Towards Precise POI Localization with Social Media

Adrian Popescu
CEA, LIST,
Vision & Content Engineering Laboratory
Gif-sur-Yvette, France
adrian.popescu@cea.fr

Aymen Shabou
CEA, LIST,
Vision & Content Engineering Laboratory
Gif-sur-Yvette, France
aymen.shabou@gmail.com

ABSTRACT

Points of interest (POIs) are a core component of geographical databases and of location based services. POI acquisition was performed by domain experts but associated costs and access difficulties in many regions of the world reduce the coverage of manually built geographical databases. With the availability of large geotagged multimedia datasets on the Web, a sustained research effort was dedicated to automatic POI discovery and characterization. However, in spite of its practical importance, POI localization was only marginally addressed. To compute POI coordinates an assumption was made that the more data were available, the more precise the localization will be. Here we shift the focus of the process from data quantity to data quality. Given a set of geotagged Flickr photos associated to a POI, close-up classification is used to trigger a spatial clustering process. To evaluate the newly introduced method against different other localization schemes, we create an accurate ground truth. We show that significant localization error reductions are obtained compared to a coordinate averaging approach and to a X-Means clustering scheme.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]

General Terms

Algorithms, Experimentation

Keywords

geographic information extraction, POI localization, Flickr

1. INTRODUCTION

The success of geographically enabled applications, such as location based services (LBS), mapping interfaces or navigation systems), is highly dependent on the coverage and quality of the points of interest (POIs) present in geographic databases. POI were classically added to databases by a large number of domain experts who visited physical locations or browsed existing resources [7]. The resulting databases are accurate but their constitution and update have a significant cost. Large geotagged multimedia datasets are a by-product of the success of social media and they were exploited in a number of geographic information extraction tasks. Representative works that leverage user-generate content for POI mining include [8], [2], [7] and are focused on the extraction of POI names, categories and/or popularity. Curiously, while POI localization imprecision drastically limits the practical utility of automatically extracted POIs in LBS, it was only marginally considered in existing works. We improve POI localization through an appropriate combination of text and image analysis. Text is used to retrieve the initial set of images associated to a POI and visual information is then exploited to filter the initial set. Our key hypotheses are that not all items are equally important for deriving POI location and that the best approximations of true POI locations are given by spatially dense and socially diverse data subsets. To test these hypotheses, we perform close-far image classification and introduce a simple but efficient spatial clustering algorithm seeded with POI close-up photos. By doing so, we add a qualitative aspect to the localization process and shift its focus from quantitative approaches that prevail in existing work. The evaluation of the approach is done using a Flickr dataset.
The evaluation of POI localization algorithms is a difficult task due to the non-existence of standardized datasets [7]. It is usually performed against POI coordinates from manually created databases, such as Geonames[1]. A close inspection of the coordinates provided by Geonames shows that no coherent rules are followed when placing the POIs and that a more accurate ground truth is needed. We create such a ground truth through manual inspection of Google Maps.

In addition to coordinates, we also estimate a POI radius that can be used as a proxy for an acceptable localization error. Experiments carried on this new dataset show that significant improvements are obtained through the use of multimedia processing.

2. RELATED WORK

To our knowledge, existing works explored the use of multimedia data to geotag individual documents but multimedia geotagging was not tested for POIs. This section discusses relevant prior work in geographic information extraction, with focus on geotagged multimedia datasets, and in

[1] http://geonames.org

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10$15.00..
http://dx.doi.org/10.1145/2502081.2502151.

image classification. Rattenbury and al. [8] pioneered the use of Flickr data to extract POI name, popularity and location using spatial burst detection. In a follow-up, Kennedy and Naaman [5] focused on using multimedia clustering to produce representative and diversified visual representation of POIs. One of their key findings is that socially representative clusters (i.e. including a large number of users) are more relevant than the others. Rae and al. [7] exploited Web search snippets to scale-up POI discovery. These works focus on POI properties such as name, category or popularity and give only limited attention to POI location. A simple coordinate averaging procedure was introduced in [8] and statistical location models were used in [7]. Typical location errors are in the hundreds of meters range (290 m in [7]), an imprecision that is too big for direct use of automatically extracted POIs in LBS.

Given its high applicative value, individual document localization received a lot of attention. Hays and Efros [4] proposed an algorithm that predicts the location of an image based on global visual descriptors. With this simple visual method, the reported precision is in the range of hundreds of kilometers. Crandall and al. [2], analyzed a 30 million geotagged image dataset using image local features (SIFTs) and textual features to estimate image location. In a simple classification setting (10 possible locations), a combination of visual and textual features achieves approximately 70% accuracy and textual features clearly outperform visual ones when exploited alone. Textual location modeling was proposed in [9] using an adaptation of language models. Image localization within 1 km from the true coordinates is successful in 14.1% of the cases. While interesting, these performances are not sufficient to localize POIs accurately.

