Video2GPS: A Demo of Multimodal Location Estimation on Flickr Videos

Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
fractor@icsi.berkeley.edu

Jaeyoung Choi
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
jaeyoung@icsi.berkeley.edu

Adam Janin
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
janin@icsi.berkeley.edu

ABSTRACT

The following article describes an approach to determining the geo-coordinates of the recording place of Flickr videos based on both textual metadata and visual cues. The underlying system has been tested on the MediaEval 2010 Placing Task evaluation data, which consists of 5091 unfiltered test videos is able to classify 14 % of the videos to within an accuracy of 10 m.

Categories and Subject Descriptors

H3.1 [**Information Storage and Retrieval**]: Indexing methods; I4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor Fusion*

General Terms

Experimentation

Keywords

Video, Tagging, Multimodal, Location Estimation, Content Analysis

1. INTRODUCTION

A multimedia content analysis task that has only recently become tractable to research 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.

This article describes an approach to determining the geo-coordinates of the place where Flickr videos were recorded based on textual metadata and visual cues. The system was tested on the MediaEval 2010 Placing Task evaluation data and is able to classify 14 % of the videos to within 10 m or less.

MM'11, November 28–December 1, 2011, Scottsdale, AZ

2. RELATED WORK

In earlier articles [7, 8, 3], the location estimation task is reduced to a retrieval problem on self-produced, location-tagged image databases. The idea is that if the image is the same then the location must be the same too. Previous work that has been carried out in the area of automatic geo-tagging of multimedia based on tags, has also been mostly carried out on Flickr images. The approach in [2] 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 [1] 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 [4]. Several notable systems participated in the evaluation, including the predecessor of the system described herein. We have made significant strides in accuracy since the Placing Task evaluation in August 2010, while at the same time using less training data and reducing the complexity of the system.

3. ALGORITHM

Our approach to location estimation is a data-driven multimodal method that uses both the textual tags as well as visual features. The input is a test video with metadata. 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. The algorithm is described as follows.

First we process the tags. 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, which is the point with latitude and longitude (40.71257011,

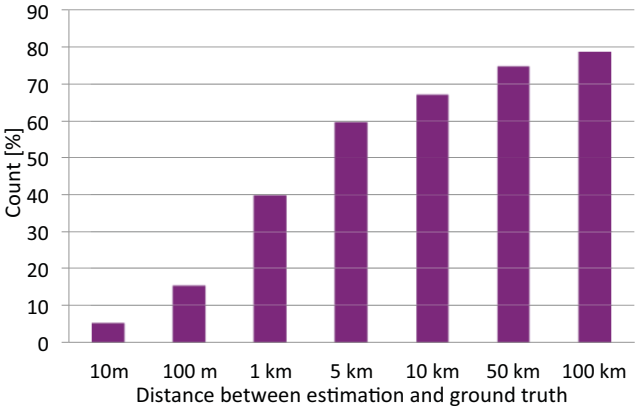

Figure 1: The resulting accuracy of the algorithm as described in Section 3.

-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").

For the visual processing step, the input is the median frame of the test video and the 1 to 3 coordinates of the previous stop. We resize the frame to 256×256 pixels and extract GIST [6] features and color histogram. The GIST descriptor is based on a 5×5 spatial resolution with each bin containing responses to 6 orientation and 4 scales. The color histograms were created based on the CIELAB transformed pixels for the frame, like in [3]. The histogram has 4 bins for L, and 14 bins for A and B, respectively. We then also adopt the matching methodology from [3]. We used Euclidean distance to compare GIST descriptors and chi-square distance for color histograms. Weighted linear combination of distances was used as the final distance between the training and test frames. The scaling of the weights was learned by using a small sample of the training set and normalizing the individual distance distributions so that each the standard deviation of each of them would be similar. We use 1-nearest neighbor matching between the test frame and the all the images in a 100 km radius around the 1 to 3 coordinates from the tag-processing step. We pick the match with the smallest distance and output its coordinates as a final result.

4. RESULTS

The MediaEval 2010 Placing Task, organized by [5], 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. The data set consists of Creative Common-licensed videos that were manually crawled from Flickr divided into training data (5091 videos) and test data (5125 videos). The evaluation of our results is performed by applying the same rules and using the same metric as in the MediaEval 2010 evaluation, i.e. by calculating the geographical distance from the actual geo-location of the video (assigned by a Flickr user, creator of the video) to the predicted geo-location (assigned by the system). While it was important to minimize the distances over all test videos, runs were compared by finding how many videos were placed within a threshold distance of 1 km, 5 km, 10 km, 50 km and 100 km. For analyzing the algorithm in greater detail, here we also show distances of below 100 m and below 10 m. The lowest distance category is about the accuracy of a typical GPS localization system in a camera or smartphone. The results are visualized in Figure 1. The results shown are superior in accuracy than any system presented in MediaEval 2010. At the same time, our approach relies on less data and its implementation seems to be the least complex compared to related work.

5. CONCLUSION

In this article we described a system for the estimation of the recording location of Flickr videos. The system uses tags as well as video content and achieves significant accuracy improvements due to the integration of the two media. The demo is a web interface to the system. It allows a user to specify the URL of a Flickr video and the system shows the results step by step, including the tags that contributed most to the estimation. 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.

6. REFERENCES

[1] G. Friedland, O. Vinyals, and T. Darrell. Multimodal Location Estimation. In *Proceedings of ACM Multimedia*, pages 1245–1251, 2010.

[2] 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.

[3] J. Hays and 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.

[4] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. 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.

[5] Mediaeval web site. *http://www.multimediaeval.org*.

[6] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[7] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–7, 2007.

[8] W. Zhang and J. Kosecka. Image based localization in urban environments. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pages 33–40, 2006.