Finding Social Points of Interest from Georeferenced and Oriented Online Photographs

BART THOMEE, Yahoo Labs
IOANNIS ARAPAKIS, Yahoo Labs
DAVID A. SHAMMA, Yahoo Labs

Points of interest are an important requirement for location-based services, yet they are editorially curated and maintained, either professionally or through community. Beyond the laborious manual annotation task, further complications arise as points of interest may appear, relocate, or disappear over time, and may be relevant only to specific communities. To assist, complement, or even replace manual annotation, we propose a novel method for the automatic localization of points of interest depicted in photos taken by people across the world. Our technique exploits the geographic coordinates and the compass direction supplied by modern cameras, while accounting for possible measurement errors due to the variability in accuracy of the sensors that produced them. We statistically demonstrate that our method significantly outperforms techniques from the research literature on the task of estimating the geographic coordinates and geographic footprints of points of interest in various cities, even when photos are involved in the estimation process that do not show the point of interest at all.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Design, Algorithms, Performance, Measurement

Additional Key Words and Phrases: Points of interest; location estimation; georeferenced photos; oriented photos; line of sight; field of view; GPS; compass; sensor accuracy; photo composition

ACM Reference Format:
Bart Thomee, Ioannis Arapakis, and David A. Shamma, 2015. Finding social points of interest from georeferenced and oriented online photographs. *ACM Trans. Multimedia Comput. Commun. Appl.* V, N, Article A (January YYYY), 28 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

The world is filled with areas of interest, vista points, and places to visit. While historically lists of interesting places have been manually aggregated by sources like the UNESCO[1], the Lonely Planet[2], and Wikipedia[3], the task is time-consuming and laborious, especially if global coverage is desired. Furthermore, points of interest (POIs) may relocate or disappear over time, and new ones may be formed, requiring the annotation process to be periodically repeated. Events are, in essence, also POIs, albeit with a lim-

[1] http://whc.unesco.org/en/list/
[2] https://www.lonelyplanet.com/
[3] https://www.wikipedia.org/

Author's addresses: B. Thomee and D. A. Shamma, Yahoo Labs, San Francisco, CA, USA; I. Arapakis, Yahoo Labs, Barcelona, Spain. All URLs mentioned in this article were last accessed in September 2015.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© YYYY ACM 1551-6857/YYYY/01-ARTA $15.00
DOI:http://dx.doi.org/10.1145/0000000.0000000

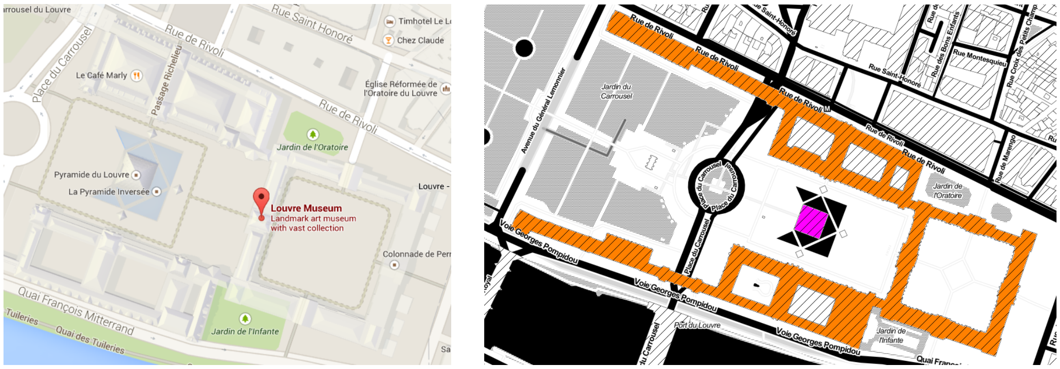


Fig. 1: The Musée du Louvre is only represented as a single geographic coordinate by Google Maps (left), whereas in reality it covers a relatively large area. Its geographic footprint on the surface of our planet (right) consists of two polygons, of which one (the pyramid, shown in magenta) is convex and the other (the palace, shown in orange) is concave and contains multiple holes.

ited lifetime. Check-ins on Foursquare, tweets sent on Twitter, and photos uploaded to Flickr reveal locations that people consider to be significant enough to create content, making social and community-based websites an ideal source for measuring location engagement. Techniques for automatic localization of POIs based on social signals can therefore assist, complement, or may even replace manual annotation.

A POI is traditionally represented by either its geographic center or by its bounding box. However, both these representations insufficiently capture the actual location of a POI. Namely, a geographical center does not give any indication of the actual landmass a POI covers, while a bounding box may greatly overstate its surface area. Furthermore, the geographic footprint of a POI is not necessarily formed by a single convex and holeless polygon (see Figure 1), such that its center may fall outside of its footprint. In this article we will focus on POI location estimation from two different perspectives:

(1) We address the problem of correctly localizing a POI on a map. We require localization techniques to produce a single coordinate per POI, where this coordinate ideally falls inside its geographic footprint, or at least is as close as possible to it.
(2) We address the problem of correctly capturing the landmass of a POI. We require localization techniques to produce one or more areas per POI, where their union ideally matches its geographic footprint, or at least overlaps it as much as possible.

People take and upload an astounding number of photos every day, covering a large portion of the globe (see Figure 2). To Facebook alone more than 250 billion photos have been uploaded, and it receives over 350 million new photos every single day on average [Facebook et al. 2013]. In recent years a variety of applications have been proposed that exploit the locations of where photos have been taken, ranging from characterizing places [Hollenstein and Purves 2010] and discovering events [Rattenbury et al. 2007] to constructing touristic itineraries [De Choudhury et al. 2010]. Considering that a substantial number of photos are travel-related [Ahern et al. 2007], many of these photos are likely to show POIs. We will leverage such photos to automatically and accurately determine the locations of POIs around the world.

Modern cameras, and in particular cameraphones, are often equipped with a GPS chip, a device that measures the position of the camera in terms of degrees longitude

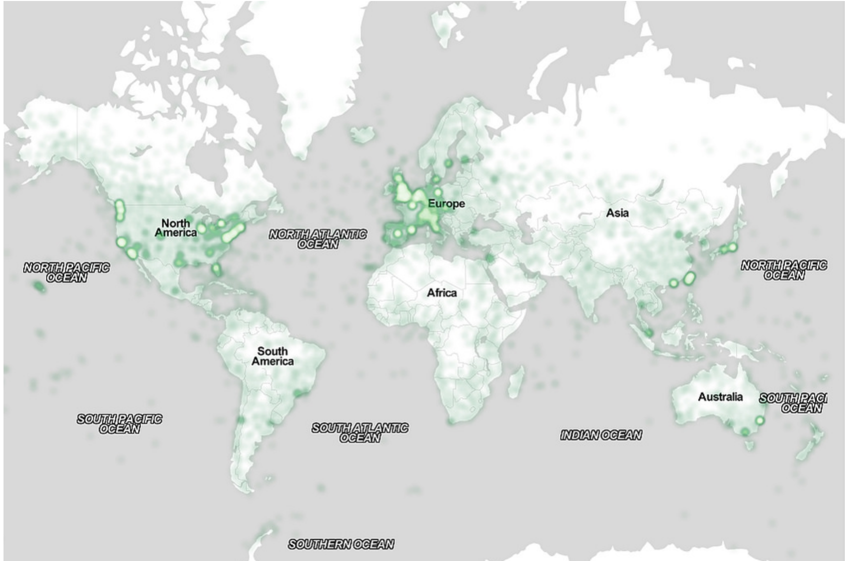


Fig. 2: Global coverage of a sample of georeferenced Flickr photos. *One Million Creative Commons Geo-tagged Photos* by David Shamma 🅭🅯⊜ https://flic.kr/p/o1Ao2o.

and latitude. Each time a photo is taken the camera stores the last known position as Exif metadata alongside the image pixels; a photo is *georeferenced* (or *geotagged*) when it is associated with a geographic coordinate. Camera position information can be used to estimate the location of a POI, e.g. by clustering the coordinates where photos were taken [Crandall et al. 2009]. The latest cameras also include a digital compass, a sensor that measures the direction the camera is facing, and upon photo capture the last known direction also gets embedded into the Exif metadata; a photo is *oriented* when it is associated with a compass direction. Together, the camera position and orientation information open up new opportunities for location estimation, e.g. by intersecting the fields of view of multiple cameras [Epshtein et al. 2007].

In this article we propose a novel method that probabilistically models the lines of sight of cameras and their intersections to estimate the location of a POI. Here, with *line of sight* we refer to the center line of the field of view, and with *field of view* to the extent of the world observable by the camera. Our method essentially first constructs an individual weight map for each camera based on its orientation and position, where each location in the map reflects how likely it was seen by the camera, and then combines all individual maps into a single collective weight map. To estimate the geographic coordinate of a POI our method then inspects the aggregated weight map to identify the location that was most often seen by all cameras, while to estimate the geographic footprint of a POI our method thresholds the aggregated weight map to define the extent of its boundaries. Our method, when modeling the lines of sight and their intersections, takes the inaccuracies of compass and GPS sensor measurements into account, as well as off-center framing of the POI by the photographer.

The remainder of this article is organized as follows. In Section 2 we first discuss related work. In Section 3 we investigate how people take photos of POIs in terms of camera position, camera orientation, and photo composition, and Section 4 we present our novel method in detail. In Section 5 we evaluate the performance of our method against several baselines from the literature on the tasks of coordinate and footprint estimation, and in Section 6 we finally conclude.

2. RELATED WORK

Gazetteers, such as GeoNames[4] and OpenStreetMap[5], contain millions of POIs contributed from various authoritative and non-authoritative data sources, ranging across landmarks, mountains, restaurants, schools and places of worship. While such gazetteers can be used to lookup where POIs are located in the world, researchers have analyzed various types of user-generated content to uncover locations that people themselves consider to be of importance [Mummidi and Krumm 2008; Rae et al. 2012]. They have further looked into separating places from events [Rattenbury et al. 2007; Quack et al. 2008; Rattenbury and Naaman 2009; Papadopoulos et al. 2011], as well as distinguishing between periodic and aperiodic events [Chen and Roy 2009].