Image depth estimation using wavelet based image indexing was introduced in Torralba and Oliva [11]. They used the scene structure to derive the absolute mean depth of the image. The problem we tackle here is somewhat similar although, given the high variability of images available on social media, we do not set out to estimate the absolute depth but rather to rank images associated to a POI in the close-far spectrum. Also relevant is indoor-outdoor classification [10], which became a standard classifier in scene understanding. However, given the similarity of inside/outside close-ups and the fact that there are no indoor photos for POIs that cannot be visited on the inside, we don't consider indoor-outdoor classification.

3. POI LOCALIZATION

The problem we address here can be formulated as: *"given a set of geotagged images associated to a POI, leverage their metadata, textual annotations and/or visual content to automatically predict the best possible POI location."*

When deriving POI location from geotagged photo datasets available on social media, there are three main types of problems that lead to inaccurate results: (1) Photos are usually placed at the point where they were taken and this point is often distant from the true POI location; (2) Geotagging errors may occur, especially when the process is manual because users often do not place the photos exactly where they were taken; (3) Given that there is no control of user generated content, only a fraction of the photos tagged with a POI name actually depict it. For instance, bulk upload was identified to have a negative effect on POI extraction [7].

We tackle the first difficulty directly by using a photo ranking based on their probability to be close-ups. The second and the third problems are indirectly tackled by the introduction of spatial clustering algorithms that favor regions that have a high density of close-up photos and that are socially diversified. Different text-based and multimedia POI localization methods were tested. The most representative of them are described in the next two subsections, along with textual and visual close-far classifiers.

3.1 Text-based POI localization

3.1.1 Text-based close-up ranking

Here we exploit textual cues to determine if a photo is a close-up. We retain 48,815 geotagged photo annotations that are taken from less than 50 meters and more than 300 meters w.r.t. the coordinates of a POI in order to form *textual close* and *textual far* classes. The distance thresholds for the two classes were empirically chosen after testing different combinations. We compute each word's *textual close* probability by dividing its *textual close* count by its total count. The photos associated to a POI are ranked using the dot product similarity between each photo's tags and title and the representation of *textual close*. Preliminary tests showed that the best results are obtained when the first 40% of the ranked photos list are retained.

3.1.2 Text-based POI localization methods

We test the following text localization methods:

* *Simple Avg* - baseline method introduced in [8]. It computes the coordinates from by averaging the coordinates of all the images available for a POI. For this type of method to work well, photos should be evenly distributed in all directions around the POI. Such a condition is seldom met because of the physical configuration of POIs and of the fact some POI parts arouse more interest than others.

* *No Bulk* - method that uses the same averaging procedure as *Simple Average* but is applied after bulk removal (i.e. removal of images uploaded by an user that have the same coordinates and the same set of tags).

* *Txt Close* - method that exploits the text-based close-up ranking. The POI location is computed by retaining only those images that were considered to be close-ups. No bulk removal is performed.

* *Txt Close Iter* - method that combines *No Bulk* to produce an initial average location and textual close-up ranking. The initial location is used to seed a second iteration that considers only a percentage of the photos that are closest to the initial location ($neigh(\%)$) and that belong to *textual close*. Supplementary iterations do not improve results.

3.2 Multimedia POI localization

3.2.1 Visual close-up classifier

Given the wide variety of available images, we build a learning base that contains 2000 diversified examples for each of the *visual close* and *visual far* classes. The visual classifier is built on top of a bag of visual words representation of images. Dense SIFTs are extracted and clustered using K-means to build a codebook of size 128. Then SIFTs are encoded over the obtained codebook by locality soft coding and aggregated using the max-pooling scheme. We use a spatial pyramid matching (1 + 2x2 + 3x3) to add global spatial contextual information to the final signature, which is obtained through the concatenation of individual signatures. *Visual close* and *visual far* classes are learned with a linear SVM which is then used to classify test images. Scores vary between 1 (close) and 0 (far). The learning set was split in half to test performances of the approach and we obtained 84% correct classifications. To further improve accuracy, only images whose close-up score is above 0.6 were retained for experiments. We illustrate classification results for the Eiffel Tower (fig. 1). There is a strong concentration of close-ups around the Eiffel Tower's true coordinates. Inversely, long-distance depictions of the Eiffel Tower are concentrated in areas that offer a good panoramic view, such as the Trocadero, Champs de Mars and Bir-Hakeim Bridge.

