The Placing Task: A Large-Scale Geo-Estimation Challenge for Social-Media Videos and Images

Jaeyoung Choi[1], Bart Thomee[2], Gerald Friedland[1], Liangliang Cao[4], Karl Ni[3], Damian Borth[1], Benjamin Elizalde[1], Luke Gottlieb[1], Carmen Carrano[3], Roger Pearce[3], and Doug Poland[3]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]Yahoo Labs, San Francisco, CA, USA
[3]Lawrence Livermore National Laboratory, Livermore, CA, USA
[4]IBM Research, Yorktown Heights, NY, USA

[1]{jaeyoung,fractor,borth,benmael}@icsi.berkeley.edu
[2]bthomee@yahoo-inc.com
[3]{karl_ni,carrano2,rpearce,poland1}@llnl.gov
[4]liangliang.cao@us.ibm.com

ABSTRACT

The Placing Task is a yearly challenge offered by the MediaEval Multimedia Benchmarking Initiative that requires participants to develop algorithms that automatically predict the geo-location of social media videos and images. We introduce a recent development of a new standardized web-scale geo-tagged dataset for Placing Task 2014, which contains 5.5 million images and 35,000 videos. This standardized benchmark with a large persistent dataset allows the research community to easily evaluate new algorithms and to analyze their performance with respect to the state-of-the-art approaches. We discuss the characteristics of this year's Placing Task along with the description of the new dataset components and how they were collected.

Keywords

Location estimation; geotagging; benchmark

1. INTRODUCTION

In recent years, the rise in widespread use of GPS sensors, in combination with the increasing availability of open geographical databases, has motivated a large volume of work on geotagging. The increased use of geotagging and improvements in geo-location support systems open up a new dimension for the description, organization, and application of multimedia data. This new dimension radically expands the usefulness of multimedia data, not just for daily users of the Internet and social networking sites, but also for experts in particular application scenarios. As a result, location-based services have gained more and more attention, from big players such as Google and Yahoo, as well as from a number of smaller start-ups.

There are many motivations for working with geotagged multimedia. For instance, geotagged photos can be used in travel-related applications. Given a system based on a rich dataset of geotagged photos, a user could provide either a photo of the desired scenery or a keyword describing the type of place they want to visit, and the system would suggest a tourism destination or a travel route. In addition, information about people's geographical locations can be used for many other services, such as restaurant recommendations, transportation planning, and targeted advertisements. More and more companies and research labs have recognized the importance of geotagged information and have therefore spent considerable effort on collecting geotagged data.

Such geotag-based applications and services are drastically more useful if they can work not only with media for which GPS information is included in the original metadata but with media that has been automatically tagged with an approximate location based on its content. Placing Task[1] is a benchmark offered by the MediaEval Multimedia Benchmarking Initiative[2] that tries to tackle exactly this problem of automatic geo-tagging of media. The task requires participants to develop algorithms that automatically estimate the geo-coordinates (latitude and longitude) of videos and images.

To address the challenges and opportunities in working with geotagged multimedia, there has been a demand for large standardized geotagged datasets in the multimedia research community. Although quite a number of online sharing communities have APIs that allow the public to download certain location-based data, this has thus far largely resulted only in limited ad hoc datasets. The newly created Plaing Task 2014 dataset provides a standard sandbox that provides web-scale geotagged data, for comparing the performance of different algorithms for the Placing Task benchmark. The dataset can be further utilized for demonstrating

GeoMM'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2661118.2661125.

[1]http://www.multimediaeval.org/mediaeval2014/placing2014/
[2]http://www.multimediaeval.org/

powerful applications that can be created with geotagged videos and images.

However, there are several challenges in creating such a large-scale dataset, including (i) *collecting* the data, because many online sources cannot freely share data due to privacy and copyright concerns, (ii) *storing* the data, which would require quite a considerable amount of data storage, and (iii) *processing* the data, which requires high-performance parallel computing facilities to accomplish in a timely fashion. In this paper we discuss how the aforementioned challenges of collecting, storing, and processing were approached for the Placing Task, and present our outlooks for the future in the context of large geotagged multimedia datasets.