Georeferenced photos have been used in a variety of research endeavors [Luo et al. 2011; Zheng et al. 2011], such as studying vernacular geography [Hollenstein and Purves 2010], visually describing the essence of cities [Doersch et al. 2012], organizing photo albums [Naaman et al. 2004], understanding where and how people travel [Popescu et al. 2009], as well as constructing travel itineraries [De Choudhury et al. 2010], to name just a few. At present, only a relatively small percentage of photos has a GPS position (≈8% based on a recent multi-million sample of public Flickr photos we inspected), and an even smaller percentage has a compass orientation (≈2%). This notwithstanding, their relative proportions are on the rise due to the addition of GPS and compass sensors to cameras, as well as due to the increased use of smartphones that have these sensors already built in. To address this lack of geographic information for the majority of photos, methods have been proposed to predict the location where a photo was taken [Serdyukov et al. 2009; Van Laere et al. 2010], as well as to estimate [Kosecká and Zhang 2002; Cham et al. 2010; Luo et al. 2010] and correct erroneous [Wang et al. 2013b] camera orientations. Early location estimation techniques achieved median errors on the order of 500km [Hays and Efros 2008], while a recent approach reduced that to just 2km [Popescu 2013]. Despite these promising advances, the predicted positions and orientations of photos are not yet reliable enough for applications that need street-level accuracy.

POIs often play a prominent role in research using georeferenced photos, where automatically detecting, recognizing and summarizing POIs by analyzing photos has been an active research topic for many years [Jaffe et al. 2006; Kennedy and Naaman 2008; Li et al. 2009; Zheng et al. 2009; Rudinac et al. 2011; Raguram et al. 2011]. To automatically detect POIs from sets of photos, clustering methods have typically been used to isolate areas of high photo density, for instance k-means [Rattenbury and Naaman 2009], X-means [Popescu and Shabou 2013], P-DBSCAN [Kisilevich et al. 2010], mean-shift [Crandall et al. 2009] and spectral [Yang et al. 2011] clustering, where each formed cluster may be considered a POI. Besides clustering, content-based analysis has also been employed to identify locations of interest, e.g. by integrating textual, visual, user, and cluster analysis [Popescu and Shabou 2013] to obtain accurate estimates.

The 3D reconstruction of scenes using georeferenced photos has attracted substantial attention in the last few years [Snavely et al. 2006, 2008; Agarwal et al. 2009; Wang et al. 2013a], where a 3D model is formed by triangulating feature correspondences between photos that were taken of the same scene from a variety of positions and orientations. By exploiting the positions of all contributing photos, the model itself can be properly positioned in the world as well, where additional data sources such as satellite imagery [Kaminsky et al. 2009] or Google Street View [Wang et al. 2013a] can further improve alignment to within a few meters of its actual location.

[4] http://www.geonames.org

[5] http://www.openstreetmap.org

While reconstructing a scene is a very computationally intensive task, recent efforts have managed to considerably speed up the process [Crandall et al. 2013; Heinly et al. 2015], with scenes containing up to 5,000 photos finishing in less than an hour using a high-end machine [Wu 2013]. Note that it is also possible perform a near-exact recovery of the actual position and orientation of a camera from a reconstructed scene; this can be done for any camera that captures a photo of the scene, whether or not it has a GPS or compass sensor itself.

In the literature we identified three articles that propose techniques exploiting the compass direction to perform POI discovery. The compass clustering method [Lacerda et al. 2012] detects POIs by clustering the intersections between the lines of sight that reflect the directions in which the photos were taken. The geo-relevance method [Epshtein et al. 2007] weights locations in the world by the frequency with which they lie within the fields of view of the cameras, after which the area with the highest frequency is identified. Two similar techniques were also presented by Hao et al. [2014], which principally focused on georeferenced and oriented video sequences. Our work also exploits the camera position and orientation, although we additionally take into account the possibility of errors introduced by the sensors that produced these measurements, as well as different styles of photo composition.

3. PHOTO CAPTURE PROPERTIES

Leading up to this article we had looked at the geographic distributions and the content of photos tagged with the name of a landmark (see Figure 3). We made three core observations, namely that (i) a substantial number of photos do not show the landmark at all, (ii) the cameras that captured the photos showing the landmark occasionally do not contain it within their fields of view according to their supposed position and orientation, and (iii) yet, the lines of sight of cameras that took landmark photos overall seem to converge to a single region of interest. The first of our observations can likely be attributed to people simply tagging all photos they took on the same day or in the same city with the name of the landmark, while the second observation suggests that field of view mismatches may be caused by position and orientation sensor errors, and/or due to how people frame landmarks in their photos. In this section we look deeper into sensor quality, as well as into photo composition rules, and perform an analysis of how landmark and non-landmark photos differ from each other.

3.1. Photo position and orientation

The precision of digital sensors supplying measurements such as longitude and latitude coordinates can vary depending on manufacturing quality, structural interference, atmospheric conditions, and signal reception. In Zandbergen [2008] the longitude and latitude error distributions of geographic coordinates were observed to be symmetrical with a peak around zero. The authors further noticed that the joint distribution was skewed with a peak around 2 meters, signifying that it was more similar to a Rayleigh distribution than to a bivariate normal distribution. This notwithstanding, the GPS position error distribution is often modeled as a normal distribution, a simplification we also make in this paper. An analysis of photos uploaded to Flickr revealed that mobile phones are frequently used as cameras[6]. According to a recent study, mobile phones provided inaccurate positioning with maximum errors exceeding 300m across a wide range of urban areas, where even with good visibility the maximum error could be greater than 100m [Paek et al. 2010]. Even though smaller positioning errors were reported in Zandbergen and Barbeau [2011], large errors were not uncommon either. The latter study also revealed that mobile phones suffer from larger

[6] http://www.flickr.com/cameras/

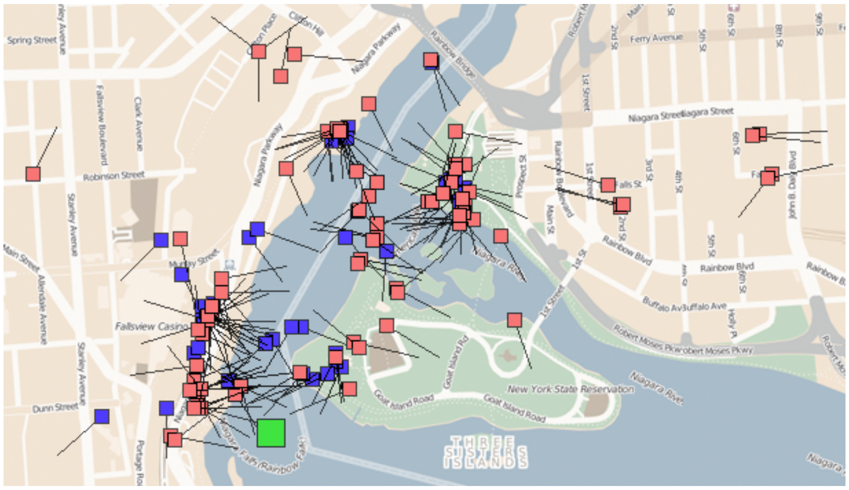


Fig. 3: Geographic distribution of photos tagged with the Horseshoe Falls. The blueish squares indicate the camera positions of the photos that actually depicted the waterfall, whereas the reddish squares indicate those that did not show it (e.g. they showed one of the other parts of the Niagara Falls). The large green square indicates the main location of the landmark. The straight lines extruding from the smaller squares indicate the direction in which the cameras were facing when the photos were taken.

positional errors than traditional consumer GPS devices. While the so-called assisted-GPS technology allows mobile phones to obtain a faster lock onto the GPS signal when the sensor is activated, the question remains whether the time between sensor activation and photo capture is sufficient for a reliable position estimate to be obtained; this has not yet been fully explored in the literature.

Compasses are sensitive to a variety of error sources that may affect the orientation measurement accuracy, such as magnetic interference, vibrations, and velocity during measurement [Ojeda and Borenstein 2000]. In our review of the literature on the accuracy of compass orientations provided by cameras and other mobile devices, we only found a single relevant study that analyzed the measurement errors of 7 mobile devices in a harsh indoor industrial environment [Hölzl et al. 2013]. While not necessarily representative for indoor and outdoor environments where photos are typically captured, in this light the analysis can be considered a worst-case scenario due to the presence of strong magnetic interference in areas of the industrial hall where the tests were conducted. The authors found that the digital compasses in the mobile phones performed comparably to a traditional magnetic compass, yielding a mean orientation error of approximately 22° and a standard deviation of 31°. The measurement errors were generally found to be less than 5°, and in 85% of the measurements the error was smaller than 20°, although the maximum error did reach up to 164°.

3.2. Photo composition

The physical characteristics of a POI and the environment in which it is situated influence the way in which photos can be taken of it. For example, the Eiffel Tower in Paris is typically better captured from far away than from up close, whereas the reverse is true for the small Manneken Pis statue in Brussels that can only be seen from at most 50m away. In a similar vein, some POIs can be seen from all directions, such as the Washington Monument, whereas others can only be seen from limited angles, such as the famous façades in the historical city of Petra in Jordan. In photography, there is a general thought that center-aligned shots are less aesthetically pleasing than aligning the focal point to the left or right third of the photo [Smith 1797]. The basic notion is that by aligning objects in the photo towards the golden ratio, one's eye will visually follow a narrative in a visual scene. While this is generally understood not to be a universal law [Field 1845], it is a widely accepted rule of thumb called the "rule of thirds," as coined by Smith [1797]. This rule states that the objects of interest should be located along the four inner intersections when cutting an image into three by three equally-sized blocks or along their inner edges, which is also evident as the grid in many camera and cameraphone viewfinders. Whether or not the rule of thirds is followed by the average photographer, we cannot assume the main point of interest to be necessarily located in the center of each photo.

