Data-Driven vs Semantic-Technology-Driven Tag-Based Video Location Estimation

Jaeyoung Choi
Department of EECS
University of California, Berkeley
Email: jaeyoung@cs.berkeley.edu

Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704
Email: fractor@icsi.berkeley.edu

***Abstract*—The following article describes two approaches to determining the geo-coordinates of the recording place of Flickr videos based on textual metadata. The systems are tested on the MediaEval 2010 Placing Task evaluation data, which consists of 5091 unfiltered test videos and metadata records. The first system is a data-driven approach that uses a heuristics based on the spatial variance of tags. The second one extends this heuristics by using semantic technologies, such as extended Wordnet and a geographical gazetteer. The performance peaks at being able to classify 14 % of the videos to within an accuracy of 10 m. The article present the two algorithms, evaluates their accuracy and discusses the advantages and disadvantages of using Semantic technologies for this task.**

I. Introduction

A multimedia retrieval task that has only recently caught the attention of the research community, is estimating the location of origin of a video recording that lacks geo-location metadata. The task is sometimes called "multimodal location estimation" or "placing". Just as a human analyst uses multiple sources of information and context to determine geo-location, it seems obvious that for location estimation, the investigation of clues across different modalities and combination with diverse knowledge sources from the web can lead to better results than investigating only one stream of sensor input (e.g. reducing the task to an to image retrieval problem).

The task has recently caught the attention of researchers in the multimedia, signal processing, and machine learning communities because of the large amount of available geo-tagged media on the Internet that can be used as training data, allowing algorithms to work on data volumes rarely seen before. In addition, the task is hard enough to require collaboration between many experts and in diverse research communities, which is a challenge on its own.

A very vivid discussion among researcher is currently whether the inclusion of information from semantic databases, such as DBPedia, Geonames, and Wordnet is helpful or not. This article therefore describes and compares the two approaches for determining the geo-coordinates of the place where Flickr videos were recorded based on textual metadata. The systems were both tested on the MediaEval 2010 Placing Task evaluation data.

The article is organized as follows. We start with a short survey of prior and related work in Section II. Section III then introduces the datasets used and the experimental setup, before Section IV discusses our two technical approaches. Section V then presents results, leading to Section VI which concludes the article with final remarks.

II. Related Work

Previous work that has been carried out in the area of automatic geo-tagging of multimedia based on tags have also been mostly carried out on Flickr images. User-contributed tags have a strong location component, as brought out by [14], who reported that over 13 % of Flickr image tags could be classified as locations using Wordnet. In [12], the geo-locations associated with specific Flickr tags are predicted using spatial distributions of tag use. A tag which is strongly concentrated in a specific location has a semantic relationship with that location. User-contributed tags are exploited for geo-tagging by [13], who use tag distributions associated with locations represented as grid cells on a map of the Earth is used to infer the geographic locations of Flickr images. The approach in [6] reports on combining visual content with user tags. However, the accuracy is only reported with a minimum granularity of 200 km. Multimodal location estimation on videos has been first defined and attempted in [5] where the authors match ambulance videos from different cities, even without using textual tags. The first evaluation on multimodal location estimation on randomly selected consumer-produced videos has been performed in the 2010 MediaEval Placing task [9]. Several notable systems participated in the evaluation [16], [8], [2], [4], [11], including the predecessor of the system described herein. The rules of the evaluation prohibit us to compare and rank the system results as of the evaluation. Please refer to the cited references for further information.

III. Datasets

A. MediaEval 2010 Dataset

The MediaEval 2010 Placing Task, organized by [10], is to automatically guess the location of the video, i.e., assign geo-coordinates (latitude and longitude) to videos using one or more of: video metadata (tags, titles), visual content, audio content, social information. Any use of open resources, such as gazetteers, or geo-tagged articles in Wikipedia is encouraged. The goal of the task is to come as close to possible to the geo-coordinates of the videos as provided by users or their GPS devices.

