axes. For estimating the survival probability of relevant geo-tagged photos by distance, the latitude distance is used as survival distance on East, West axis. Similarly, the longitude distance is used as survival distance on North, South axis. Based on the survival probability, we specify the first distance on each axis where the survival probability of relevant geo-tagged photos is not considerable. To this end, we define a threshold denoted as γ and find the first survival distance on each axis that the survival probability less than the threshold. We denote these distances as d_E, d_W, d_N and d_S on East, West, North and South axis respectively. When the distances are specified on the four axes, the MBR of the POI is formed. Figure 3 illustrates the estimated MBR for the POI after specifying the distances d_E, d_W, d_N and d_S.

VI. PERFORMANCE EVALUATION

We evaluate our proposed method in terms of the accuracy of the bounding boxes as MBRs of POIs which are estimated by our approach. We carry out the validation for the proposed method with the POIs, the geo-tagged photos dataset and the relevant geo-tagged photos for the POIs as mentioned in Section III. The information of POIs is presented in Table 1. The POIs are located in San Francisco City, California, USA. The statistics of relevant geo-tagged photos for the POIs are shown in Table 2.

A. EXPERIMENTAL SETUPS

In our survival analysis model, we set distance interval Δr as 10 meters, which is a reasonable distance interval for examining the variance of the survival probability of relevant geo-tagged photos by distance for POIs. The threshold parameter γ is set equal 0.2 when comparing with competing methods. We also examine the changes in performance of the proposed approach by varying the threshold parameter with different settings $\gamma \in \{0.25, 0.2, 0.15, 0.1, 0.05\}$.

B. EVALUATION MEASURE

The ground truth for our evaluation is obtained by querying OpenStreetMap (OSM). OSM has been used similarly in the literature [46, 47, 48, 49, 50]. For each POI, we retrieve the ground truth MBR for the POI by using the reverse Geocoding API provided by OSM.[2] This API function returns a bounding box as MBR for a POI by passing latitude and longitude of the POI location. The MBR of a POI is a rectangle defined by two longitudes and two latitudes. The standard format of MBR is: [min Longitude, min Latitude, max Longitude, max Latitude]. We measure the accuracy of the estimated MBR of a POI by comparing against the MBR of the POI retrieved from OSM. In the evaluation, we compute the approximate area of each MBR by calculating the Haversine distance between endpoints to infer its length and width and then using this information to compute the area. We note that each MBR is a curvilinear rectangle on the spherical surface of earth. In here, the approximation for MBR area is justified since the calculation for the area of the estimated MBR and the area of MBR retrieved from OSM is the same.

The accuracy of estimated MBR is measured in terms of the overlapped area to the ground truth MBR provided by OSM. For an estimated MBR denoted as E and a ground truth MBR provided by OSM denoted as G, we use the following measures of accuracy:

- Area Overlap Accuracy (AOA): The ratio of the area which is the intersection between E and G to the area which is the union between E and G.
- False Negative (FN) for area overlap: The ratio of the area which is the area of G not covered by E to the area of G.
- False Positive (FP) for area overlap: The ratio of the area which is the area of E not part of G to the area of E.

C. ESTIMATION PERFORMANCE

This section presents the evaluation results for the POIs. The goal of our evaluation is to examine the accuracy of the estimated MBRs by the proposed method and to uncover the affect of the threshold parameter. To describe for the process of estimating bounding boxes as MBRs of the POIs, we take an example of California Academy of Sciences. Figure 4 presents the survival probability by distance on four axes of four geographic directions including North, South, East, and West from the POI location of California Academy of Sciences. When the threshold γ is set as 0.2, the first distance from the POI location on North axis that survival probability is less than 0.2 is 60 meters. Similarly, the first distances on South, East and West axis are 50 meters, 40 meters, 40 meters respectively. Based on these specified distances,