3.3. Analysis

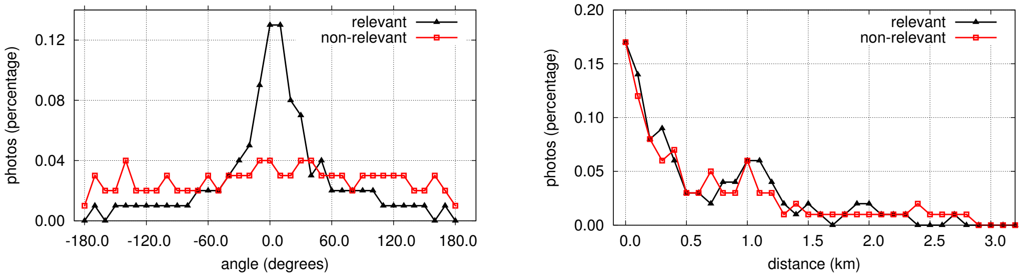
To investigate the interplay between the composition of a photo and its location and orientation measurements in the context of POIs, we queried the Flickr API for georeferenced and oriented photos taken within 2.5km of 20 famous landmarks, such as the St. Peter's Basilica in Vatican City and the Louvre in Paris, where at least one of the tags associated with a photo matched the landmark's official or alternative names in various languages (ignoring capitalization, diacritics and whitespace) as available in OpenStreetMap; we used the same 20 landmarks as in our preliminary work on this topic [Thomee 2013]. We manually inspected the downloaded photos and separated out those that clearly depicted the landmark from those that clearly did not show it. In order to mitigate the bias introduced by photographers who took many photos and by the characteristics of the cameras they used, we restricted the number of photos per photographer to at most 10 per landmark. We ended up with 1,215 (relevant) photos that showed a landmarks and with 1,018 (non-relevant) photos that did not.

To understand the joint effect of the compass orientation and photo composition on the positioning of the landmark, for each of the photos we measured the angle between the camera's actual orientation and the orientation in which it should have been held to capture the landmark in the center of the photo, see Figure 4(a). We see that the orientation difference distribution of non-relevant photos is almost flat, whereas the distribution of the relevant photos indicates that a large proportion of them points towards the landmark, about half of them within 20°. Still, almost 30% of the relevant photos have an orientation difference of more than 60°; given that the field of view of camera lenses typically ranges between 40° and 60°, larger differences therefore are unlikely to have been the result of photo composition alone.

We also inspected the distances between the camera positions and the geographical center of the landmarks, see Figure 4(b). As can be observed, most photos appear to be taken within close range of the landmarks, although those taken 1–3 kilometer away are not uncommon either. Given that the effect of any orientation error increases over distance—to illustrate, an error of 10° would lead to a deviation of about 17m for every 100m traveled—it may thus be worthwhile to place a limit on the distance beyond which a captured photo has no influence. The distance distributions of the relevant and non-relevant photos are, however, comparable, and do not appear to yield a useful signal that would allow us to easily distinguish between the two.

4. METHOD

In this section we present a robust method for accurately estimating the location of a POI given a set of photos, solely by using the position and orientation measurements that were embedded into the Exif metadata by the cameras that took the photos. The



(a) (b)

Fig. 4: Distribution of the relative camera–POI orientation angles (a) and distances (b) for relevant and non-relevant photos.

design of our method has been principally driven by the insights we obtained in Section 3, for it to successfully cope with erroneous sensor data and off-center framing.

4.1. Definitions

We define the set of photos for a certain POI as P, and the bounding box that encloses these photos as B. This bounding box could in an extreme case cover the entire surface of the planet, but without loss of generality we assume that it is much smaller; in particular since the visibility of a POI by a camera is limited due to the curvature of the Earth, as well as due to possible structural interference from natural elements and manmade objects. Each photo $p \in P$ is represented by a tuple (υ_p, θ_p), where υ_p is the camera position of the photo, measured in radians longitude $\upsilon_p^\lambda \in [-\pi, \pi)$ and latitude $\upsilon_p^\phi \in [-\frac{\pi}{2}, \frac{\pi}{2})$, and $\theta_p \in [0, 2\pi)$ is the camera orientation of the photo, measured in radians clockwise from true north[7]. With respect to a certain photo p, any point $\upsilon_b \in B$ can be associated with an orientation θ_b, which is the compass direction in which the camera should have been held in order to frame the point in the center of the photo. We further define the distance in kilometers between the positions υ_p and υ_b as $\delta(\upsilon_p, \upsilon_b)$, and the angle between the orientations θ_p and θ_b as $\alpha(\theta_p, \theta_b)$, as is illustrated in Figure 5. We consider the shape of our planet as an oblate ellipsoid according to the WSG84 specification and apply Karney's formula [Karney 2013] to compute $\delta(\upsilon_p, \upsilon_b)$ as the length of the shortest path over the planet's surface. Applying Karney's formula yields the orientation θ_b as a byproduct of computing $\delta(\upsilon_p, \upsilon_b)$. To obtain the angle $\alpha(\theta_p, \theta_b) \in [0, \pi]$ we then only need to consider the smallest angular difference between θ_p and θ_b, given by

$$\alpha(\theta_p, \theta_b) = \pi - \big|\, |\theta_b - \theta_p| - \pi \big| \qquad (1)$$

4.2. Coordinate-based location estimation

We can trivially produce an accurate coordinate estimate for a POI when all photos in the set have precise location and orientation measurements. Namely, from each position where a photo was captured we can first trace a line of sight along the direction

[7] The compass sensor embedded in a photo camera provides orientations that refer either to true north or to magnetic north. To avoid incompatible orientations due to the effect of magnetic declination (the angle between true north and magnetic north, which varies across location and time) we converted magnetic north orientations to true north using the 11th generation International Geomagnetic Reference Field (IGRF) model (http://www.ngdc.noaa.gov/IAGA/vmod/igrf.html).

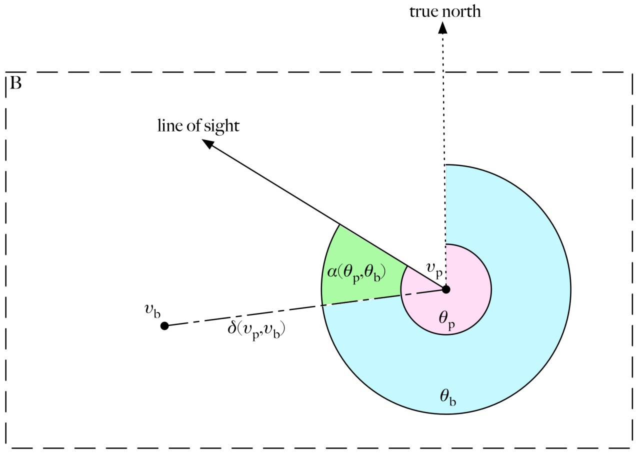


Fig. 5: Within bounding box B the camera is positioned at υ_p and is oriented θ_p radians clockwise from true north (represented by the pinkish area), as is also reflected by its line of sight. There is another point shown, having position υ_b and orientation θ_b (represented by the blueish area), where the latter reflects the orientation the camera should have had for it to look straight at the point. The distance $\delta(\upsilon_p, \upsilon_b)$ is computed between their positions, and measures how far away the camera and the point are from each other. The angle $\alpha(\theta_p, \theta_b)$ (represented by the greenish area) is computed between their orientations, and measures how much the camera should rotate from its current orientation to focus on the point.

in which the camera was facing, and the location estimate is then the geographic coordinate where all lines of sight intersect, as is illustrated in Figure 6(a). However, as we discovered in Section 3, position and orientation measurements may not be accurate, and photographers do not necessarily frame a POI in the center of a photo, if at all. The lines of sight may thus not coincide at a single location, as is illustrated in Figure 6(b), making it non-trivial to produce an accurate location estimate under real-world conditions. To address this, our method probabilistically models the lines of sight and their intersections by taking the variance of sensor measurements and the variability of photo composition into account. The underlying idea is that, even when the sensor measurements or the composition of individual photos may not be completely accurate or may even be completely wrong, the probabilistically modeled lines of sight will still collectively converge to a single location. Our technique effectively creates a weight map, where each point in the map reflects the probability of the POI being located there, such that its most likely location is the point with the highest weight.

Our approach consists of the following five steps:

1. Orientation modeling: For each photo p, we apply a weight to each point $\upsilon_b \in B$ depending on the angle $\alpha(\theta_p, \theta_b)$ in order to account for errors introduced by the sensor that provides the orientation measurement, as well as any displacement due to off-center framing of the POI by the photographer. We model the orientation measurement error using a one-dimensional half-normal distribution with standard deviation σ_θ as

$$G_\theta(\theta_p, \theta_b; \sigma_\theta) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma_\theta} e^{-\frac{\alpha(\theta_p,\theta_b)^2}{2\sigma_\theta{}^2}} \tag{2}$$

which effectively applies larger weights to points closer to the line of sight than to those further away, as is shown in Figure 7(a). As a special case for sensors that pro-

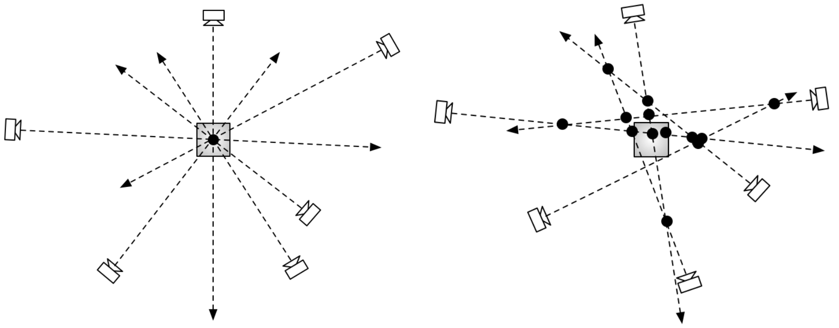


Fig. 6: Estimating the location of a POI by intersecting the lines of sight from multiple cameras. This can be done accurately when the camera position and orientation measurements are precise and the POI is framed in the center of each photo (a), but it cannot be done accurately otherwise (a).

duce perfect orientation measurements, i.e. $\sigma_\theta = 0$, we substitute the half-normal distribution in Equation 2 by the Kronecker delta, using Iverson bracket notation, $G_\theta(\theta_p, \theta_b; \sigma_\theta) = [\alpha(\theta_p, \theta_b) = 0]$, such that only points along the line of sight receive a non-zero weight. Applying Equation 2 to each point $\upsilon_b \in B$ for a given photo p produces a two-dimensional weight map W_{θ_p}, where we place its origin at υ_p.