978-0-7695-4492-2/11 $26.00 © 2011 IEEE
DOI 10.1109/ICSC.2011.37

The data set consists of Creative Common-licensed videos that were manually crawled from Flickr. The videos are in MPEG-4 format and include the Flickr metadata in XML format. The meta-data for each video includes user-contributed title, tags, description, comments and also information about the user who uploaded the videos. Additionally, the metadata records also include information about the user's contacts, favorites, and all videos uploaded in the past. The data set was divided into training data (5091 videos) and test data (5125 videos).

According to [9], videos were selected both to provide a broad coverage of users, and also because they were geo-tagged with a high accuracy at the "street level". An accuracy field indicates the zoom level the uploader used when placing the photo or video on the map. There are 16 zoom levels, and these correspond to 16 accuracy levels (e.g., "region level", "city level", "street level"). The sets of users from the test and the training collections were disjoint in order to not introduce a user-specific bias. This bias will be discussed further in Section V.

B. Characteristics of the Data

Flickr requires that an uploaded video must be created by its uploader (if a user violates this policy, Flickr sends a warning and removes the video). Manual inspection of the data set lead us initially to conclude that most of visual/audio contents lack reasonable evidence to estimate the location without textual metadata. For example, many videos were recorded indoors or in a private space such as a backyard of a house, which make the Placing Task nearly impossible if we examine only the visual and audio contents. This indicates that the videos are not pre-filtered or pre-selected in any way to make the data set more relevant to the task, and are therefore likely representative of videos selected at random.

However, metadata provided by the user often provides direct and sensible clues for the task. 98.8 % of videos in the training set were annotated by their uploaders with at least a title, tags, or description, often including location information. For a human, it is a fairly straightforward task to determine from the metadata which keyword or keywords combination indicates the smallest and most accurate geographical entity. However, for a machine, extracting a list of toponym candidate keywords and further choosing a correct single keyword or combination of keywords is a challenging task. Misspelled or compound words concatenated without spaces are commonly found in user-annotated metadata and these add more difficulty to the task. For example, "my trip to fishermanswharf san francisco" should resolve to the "Fisherman's Wharf" in "San Francisco".

Furthermore, partly because of social, political, and economical reasons, in current online video databases (e.g. Flickr and YouTube), videos are not equally distributed over the earth. Therefore downloading a random sample, as performed for MediaEval 2010, leads to a large bias towards certain locations. Figure 1 shows the distribution of the MediaEval training set. While it will always be difficult to find videos from certain countries or remote places, a training set that is more equally distributed is desirable for improving global retrieval precision and recall.

C. Additional Data

Because of the non-uniformity of the MediaEval2010 training and test set, we used additional data to make the training data more equally distributed over the earth. In addition to the MediaEval 2010 data, we also included the data used for the experiments described in [7]. The data originally consists of 6.4 million images from Flickr categorized into countries and states (in case of US). We sampled pictures from each region and used their unique Flickr photo ID to download the metadata from Flickr. 759,249 metadata records were collected in this way. Furthermore, we collected additional photos from Flickr by dividing the area of the earth into 1 km grid cells, counting the number of photos for each grid cell. If the cell contained more than 15 photos, we sampled 15 % of photos. This resulted in about 1,131,698 new metadata records and photos. All metadata was collected and saved in the same format as the MediaEval photo dataset UserID, PhotoID, HTML link to photo, latitude and longitude, tags, date taken, and date uploaded. Again, we ensured that the user set stays disjoint between training and test set.

IV. Technical Approaches

A. Data-Driven Approach

Our first approach to location estimation is a data-driven method. The input is the metadata of a test video. From the metadata, we only use the user-annotated tags (not the title, or descriptions) that are included in the metadata record for each Flickr video or photo. We also experimented with using title and descriptions but the results were significantly worse than only using the tags. Furthermore, 2601 of the 5125 videos in the test data did not contain a description. The algorithm is described as follows.

