The 2010 ICSI Video Location Estimation System

Jaeyoung Choi
ICSI
1947 Center St. Suite 600
Berkeley, CA 94704, USA
jaeyoung@icsi.berkeley.edu

Adam Janin
ICSI
1947 Center St. Suite 600
Berkeley, CA 94704, USA
janin@icsi.berkeley.edu

Gerald Friedland
ICSI
1947 Center St. Suite 600
Berkeley, CA 94704, USA
fractor@icsi.berkeley.edu

ABSTRACT

In this paper, we describe the International Computer Science Institute's (ICSI's) video location estimation system for the MediaEval 2010 Placing Task. We propose two approaches: using the prior distribution of training dataset; and using GeoNames[1], a geographical gazetteer, for toponym (placename) resolution. Both approaches use textual metadata only. We show that the location of a video can be estimated with reasonable accuracy using our system.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Algorithms, Performance, Experimentation

Keywords

video localization, placing task, MediaEval

1. INTRODUCTION

MediaEval's 2010 Placing Task was to automatically guess the location (latitude and longitude) of each test videos using any or all of metadata, visual/audio content, and/or social information. Metadata contains user annotated title, tags, description, and the user's home location among other information. Herein, we only considered metadata.

2. SYSTEM DESCRIPTION

First, we manually examined random videos from the training set. We found that most lacked reasonable visual or audio cues to estimate their location, even for a human. For instance, it was nearly impossible to estimate the location of videos recorded indoors using visual and audio content alone.

On the other hand, 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 one 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".

2.1 Using the prior distribution of tags

Our first approach was a data-driven method based on picking the keyword with the smallest spatial variance in its prior distribution. The algorithm was as follows: For each keyword in a given video, search for all videos in the dataset that contain an exact match. For each such matching video, plot its location on a map. Pick the one coordinate that has the largest count of other points within a given radius[2]. Divide this count by the total number of matches to normalize. After doing this for all keywords, pick a geo-coordinate with the largest normalized count of neighbors. The intuition was to pick a keyword whose prior distribution on the world map had the smallest spatial variance and to choose the most central location in the distribution.

Instead of searching for a matching keyword in *all* videos within the dataset, we may limit the search to just the videos uploaded by the same user. Since each person has his own idiosyncratic method of choosing a keyword for certain events, this scheme has a higher likelihood of finding geographically related videos. If the user was never seen before, we used all videos from all users.

2.2 Using GeoNames for toponym resolution

Our second approach was a supervised resolution of toponyms using GeoNames, a geographical gazetteer. GeoNames covers all countries and contains 8 million entries. It provides a web based search engine 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 was crucial to find a combination of keywords that minimizes ambiguity if possible. A pricey but effective way to do this is to search the GeoNames DB exhaustively for every possible combination of keywords. To reduce the running 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

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

[2] 1 km resulted in the best accuracy over the training set.

MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy

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 removed by the GeoNames search engine. For this approach, the title, tag and description were used. Tags were concatenated into a string in their original order. Order was preserved to handle the context within compound words such as "San Francisco" or "Washington DC".

One problem with this was that GeoNames included generic words such as "video" and "vacation", since there is a city of Video in Brazil and a Vacation Island in San Diego. We filtered out these common nouns by looking at their part-of-speech tag retrieved using Augmented-WordNet (WordNet[3] is a freely available online lexical database of English which contains a network of semantic relationships between words; Augmented-WordNet[4] is an extended version of WordNet with more data).

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 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 provides country code, code of administrative subdivision (typically the city), and feature class parameters that we use to resolve the containment problem. We gave more weight to the smaller entity by removing the larger entity and adding the smaller entity twice to the list, thereby doubling the weight.

Choosing the best guess among the list of candidates was similar to the method we used in Section 2.1 — plotting all candidate entities on a map, and picking the one that has the largest count of neighbors within a given radius. If there was a tie, the coordinate that was closest to the user's home location was picked.

If there was no matching entity for all keywords in the metadata of a given video, we applied two backup steps. First, we returned the geo-coordinate of the user's home location. This is better than blind guess based on priors, since our observations found that people tend to under-annotate videos about their ordinary everyday life, and these tend happen close to where they live. If a video did not contain the user's home location, we used the location with the highest concentration in the prior, which in our case was the east coast of the US.

3. EXPERIMENTS

We performed cross-validation on the training set using the prior distribution of tags of same-user, and it gave us an unexpectedly good result as can be seen in Figure 1. 42.1% of videos were guessed right on top of their ground truth, and 85.2% were within the 100 km range of ground truth.

However, videos in the test dataset were chosen to have a disjoint set of users from the training dataset. Thus, this approach did not have any effect in our official run. The best result of our official run can be seen in Figure 2. 10.7% of estimated points lie within 1 km radius of the ground truth and about 60% lie within 100 km radius of ground truth. 1 km was used as the radius value for counting the neighbors around each candidate point.

[3]http://wordnet.princeton.edu
[4]http://ai.stanford.edu/~rion/swn

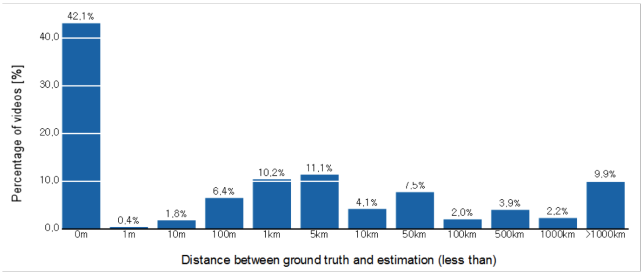

Figure 1: Cross-validation result of unofficial run on training dataset using prior distribution approach.

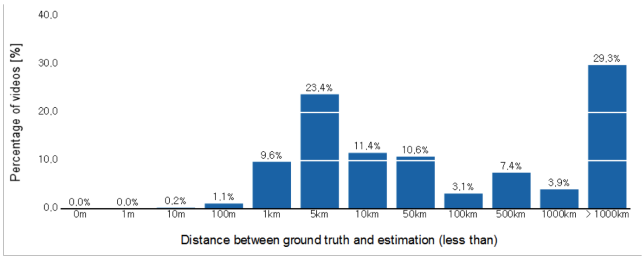

Figure 2: Result of official run on test dataset using the mixture of GeoNames and prior distribution approaches, followed by two backup methods.

4. DISCUSSION

When the user provides sufficient tags in the metadata, and the tags are location-specific to where the video was taken, our approach has the potential to return the location very accurately. However, there are some videos which confusingly contain toponyms in their metadata to describe an incident or an object which is not proximal to where the video was recorded (e.g. "Goodbye Oregon, hello San Francisco"). These cases are much more difficult, as the machine needs to "understand" the context of metadata to pick the correct toponym.

The future work is to explore robust data-driven probabilistic models instead of the current rule-based approach which relies heavily on the coverage of GeoNames. Utilizing visual and audio cue such as [2] and designing a novel framework for putting all available cues together are another challenging tasks to address.

5. ACKNOWLEDGMENTS

We would like to thank Michael Ellsworth for his advice on filtering issue. This research is partly funded by NGA NURI grant.

6. REFERENCE

[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. Hayes, and A. Efros. IM2GPS: estimating geographic information from a single image. *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.