2. Position modeling: We also apply a weight to each point $\upsilon_b \in B$ depending on the distance $\delta(\upsilon_p, \upsilon_b)$ in order to account for errors introduced by the sensor that provides the camera position measurement. We also model the position measurement error using a one-dimensional half-normal distribution with standard deviation σ_υ as

$$G_\upsilon(\upsilon_p, \upsilon_b; \sigma_\upsilon) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma_\upsilon} e^{-\frac{\delta(\upsilon_p, \upsilon_b)^2}{2\sigma_\upsilon{}^2}} \tag{3}$$

which effectively applies a weight to points depending on how far away they are from the camera position, with larger weights applied to points closer to the camera position than to those farther away, see Figure 7(b). As a special case for sensors that produce perfect position measurements, i.e. $\sigma_\upsilon = 0$, we substitute the half-normal distribution in Equation 3 by the Kronecker delta $G_\upsilon(\upsilon_p, \upsilon_b; \sigma_\upsilon) = [\delta(\upsilon_p, \upsilon_b) = 0]$, using Iverson bracket notation, such that only the camera position itself receives a non-zero weight. Applying Equation 3 to each point $\upsilon_b \in B$ for a given photo p produces a two-dimensional weight map W_{υ_p}, where we place its origin at υ_p.

3. Weight convolution: We then obtain a combined weight map W_p for a particular photo p by convolving the orientation error weights with the position error weights, i.e.

$$W_p = W_{\theta_p} * W_{\upsilon_p} \tag{4}$$

which ensures each point $\upsilon_b \in B$ is affected by both the camera orientation and position measurement errors, see Figure 7(c). While the convolution produces an output map larger than both input maps, we trim its edges such that W_p ends up the same size as W_{θ_p} and W_{υ_p}, and thus exactly covers the same area as the bounding box B.

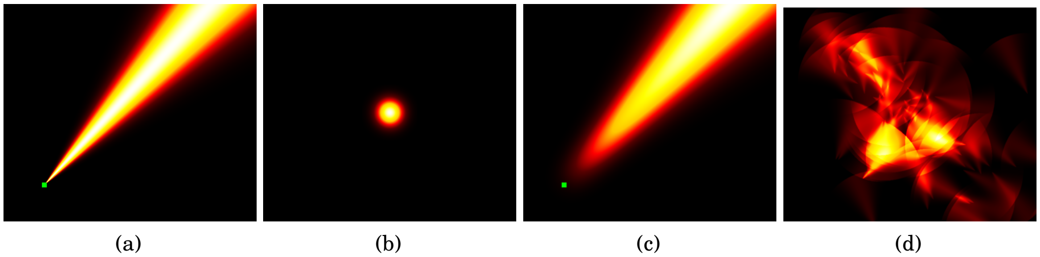


Fig. 7: Modeling the camera orientation measurement error (a) and camera position measurement error (b) of a photo using half-normal distributions, after which they are combined through convolution (c). By aggregating the resulting weights over all photos a weight map is obtained that captures the likelihood of a POI being present at each point in the map (d). The aggregated weight map shown here is generated from 82 photos tagged with the name of the Petronas Twin Towers in Kuala Lumpur, where the colors represent weight magnitudes, ranging from low (black, red) to high (yellow, white). The two dense areas near the center each refer to a tower, while the less dense area in the top-left refers to another tall building, the Menara Public Bank, of which also photos were taken (despite being tagged with the name of the Petronas towers).

4. Weight aggregation: We sum the combined weights computed for each point $\upsilon_b \in B$ across all photos $p \in P$ to obtain the final weight map $\mathbb{W}_P$, where the weight associated with each coordinate in $\mathbb{W}_P$ ultimately reflects the likelihood of a POI being situated at that location, as is shown in Figure 7(d).

5. Coordinate estimation: Given that we only consider a set of photos that refer to a specific POI, the optimal location estimate $\hat{\upsilon}_P$, is to select the point in the aggregated weight map with the highest weight. When multiple points share the highest weight a possible solution is to pick their average as the location estimate, which is what we do in this article. More formally, we define the set of maxima as

$$\Upsilon_P = \arg\max_{\upsilon} \mathbb{W}_P(\upsilon) \tag{5}$$

and then find the average location from this set, given by

$$\hat{\upsilon}_P = \frac{\sum_{\upsilon \in \Upsilon_P} \upsilon}{|\Upsilon_P|} \tag{6}$$

where the averaging is applied to the longitude and latitude coordinates separately. To produce a meaningful location estimate our method requires at least two lines of sight to intersect, although we can still produce an estimate when no intersections occur by averaging the points that share the highest weight (i.e. those that lie along all lines of sight), provided the lines of sight are constrained to be of finite length.

The parameters σ_θ and σ_υ essentially affect the amount of orientation and location smoothing applied to the sensor measurements. A threshold χ can further be applied to the length of the line of sight to prevent weights being computed for distant points where a POI is unlikely to be present or visible, i.e. each photo p is effectively limited to its own local bounding box $B_p = \{\upsilon_b : \upsilon_b \in B, \delta(\upsilon_p, \upsilon_b) \in [0, \chi]\}$ rather than the global bounding box B. We compute the optimal values of the parameters σ_θ, σ_υ and χ in Section 5. Note that the only situation in which our method fails is when a camera is exactly positioned at the North Pole, since at that location θ_p is undefined.

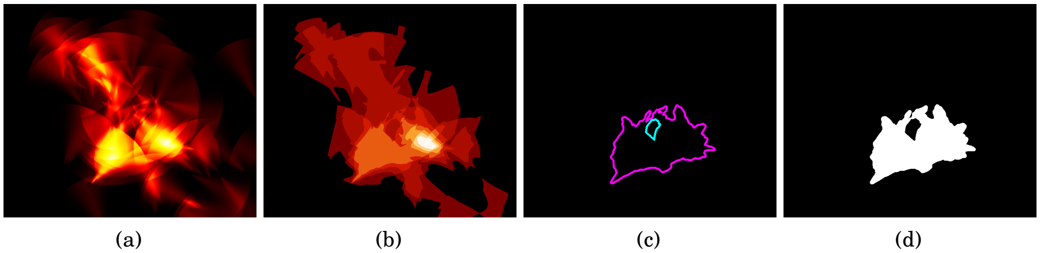


Fig. 8: We convert a normalized weight map (a) into a binary map through thresholding. To illustrate the effect of thresholding we show a composite image that contains the binary maps produced for various thresholds (b). Given the binary map produced for a certain threshold we then trace any outer contours (shown in magenta) and inner contours (shown in cyan) (c). The final footprint estimate is formed by subtracting the inner areas from the outer areas (d); the footprint shown here reveals two connected areas, each referring to a tower of the Petronas Twin Towers.

4.3. Footprint-based location estimation

To produce an estimate of the landmass occupied by a POI, we analyze the aggregated weight map to identify one or more areas that have sufficient support for the presence of the POI, where these areas together form its estimated footprint. Specifically, we want to apply a suitable threshold ω to the weight map in order to only retain the high-density areas bounded by the level curve $\mathbb{L}_P(\omega) = \{v : v \in \mathbb{W}_P, \mathbb{W}_P(v) = \omega\}$ that together ideally match the actual footprint of the POI as closely as possible. To achieve this, we apply the following three steps in our implementation:

1. Binary thresholding: The range of the weights in $\mathbb{W}_P$ depends on the number of photos in P and the extent of which their lines of sight intersect. In order to be able to apply a threshold from a fixed range $\omega \in [0, 1]$ we first normalize the weight map. We create the binary map $\mathbb{B}_P$ that separates the high density spatial activity found in $\mathbb{W}_P$ from the low density activity, according to

$$\mathbb{B}_P(\omega) = \left\{ \left[\frac{\mathbb{W}_P(v)}{\max(\mathbb{W}_P)} \geq \omega \right] : v \in \mathbb{W}_P \right\} \tag{7}$$

using Iverson bracket notation. We illustrate this in Figure 8(b) for varying thresholds.

2. Area isolation: We apply contour tracing [Chang et al. 2004] to the binary map to extract the outer contours $\mathbb{C}^+_P(\omega)$ and any inner contours $\mathbb{C}^-_P(\omega)$ that delineate the areas where the weight crosses the threshold, as is shown in Figure 8(c). These outer and inner contours envelope the areas $\mathbb{A}^+_P(\omega)$ and $\mathbb{A}^-_P(\omega)$, respectively. To exclude spurious high-density areas that are unlikely to cover any actual landmass of the POI, we omit any outer areas that do not include at least one of the points at which the maximum weight in the aggregated weight map was observed. No areas are produced when all weights are below the threshold.