For each given tag in the test video record, we determine the spatial variance by searching the training data for an exact match of the tag and creating a list of the geo-locations of the matches. If only one location is found, the spatial variance is trivially small. We pick the centroid location of the top-3 tags with the smallest spatial variance. This results in 0 to 3 coordinates. In the case of 0 coordinates (e.g. because the video is not tagged or no tags match), we assume the most likely geo-coordinate based on the prior distribution of the MediaEval training set (see Figure 1), which is the point with latitude and longitude (40.71257011, -74.10224). A place close to New York City. For example, if a test video's metadata contains the tags "Campanile", "Berkeley", and "California", the system would match all training videos that contain any of those tags. We then plot the GPS coordinates of the training videos containing the tags "Campanile", "Berkeley", and "California" and select the centroid of the tag with the smallest spacial extent (in this case, "Campanile") as our final location.

Fig. 1. Distribution of the videos of the MediaEval 2010 Placing Task development set. As discussed in Section III, randomly sampling videos from Flickr results in a non-uniform geographical distribution.

B. Semantic Approach

As discussed in Section II, related work has tried using databases of named location information, known as gazetteers, to increase the robustness of the search. Also, Flickr provides the home location of the user of an uploaded video which could be treated as an equivalent to a user-based gazetteer as every user can be mapped to a place on earth. We therefore performed experiments to see if the incorporation of this type of information would be useful. We used the open service *Geonames.org*. GeoNames covers all countries and contains 8 million entries of place names and corresponding geo-coordinates. It provides a web-based search engine and an API which returns a list of matching entries ordered by their relevance to the query. A single keyword may cause ambiguity by representing multiple entities (e.g. "Paris Texas" vs. "Paris France"). Thus it is crucial to find a combination of keywords that minimizes ambiguity if possible. A computationally inefficient but effective way to do this, is to query the Geonames database exhaustively for every possible combination of keywords. To reduce the run time of the search, we filtered the keywords using a Bloom Filter [1] built over the downloaded database of Geonames. In this method, all compound keywords of every length were tested (e.g. "sanfrancisco" and "San Francisco" were both in the Bloom Filter). If the Bloom Filter returned positive, they were added to a candidate list. The Bloom Filter may sometimes return false positives, but these were assumed to be removed by the Geonames search engine. Tags were concatenated into a string in their original order. The order is preserved to handle the context within compound words such as "San Francisco" or "Washington DC".

One problem with using a gazetteer is that it has no background model of words that are likely to appear in regular language, i.e. it does give positive results on words such as "video" and "vacation" because there is a city of Video in Brazil and a Vacation Island in San Diego. Therefore we filtered out common nouns by using Augmented-WordNet [15]. Augmented-WordNet is an extended version of WordNet [3], that among other things includes annotation for geographical entities. WordNet is a freely available online lexical database of English which contains a network of semantic relationships between words. Note that Flickr videos and photos are annotated in any language so this approach only helped for the English subset.

After filtering, we passed the query to the Geonames search engine and retrieved the list of possible matches. We added the entity with the highest relevance (the first entity in the response list) to the list of candidate entities. Once we obtained the list of candidate entities, we resolved the containment problem (e.g. "Fisherman's Wharf, San Francisco, CA"): Geonames entities provide country code, code of administrative subdivision (typically the city), and feature class parameters. We gave higher priority to entities representing a smaller region (as of Geonames) by removing larger entities containing the smaller entities.

Choosing the best match among the list of candidates is similar to the method we used in Section IV-A. We plot all candidate entities on a map and pick the one that has the largest count of neighbors with lowest spatial variance. If there is a tie, the coordinate that is closest to the user's home location is picked (as described in the videos' metadata). If there is no matching entity for all keywords in the metadata of a given video, we apply two backup steps. First, we return the geo-coordinate of the user's home location. This is better than a blind guess on the prior, since our observations found that people tend to under-annotate videos about their ordinary