[2] https://nominatim.openstreetmap.org/reverse

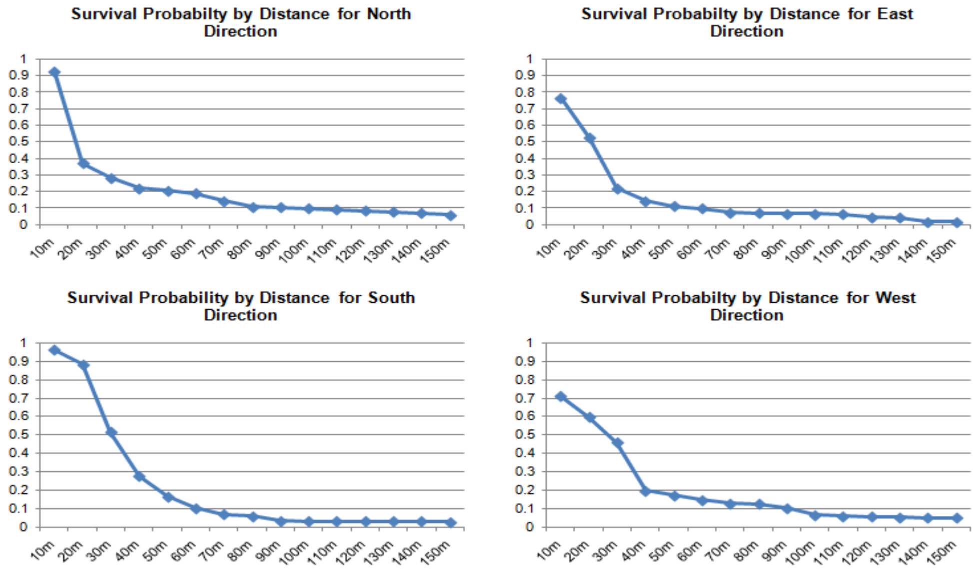


FIGURE 4. Survival probability by distance from the POI location on each axis for California Academy of Sciences.

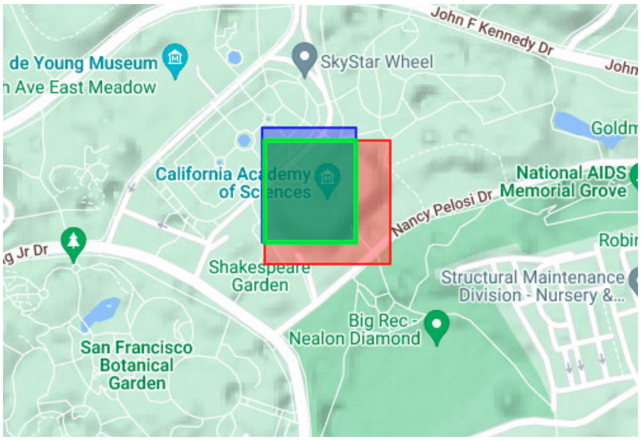


FIGURE 5. The estimated MBR in blue and the ground truth MBR in red of California Academy of Science on Google Maps.

the estimated MBR of California Academy of Sciences is constructed.

For better understanding, the MBR of California Academy of Sciences estimated by the proposed method is drawn on Google Maps and presented as in Figure 5; the red rectangle is the ground truth MBR; the blue rectangle is the estimated MBR; and the green rectangle is the overlapping rectangle between the ground truth MBR and the estimated MBR.

The performance results of the proposed method are shown according to the evaluation measure. To verify the superiority of the proposed approach, we compare the performance of our proposed approach with baseline methods as presented below.

1) COMPARISON WITH COMPETING METHODS

We compare our proposed method with three baseline methods, namely DBSCAN [30], mean shift [35] and P-DBSCAN [41], which are popularly used for POI identification. The baseline methods use the geo-tagged photo dataset for identifying POIs. Due to the fact that clustering algorithms including DBSCAN, mean shift and P-DBSCAN return arbitrarily-shaped POI clusters, we need to find MBRs containing the POI clusters in order to evaluate these methods under our problem setting. For the evaluation, we compute the MBR for each POI cluster as follows. The north latitude and east longitude of the MBR are given by the maximum latitude and maximum longitude respectively between the latitudes and the longitudes of geo-tagged photos in the POI cluster. Similarly, the south latitude and west longitude of the MBR are given by the minimum latitude and minimum longitude respectively between the latitudes and the longitudes of geo-tagged photos in the POI cluster. Finally, we present performance comparison among our approach and three competing methods in terms of measures of accuracy for MBR. In the following, we briefly describe the basic mechanism and parameter settings of each base line method.