3. Footprint estimation: The POI footprint estimate $\hat{\mathbb{F}}_P$ is finally formed by subtracting the inner areas from the outer areas, i.e.

$$\hat{\mathbb{F}}_P(\omega) = \mathbb{A}^+_P(\omega) \setminus \mathbb{A}^-_P(\omega) \tag{8}$$

as is illustrated in Figure 8(d).

This approach can produce POI footprints of arbitrary shape, which may contain holes. In Section 5, after having computed the optimal values for σ_θ, σ_υ and χ, we will determine the threshold ω that overall generates the best fitting footprint for a POI.

4.4. Continuous vs. discrete representation of the world

Up until now we have presented our method from a theoretical point of view that operates on the world as we know it, in which geographic coordinates are represented as a bounded function of continuous longitude and latitude values. However, in practice they are expressed in the discrete domain, in particular because digital global positioning systems are discrete by nature. In this article we represent our bounding box B as a two-dimensional histogram along longitude and latitude using finite addressability, where each cell in the histogram measures $5\text{E-}5^\circ$ longitude by $5\text{E-}5^\circ$ latitude, which at the equator is approximately 5.5m by 5.5m. Each $\upsilon_b \in B$ refers to a unique cell, where its position υ_b is represented by the geographic center of the cell. Each photo $p \in P$ is assigned to the cell containing its longitude and latitude coordinates.

5. EXPERIMENTS

We used a semi-automatic approach to create a collection of POIs for evaluation. In contrast with existing work we do not pick a single geographic coordinate to represent the location of a POI, since doing so is not always meaningful, e.g. the geographic center of Blenheim Palace lies outside the building, while the Golden Gate Bridge spans a large geographic area. Instead, we extracted the spatial (polygonal) footprints of buildings from a 2013 database dump of OpenStreetMap, discarding those that had either broken polygons or were represented by only a single coordinate, yielding a total of about 1 million footprints. We manually readded the footprints of a number of famous landmarks that had been discarded earlier. We then processed a snapshot of over 1 billion public Flickr photos and extracted all georeferenced and oriented photos taken within 2.5km of one of the buildings and tagged with its name. Since GPS signals tend to not reach into buildings, the geographic location reported by a camera is often inaccurate for photos taken indoors. We therefore used a deep learning approach to automatically determine whether a photo was taken indoors or outdoors[8], and only kept those that had been confidently classified as being both outdoors and non-indoors. We excluded the photos we had used earlier and once more restricted the number of photos per photographer to at most 10, while ensuring that photos of at least 4 different photographers were included per building. We ultimately ended up with 105 landmarks located all around the world, see Online Appendix A for a complete graphical listing. The landmarks have varying characteristics in terms of area size, height and shape, which makes it challenging to accurately estimate their coordinates and footprints. The total number of photos collected per landmark ranges between 15 and 1913, with an average of 321 photos.

5.1. Parameter exploration

In order to account for camera position and orientation measurement errors, and the variety in photo composition styles on the coordinate and footprint estimates produced by our method, we perform a two-stage parameter exploration. In the first stage we

[8] We applied an off-the-shelf deep convolutional neural network [Krizhevsky et al. 2012] with 7 hidden layers, 5 convolutional layers and 2 fully connected ones. The penultimate layer of the convolutional neural network output was employed as the image feature representation to train the visual concept classifiers. We used Caffe [Jia et al. 2014] to train an indoor and an outdoor classifier, each being a binary SVM, using photos taken from the entire Flickr corpus; 50,000 positive examples were crowd labeled or handpicked based on targeted search/group results, while the same number of negative examples were drawn from a general pool. We tuned each classifier such that it achieved at least 90% precision on a held-out test set.

Table I: The explored orientation error (σ_θ), position error (σ_υ), line of sight length (χ), and normalized density threshold (ω) parameter values.

	Explored parameter values
σ_θ	0.00, 0.05, 0.10, 0.20, 0.30
σ_υ	0.00, 0.50, 1.00, 2.50, 5.00 ($\times 10^{-4}$)
χ	0.05, 0.10, 0.20, 0.50, 1.00, 1.50, 2.00, 2.50, 5.00
ω	0.02, 0.04, 0.06, . . . , 0.96, 0.98, 1.00

Table II: Number of randomly sampled photos (**P**) and landmarks (**L**) included in each subset bin for the coordinate and the footprint parameter exploration stages, as well as for the performance evaluation against the baseline methods. We repeat the random sampling (**R**) five times for each bin except for the bins that contain all photos of all landmarks. The weights (**W**) are used during rank aggregation in both parameter exploration stages, and represent the relative proportion of estimates in each bin.

		Subset bins									
Coordinate-based parameter exploration	**P**	5	10	20	30	40	60	80	100	200	all
	L	50	50	49	48	43	34	28	21	11	50
	R	5	5	5	5	5	5	5	5	5	1
	W	0.15	0.15	0.14	0.14	0.13	0.10	0.08	0.06	0.03	0.03
Footprint-based parameter exploration	**P**	5	10	20	30	40	60	80	100	200	all
	L	30	30	29	26	25	18	16	16	7	30
	R	5	5	5	5	5	5	5	5	5	1
	W	0.15	0.15	0.14	0.13	0.12	0.09	0.09	0.08	0.03	0.03
Performance evaluation	**P**	5	10	20	30	40	60	80	100	200	all
	L	25	25	25	24	22	17	11	10	7	25
	R	5	5	5	5	5	5	5	5	5	1

determine the optimal parameter values for σ_θ, σ_υ, and χ, and aim to understand their interactions with each other and their influence on the accuracy of the coordinate location estimate for a landmark. In the second stage we do the same for ω in terms of the footprint produced for a landmark. Guided by our earlier observations in Section 3, we explore the parameter values listed in Table I. The values we ultimately select for the four parameters ideally produce an accurate location estimate for any given landmark, irrespective of the number of photos that were taken of it. This notwithstanding, we expect our method to be more forgiving (large σ_θ, large σ_υ, small χ) when only few photos are available, since in that case any sensor errors may have a disproportionally strong impact on the accuracy of the location estimate. We therefore also investigate how the optimal parameter values differ for varying numbers of photos.

We randomly split our collection of 105 landmarks into 50 for the first stage, 30 for the second stage, and the remaining 25 will be used for assessing the accuracy of both our coordinate and footprint location estimation methods using the optimal parameter values. We create random subsets (bins) ranging between 5 and 200 photos for each landmark; a landmark is omitted from a bin when it does not have a sufficient number of photos available. We repeat the random sampling five times for each bin. This grouping reduces the sources of variability in our dataset and the likelihood that the observed effects are due to confounding factors, and thus leads to greater accuracy. We further include an additional bin that contains all photos available for all landmarks. Essentially, each bin $s \in S$ contains a number of landmarks $l \in L_s$ for which we created one or more random samplings $r \in R_s$ from the set of photos that were tagged with its name. We show an overview in Table II.

Table III: Correlation analysis of the rankings produced in the coordinate parameter exploration stage for each bin using the Kendall's tau (τ) coefficient test.

	5	**10**	**20**	**30**	**40**	**60**	**80**	**100**	**200**	**all**
5	1	.399***	.280***	.299***	.329***	.280***	.144**	-.083	-.313***	.134**
10		1	.353***	.309***	.278***	.325***	.160***	-.084	-.221***	.141**
20			1	.322***	.273***	.211***	.214***	.021	-.140**	.067
30				1	.295***	.350***	.137**	.023	-.190***	.101*
40					1	.212***	.223***	.008	-.175***	.112*
60						1	.152***	.034	-.148***	.167***
80							1	.035	-.091*	.101*
100								1	.132**	.030
200									1	-.085

*. Correlation is significant at the .05 level (2-tailed).
**. Correlation is significant at the .01 level (2-tailed).
***. Correlation is significant at the .001 level (2-tailed).

5.1.1. Coordinate parameter exploration stage. We perform **subset ranking** in which we evaluate all 225 $(\sigma_\theta, \sigma_\upsilon, \chi)$ parameter configurations for each of the 10 different coordinate subset bins $S_c \subset S$. Our method produces a coordinate estimate $\hat{\upsilon}_{lr}$ for each random sampling of photos for a landmark in a bin. Given that our main objective is to estimate the locations of the landmarks in the test set with minimal error, we measure the performance of our method for a given parameter configuration by computing the mean absolute error (MAE) over all the coordinate estimates in bin $s \in S_c$, given by

$$MAE_s = \frac{1}{|L_s||R_s|} \sum_{l \in L_s} \sum_{r \in R_s} \min_{\upsilon_l \in \mathbb{F}_l} \delta(\hat{\upsilon}_{lr}, \upsilon_l) \tag{9}$$

where for each estimate we compute the distance to its closest coordinate on the footprint of the landmark $\mathbb{F}_l$. For each bin we produce a different ranking of parameter configurations, ordered by the MAE.

To determine to what extent the parameter configurations are specific to a bin, we perform pairwise comparisons and compute the similarity of the rankings using the Kendall's tau (τ) coefficient. In Table III we report several statistically significant correlations between the rankings computed for each bin. Of interest is the inverse relationship observed between the correlation strength and the number of photos used for computing the location estimates. This relationship indicates that the number of photos available is a determining factor for whether a particular parameter configuration will likely perform well. This notwithstanding, for the bins containing 60 or less photos we observe several positive, medium-size correlations that indicate similarities in the rankings, which suggests that certain parameter configurations are more tolerant to the availability of photos and are capable, to some extent, of performing well irrespective of that. When we assume a qualitative approach and compare the top configurations produced for each bin, we observe that as the number of photos in a bin changes, better performing configurations tend to have the following parameters:

(i) $\sigma_\theta = 0.20$ to 0.30 for 5 photos vs. $\sigma_\theta = 0.00$ to 0.10 for $\geq$10 photos.
(ii) $\sigma_\upsilon = 0.00$ to 1.00 for $\leq$10 photos vs. $\sigma_\upsilon = 2.50$ to 5.00 for $\geq$20 photos.
(iii) $\chi = 0.10$ to 0.50 for $\leq$10 photos vs. $\chi = 1.00$ to 5.00 for $\geq$60 photos.