In Section 2, we introduce relevant work in research areas that rely on geotagged data. We further describe the Placing Task in Section 3 and the dataset in Section 4. In Section 5, we finally conclude with a discussion of the future of Placing.

2. RELATED WORK

A fundamental problem in geotag-based research is how to estimate the geographical location where an image or video was shot by analyzing its content, especially when its geotags are not available or unreliable. The success of large-scale classification and retrieval suggests a data-driven approach. Existing work in this domain has mostly been carried out on Flickr images and focused on their associated tags. For instance, geo-locations associated with Flickr tags were predicted using the spatial distributions of where they were used, where a tag that was found to be strongly concentrated in a specific place was considered to have a semantic relationship with that location [15]. Also, user-contributed tags were exploited for geotagging by associating tag distributions with locations that were represented as grid cells on a map of the earth, which were then used to infer the geographic locations of where Flickr images were taken [16].

Hays and Efros [8] were among the first to consider the problem of estimating the location of an image using only its visual content. They collected millions of geotagged Flickr images. Using a comprehensive set of visual features, they employed nearest-neighbor search to locate an image with respect to the reference set. This approach was able to locate about a quarter of the test images to within approximately 750 km of their true location—about the width of a small country.

Landmarks depicted in videos and images have received a large amount of attention in the research community as well. A web-scale landmark-recognition engine called "Tour the World" was built by using 20 million GPS-tagged photos of landmarks cross-referenced with online tour guide web pages [18]. Experiments carried out using the system demonstrated that such an engine can deliver satisfactory recognition performance with high efficiency. However, it is still an open question whether it is possible to recognize non-landmark locations reliably.

Multimodal location estimation on videos that utilize audio was first attempted by Friedland et al. [6], where the authors assigned videos to different cities by matching the embedded audio against the typical sounds of ambulance sirens; textual tags were not used. Audio tracks from the Placing Task 2011 dataset videos were also used to train a city-level location estimation system with a reasonable performance [12].

A human baseline for location estimation for three different combinations of modalities (audio only, audio + video, audio + video + textual metadata) was collected and compared with the machine algorithm's performance [4]. The study demonstrated cases when humans could effectively identify audio cue for estimating video's location when the machine algorithm failed. Humans were also effective at inferring the location by combining visual and audio cues when visual cue alone does not carry enough information for the location estimation. This work suggests the potential benefit of utilizing audio, which is often overlooked.

Evaluations on multimodal location estimation on randomly selected consumer-produced videos were carried out in the 2010, 2011, 2012, and 2013 MediaEval Placing tasks. One of the participants notably used a combination of language models and similarity search to geo-tag the videos using their associated tags [17]. Many participants tried to utilize both visual and textual features for their location estimations. One such approach augmented the data from the visual and textual modalities with external geographical knowledge bases, by building a hierarchical model that combined data-driven and semantic methods to group visual and textual features together within geographical regions [9]. As a result, the proposed method successfully located 40% of the videos in the MediaEval 2010 Placing Task test set within a radius of 100m.

A novel model, logistic canonical correlation regression, explored the canonical correlations between geographical locations, visual content, and community tags [2]. In contrast with existing work (e.g. [8]), however, the authors argued that it is difficult to estimate the exact location at which a photo was taken, and therefore rather focused on accurately estimating the most probable spatial region where it was captured. Their experiments demonstrated that inferring coarse locations can lead to accurate annotations. A similar method also focused on accurately estimating the approximate location of a novel photo [5].