2) DBSCAN (BASELINE #1)

As mentioned in many studies, DBSCAN is one of the most common methods applied for POI identification [31], [32], [33], [34]. Thus, it can be used as a baseline method in our study. For using DBSCAN, we need to set two parameters *MinPts* and ϵ, where *MinPts* is the minimum number of points required to form a dense region within a distance threshold ϵ [30]. In our evaluation, we set $\epsilon = 50$ meters and $MinPts = 100$ similarly as in Sun et al. [32].

3) MEAN SHIFT (BASELINE #2)

Mean shift [35] is a popular clustering method for POI identification. It has been applied to identify POI from social media geo-tagged photos in several studies [36], [37], [38], [39].

TABLE 3. Comparison with competing methods.

POI Name	AOA				FN				FP			
	Our	DBSCAN	Mean Shift	P-DBSCAN	Our	DBSCAN	Mean Shift	P-DBSCAN	Our	DBSCAN	Mean Shift	P-DBSCAN
CAS	**0.542**	0.343	0.101	0.152	0.405	0.081	0.002	0.066	0.141	0.688	0.899	0.846
DYM	**0.648**	0.395	0.107	0.121	0.185	0.001	0.002	0.002	0.239	0.603	0.893	0.879
MDP	**0.767**	0.332	0.286	0.610	0.100	0.668	0.002	0.240	0.161	0.000	0.714	0.245
MMA	**0.589**	0.103	0.043	0.014	0.400	0.101	0.002	0.002	0.028	0.894	0.957	0.986
JTG	**0.693**	0.527	0.104	0.105	0.198	0.002	0.002	0.002	0.164	0.472	0.896	0.895
AS	**0.759**	0.366	0.156	0.366	0.171	0.546	0.002	0.546	0.099	0.345	0.844	0.345
FB	**0.713**	0.229	0.057	0.092	0.010	0.002	0.002	0.002	0.184	0.771	0.943	0.908
AAM	**0.669**	0.034	0.020	0.044	0.093	0.671	0.106	0.348	0.281	0.963	0.980	0.955
P39	**0.684**	0.153	0.137	0.060	0.249	0.002	0.319	0.002	0.115	0.847	0.853	0.940
CJM	**0.667**	0.052	0.011	0.006	0.236	0.002	0.000	0.002	0.188	0.948	0.989	0.994

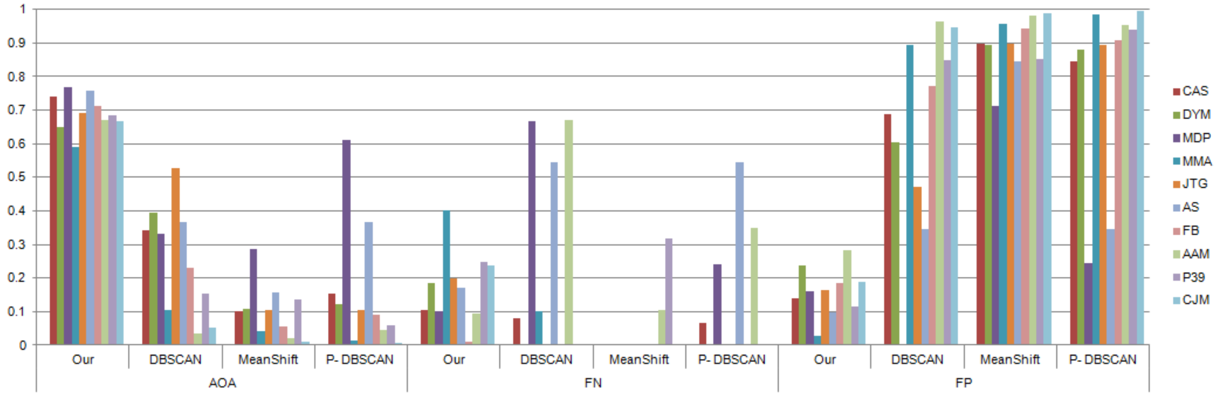


FIGURE 6. Comparision between the proposed method (our) and others.