The results are mostly in concordance with our earlier hypothesis that our coordinate-based estimation method needs to be more flexible (large σ_θ, small χ) when fewer photos are available, while it can be more strict (small σ_θ, large χ) when more photos are available. We surmise that a large value of σ_θ already causes enough smoothing to not additionally require a large value of σ_υ, and vice versa.

In order to pick a single overall best configuration we first perform **rank aggregation** to select the parameter configuration that is ranked highly in as many bins as possible. The problem of selecting a parameter configuration that is optimal for the general case of estimating the location of a landmark is a typical instance of a *multiple-winner voting problem*. Although the *Borda Count* and the *Condorcet* methods are the most widely known and studied techniques for such problems, we choose rank aggregation [Dwork et al. 2001], which has ties with the aforementioned methods, but is better at filtering out noise. For rank aggregation we use an implementation[9] for R with the settings recommended by Pihur et al. [2007]. We can formalize our goal within the framework of the following minimization problem. Find κ^* such that

$$\kappa^* = \arg\min_{\kappa} \sum_{m \in M} d(\kappa, E_m) \tag{10}$$

where E_m is an ordered list of objects produced by a measure $m \in M$, d an appropriate distance function, and the minimization is carried out over all possible ordered lists κ of size $|E_m|$. In our case, objects are the parameter configurations and each ranked list κ is a bin. We rank the lists in ascending order according to the MAE that each parameter configuration achieved, and subsequently produce their aggregated ranking using the Monte Carlo cross-entropy algorithm [de Boer et al. 2005; Rubinstein and Kroese 2005]. We apply a weighted aggregation that optimizes a distance criterion, e.g. Kendall's tau and Spearman's Footrule distance, and allows for a far more objective and automated assessment of the results. We use the relative proportion of estimates in each bin (see Table II) as weights; since the locations are estimated for many more landmarks in the smaller bins than in the larger bins, we assign the former more importance during rank aggregation than the latter. The result is a list with the top positions granted to the parameter configurations that overall performed best across all bins. We finally select the one that has the lowest average MAE across all bins, which has as parameter values $\sigma_\theta = 0.00$, $\sigma_\upsilon = 5.00$ and $\chi = 0.50$. In Figure 9 we illustrate the performance of our method using these values on the training set of 50 landmarks; the actual evaluation using the test set of 25 landmarks is presented in Section 5.3.

5.1.2. Footprint parameter exploration stage. To determine the optimal value of the normalized density threshold w we again perform subset ranking, this time on all 50 $(\omega; \sigma_\theta = 0.00, \sigma_\upsilon = 5.00, \chi = 0.50)$ parameter configurations for each of the 10 different footprint subset bins $S_f \subset S$. Our method produces a footprint estimate $\hat{\mathbb{F}}_{lr}$ for each random sampling of photos for a landmark in a bin. We compute the mean absolute error over all the footprint estimates in bin $s \in S_f$, given by

$$MAE_s = \frac{1}{|L_s||R_s|} \sum_{l \in L_s} \sum_{r \in R_s} J(\hat{\mathbb{F}}_{lr}, \mathbb{F}_l) \tag{11}$$

where $J(\hat{\mathbb{F}}_{lr}, \mathbb{F}_l)$ is the Jaccard index between the footprint estimate and the footprint of the landmark. After performing subset ranking and rank aggregation based on the MAE, we obtain our overall best footprint-based location estimation parameter configuration with the value $\omega = 0.8$, which achieves an average of 13% overlap with the footprints of the POIs in the training set, see Figure 9.

[9] http://cran.r-project.org/web/packages/RankAggreg/

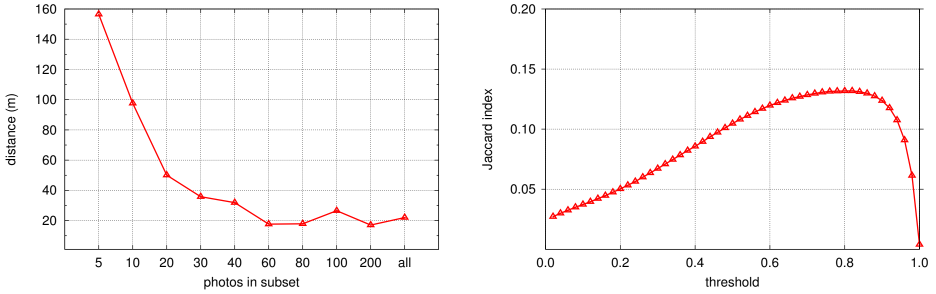


Fig. 9: Distance (MAE) between the footprints of the 50 POIs in the first training set and the coordinate estimates produced by our method for different photo subsets (left). Overlap (MAE) between the footprints of the 30 POIs in the second training set and the footprint estimates produced by our method for varying thresholds ω (right). Our methods used the overall best parameter configuration ($\sigma_\theta = 0.00$, $\sigma_\upsilon = 5.00$, $\chi = 0.50$).

5.2. Baselines

To place the actual location estimates produced by both our methods into context, we compare against several baselines from the research literature. To select our baselines, we predominantly focused on techniques for which good parameters (if needed) have been reported in the literature in the context of POI detection and for which verified open source code was available, or either implemented in open source toolkits such as for instance Weka; most methods proposed in the literature are based on clustering. Given that we aim to estimate POI locations solely based on capture metadata, techniques that required analyzing and classifying visual content (e.g. [Popescu and Shabou 2013]) fall outside the scope of the current paper.

5.2.1. Averaging-based technique. The **averaging** method considers the average location from where photos were taken as the location estimate of the POI. This method assumes that on average people will take photos from all positions around a POI. Note that the averaging method cannot produce a footprint estimate.

5.2.2. Clustering-based techniques. Clustering-based techniques have an advantage over the averaging baseline in that they incorporate a mechanism for automatically detecting the number of clusters present in the photo distributions, which enables them to more easily identify supposed outliers and to focus only on the photos that are more likely to be actually taken of the POI. For each of the following baselines we retain only the cluster to which the highest number of photos has been assigned. The center of this cluster is then considered as the estimate of the location of the POI, and the surface area covered by the cluster as its footprint. In case multiple clusters share the highest number of photos, the coordinate estimate is formed by averaging their centers, while the footprint estimate is formed by the union of their surface areas.

EM [Dempster et al. 1977] starts with a single cluster and randomly splits the dataset into a number of folds. The approach then computes a probability distribution for each data point within each fold that indicates to what extent it belongs to each cluster. If the average log-likelihood of the data across all folds increases with respect to the previous iteration, the number of clusters is increased and the process repeats.

X**-means** [Pelleg and Moore 2000] is an extension of k-means. The approach iteratively applies k-means to cluster the data and then determines which cluster(s), if any, should be split into two to improve the local structure using the Bayesian Information Criterion (BIC). We initialize the clustering process with 1 cluster and let it run until termination. The final clustering is the one with the highest global BIC score.

DBSCAN [Ester et al. 1996] forms clusters of points where each is surrounded by a sufficient number of nearby other points, unless the point itself is a border point. In the same way as for our own methods, we determine the optimal parameters by exploring a range of values reported in the literature[10], yielding the values $MinPts = 3$ and $Eps = 0.002°$.

P-DBSCAN [Kisilevich et al. 2010] is an extension of DBSCAN and is specifically designed for clustering georeferenced photos. The number of unique users in a cluster plays an important role, while the method also takes the local density around a photo into account for deciding whether to add nearby photos to a cluster. Our parameter exploration[11] yielded the values $MinOwners = 3$, $Eps = 0.005°$ and $Addt = 0.10$.

5.2.3. Line of sight-based techniques. In the literature we encountered two techniques that exploited the direction in which the camera was facing. One of these techniques is compass clustering [Lacerda et al. 2012; Hao et al. 2014], although both papers leave out vital information regarding the underlying clustering technique used and its parameterization, making this technique difficult to reproduce. The other method, described below, was not formalized in its respective paper nor did it evaluate its performance. Nonetheless, we included it in our evaluations because it is closely related to our method and it was straightforward to implement.

Geo-relevance [Epshtein et al. 2007] weights locations by the frequency of which they lie within the fields of view of the cameras that took the photos. A field of view (FOV) depends on several factors, such as camera model and type of lens, and is not straightforward to compute. To approximate the FOV we extracted the focal length f from the Exif metadata and looked up the camera's CCD/CMOS sensor dimensions d in order to compute the FOV according to $fov = 2\,\mathrm{atan}\,\frac{d}{2f}$. When the focal length or sensor dimensions could not be determined, we assumed a default FOV of $58.72°$. The method yields a coordinate estimate by averaging all locations sharing the highest weight, while its footprint estimate is produced by applying connected components analysis to these locations, where we treat each resulting component as a cluster. We used the same discrete representation of the world as our own method (see Section 4.4) to compute the coordinate and footprint estimates.

5.3. Evaluation

To perform an unbiased evaluation, we compare the performance of the methods on the tasks of coordinate and footprint location estimation. We evaluate the performance for all methods according to the MAE. On the task of estimating the coordinates of the landmarks in the test set, see Figure 10(a), we observe that our method overall outperforms all baselines by a large margin. While for the subsets containing up to 20 photos the performance of our method and DBSCAN is comparable, for the other subsets the performance differences are more pronounced. For example, for 200 photos our method achieves a MAE of only 6.7m, whereas the best baseline for this subset, EM, has a MAE of 18.8m; when considering all photos our method obtains a MAE of 19.6m, while DBSCAN has a MAE of 42.5m. On the task of estimating the footprints, see Figure 10(b), our method achieves a higher overlap than all baselines for each

[10] We explored all combinations of $MinPts \in \{3, 4, 5, 10\}$, $Eps \in \{0.001, 0.002, 0.005\}$.

