CEA LIST's Participation at MediaEval 2012 Placing Task

Adrian Popescu
CEA, LIST,
Vision & Content Engineering Laboratory,
91190 Gif-sur-Yvette, France.
adrian.popescu@cea.fr

Nicolas Ballas
CEA, LIST,
Vision & Content Engineering Laboratory,
91190 Gif-sur-Yvette, France.
nicolas.ballas@cea.fr

ABSTRACT

Multimedia content geotagging is potentially useful in a wide variety of application. Although an increasing number of existing devices include geotagging options, a wide majority of online content is still not geotagged and methods for efficient automatic geotagging are needed. Here we describe CEA LIST's participation to the MediaEval 2012 Placing Task [3]. We submitted runs that exploit either textual or visual information to place videos on the map and we will briefly introduce the methods used. The main innovation with respect to state of the art methods was to create user geotagging models and to rely on them if useful textual annotation are missing. Exploiting user models proves efficient since geotagging accuracy is significantly improved compared to the use of generic language models and/or gazetteers.

Keywords

Geotagging,Video Annotation, Gazetteer, User records

1. INTRODUCTION

Location is one important feature associated to multimedia content and there is growing interest in developing methods for efficient content placing. Early works include [2] for visual-based and [4] for text-based placing methods and reported results are very promising. However, existing methods assume, in a large majority of cases, that data flows are generated by unknown sources and disregard user information. We build on state of the art algorithms but also introduce a user-oriented approach based on her past behavior. We produced runs that exploit different data sources and discuss text- and visual-based placing separately since they were not combined.

2. TEXT-BASED VIDEO PLACING

2.1 Language Models Based Placing

We propose three textual runs, out of which two are inspired by language models described in [4] and the third also exploits user geotagging models. The surface on the Earth is split in (nearly) rectangular cells characterized by a set of tags and their probability of occurrence in that cell. Given that the smallest location precision tested during the task is 1 km, we chose to create cells that have approximately this size and split the map in rectangles of size 0.01 of latitude and longitude degree. Flickr users are free to tag their content with any textual string they want and many of these annotations are useful only to them or subjective. In order to select socially relevant tags and to keep the size of the tag vocabulary easily tractable, we selected only tags that were used by at least two users. This filtering process generates a vocabulary of around 220,000 tags that is used for all subsequent processing. Cell tag probability is computed as the number of different users that used the tag in the cell divided by the overall tag's user count. This user-based probability avoids the creation of skewed models due to bulk tagging and to intensive use of a tag by a single user. Not all tags have the same importance in the geographic domain and we want to favor terms that are geographically discriminant. Classically, lists of toponyms are used to select possible locations of a document but toponyms are often ambiguous and we propose to exploit spatially unambiguous pairs of toponyms instead. Preliminary tests showed that the use of unambiguous pairs to emphasize the importance potential places improves results compared to plain language models.

RUN 1 exploits only internal data provided for the task and unambiguous pairs of potential toponyms were extracting by selecting all pairs of tags whose probability of occurrence within a radius of 50 km is higher than 0.95. These values were empirically chosen so as to have a reasonable number of discovered pairs - 600,000. RUN 2 exploits Gazetiki [1], an publicly available gazetteer, and unambiguous pairs are formed of local POIs and of encompassing regions (cities, regions, countries). It also exploits unambiguous location names that are not frequent words from the general vocabulary.

To find the cells that are the most probable locations of a video, we first test whether an unambiguous pair is found. We then compute the dot product between the target's tags and all cells models either around the unambiguous pair or over the entire set of cells. The top cell is sometimes an isolated one and we hypothesize that a spatial clustering that discovers dense regions among the most similar cells is helpful. The top 5 cells are retained as potential locations

MediaEval 2012 Workshop October 4-5, 2012, Pisa, Italy
.

[1] http://georama-project.labs.exalead.com/gazetiki.htm

and we search for the number of neighbors (within a radius of 10 km) that are also present among the top 150 neighbors. The number of seed cells and the size of the similar cells set were chosen after testing a large number of values on the learning set. In addition, we place all videos that do not have textual metadata associated at the center of the cells that have the highest number of associated geotagged videos.