Mean shift clustering can generate arbitrary number of clusters with different sizes. Since the result of mean shift clustering strongly depends on a kernel function rather than some parameters, mean shift is regarded as a non-parametric feature-space technique. But, in fact, mean shift clustering is not a completely parametric-free because it requires setting a neighborhood bandwidth. The clustering result is influenced by setting the bandwidth parameter. In our evaluation, we set the bandwidth as 100 meters similarly as in Crandall et al. [36].

4) P-DBSCAN (BASELINE #3)

A new and more advanced version of DBSCAN, namely P-DBSCAN, was invented for POI identification [41]. P-DBSCAN was proposed for clustering geo-tagged photos by taking into account information about the photo owners. P-DBSCAN is a common clustering method for identifying POIs and being used in several studies [40], [42], [43], [44], [45]. In our evaluation, we set parameters similarly as in Kisilevich et al. [41]. We apply $MinOwners = 50$ for the minimum number of photo owners in a cluster, $\epsilon = 50$ meters for the distance threshold, and 10% for the density drop threshold.

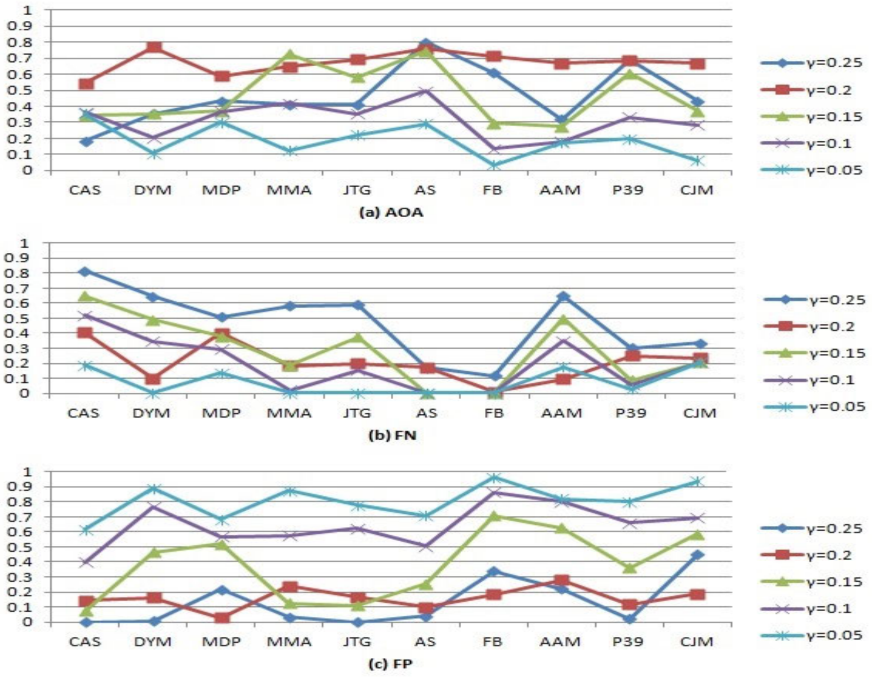
5) PERFORMANCE COMPARISON

We present performance comparison among our proposed approach and three competing methods in terms of AOA, FN, FP as in Table 3. The comparison is also visualized as in Figure 6. As can be seen from Figure 6, the proposed approach outperforms other methods in terms of area overlap accuracy for all cases. Since the area overlap accuracy presents for the quality or accuracy of estimated MBR, our propose method gives the best quality of estimated MBR in comparison with the competing methods. In addition, the false positive rate of our proposed approach is much lower than other methods in most cases of POIs. To examine the affect of the threshold parameter, we vary the threshold γ and observe the performance results with different settings of the threshold parameter including 0.25, 0.2, 0.15, 0.1 and 0.05.