[11] We explored all combinations of $MinOwners \in \{3, 4, 5, 10\}$, $Eps \in \{0.001, 0.002, 0.005\}$, $Addt = 0.10$.

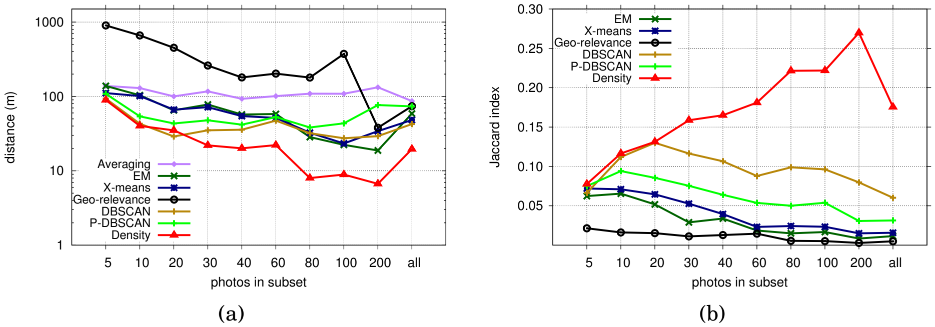


Fig. 10: Distance (in log-scale) achieved by all methods for estimating the landmark coordinates on different test subsets, where lower is better (a). Overlap between the footprints of the landmarks in the test set and the footprint estimates produced by all methods except the Averaging method, where higher is better (b). All methods used their overall best parameter configurations, if applicable.

subset; while the overlap of the baselines actually decreases for the larger subsets, it keeps increasing for our method. When using all photos per POI photos our method obtains an average overlap of 17.6%, while the best baseline, DBSCAN, achieves 6.0%. We show some success and failure cases of our method in Online Appendix B.

We further look at the pairwise differences of the computed error estimates and evaluate the significance of the observed improvements. To choose an appropriate statistical test, we first examine the distribution of our data using the Anderson-Darling and Cramer-von Mises tests. These tests are known to perform better compared to the Kolomorov-Smirnov test [Stephens 1974; Thode 2002], although in large samples they tend to be significant even for scores that are marginally different from a normal distribution; we thus interpret them in conjunction with Q-Q plots, while also accounting for the skew and kurtosis values. Since in all cases we observe a non-normal distribution in the absolute differences of the estimate errors, we opt for the Mann-Whitney test and report our results at an α level of .05. Finally, to take an appropriate control of Type I errors in multiple pairwise comparisons we apply the Bonferroni correction.

Table IV shows the Mann-Whitney test results for all comparisons between our method and the baselines, for the coordinate and footprint methods. In the case of the coordinate method, our method achieved a significantly smaller estimate error compared to the other baselines. In most cases, this significant difference represents a medium to large effect, with the smallest improvement being over the DBSCAN method and the largest improvement being over the Averaging method. With respect to the footprint method, our method attained a significantly higher overlap with the POIs over all baselines. This finding represents a medium to huge effect, with the smallest improvement being over the DBSCAN method and the largest improvement being over the Geo-relevance method. The Mann-Whitney test results indicate that, overall, our method significantly outperforms all baselines, at varying degrees.

While error assessment measures such as the MAE allow to evaluate the prediction quality of the methods, they do not quantify the relative difference of their performance. Therefore, we also compute the coefficient of determination R^2, which is a measure of the relative improvement of our prediction method over the other baseline methods. Negative R^2 values are obtained whenever our method underperforms compared to a baseline. Note that for the footprint estimates we use the Jaccard distance instead of the Jaccard index, since R^2 expects error measurements. The performance

Table IV: Mann-Whitney significance results of the coordinate and footprint estimate differences between the Density method and all baselines.

(Coordinate) (Mdn)	**Averaging** (74.16)	**EM** (40.26)	**X-means** (35.46)	**Geo-relevance** (27.79)	**DBSCAN** (14.64)	**P-DBSCAN** (33.42)
Density (2.44)	$U = 1.62$E+5 $z = -20.16$ $p < .001$ $r = -0.68$	$U = 2.39$E+5 $z = -12.65$ $p < .001$ $r = -0.43$	$U = 2.51$E+5 $z = -11.43$ $p < .001$ $r = -0.39$	$U = 2.49$E+5 $z = -11.70$ $p < .001$ $r = -0.40$	$U = 3.07$E+5 $z = -5.90$ $p < .001$ $r = 0.20$	$U = 2.59$E+5 $z = -10.65$ $p < .001$ $r = -0.36$
(Footprint) (Mdn)		**EM** (.0108)	**X-means** (.0176)	**Geo-relevance** (.0004)	**DBSCAN** (.0833)	**P-DBSCAN** (.0419)
Density (.1446)		$U = 1.51$E+5 $z = -21.01$ $p < .001$ $r = -.71$	$U = 1.70$E+5 $z = -19.09$ $p < .001$ $r = -0.65$	$U = 8.32$E+5 $z = -27.96$ $p < .001$ $r = -.95$	$U = 2.69$E+5 $z = -9.40$ $p < .001$ $r = -.32$	$U = 2.11$E+5 $z = -15.08$ $p < .001$ $r = -.51$

Table V: Performance of our overall best method against all baselines. The inside column indicates the percentage of coordinate estimates that were within the footprint of the landmark. The time columns indicate how much time each method needed for processing all photos for the 105 landmarks, averaged per photo, computed with a 2013 MacBook Pro using implementations in Java of each method.

	Coordinate estimates				**Footprint estimates**		
	Inside (%)	MAE ± SD	R^2	Time (ms)	MAE ± SD	R^2	Time (ms)
Averaging	14.79	113.58 ± 127.29	72.37%	2.84E-2			
EM	24.32	74.58 ± 116.36	57.89%	8.03E+1	.04 ± .06	22.19%	1.02E+2
X-means	20.86	69.19 ± 102.88	47.68%	8.14E-2	.05 ± .07	20.45%	4.67E-1
Geo-relevance	25.74	411.69 ± 712.21	98.81%	1.87E+3	.01 ± .03	26.23%	1.87E+3
DBSCAN	31.36	44.50 ± 77.70	.30%	1.05E-1	.10 ± .09	11.01%	5.29E-1
P-DBSCAN	24.37	57.69 ± 82.54	20.70%	1.14E-1	.07 ± .08	16.62%	5.32E-1
Density	46.21	33.81 ± 83.06	-	6.48E-1	.16 ± .11	-	1.13E+0

scores shown in Table V reveal that our method significantly outperforms the baseline methods on both the coordinate and footprint estimation tasks, with the exception of DBSCAN that performs comparably on coordinate estimation. Yet, the percentage of correctly estimated coordinates (i.e. those falling inside the POI footprints) achieved by our method is 46.2%, substantially outperforming the 31.4% obtained by DBSCAN.

The average time our method needed to process a photo was around eight times as slow as X-means, and six times as slow as DBSCAN and P-DBSCAN on the coordinate task and about twice as slow compared to all three on the footprint task. In contrast, the EM method achieved about half the accuracy in the coordinate estimation task and about a quarter of the overlap in the footprint estimation task, while being a factor 100 slower to compute compared to our method. The Geo-relevance method did not perform as well as the other methods, because it gives equal weight to all locations in the fields of view of infinite length. This results in a large geographic area where each location has an equal weight, producing coordinate estimates far away from the POIs and footprint estimates many times too big.

6. CONCLUSIONS

We proposed a novel algorithm for the localization of POIs in terms of their geographic coordinates and their footprints. Our method exploited the geographic positions and compass orientations measurements supplied by modern cameras. We took the possible errors produced by the sensors into account, as well as different styles of photo composition, and modeled the world as a probabilistic map where each coordinate in the map ultimately reflected the likelihood of a POI being present at that location. We extensively analyzed and evaluated our method on a set of 105 POIs, estimating their coordinates and footprints significantly more accurately than state of the art baselines from the research literature. Our coordinate estimation method produced correct estimates 46.2% of the time, while achieving an average distance error of less than 100m with just 5 photos per POI and an error of less than 10m when using 80+ photos. Our footprint estimation method reached an average overlap of 16% with the actual footprint of a POI, and even achieved an overlap of 27% when using 200 photos. The results suggest that even with few geotagged and oriented photos referring to an unknown POI our method is able to pinpoint its location within a small distance of where it is really located, where the coordinate estimate gets more accurate with additional photos. The estimated footprint further tends to overlap its actual footprint, where the footprint estimate also gets more accurate with more photos. Our work opens doors to automatically discovering POIs by analyzing all photos (even those without tags) taken within any geographic area by inspecting the produced density maps for local maxima, as well as automatically detecting local events by analyzing streams of recently taken and uploaded photos (see Online Appendix C for examples). We aim to address such problems in future work.

REFERENCES

S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. 2009. Building Rome in a day. In *Proc. IEEE Intl. Conf. on Computer Vision*. 72–79.

S. Ahern, M. Naaman, R. Nair, and J. Yang. 2007. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proc. ACM/IEEE-CS Joint Conf. on Digital Libraries*. 1–10.

T. Cham, A. Ciptadi, W. Tan, M. Pham, and L. Chia. 2010. Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. 366–373.

F. Chang, C. Chen, and C. Lu. 2004. A linear-time component-labeling algorithm using contour tracing technique. *Computer Vision and Image Understanding* 93, 2 (2004), 206–220.

L. Chen and A. Roy. 2009. Event detection from Flickr data through wavelet-based spatial analysis. (2009), 523–532.

D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. Mapping the world's photos. In *Proc. IW3C2 Intl. Conf. on World Wide Web*. 761–770.

D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. 2013. SfM with MRFs: discrete-continuous optimization for large-scale structure from motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35, 12 (2013), 2841–2853.

P. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research* 134, 1 (2005), 19–67.