All of the approaches described above have the common feature of processing each query photo or video independently using a geo-tagged training database. Clearly, the performance of these systems therefore largely depends on the size and quality of the training database. However, data sparsity is one of the major issues that can adversely affect the performance of these systems. A novel approach therefore *jointly* estimated the geo-locations of all of the input query images [3], where each query image added to the database enhanced the quality of the database by acting as "virtual" training data and consequently boosted the performance of the algorithm.

3. THE MEDIAEVAL PLACING TASK

3.1 Overview

The Placing Task is a yearly benchmark challenge offered by MediaEval. The task, launched in 2010, requires participants to develop algorithms that automatically estimate the geo-coordinates (latitude and longitude) of videos and images in datasets drawn from Flickr[3]. Participating in this benchmark enables researchers to collaboratively address the challenge, exploring the effectiveness of their algo-

[3]`https://www.flickr.com/`

rithms and evaluating their results as compared to the state of the art on a single representative dataset.

The parameters of the Placing Task differ from those of previous work on automatic geo-estimation in many aspects, due in part to characteristics of the dataset and in part to the evaluation criteria:

- Multiple modalities: Location estimation algorithms should effectively exploit the complementary information provided by each of the modalities in the media. As discussed in [4], the contribution of each modality should enhance the ability of an algorithm to distinguish location.

- Scalablity: The vidoes and images are from the entire world, so algorithms must be able to build location models that can successfully place media recorded in any location on the globe. The ability of an algorithm to accurately predict a location thus rests on its capacity to process and exploit the large amounts of available social-multimedia data from all over the world, which amounts to, at the least, tens of millions of documents.

- Noisy, unfiltered dataset: Placing algorithms must be robust, to handle the noise and uncertainty associated with social multimedia. Videos and images collected for the benchmark datasets are not filtered by content. They are sampled randomly, limited only with respect to videos/images per user, to prevent the photo/video streams of a few users from flooding the dataset and potentially introducing user bias.

- Location bias and sparsity: The distribution of geotagged training data is uneven. Some highly populated areas are represented by large numbers of videos and images, taken by both locals and tourists, while other locations are represented by little to no data, creating a bias and sparsity problem.

These unique parameters have inspired approaches that might not have been tried with a more controlled dataset. For example, participants have attempted to tackle this challenge by exploiting semantic data, such as geographical gazetteers (e.g., GeoNames[4]), or by using graphical framework approaches.

The composition of the dataset and the resources offered to participants have changed as the Placing Task evolved, driven in part by a survey conducted yearly among the participant community to determine which additional resources are most in demand. In response to community requests, pre-computed visual and audio features commonly used in multimedia analysis were provided for each of the items. In addition to visual features released by organizers in the previous years (2010-2013), we added two more visual features that are widely used in practice and demanded by the participants: Gist descriptor [14] and SIFT feature [13]. Audio features were added to encourage participants to utilize audio tracks in the video as they have proved to be useful in location estimation [6, 12]. More on the history and evolution of the Placing Task can be found in [10].

[4]`http://www.geonames.org/`

3.2 Evaluation Metric

In the previous years' Placing Task evaluation, the error distance between the groundtruth and estimated location was calculated, and the number of those placed within $1km$, $10km$, $100km$, $1000km$, and $5000km$ from the groundtruth were counted. The performance of the systems of the Placing Task participants has improved drastically in the last four years [10], and thus the analysis of the algorithms should address the performance of estimating the location in finer granularity than $1km$. This year, $10m$ and $100m$ buckets were added to the evaluation metric. According to real-world GPS accuracy analysis in [1], GPS receivers typically have a horizontal accuracy of better than within 3 meters. Also, modern smartphones, which comprise a large proportion of the devices used to produce the media on Flickr, use a number of methods (e.g., triangulation between cell towers and known WiFi access points) to enhance the accuracy of their geo-location. For the hand-labeled media, since the Placing Task started in 2010, all videos and images chosen for the Placing Task dataset have a groundtruth location resolution of "street level", i.e., the user used the highest zoom level when placing the photo or video on the map. There are 16 zoom levels, corresponding to 16 accuracy levels (e.g., "region level", "city level", "street level"). Thus, for the videos and images that have device or hand-labeled ground-truth locations, the finest granularity used for the evaluation, $10m$, is considered reasonable.