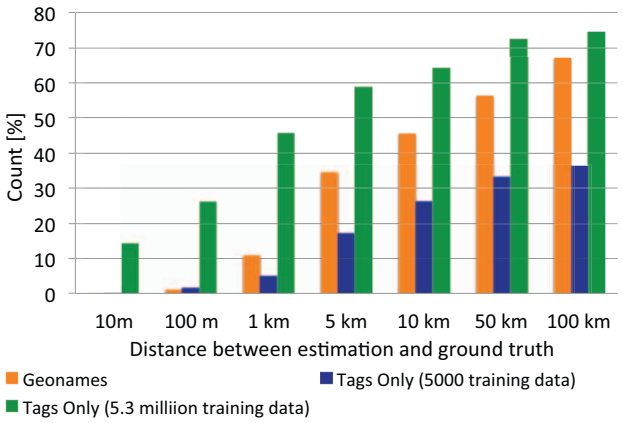

Fig. 2. Comparing the use of a geographical gazetteer versus the data-driven approach in Section IV-A with different training data volumes. See also discussion in Section V.

everyday life, which tend to have been recorded close to where they reside. If a video did not contain the user's home location, we used the default location close to New York City, as explained in Section IV-A.

V. Results

We found that incorporating gazetteer information can help significantly with sparse datasets. However, with enough sample records, tag matching as described in Section IV-A outperforms the gazetteer approach, even when incorporating the Flickr-specific home location as described above. Figure 2 shows the results comparing tag matching and using Geonames plus a user's home location. To compare two sparseness levels, we tried the data-driven approach using the basic MediaEval dataset and the full dataset (as described in Section III).

VI. Conclusion

In this article we described two systems for the estimation of the recording location of Flickr videos based on tags. One system is a purely data-driven algorithm and the other system uses semantic technologies. We found that the use of gazetteer data and other semantic technologies is helpful, but mostly in situations were not enough training data is available. We therefore conclude that the usefulness of semantic databases for this task depends on the application scenario. Semantic approaches should be especially interesting for forensic and intelligence use cases as the reliance on publicly available videos might not be optimal for these kinds of applications. However, when sorting and organizing a collection of touristic photos and videos for private use, data-driven approaches might probably give better accuracy.

Further information about the project can be found at http://mmle.icsi.berkeley.edu.

Acknowledgments

This research is supported by an NGA NURI grant #HM11582-10-1-0008.

References

[1] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.
[2] J. Choi, A. Janin, and G. Friedland. The 2010 ICSI Video Location Estimation System. In *Proceedings of MediaEval*, October 2010.
[3] C. Fellbaum. *WordNet: An electronic lexical database*. The MIT press, 1998.
[4] D. Ferres and H. Rodriguez. TALP at MediaEval 2010 Placing Task: Geographical Focus Detection of Flickr Textual Annotations. In *Proceedings of MediaEval*, October 2010.
[5] G. Friedland, O. Vinyals, and T. Darrell. Multimodal Location Estimation. In *Proceedings of ACM Multimedia*, pages 1245–1251, 2010.
[6] A. Gallagher, D. Joshi, J. Yu, and J. Luo. Geo-location inference from image content and user tags. In *Proceedings of IEEE CVPR*. IEEE, 2009.
[7] J. Hays and A.A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8, 2008.
[8] P. Kelm, S. Schmiedeke, and T. Sikora. Video2GPS: Geotagging using collaborative systems, textual and visual features: MediaEval 2010 Placing Task. In *Proceedings of MediaEval*, October 2010.
[9] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and Gareth J.F. Jones. Automatic Tagging and Geo-Tagging in Video Collections and Communities. In *ACM International Conference on Multimedia Retrieval (ICMR 2011)*, page to appear, April 2011.
[10] Mediaeval web site. *http://www.multimediaeval.org*.
[11] J.M. Perea-Ortega, M.A. Garcia-Cumbreras, L.A. Urena-Lopez, and M. Garcia-Vega. SINAI at Placing Task of MediaEval 2010. In *Proceedings of MediaEval*, October 2010.
[12] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1), 2009.
[13] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *ACM SIGIR*, pages 484–491, 2009.
[14] B. Sigurbjoernsson and R. Van Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *ACM WWW*, pages 327–336, April 2008.
[15] R. Snow, D. Jurafsky, and A.Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 801–808. Association for Computational Linguistics, 2006.
[16] O. Van Laere, S. Schockaert, and B. Dhoedt. Ghent University at the 2010 Placing Task. In *Proceedings of MediaEval*, October 2010.