2.2 User Based Video Placing

Existing video placing methods, based on generic language models, hypothesize that location tagging probabilities are the same for all users. However most users take photos in a limited number of locations and a majority of them geotag a large amounts of content around their home location. To account for user behavior, we create user models that are based on their past geotagging behavior. We download up to 3000 geotagged metadata pieces per user and compute the probability for a user to tag in a given cell (photos in a cell divided by total number of photos). We avoid learning on test data by discarding all photos that were taken or uploaded on the same day as the photos present in the test set. We computed the proportion of images found in most populous 1 km cells of their models to illustrate user behavior: 24.2% of the photos are found in the top cell, 42.2% in the top 3 cells and 51.9% in the top 5 cells.

The creation of user models based on past experience operate under a closed world assumption as it is impossible to add new locations to the model without the user's explicit intervention. To avoid such problems, RUN 3 exploits generic language models whenever location metadata (unambiguous pairs or unambiguous locations used for RUN 2 are present) and user models for the other videos.

3. CONTENT-BASED VIDEO PLACING

We submitted a fourth run (RUN 4) that exploits uniquely video content. We investigate two feature types that capture either local image properties or motion information. Nearest-neighbor search between the video features is then performed to place a target video with respect to a ground truth set.

We use the video samples provided by the organizers to compute SURF features using a dense sampling approach. A bag-of-words model is then used to transform the local SURF-descriptor associated with the video frames into a global video representation. A dictionary of size 4096, learned by k-means on the training dataset, is used to construct the bag-of-words model. The test set formed of approximately 1 million geotagged Flickr images that are also tagged with POI names.

Dense point trajectories have also been investigated to describe videos. Keypoints are densely sampled at multiple spatial scales in each video frame. Dense optical flow is then used to match a point frame to frame. The accumulation of frame to frame point correspondences forms a trajectory. Trajectory descriptors are computed using the approach described in [1]. Finally the local motion trajectory descriptors are aggregated using bag-of-words model with a dictionary of size 1000. The test set is comprises the geotagged videos provided as ground truth by the organizers.

We consider the 50 nearest neighbors of a target video to perform spatial clustering and compute the number of nearest neighbors falling within a specific distance d (here $d = 5$ km) of the seeds. The test video is associated with location that contains the maximum number of nearest neighbors. We observed that the motion features have lower performance and RUN 4 was produced using SURFs.

Table 1: Videos placed accurately at different scales

Run name	1 km	10 km	100 km	1000 km
RUN 1	452	994	1264	1773
RUN 2	447	990	1260	1771
RUN 3	1166	2008	2571	3070
RUN 4	3	8	32	483

4. RESULTS AND DISCUSSION

The results obtained with the textual and visual approaches are presented in Table 1. Text based runs have much better performances because place recognition from visual content only is a very difficult problem. The number of potential locations is very high and their visual appearance very diversified. There is also a problem due to the difference in quality between video keyframes and photos in the ground truth.

RUN 1 and RUN 2 are of equivalent quality and it is surprising that the use of a rich gazetteer does not improve results compared to the use of unambiguous place descriptions extracted directly from data. RUN 3 shows that the introduction of user models has a significant positive effect on results. Such an approach is useful if the data source is known, which is usually the case for social media, and if a user model can be derived from past contributions. As with other activities, geotagging is closely related to our daily habits and their proper modeling is beneficial for user characterization.

Future work includes an improvement of more semantically motivated cells than rectangular ones and tighter integration of user models and of generic language models. Also further tests are needed to understand why the gazetteer based method does not improve results compared to the approach described for RUN 1. For visual placing we will test more advanced visual descriptors and will increase the size of the ground truth to have improved coverage of the Earth's surface.

5. REFERENCES

[1] N. Ballas, B. Delezoide, and F. Preteux. Trajectories based descriptor for dynamic events annotation. In *Proc. of the 2011 joint ACM workshop on Modeling and representing events.*

[2] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Proc. of CVPR 2008.*

[3] A. Rae and P. Kelm. Working notes for the placing task at mediaeval 2012. In *MediaEval Multimedia Evaluation Workshop 2012*, 2012.

[4] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *Proc. of SIGIR 2009.*