The Placing Task at MediaEval 2015

Jaeyoung Choi[1/2], Claudia Hauff[2], Olivier Van Laere[3], and Bart Thomee[4]
[1]International Computer Science Institute, Berkeley, USA
[2]Delft University of Technology, the Netherlands
[3]Blueshift Labs, San Francisco, USA
[4]Yahoo Labs, USA
jaeyoung@icsi.berkeley.edu, c.hauff@tudelft.nl, oliviervanlaere@gmail.com, bthomee@yahoo-inc.com

ABSTRACT

The sixth edition of the Placing Task at MediaEval introduces two new sub-tasks: (1) , which emphasizes the need to move away from an evaluation purely based on latitude and longitude towards an entity-centered evaluation, and (2) , which addresses predicting missing locations within a sequence of movements; the latter is a specific real-world use case that so far has received little attention within the research community. Two additional changes over the previous years are the introduction of open source for both sub-tasks shortly after the official data release, and the implementation of a , which allows the participants to gain insights into the effectiveness of their approaches compared to the official baselines and in relation to each other at an early stage, before the actual run submissions are due.

1. INTRODUCTION

The Placing Task challenges participants to develop techniques to automatically annotate photos and videos with their geolocation using their visual content and/or textual metadata. In particular, we wish to see those taking part to extend and improve upon the contributions of participants from previous editions, as well as of the research community at large, e.g. [7, 10, 3, 1, 5, 8]. Although the Placing Task has indeed been shown to be a "research catalyst" [6] for geoprediction of social multimedia, with each edition of the task it becomes a greater challenge to alter the benchmark sufficiently to allow and motivate participants to make substantial changes to their frameworks and systems instead of small technical ones—this year's introduction of organizer baselines, a leaderboard, as well as novel sub-tasks were driven by this consideration.

2. DATA

This year's edition of the Placing Task was based on the `YFCC100M`[1] [9], which to date is the largest social multimedia collection that is publicly and freely available. The full dataset consists of 100 million Flickr[2] Creative Commons[3] licensed photos and videos with associated metadata. Similar to last year's edition [2], we sampled a subset of the `YFCC100M` for training and testing, see Table 1. The need for two separate datasets arose from the task requirements (described in Section 3). No user appeared both in the training set and in the test set, and to minimize user and location bias, each user was limited to contributing at most 250 photos and 50 videos, where no photos/videos were included that were taken by a user less than 10 minutes apart. The rather uncontrolled nature of the data (sampled from longitudinal, large-scale, noisy and biased raw data) confronts participants with additional challenges. To lower the entrance barrier, we precomputed and provided participants with fifteen visual, and three aural features commonly used in multimedia analysis for each of the media objects including SIFT, Gist, color and texture histograms for visual analysis, and MFCC for audio analysis [2].

Training		Testing	
#Photos	#Videos	#Photos	#Videos
4; 672; 382	22; 767	931; 573	18; 316
148; 349	0	33; 026	0

Table 1: Overview of training and test sets for both sub-tasks.

3. TASKS

Locale-based sub-task: In this sub-task, participants were given a hierarchy of places across the world, ranging across neighborhoods, cities, regions, countries and continents. For each photo and video, they were asked to pick a node (i.e. a place) from the hierarchy in which they most confidently believe it had been taken. While the ground truth locations of the photos and videos were associated with the most accurate nodes (i.e. the leaves) in the hierarchy, the participants could express a reduced confidence in their location estimates by selecting nodes at higher levels in the hierarchy. If their confidence was sufficiently high, participants could naturally directly estimate the geographic coordinate of the photo/video instead of choosing a node from the hierarchy.