To evaluate the performance of the algorithms, the geodesic distance between the ground truth geo-coordinates (latitude/longitude pair) and those of the outputs from the algorithms were compared. The Haversine distance was used to take the geographic nature of the evaluation into account. This measure is calculated as:

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

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

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 6,378.137 km was used).

4. THE 2014 PLACING DATASET

In this section, we discuss the composition of the dataset used for the MediaEval Placing Task in 2014—which we refer to as MP2014—and its components, such as metadata and features.

4.1 Composition

The MP2014 dataset is drawn from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset,[5] which contains the metadata for 99.2 million photos and 0.8 million videos that have been uploaded to Flickr and assigned a Creative Commons license by the uploader.[6] About half the YFCC100M dataset is geotagged; a subset of these geotagged videos were used to create the MP2014 dataset. From the geotagged media items, we semi-randomly selected 5 million images and 25,000 videos for the training set, and

[5]`http://bit.ly/yfcc100md`

[6]Flickr requires that an uploaded video or image must be created by its uploader.

Metadata	Photo/video identifier, Date taken, Date uploaded, Capture device, Title, Description, User tags, Machine tags, Longitude, Latitude, Accuracy
Visual Features	Auto Color Correlogram, BasicFeatures, CEDD (Color and Edge Directivity Descriptor), Color Layout, Edge Histogram, FCTH (Fuzzy Color and Texture Histogram), Fuzzy Opponent Histogram, Gabor, Joint Histogram, Joint Opponent Histogram, Scalable Color, Simple Color Histogram, Tamura, Gist, SIFT
Audio Features	MFCC20 (20 lowest Mel-frequency Cepstral Coefficents), Kaldi pitch features, SAcC pitch (Subband Autocorrelation Classification pitch tracker)

Table 1: List of metadata and computed features (visual/audio) included in the 2014 Placing Task dataset.

500,000 images and 10,000 videos for the test set. In total, more than 223,000 unique users contributed to the dataset. To ensure that our dataset would be sufficiently challenging, we constrained the selection such that each user only contributed at most 250 images and 50 videos, and that the recordings for a given user were all made more than 10 minutes apart from each other. None of the users who contributed videos or images to the training set also contributed to the test set, and vice versa. In addition, 80,000 videos and 1 million images were reserved to be used as test sets for the Placing Task in the years 2015 and 2016.

In the research community including multimedia analysis researchers, there has been an increasing divide between researchers who have access to high-performance computing facilities and those that do not. To enable everyone to participate, we therefore created several versions of the test set, with the larger sets being supersets of the smaller ones. This scheme allows each participant in the Placing Task to solve a test set of a size their facilities can comfortably handle. Similarly, participants can choose to use only a subset of the large training set for training and evaluating their models (though, unlike the test set, it is not standardized and thus participants may form their own subsets).

4.2 Metadata

The YFCC100M provides pertinent metadata for all of the indexed videos and images, as listed in Table 1, which includes user-supplied textual metadata as well as geotags. This metadata was released under Yahoo's Webscope license. For images, geo-coordinates are often associated automatically by the capture device, although users may also manually place them using a map interface. Geotagging of Flickr videos differs from that of images in that they must usually be geotagged manually by the user, and thus far fewer videos than images have been geotagged.