The performance results of the proposed method according to different values of the threshold parameter $\gamma \in \{0.25, 0.2, 0.15, 0.1, 0.05\}$ are shown in Table 4 where POI names are presented in abbreviation. We visualize the performance results in term of area overlap accuracy, false negative, false positive as in Figure 7(a), Figure 7(b) and Figure 7(c) respectively with different settings of the threshold parameter $\gamma \in \{0.25, 0.2, 0.15, 0.1, 0.05\}$. As can be seen from Figure 7(a),

TABLE 4. The performance results when varying the threshold parameter.

POI	AOA					FN					FP				
Name	γ=0.25	γ=0.2	γ=0.15	γ=0.1	γ=0.05	γ=0.25	γ=0.2	γ=0.15	γ=0.1	γ=0.05	γ=0.25	γ=0.2	γ=0.15	γ=0.1	γ=0.05
CAS	0.183	**0.542**	0.342	0.364	0.353	0.817	0.405	0.648	0.521	0.186	0	0.141	0.071	0.396	0.616
DYM	0.353	**0.767**	0.354	0.206	0.109	0.646	0.1	0.49	0.346	0.002	0.007	0.161	0.464	0.769	0.891
MDP	0.432	**0.589**	0.372	0.367	0.302	0.509	0.4	0.378	0.294	0.136	0.216	0.028	0.52	0.567	0.683
MMA	0.413	0.648	**0.728**	0.42	0.124	0.581	0.185	0.188	0.021	0.002	0.032	0.239	0.124	0.576	0.876
JTG	0.411	**0.693**	0.581	0.354	0.221	0.589	0.198	0.374	0.154	0.002	0	0.164	0.11	0.622	0.779
AS	**0.801**	0.759	0.745	0.494	0.29	0.171	0.170	0.002	0.002	0.002	0.041	0.099	0.254	0.506	0.71
FB	0.609	**0.713**	0.293	0.136	0.036	0.118	0.01	0.002	0.002	0.002	0.337	0.184	0.707	0.864	0.964
AAM	0.32	**0.669**	0.273	0.18	0.175	0.648	0.093	0.499	0.35	0.176	0.22	0.281	0.626	0.801	0.818
P39	**0.689**	0.684	0.605	0.334	0.197	0.3	0.249	0.086	0.056	0.027	0.022	0.115	0.359	0.66	0.802
CJM	0.432	**0.667**	0.372	0.285	0.062	0.335	0.236	0.206	0.206	0.206	0.448	0.188	0.588	0.693	0.937



(a) AOA

(b) FN

(c) FP

FIGURE 7. The performance results according to AOA, FN and FP with different settings of the threshold parameter.

the area overlap accuracy when the threshold $\gamma = 0.2$ gives the best performance result in most cases of POIs. In contrast, it gives the worst performance result in most cases of POIs when the threshold $\gamma = 0.05$.

In addition, the area overlap accuracy tends to decline when the value of the threshold decreases from 0.2 to 0.05 in most cases of POIs. As in Figure 7(b), the performance result in term of false negative is highest when the threshold $\gamma = 0.25$ for all cases and lowest when the threshold $\gamma = 0.05$ in most cases. False negative measure is expressed by the ratio of the area which is the area of ground truth MBR not covered by the area of estimated MBR to the area of ground truth MBR. Since the area of ground truth MBR for a particular POI is a constant, the high value of false negative means the high value of the area of ground truth MBR not covered by the area of estimated MBR. The high value of false negative shows the low accuracy of the estimated MBR. From Figure 7(c), the performance result in term of false positive is highest when the threshold $\gamma = 0.05$ for all cases, while the false positive rate when the threshold $\gamma = 0.25$ is lowest in many but not all cases. The high false positive rate implies that the accuracy of estimated MBR is low.

