The Placing Task at MediaEval 2016

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

ABSTRACT

The seventh edition of the Placing Task at MediaEval focuses on two challenges: (1) *estimation-based placing*, which addresses estimating the geographic location where a photo or video was taken, and (2) *verification-based placing*, which addresses verifying whether a photo or video was indeed taken at a pre-specified geographic location. Like the previous edition, we made the *organizer baselines* for both subtasks available as open source code, and published a *live leaderboard* that 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 determine where in the world photos and videos were captured based on analyzing their visual content and/or textual metadata, optionally augmented with knowledge from external resources like gazetteers. In particular, we aim to see those taking part to improve upon the contributions of participants from previous editions, as well as of the research community at large, e.g. [8, 11, 4, 2, 6, 9]. Although the Placing Task has indeed been shown to be a "research catalyst" [7] for geo-prediction 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. The introduction of the verification sub-task this year was driven by this consideration, as it requires participants to integrate a notion of confidence in their location predictions to decide whether or not a photo or video was taken in a particular country, state, city or neighborhood.

2. DATA

This year's edition of the Placing Task was once again based on the `YFCC100M` [10], which to date is the largest publicly and freely available social multimedia collection, and which can be obtained through the Yahoo Webscope program[1]. 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 [1], we sampled a subset of the `YFCC100M` for training and testing, see Table 1. 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. We included both test sets used in the Placing Tasks of 2014 and 2015 in this year's test set, allowing us to assess how the location estimation performance has improved over time.

Training		Testing	
#Photos	#Videos	#Photos	#Videos
4,991,679	24,955	1,497,464	29,934

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

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 [3], which together with the original photo and video content are publicly and freely available through the Multimedia Commons Initiative[4]. In addition, several expansion packs have been released by the creators of the `YFCC100M` dataset, such as detected visual concepts and Exif metadata, which could prove useful for the participants.

3. TASKS

Estimation-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 their actual coordinates and thus in essence 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.

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

MediaEval 2016 Workshop, Oct. 20–21, 2016, Hilversum, Netherlands

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 the *Places expansion pack* of the `YFCC100M` dataset, in which each geotagged photo and video is geotagged to its corresponding place, which follows a variation of the general hierarchy:

Country→State→City→Neighborhood

Due to the use of the hierarchy, only photos and videos that were successfully reverse geocoded were included in this sub-task, and thus media captured in or above international waters were excluded.

Verification-based sub-task: In this sub-task, participants were given a photo or video and a place from the hierarchy, and were asked to verify whether or not the media item was really captured in the given place. In the test set, we randomly switched the locations of 50% of the photos and videos, where we required that those switched were at least taken in a different country. Then, for 25% of the media items we removed the neighborhood level and below, for 25% the city level and below, and for 25% the state level and below, enabling us to assess how the level of the hierarchy affects the verification quality of the participants' systems.

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.

5. EVALUATION

For the *estimation-based* sub-task, the evaluation metric is based on the geographic distance between the ground truth coordinate and the predicted coordinate or place from the hierarchy. Whenever a participant estimates a place from the hierarchy, we substitute it by its geographic centroid. We measure geographic distances with Karney's formula [5]; 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. For the *verification-based* sub-task, we measure the classification accuracy.

6. BASELINES & LEADERBOARD

As task organizers, we provided two open source baselines[5] to the participants, one for the estimation sub-task and one for the verification 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 the, but not the complete, test set).

[5] `http://bit.ly/2dnggcg`

7. REFERENCES

[1] J. Choi, C. Hauff, O. Van Laere, and B. Thomee. The Placing Task at MediaEval 2015. In *Working Notes of the MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2015.

[2] 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 *Proceedings of the ACM International Conference on Multimedia*, pages 867–876, 2013.

[3] 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 *Proceedings of the ACM International Workshop on Geotagging and Its Applications in Multimedia*, pages 27–31, 2014.

[4] C. Hauff and G. Houben. Placing images on the world map: a microblog-based enrichment approach. In *Proceedings of the ACM Conference on Research and Development in Information Retrieval*, pages 691–700, 2012.

[5] C. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013.

[6] 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 *Proceedings of the ACM International Workshop on Geotagging and Its Applications in Multimedia*, pages 7–12, 2013.

[7] 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 *Multimodal Location Estimation of Videos and Images*. 2014.

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

[9] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *Proceedings of the ACM Conference on Research and Development in Information Retrieval*, pages 484–491, 2009.

[10] B. Thomee, D. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[11] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 1–8, 2013.