Working Notes for the Placing Task at MediaEval 2012*

Adam Rae
Yahoo! Research
adamrae@yahoo-inc.com

Pascal Kelm
Technische Universität Berlin
kelm@nue.tu-berlin.de

ABSTRACT

This paper provides a description of the MediaEval 2012 Placing Task. The task requires participants to automatically assign latitude and longitude coordinates to each of the provided test videos. This kind of geographical location tag, or geotag, helps users localise videos, allowing their media to be anchored to real world locations. Currently, however, most videos online are not labelled with this kind of data. This task encourages participants to find innovative ways of doing this labelling automatically. The data comes from Flickr—an example of a photo sharing website that allows users to both encode their photos and videos with geotags, as well as use them when searching and browsing. This paper describes the task, the data sets provided and how the individual participants results are evaluated.

Keywords

Geotags, Location, Video Annotation, Benchmark

1. INTRODUCTION

This task invites participants to propose new and creative approaches to tackling the problem of automatic annotation of video with geotags and to extend the current state of the art. These tags are usually added in one of two ways: by the photo device (e.g. camera or camera-equipped mobile phone) or manually by the user. An increasing number of devices are available that can automatically encode geotags, using satellite-based positioning systems, mobile cell towers or look-up of the coordinates of local Wi-Fi networks. Users are also becoming more aware of the value of adding such data manually, as shown by the increase in photo management software and websites that allows users to annotate, browse and search according to location (e.g. Flickr, Apple's iPhoto and Aperture, Google Picasa WebAlbums).

However, newly uploaded digital media and videos in particular, with any form of geographical data, are still relatively rare compared to the total quantity uploaded. There is also a significant amount of data that has already been uploaded that does not currently have geotags.

*This work was supported by the European Commission under contract FP7-248984 GLOCAL and FP7-261743 VideoSense.

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

This task challenges participants to develop techniques to automatically annotate videos using their visual content and some selected, associated textual metadata. In particular, we wish to see those taking part extend and improve upon the work of previous tasks at MediaEval and elsewhere in the community [5, 2, 1, 3, 6].

2. DATA

The data set is an extension of the MediaEval 2010 Placing Task data set [3] and contains a set of geotagged Flickr videos as well as the metadata for geotagged Flickr images. A set of basic visual features extracted for all images and for the frames of the videos is provided to participants. All selected videos and images are shared by their owners under the Creative Commons license.

2.1 Development data

Development data is the combination of the development and test data from the MediaEval 2010 and 2011 Placing Tasks. The sets are pooled to form the 2012 development set. We provide as much publicly accessible metadata as possible to participants, giving them a variety of information sources for use when predicting locations. This includes the title, tags (labelled *Keywords* in the provided metadata files), description and comments. We also include information about the user who uploaded the videos and about his/her contacts, his/her favourite labelled images and the list of all videos she/he has uploaded in the past. It should be emphasised that the task requires the participants to predict the latitude and longitude for each video. The prediction of the names of locations or other geographic context information is outside the scope of this task.

The development set comes with the ground truth values for each video. This information is contained in the metadata in the field <Location>. Video keyframes are extracted at 4 second intervals from the videos and saved as individual JPEG-format images, using the freely available *ffmpeg*[1] tool. For development purposes, we distribute metadata for 3,185,258 Flickr photos uniformly sampled from all parts of the world, using geographic bounding boxes of various sizes via the Flickr API[2]. Whilst the images themselves are not distributed in this task, they are publicly accessible on Flickr (if they have not been removed since the data set was gathered) and the provided metadata contains links to the source images.

[1] `http://www.ffmpeg.org/`
[2] `http://www.flickr.com/services/api/`

From these images, their existing metadata is extracted. Most, but not all, photos have textual tags. All photos have geotags of at least region level accuracy. The accuracy attribute encodes at which zoom level the uploader used when placing the photo on a map. There are 16 zoom levels and hence 16 accuracy levels (e.g., 3 - country level, 6 - region level, 12 - city level, 16 - street level).

While these images and their metadata are potentially helpful for development purposes, the evaluation test set, however, only includes videos.

We also generated visual feature descriptors for the extracted video keyframes and training images, using the open source library *LIRE* [4] available online[3], with the default parameter settings and the default image size of 500 pixels on the longest side. This feature set comprises of the following:

- Colour and Edge Directivity Descriptor
- Gabor Texture
- Fuzzy Colour and Texture Histogram
- Colour Histogram
- Scalable Colour
- Auto Colour Correlogram
- Tamura Texture
- Edge Histogram
- Colour Layout

The Scalable Colour Edge Histogram and Colour Layout features are implemented as specified in the MPEG-7 schema.

3. GROUND TRUTH AND EVALUATION

The geo-coordinates associated with the Flickr videos will be used as the ground truth. Since these do not always serve to precisely pinpoint the location of a video, the evaluation will be carried out at each of a series of widening circles: 1km, 10km, 100km, 1000km, 10000km. If a reported location is found within a given circle radius, it is counted as correctly localised. The accuracy over each circle will be reported. To evaluate the performance of each technique, the geodesic distance between the ground truth coordinates and those of the output from a participants system were compared. To take into account the geographic nature of the task, the Haversine distance is used. This measure is calculated thus:

$$d = 2 \cdot r \cdot \arcsin\left(\sqrt{h}\right) \tag{1}$$

$$h = sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1)cos(\phi_2)sin^2\left(\frac{\psi_2 - \psi_1}{2}\right) \tag{2}$$

where d is the distance between points 1 and 2 represented as latitude (ϕ_1, ϕ_2) and longitude (ψ_1, ψ_2) and r is the radius of the Earth (in this case, the WGS-84 standard value of 6,378.137km is used).

4. TASK DETAILS

Participants may submit between three and five runs. They can make use of image metadata and audio and visual features, as well as external resources, depending on the run. A minimum of one run that uses only audio/visual features is required. The other two required runs allow for the free use of the provided data (but no other), with either the option of using a gazetteer or not. Participants may submit an optional additional run that uses a gazetteer, as well as a optional run that allows for the crawling of additional material from outside of the provided data (the *general* run).

Participants are not allowed to re-find the provided videos on-line and use actual geotags (or other related data) for preparing their runs. This is to ensure that participants help contribute to a realistic and sensible benchmark in which all test videos as "unseen". The participants are also asked to not crawl Flickr for any additional videos or images and use only those provided in the data sets (with exception made for the optional *general* run).

The runs that can be submitted for evaluation are as follows:

1. run (required): Anything is accepted, except for crawling additional web material or using a gazetteer.
2. run (required): Anything is accepted, except for crawling additional web material (gazetteer permitted).
3. run (optional): Anything is accepted, except for crawling additional web material (gazetteer permitted).
4. run (required): Using only audio-vased and visual features is accepted.
5. run (optional): Anything is accepted, except for crawling the exact items contained in the test set (crawling additional material, gazetteer permitted)

[3] `http://www.semanticmetadata.net/lire/`

5. REFERENCES

[1] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[2] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, multi-resource methods for placing flickr videos on the map. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 52:1–52:8, New York, NY, USA, 2011. ACM.

[3] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. Jones. Automatic tagging and geotagging in video collections and communities. In *ACM International Conference on Multimedia Retrieval (ICMR 2011)*, 2011.

[4] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In *Proceeding of the 16th ACM international conference on Multimedia*, MM '08, pages 1085–1088, New York, NY, USA, 2008. ACM.

[5] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 484–491, New York, NY, USA, 2009. ACM.

[6] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of flickr resources using language models and similarity search. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 48. ACM, 2011.