4.3 Features

The MP2014 included a set of extracted audio and visual features that are commonly used in multimedia analysis, to save each group of participants from having to compute them individually—thus allowing more attention to be spent on finding innovative ways to use those features. Table 1 lists the metadata and the visual and audio features that were included in the dataset. All of the data, including the metadata and the pre-computed features as well as the original videos and images, were made available online in the cloud[7], to ensure high data availability and fast downloading for the participants. The Gist features [14] were extracted using Lear's GIST implementation[8]. The SIFT descriptors [13] were computed using OpenCV[9]. The remainder of the visual features listed in Table 1 were extracted using the content-based image-retrieval tools in the open-source LIRE library[10]. Among the possible audio features for which extraction tools are available, the ones the audio-analysis community agrees that the most useful are MFCCs (Mel-frequency cepstral coefficients) and pitch features. We chose Kaldi pitch [7] because it is readily available and does well with clean audio and SAcC [11] because it does well with noisy audio.

We used a Cray Catalyst supercomputer to overcome the computational challenges in extracting most of the features. All of the features were released under the Creative Commons 0 license[11], i.e., they are in the public domain.

4.4 Original Videos/Images

We understand that researchers may need original videos and images for many purposes, such as extracting their own features, or supplying pixel data to neural networks. We therefore also provided original videos and images for the MP2014 dataset.

5. THE FUTURE OF PLACING

While work on Placing (geo-estimation) was introduced relatively recently, work in this new field is nevertheless stimulating progress in many related areas of multimedia research. Cues used to estimate location can be extracted using methods derived from current research areas. Placing work tends to deal with much larger test and training sets than traditional multimedia content analysis tasks, since it uses user-generated data available on the Internet; in addition, the data is more diverse, as the recording sources and locations differ greatly. This offers the potential to create machine-learning algorithms of higher generality. In fact, Placing is the multimedia task with the largest amount of ground-truth data available, and can therefore be regarded as the largest big-data task in current multimedia computing. Overall, we believe that the Placing task has the potential to advance many fields—some of which we don't yet even know of, as new fields will be created based on user demand for new applications. We are therefore very much looking forward to seeing what interesting developments may occur in the next few years.

Acknowledgements

This work was partially supported by funding provided to ICSI through National Science Foundation grant IIS:1251276

[7] `http://dataset.icsi.berkeley.edu`

[8] `http://lear.inrialpes.fr/software/`

[9] `http://opencv.org/`

[10] `http://www.semanticmetadata.net/lire/`; we used version 0.9.3 with default parameter settings.

[11] `https://creativecommons.org/publicdomain/zero/1.0`

("BIGDATA: Small: DCM: DA: Collaborative Research: SMASH—Scalable Multimedia content AnalysiS in a High-level language"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the National Science Foundation. Portions of this work were performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

We thank Yahoo for releasing the YFCC100M dataset and Lawrence Livermore National Laboratory for providing the massive computing resources that enabled the release of the large feature set. We thank Julia Bernd at ICSI for her valuable feedback.

6. REFERENCES

[1] Global positioning system standard positioning service performance standard. Technical report, Department of Defense, September 2008.

[2] L. Cao, J. Yu, J. Luo, and T. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *Proceedings of the ACM International Conference on Multimedia*, pages 125–134, 2009.

[3] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: dealing with sparse training data. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 43–48, 2012.

[4] 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.

[5] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the IW3C2 International Conference on World Wide Web*, pages 761–770, 2009.

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

[7] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2494–2498, 2014.

[8] J. Hays and A. Efros. IM2PGS: estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] 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.

[10] 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 bnchmark as a research catalyst: charting the progress of geo-prediction for social multimedia. In J. Choi and G. Friedland, editors, *Multimodal Location Estimation of Videos and Images*. Springer, 2014.

[11] B. Lee and D. Ellis. Noise robust pitch tracking by subband autocorrelation classification. In *Interspeech*, 2012.

[12] H. Lei, J. Choi, and G. Friedland. Multimodal city-verification on Flickr videos using acoustic and textual features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2273–2276, 2012.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[15] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1), 2009.

[16] 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.

[17] O. van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.

[18] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. Chua, and H. Nevens. Tour the World: building a web-scale landmark recogntion engine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.