3.2.2 Multimedia POI localization methods

We test the following multimedia localization methods:
* *Vis Close* - Method that is similar to *Txt Close* but exploits *visual close* class instead of *textual close.*
* *Vis Close Iter* - Method that is similar to *Txt Close Iter* but exploits *visual close* class instead of *textual close.*
* *Density Clustering* - method that retrieves dense regions in a POI's spatial footprint. As illustrated in 1 the intuition that supports this approach is that regions that include a lot of close-up photos uploaded by a large number of users are likely to be close to the real POI location. First, bulk removal is applied and and photos are ranked with *visual close* to select close-ups. Second, the top k photos from the *visual close* class are used as seeds and clusters are obtained by retaining all close-ups within a radius *rad* of the seed. Finally, clusters are ranked based on the number of distinct users that contributed to them. Ties are broken using by favoring clusters that contain larger numbers of *visual close* images. The final POI coordinates are computed by averaging the coordinates of images from the top ranked cluster.

The cluster ranking based on the number of contributors exploits an insight from [5], who show that user frequency based landmark clustering outperforms term based clustering due to the negative influence of bulk upload. Our algorithm is related to OPTICS [1], the main difference being that we introduce a global radius to cluster points instead of a maximum local radius that is used to retain new points.
* *Spatial X-Means* - X-Means [6] is an extension of K-Means that estimates the optimal number of clusters automatically for a dataset. The Weka 3 implementation of the X-Means [3] is used here to assess the performances of standard clustering algorithms in a POI localization task. The results of *Vis Close Iter* are used as input for *Spatial X-Means* and the centroid of the largest cluster is predicted as POI location.

4. EVALUATION

We first present a new POI localization ground truth, then describe the test dataset used and discuss localization results obtained with the methods presented in section 3.

4.1 POI Localization Ground Truth Creation

Precise POI localization is a difficult task since in involves mapping POIs that have dimensions in the range of tens or hundreds of meters to point-based representations. A close inspection of Geonames shows that the proposed POI locations are not consistently attributed. Coordinates are often placed at the center of the POI but also in other parts of the area covered by it or even outside the POI area. To create a precise ground truth, we have selected 1200 POIs that are represented by at least 250 images in Flickr, are situated in various regions of the world, correspond to different geographic categories and have a variable spatial extent. We inspected Google Maps to determine the coordinates of the center of the POI. It was possible to precisely find the point-based location and the radius for 704 POIs that are used in further experiments. The radius was estimated only for shapes that allow such an estimation (rectangular, circles, square, etc.). It is used as a proxy for the acceptable localization error if we consider as acceptable all localizations that fall within the POI area. The average Geonames localization error is 18.2 m and 84.8% of Geonames POIs are placed within the surface covered by the POI. There is no publicly available accurate ground truth for evaluating POI localization and we release the one created here[2].

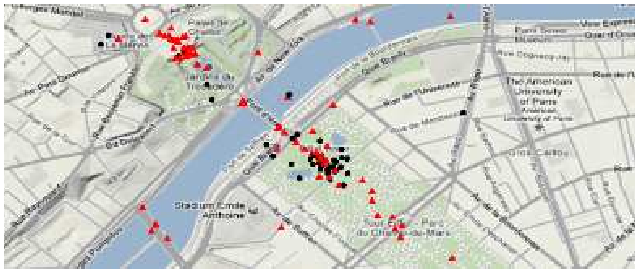

Figure 1: Close/far classification results - circles/triangles Eiffel Tower (centered).

4.2 Test Dataset

Using the Flickr API, we downloaded 250 images and associated metadata (tags, titles, geotags) for each POI represented in the ground truth. For each POI, textual annotations were ranked with respect to their similarity to *textual close.* Similarly, images were indexed and classified based on their similarity to *visual close.* To facilitate reproducibility, the raw textual annotations and image URLs of the test dataset are distributed with the ground truth.

4.3 Evaluation Results

We first test the influence of the number of available images on POI localization error and present results obtained using *Simple Avg* with 10 to 250 images, with a step of 10 in fig. 2. The number of available photos has an important effect up to 50 available photos and the performance improvement is weaker beyond. For instance, the average

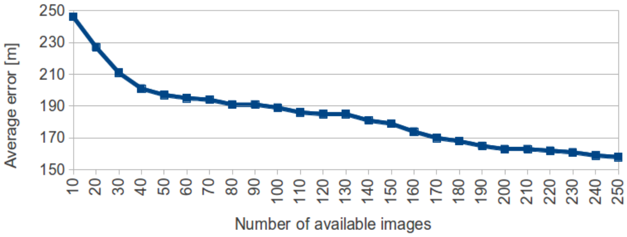

Figure 2: Evolution of the average localization error with the number of available images for *Simple Avg.*

[2] http://comupedia.org/poiLocation/

error is 163 m for 200 photos and 158 for 250. These results show the limits of purely quantitative POI localization methods and support the introduction of qualitative methods to improve performances.