As our place hierarchy we used version 2.0 of the open source GADM database[4], which contains the spatial boundaries of the world's administrative areas. As the GADM only

[1]`https://bit.ly/yfcc100md`
[2]`https://www.flickr.com`
[3]`https://www.creativecommons.org`
[4]`http://www.gadm.org`

Sept. 14–15, 2015, Wurzen, Germany

contains data up to city level, we manually supplemented it with neighbourhood data for several cities obtained from the geo-game ClickThatHood[5]. In total, the hierarchy contains 221,458 leaf nodes that are spread across 253 countries. The hierarchy has a maximum depth of 7 and an average depth of 4.33, with each place being a variation of the general hierarchy:

Country→State→Province→County→City→Neighborhood

Due to the use of the hierarchy, only photos and videos taken within any of the GADM boundaries were part of this subtask, and thus media captured in or above international waters were excluded.

Mobility-based sub-task: In this sub-task, participants were given a of photos taken in a certain city by a specific user, of which not all photos were associated with a geographic coordinate (e.g. the user took some photos when GPS was temporarily unavailable). The participants were asked to predict the locations of those photos with missing coordinates. The nearly 150K training photos of this sub-task were divided into 23,116 sequences, while the approximately 33K test photos were separated into 5,119 sequences. From each sequence in the test set about 30% of the coordinates were missing, which are the ones that needed to be predicted.

4. RUNS

Participants may submit up to five attempts ('runs') for each sub-task. They can make use of the provided metadata and precomputed features, as well as external resources (e.g. gazetteers, dictionaries, Web corpora), depending on the run type. We distinguish between the following five run types:

Run 1: Only provided textual metadata may be used.

Run 2: Only provided visual & aural features may be used.

Run 3: Only provided textual metadata, visual features and the visual & aural features may be used.

Run 4–5: Everything is allowed, except for crawling the exact items contained in the test set, or any items by a test user taken within 24 hours before the first and after the last timestamp of a photo sequence in the mobility test set.

5. EVALUATION

For the sub-task, the evaluation metric is based on a hierarchical distance between the ground truth node and the predicted node or coordinate in the place hierarchy. The sub-task is evaluated according to the familiar geographic distance-based metric, where for each test item the distance is computed between the ground truth coordinate and the estimated coordinate. One important difference with past editions is that this year we measure geographic distances with Karney's formula [4]; this formula is based on the assumption that the shape of the Earth is an oblate spheroid, which produces more accurate distances than methods such as the great-circle distance that assume the shape of the Earth to be a sphere.

[5]http://www.click-that-hood.com/

6. BASELINES & LEADERBOARD

As task organizers, we provided two open source baselines to the participants, one for the locale[6] sub-task and one for the mobility[7] sub-task. Additionally, we implemented a live leaderboard that allowed participants to submit runs and view their relative standing towards others, as evaluated on a representative development set (i.e. part of, but not the complete, test set).

7. REFERENCES

[1] J. Choi, H. Lei, V. Ekambaram, P. Kelm, L. Gottlieb, T. Sikora, K. Ramchandran, and G. Friedland. Human vs machine: establishing a human baseline for multimodal location estimation. In , pages 867–876, 2013.

[2] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, et al. The Placing Task: a large-scale geo-estimation challenge for social-media videos and images. In , pages 27–31, 2014.

[3] C. Hauff and G. Houben. Placing images on the world map: a microblog-based enrichment approach. In , pages 691–700, 2012.

[4] C. Karney. Algorithms for geodesics. , 87(1):43–55, 2013.

[5] P. Kelm, S. Schmiedeke, J. Choi, G. Friedland, V. Ekambaram, K. Ramchandran, and T. Sikora. A novel fusion method for integrating multiple modalities and knowledge for multimodal location estimation. In , pages 7–12, 2013.

[6] M. Larson, P. Kelm, A. Rae, C. Hauff, B. Thomee, M. Trevisiol, J. Choi, O. van Laere, S. Schockaert, G. Jones, P. Serdyukov, V. Murdock, and G. Friedland. The benchmark as a research catalyst: charting the progress of geo-prediction for social multimedia. In . 2014.

[7] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval 2012, 2012.

[8] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In , pages 484–491, 2009.

[9] B. Thomee, D. Shamma, B. Friedland, G.and Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. , 2015. To appear.

[10] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In , pages 1–8, 2013.

[6]http://bit.ly/1gsrmvx

[7]http://bit.ly/1K8vUy8