VII. CONCLUSION

Bounding boxes are used broadly for approximating the spatial extents for POIs. The bounding boxes of POIs are benificial for many location-based applications and especially useful for addressing complex problems in geographical information science. This paper presents a novel approach to estimate bounding boxes as MBRs for POIs using social media geo-tagged photos. The evaluation results reveal that the proposed approach can estimate the MBRs of the POIs effectively, which yeilds a significant improvement over the baselines. As the future work, we can extend our approach to estimate MBRs for POIs by using different kinds of geo-tagged data such as geo-tagged tweets. In another extension, we can enhance our method by using geo-tagged photos combined from many social media platforms to address the data sparsity problem and help to estimate more robust MBRs for POIs.

ACKNOWLEDGMENT

The author sincerely thanks anonymous reviewers for useful comments and suggestions.

REFERENCES

[1] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores," *J. Spatial Inf. Sci.*, vol. 1, pp. 21–48, Jul. 2010.

[2] J. Chen and S. L. Shaw, "Representing the spatial extent of places based on Flickr photos with a representativeness-weighted kernel density estimation," in *Proc. Annu. Int. Conf. Geographic Inf. Sci.* Cham, Switzerland: Springer, 2016, pp. 130–144.

[3] J. K. Parker and J. A. Downs, "Footprint generation using fuzzy-neighborhood clustering," *GeoInformatica*, vol. 17, no. 2, pp. 285–299, Apr. 2013.

[4] H. Alani, C. B. Jones, and D. Tudhope, "Voronoi-based region approximation for geographical information retrieval with gazetteers," *Int. J. Geographical Inf. Sci.*, vol. 15, no. 4, pp. 287–306, Jun. 2001.

[5] M. J. Somodevilla and F. E. Petry, "Fuzzy minimum bounding rectangles," in *Spatio-Temporal Databases*. Cham, Switzerland: Springer, 2004, pp. 237–263.

[6] A. Popescu, G. Grefenstette, and P. A. Moëllic, "Gazetiki: Automatic creation of a geographical gazetteer," in *Proc. 8th ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL)*, Jun. 2008, pp. 85–93.

[7] M. G. D. Oliveira, C. E. C. Campelo, C. D. S. Baptista, and M. Bertolotto, "Gazetteer enrichment for addressing urban areas: A case study," *J. Location Based Services*, vol. 10, no. 2, pp. 142–159, Apr. 2016.

[8] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing Flickr photos on a map," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2009, pp. 484–491.

[9] G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris, "Geotagging social media content with a refined language modelling approach," in *Proc. Pacific–Asia Workshop Intell. Secur. Inform.*, 2015, pp. 21–40.

[10] T.-H. Bui, Y.-J. Han, S.-B. Park, and S.-Y. Park, "Detection of POI boundaries through geographical topics," in *Proc. Int. Conf. Big Data Smart Comput. (BIGCOMP)*, Feb. 2015, pp. 162–169.

[11] C. Tran, D. D. Vu, and W.-Y. Shin, "An improved approach for estimating social POI boundaries with textual attributes on social media," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106710.

[12] T. H. V. M. Moura and C. A. Davis, "Integration of linked data sources for gazetteer expansion," in *Proc. 8th Workshop Geographic Inf. Retr.*, Nov. 2014, pp. 1–8.

[13] D. D. Vu, H. To, W.-Y. Shin, and C. Shahabi, "GeoSocialBound: An efficient framework for estimating social POI boundaries using spatio-textual information," in *Proc. 3rd Int. ACM SIGMOD Workshop Manag. Mining Enriched Geo-Spatial Data*, Jun. 2016, pp. 1–6.

[14] D. R. Montello, "Where's downtown?: Behavioral methods for determining referents of vague spatial queries," in *Spatial Cognition and Computation*. London, U.K.: Psychology Press, 2003, pp. 185–204.

[15] Y. Hu, H. Mao, and G. McKenzie, "A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements," *Int. J. Geograph. Inf. Sci.*, vol. 33, no. 4, pp. 714–738, Apr. 2018.

[16] P. D. Smart, C. B. Jones, and F. A. Twaroch, "Multi-source toponym data integration and mediation for a meta-gazetteer service," in *Proc. Int. Conf. Geographic Inf. Sci.*, 2010, pp. 234–248.