M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. 2010. Automatic construction of travel itineraries using social breadcrumbs. In *Proc. ACM Conf. on Hypertext and Hypermedia*. 35–44.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Jrnl. of the Royal Statistical Society, Series B* 39, 1 (1977), 1–38.

C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. 2012. What makes Paris look like Paris? *ACM Trans. on Graphics* 31, 4 (2012), 1–9.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. 2001. Rank aggregation methods for the Web. In *Proc. IW3C2 Intl. Conf. on World Wide Web*. 613–622.

B. Epshtein, E. Ofek, Y. Wexler, and P. Zhang. 2007. Hierarchical photo organization using geo-relevance. In *Proc. ACM Intl. Conf. on Geographic Information Systems*.

M. Ester, H. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. AAAI Intl. Conf. on Knowledge Discovery and Data Mining*. 226–231.

Facebook, Ericsson, and Qualcomm. 2013. *A focus on efficiency*. Technical Report.

G. Field. 1845. *Chromatics; or, the analogy, harmony, and philosophy of colours*. D. Bogue.

J. Hao, G. Wang, B. Seo, and R. Zimmermann. 2014. Point of interest detection and visual distance estimation for sensor-rich video. *IEEE Trans. on Multimedia* 16, 7 (2014), 1929–1941.

J. Hays and A.A. Efros. 2008. IM2GPS: estimating geographic information from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. 1–8.

J. Heinly, J. L. Schönberger, E. Dunn, and J. Frahm. 2015. Reconstructing the world in six days. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. 3287–3295.

L. Hollenstein and R. S. Purves. 2010. Exploring place through user-generated content: using Flickr tags to describe city cores. *Jrnl. of Spatial Information Science* 1 (2010), 21–48.

M. Hölzl, R. Neumeier, and G. Ostzermayer. 2013. Analysis of compass sensor accuracy on several mobile devices in an industrial environment. In *Proc. Intl. Conf. on Computer Aided Systems Theory*. 381–389.

A. Jaffe, M. Naaman, T. Tassa, and M. Davis. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proc. ACM Intl. Wkshp. on Multimedia Information Retrieval*. 89–98.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. 2014. Caffe: convolutional architecture for fast feature embedding. In *Proc. ACM Intl. Conf. on Multimedia*. 675–678.

R. S. Kaminsky, N. Snavely, S. M. Seitz, and R. Szeliski. 2009. Alignment of 3D point clouds to overhead images. In *Proc. IEEE Wkshp. on Computer Vision and Pattern Recognition*. 63–70.

C. F. F. Karney. 2013. Algorithms for geodesics. *Jrnl. of Geodesy* 87, 1 (2013), 43–55.

L. S. Kennedy and M. Naaman. 2008. Generating diverse and representative image search results for landmarks. In *Proc. IW3C2 Intl. Conf. on World Wide Web*. 297–306.

S. Kisilevich, F. Mansman, and D. A. Keim. 2010. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In *Proc. Intl. Conf. and Exhibition on Comp. for Geospatial Research & Application*.

J. Kosecká and W. Zhang. 2002. Video compass. In *Proc. Europ. Conf. on Computer Vision*. 476–490.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proc. of Ann. Conf. on Advances in Neural Information Processing Systems*. 1097–1105.

Y. A. Lacerda, R. G. F. Feitosa, G. A. R. M. Esmeraldo, C. de Souza Baptista, and L. B. Marinho. 2012. Compass clustering: a new clustering method for detection of points of interest using personal collections of georeferenced and oriented photographs. In *Proc. Brazilian symposium on Multimedia and the Web*. 281–288.

Y. Li, D. J. Crandall, and D. P. Huttenlocher. 2009. Landmark classification in large-scale image collections. In *Proc. IEEE Intl. Conf. on Computer Vision*. 1957–1964.

J. Luo, D. Joshi, J. Yu, and A. C. Gallagher. 2011. Geotagging in multimedia and computer vision - a survey. *Multimedia Tools and Applications* 51, 1 (2011), 187–211.

Z. Luo, H. Li, J. Tang, R. Hong, and T. Chua. 2010. Estimating poses of world's photos with geographic metadata. In *Advances in Multimedia Modeling*. 695–700.

L. Mummidi and J. Krumm. 2008. Discovering points of interest from users' map annotations. *GeoJrnl.* 72, 3-4 (2008), 215–227.

M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. 2004. Automatic organization for digital photographs with geographic coordinates. In *Proc. ACM/IEEE-CS Joint Conf. on Digital Libraries*. 53–62.

L. Ojeda and J. Borenstein. 2000. Experimental results with the KVH C-100 fluxgate compass in mobile robots. In *Proc. IASTED Intl. Conf. on Robotics and Applications*.

J. Paek, J. Kim, and R. Govindan. 2010. Energy-efficient rate-adaptive GPS-based positioning for smartphones. In *Proc. ACM Intl. Conf. on Mobile Systems, Appl., and Services*. 299–314.

S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. 2011. Cluster-based landmark and event detection for tagged photo collections. *IEEE Multimedia* 18, 1 (2011), 52–63.

D. Pelleg and A. W. Moore. 2000. X-means: extending K-means with efficient estimation of the number of clusters. In *Proc. Intl. Conf. on Machine Learning*. 727–734.

V. Pihur, S. Datta, and S. Datta. 2007. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* 23, 13 (2007), 1607–1615.

A. Popescu. 2013. CEA LIST's participation at MediaEval 2013 placing task. In *Working Notes of the MediaEval Benchmarking Initiative for Multimedia Evaluation*.

A. Popescu, G. Grefenstette, and P. Moëllic. 2009. Mining tourist information from user-supplied collections. In *ACM Intl. Conf. on Knowledge and Information Management*. 1713–1716.

A. Popescu and A. Shabou. 2013. Towards precise POI localization with social media. In *Proc. ACM Intl. Conf. on Multimedia*. 573–576.

T. Quack, B. Leibe, and L. J. Van Gool. 2008. World-scale mining of objects and events from community photo collections. In *Proc. ACM Conf. on Image and Video Retrieval*. 47–56.

A. Rae, V. Murdock, A. Popescu, and H. Bouchard. 2012. Mining the web for points of interest. In *Proc. ACM Intl. Conf. on Research and Development in Information Retrieval*. 711–720.

R. Raguram, C. Wu, J. Frahm, and S. Lazebnik. 2011. Modeling and recognition of landmark image collections using iconic scene graphs. *Intl. Jrnl. of Computer Vision* 95, 3 (2011), 213–239.

T. Rattenbury, N. Good, and M. Naaman. 2007. Towards automatic extraction of event and place semantics from Flickr tags. In *Proc. ACM Intl. Conf. on Research and Development in Information Retrieval*. 103–110.

T. Rattenbury and M. Naaman. 2009. Methods for extracting place semantics from Flickr tags. *ACM Trans. on the Web* 3, 1 (2009).

R. Y. Rubinstein and D. P. Kroese. 2005. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*. Springer.

S. Rudinac, A. Hanjalic, and M. Larson. 2011. Finding representative and diverse community contributed images to create visual summaries of geographic areas. In *Proc. ACM Intl. Conf. on Multimedia*. 1109–1112.

P. Serdyukov, V. Murdock, and R. van Zwol. 2009. Placing Flickr photos on a map. In *Proc. ACM Intl. Conf. on Research and Development in Information Retrieval*. 484–491.

J. T. Smith. 1797. *Remarks on rural scenery; with twenty etchings of cottages, from nature; and some observations and precepts relative to the picturesque*. N. Smith and I.T. Smith.

N. Snavely, S. M. Seitz, and R. Szeliski. 2006. Photo tourism: exploring photo collections in 3D. *ACM Trans. on Graphics* 25, 3 (2006), 835–846.

N. Snavely, S. M. Seitz, and R. Szeliski. 2008. Modeling the world from internet photo collections. *Intl. Jrnl. of Computer Vision* 80, 2 (2008), 189–210.

M. A. Stephens. 1974. EDF statistics for goodness of fit and some comparisons. *Jrnl. of the American Statistical Association* 69, 347 (1974), 730–737.

H. C. Thode. 2002. *Testing for normality*. CRC Press.

B. Thomee. 2013. Localization of points of interest from georeferenced and oriented photographs. In *Proc. ACM Intl. Wkshp. on Geotagging and Its Applications in Multimedia*. 19–24.

O. Van Laere, S. Schockaert, and B. Dhoedt. 2010. Towards automated georeferencing of Flickr photos. In *Proc. ACM Wkshp. on Geographic Information Retrieval*.

C. Wang, K. Wilson, and N. Snavely. 2013a. Accurate georegistration of point clouds using geographic data. In *Proc. Intl. Conf. on 3D Vision*. 33–40.

G. Wang, Y. Yin, B. Seo, R. Zimmermann, and Z. Shen. 2013b. Orientation data correction with georeferenced mobile videos. In *Proc. ACM Intl. Conf. on Advances in Geographic Information Systems*. 390–393.

C. Wu. 2013. Towards linear-time incremental structure from motion. In *Proc. IEEE Intl. Conf. on 3D Vision*. 127–134.

Y. Yang, Z. Gong, and L. Hou. 2011. Identifying points of interest by self-tuning clustering. In *Proc. ACM Intl. Conf. on Research and Development in Information Retrieval*. 883–892.

P. A. Zandbergen. 2008. Positional accuracy of spatial data: non-normal distributions and a critique of the national standard for spatial data accuracy. *Trans. in GIS* 12, 1 (2008), 103–130.

P. A. Zandbergen and S. J. Barbeau. 2011. Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phone. *Jrnl. of Navigation* 64, 3 (2011), 381–399.

Y. Zheng, Z. Zha, and T. S. Chua. 2011. Research and applications on georeferenced multimedia: a survey. *Multimedia Tools and Applications* 51, 1 (2011), 77–98.

Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. S. Chua, and H. Neven. 2009. Tour the world: building a web-scale landmark recognition engine. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. 1085–1092.