The POI localization methods described in section 3 were applied to the test dataset and results are presented in table 1. We performed a grid search for parameter optimization and obtained the best performances with: $neigh = 60\%$ for *Txt Close Iter*, *Vis Close Iter* and *Density Clustering*; $k = 10$ seeds for *Density Clustering*; $rad = 100$ m for *Density Clustering*; $x_{min} = 3$ and $x_{max} = 10$ for *Spatial X-Means*. The smallest localization error is obtained with *Density Clustering* (average error - 65.6 m, 10.5 m standard deviation), followed by *Vis Close Iter* (77.1 m, 11.9 m) and *Spatial X-Means* (78.1 m, 12.3 m). A t-test indicates that the difference between *Density Clustering* and the other methods is statistically significant at $p < 0.01$.

These results show that that multimedia POI localization outperforms all text based localization and that *Density Clustering*, seeded with the visual close-ups, is better than *Spatial X-Means*, a standard clustering and than *Vis Close Iter*, an iterative global localization procedure. The worst results are obtained with *Simple Avg*, the baseline method. Bulk removal, the use of *textual close* and of *visual close* cues bring a small improvement compared to *Simple Avg.* A larger improvement is obtained for *Txt Close Iter* and *Vis Close Iter*, the two global ranking methods. The comparison of similar text and visual localization methods indicates that visual cues have a more important effect than textual cues. Results for the percentage of automatic locations situated inside the POI area show similar performances to average error. *Density Clustering* places the largest number of POIs within their area.

The performances of *Density Clustering* are globally better than those of the other methods tested. However, there are configurations in which this method fails. For example, if the photos are evenly distributed around the true coordinates but are all taken from a certain distance (for instance, when the POI is not physically accessible), averaging methods have better performances. Another problematic case is that when the densest cluster of close-ups is away from the true coordinates and photos that are closer to these coordinates are discarded. Here the problem can be caused either by irrelevant POI tagging and/or by a small number of users that contribute to the POI representation. Yet another problematic case is that of elongated POIs, such as bridges for which it is difficult to compute accurate coordinates using density based methods. However, computing coordinates for elongated objects is problematic for all methods since it is rare that these objects' photos are evenly distributed around their central point.

5. CONCLUSIONS

We presented different methods for automatic POI localization and showed that significant improvements are obtained compared to a coordinate averaging approach [8]). *Density Clustering* equally outperforms *Spatial X-Means*, a standard clustering algorithm, a result that supports the importance of an appropriate seeding in the POI localization process. Also, visual cues are more important in the localization process than textual cues. Even if the best automatic results still lag behind manual POI coordinates assignment, the gap was reduced. The average error difference compared to manual POI geotagging remains consequent (65.6 m for *Density Clustering* vs. 18.23 m for Geonames). However, the percentage of POI localizations that lie inside the POI area is comparable (76.7% for *Density Clustering* and 84.9% for Geonames). The last measure is more important in practice since it allows one to correctly indicate POI locations on a map and to use them for LBS services, such as tourist guiding. To facilitate result reproducibility, we share the ground truth along with the dataset used in the experimental section.

Future work will focus on a dynamic tuning of *Density Clustering* in function of the POI characteristics (POI footprint proportion of close-ups, etc.). Here we considered all photos tagged with the POI name, regardless of the relevance of the annotation, and this choice might hurt overall performances. Consequently, another line of work will concern the use of image reranking to remove potentially irrelevant clusters.

Table 1: POI localization errors with different methods. Avg. err. is the average error expressed in meters. St. dev. is the standard deviation in meters. Inside is the percentage of localizations that fall within the surface covered by the POI

Method Name	Avg. err.[m]	St. dev[m]	Inside[%]
Simple Avg	158	21.7	43.7
No Bulk	151.2	19.4	39.9
Txt Close	151	19.1	39.8
Txt Close Iter	89.7	13.7	64.1
Vis Close	140.7	18.8	46
Vis Close Iter	77.1	11.9	72.2
Density Clust.	**65.6**	**10.5**	**76.7**
Spatial X-Means	78.1	12.3	71.2

6. REFERENCES

[1] M. Ankerst and al. Optics: ordering points to identify the clustering structure. In *SIGMOD 1999*.

[2] D. J. Crandall and al. Mapping the world's photos. In *WWW2009*, pages 761–770, 2009.

[3] M. Hall and al. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[4] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR 2008*.

[5] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *WWW 2008*.

[6] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML 2000*.

[7] A. Rae and al. Mining the web for points of interest. In *SIGIR 2012*.

[8] T. Rattenbury and al. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR 2007*, pages 103–110, 2007.

[9] P. Serdyukov and al. Placing flickr photos on a map. In *SIGIR 2009*, pages 484–491, 2009.

[10] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *CAIVD 1998*.

[11] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE PAMI*, 24(9):1226–1238, 2002.