[17] C. E. C. Campelo and C. S. Baptista, "A model for geographic knowledge extraction on web documents," in *Proc. Int. Conf. Conceptual Modeling*, 2009, pp. 317–326.

[18] M. G. de Oliveira, C. E. Campelo, C. de Souza Baptista, and M. Bertolotto, "Leveraging VGI for gazetteer enrichment: A case study for geoparsing Twitter messages," in *Proc. Int. Symp. Web Wireless Geographical Inf. Syst.*, 2015, pp. 20–36.

[19] J. Gelernter, G. Ganesh, H. Krishnakumar, and W. Zhang, "Automatic gazetteer enrichment with user-geocoded data," in *Proc. 2nd ACM SIGSPATIAL Int. Workshop Crowdsourced Volunteered Geographic Inf.*, Nov. 2013, pp. 87–94.

[20] C. Keßler, P. Maué, J. T. Heuer, and T. Bartoschek, "Bottom-up gazetteers: Learning from the implicit semantics of geotags," in *Proc. Int. Conf. GeoSpatial Sematics*, 2009, pp. 83–102.

[21] C. B. Jones, R. S. Purves, P. D. Clough, and H. Joho, "Modelling vague places with knowledge from the web," *Int. J. Geographical Inf. Sci.*, vol. 22, no. 10, pp. 1045–1065, Oct. 2008.

[22] C. Grothe and J. Schaab, "Automated footprint generation from geotags with kernel density estimation and support vector machines," *Spatial Cognition Comput.*, vol. 9, no. 3, pp. 195–211, Aug. 2009.

[23] S. Intagorn and K. Lerman, "Learning boundaries of vague places from noisy annotations," in *Proc. 19th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2011, pp. 425–428.

[24] L. Li and M. F. Goodchild, "Constructing places from spatial footprints," in *Proc. 1st ACM SIGSPATIAL Int. Workshop Crowdsourced Volunteered Geographic Inf.*, Nov. 2012, pp. 15–21.

[25] D. R. Cox, "Regression models and life tables," *J. Roy. Stat. Soc., Ser. B, Methodol.*, vol. 34, pp. 189–220, Jan. 1972.

[26] E. A. Gehan, "Estimating survival functions from the life table," *J. Chronic Diseases*, vol. 21, nos. 9–10, pp. 629–644, Feb. 1969.

[27] X. Liu, *Survival Analysis: Models and Applications*. Hoboken, NJ, USA: Wiley, 2012.

[28] F. Emmert-Streib and M. Dehmer, "Introduction to survival analysis in practice," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 1013–1038, Sep. 2019.

[29] D. Ahlers, "Assessment of the accuracy of GeoNames gazetteer data," in *Proc. 7th Workshop Geographic Inf. Retr.*, Nov. 2013, pp. 74–81.

[30] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

[31] I. Lee, G. Cai, and K. Lee, "Exploration of geo-tagged photos through data mining approaches," *Expert Syst. Appl.*, vol. 41, no. 2, pp. 397–405, Feb. 2014.

[32] Y. Sun, H. Fan, M. Bakillah, and A. Zipf, "Road-based travel recommendation using geo-tagged images," *Comput. Environ. Urban Syst.*, vol. 53, pp. 110–122, Sep. 2015.

[33] W. Höpken, M. Müller, M. Fuchs, and M. Lexhagen, "Flickr data for analysing tourists' spatial behaviour and movement patterns: A comparison of clustering techniques," *J. Hospitality Tourism Technol.*, vol. 11, no. 1, pp. 69–82, 2020.

[34] C. M. Lee and J. J. Thomas, "Travel route recommendation based on geotagged photo metadata," in *Proc. Int. Vis. Informat. Conf.*, 2017, pp. 297–308.

[35] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[36] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. World Wide Web*, Apr. 2009, pp. 761–770.

[37] J. Zhang, S. Wang, and Q. Huang, "Location-based parallel tag completion for geo-tagged social image retrieval," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 355–362.

[38] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 43–56, Mar. 2016.

[39] Y.-T. Wen, P.-R. Lei, W.-C. Peng, and X.-F. Zhou, "Exploring social influence on location-based social networks," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 1043–1048.

[40] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen, "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos," *Data Knowl. Eng.*, vol. 95, pp. 66–86, Jan. 2015.

[41] S. Kisilevich, F. Mansmann, and D. Keim, "P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in *Proc. 1st Int. Conf. Exhib. Comput. Geospatial Res. Appl.*, Jun. 2010, pp. 1–4.

[42] C.-L. Kuo, T.-C. Chan, I.-C. Fan, and A. Zipf, "Efficient method for POI/ROI discovery using Flickr geotagged photos," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 3, p. 121, Mar. 2018.

[43] T. H. Bui, "Discovering shopping visitors' behavior and preferences using geo-tagged social photos: A case study of Los Angeles City," *J. Marketing Analytics*, vol. 9, no. 2, pp. 127–143, 2021.

[44] D. Lyu, L. Chen, Z. Xu, and S. Yu, "Weighted multi-information constrained matrix factorization for personalized travel location recommendation based on geo-tagged photos," *Int. J. Speech Technol.*, vol. 50, no. 3, pp. 924–938, Mar. 2020.

[45] T. Ameen, L. Chen, Z. Xu, D. Lyu, and H. Shi, "A convolutional neural network and matrix factorization-based travel location recommendation method using community-contributed geotagged photos," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 8, p. 464, Jul. 2020.

[46] Y. Yang, Z. Gong, and L. H. U, "Identifying points of interest by self-tuning clustering," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2011, pp. 883–892.

[47] Y. A. Lacerda, R. G. F. Feitosa, G. Á. R. M. Esmeraldo, C. D. S. Baptista, and L. B. Marinho, "Compass clustering: A new clustering method for detection of points of interest using personal collections of georeferenced and oriented photographs," in *Proc. 18th Brazilian Symp. Multimedia Web*, Oct. 2012, pp. 281–288.

[48] Y. Yang, Z. Gong, and L. Hou, "Identifying points of interest using heterogeneous features," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4, pp. 1–27, Jan. 2015.

[49] T.-H. Bui and S.-B. Park, "Point of interest mining with proper semantic annotation," *Multimedia Tools Appl.*, vol. 76, no. 22, pp. 23435–23457, Nov. 2017.

[50] J. Sun, T. Kinoue, and Q. Ma, "A city adaptive clustering framework for discovering POIs with different granularities," in *Proc. Int. Conf. Database Expert Syst. Appl.*, 2020, pp. 425–434.

[51] Y. Yang, Z. Gong, Q. Li, L. H. U, R. Cai, and Z. Hao, "A robust noise resistant algorithm for POI identification from Flickr data," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3294–3300.

[52] E. Pla-Sacristán, I. González-Díaz, T. Martínez-Cortés, and F. Díaz-de-María, "Finding landmarks within settled areas using hierarchical density-based clustering and meta-data from publicly available images," *Expert Syst. Appl.*, vol. 123, pp. 315–327, Jun. 2019.

[53] E. Kamalloo and D. Rafiei, "A coherent unsupervised model for toponym resolution," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1287–1296.

[54] W. Zhang and J. Gelernter, "Geocoding location expressions in Twitter messages: A preference learning method," *J. Spatial Inf. Sci.*, no. 9, pp. 37–70, 2014.

[55] J. Y. Rafiei and D. Rafiei, "Geotagging named entities in news and online documents," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 1321–1330.

THANH-HIEU BUI received the B.S. degree in computer science from the University of Technology—a member of Vietnam National University Ho Chi Minh City and the Ph.D. degree in computer science and engineering from Kyungpook National University, Republic of Korea. He is currently an Assistant Professor with the Department of Business Information Technology, College of Technology and Design, University of Economics Ho Chi Minh City, Vietnam. His research interests include machine learning, social media data mining, and business intelligence.