Real-time Analysis and Visualization of the YFCC100m Dataset

Sebastian Kalkowski [1] Christian Schulze [2] Andreas Dengel [1, 2] Damian Borth [2]

[1]University of Kaiserslautern, Germany
{s_kalkowsk11}@cs.uni-kl.de

[2]German Research Center for Artificial Intelligence (DFKI), Germany
{christian.schulze, andreas.dengel, damian.borth}@dfki.de

ABSTRACT

With the *Yahoo Flickr Creative Commons 100 Million* (YFCC100m) dataset, a novel dataset was introduced to the computer vision and multimedia research community. To maximize the benefit for the research community and utilize its potential, this dataset has to be made accessible by tools allowing to search for target concepts within the dataset and mechanism to browse images and videos of the dataset. Following best practice from data collections, such as ImageNet and MS COCO, this paper presents means of accessibility for the YFCC100m dataset. This includes a global analysis of the dataset and an online browser to explore and investigate subsets of the dataset in real-time. Providing statistics of the queried images and videos will enable researchers to refine their query successively, such that the users desired subset of interest can be narrowed down quickly. The final set of image and video can be downloaded as URLs from the browser for further processing.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

yfcc100m; dataset; search; browser; visualization

1. INTRODUCTION

Over the last years different visual recognition tasks in computer vision and multimedia research have been introduced: object classification and detection [6, 5], semantic segmentation [10], concept detection [12], multimedia event detection [3], affective or emotional categorization [15], or visual sentiment analysis [2]. Although these tasks all target different sub-disciplines in the research community, they have one thing in common, they employ supervised machine learning and therefore require datasets to train classifiers or detectors [9]. The availability of datasets such as ImageNet [5] or MS COCO [10] is helping the community to explore new approaches and to make progress in improving visual recognition performance.

MMCommons'15, October 30, 2015, Brisbane, Australia.

ACM 978-1-4503-3744-1/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2814815.2814820 .

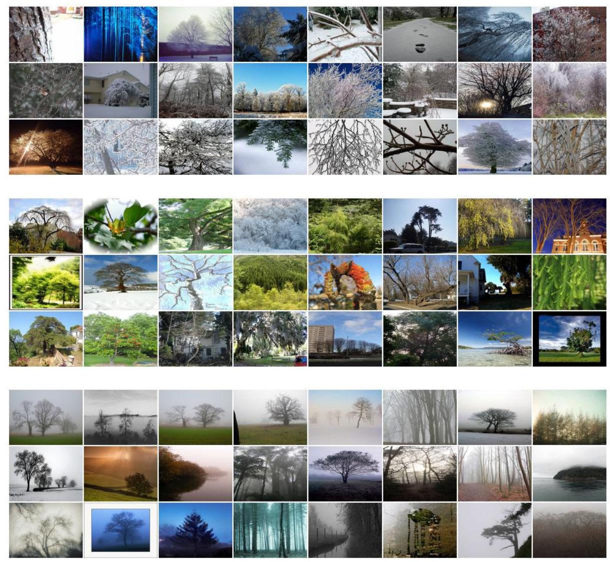

Figure 1: Samples of images tagged by "tree" illustrating different upload times and geo-location of the tree images.

One recently released dataset lining up in this context is the *Yahoo Flickr Creative Commons 100 Million* (YFCC100m) dataset [14]. The dataset consists of 99.2 million Flickr photos and 0.8 million Flickr videos, all of which carry some type of a Creative Commons license. It provides a text file containing all images and video with their associated meta data, as available on Flickr. In particular, this includes also the URL for direct download of the images and videos itself. Its vast size allows an in-depth analysis of how user generated media content is shared and annotated and its richness serves as a large-scale source for training of statistical methods [11] or the creation of specialized subsets [1, 4]. However, to make this dataset accessible for a broader audience, it requires further tools and mechanisms than the already available ones.

For example consider the following case: a computer vision researcher wants to train a classifier for"trees" (Fig. 1). Does the YFCC100m dataset contain a proper set of tree images, which can be used for such a training? Well, the dataset contains 346,594 media items tagged by "trees". How many items are images? How many items are videos? Probably the researcher wants also to identify a subset of these trees with respect to a broad variability to prevent overfitting i.e. they should not all come from the same user (a dataset of trees pictures from a single user might be biased). Further, the researcher might be interested in trees from a particular region, such as Australia or in a particular season of the year, easily inferred from the geo- and time data provided with the dataset. It is obvious that a quick evaluation of the dataset for a given query as outlined in the example is very helpful for a researcher to make the decision to download the dataset and all of its 100 Million images and videos.

This paper presents access to the YFCC100m dataset in form of an online website[1] providing the following features:

- **Global Statistics:** about the dataset, illustrating general characteristics of the dataset to evaluate its fit for pre-defined tasks in the computer vision and multimedia research community.
- **Dataset Browser:** allowing to grasp the amount of images and videos in the dataset for a particular user defined query and browse these images on the fly.
- **Query Statistics:** providing statistics for a user generated query such as co-occuring tags, user distribution, time distribution, and geo-location.
- **Query Refinement:** to modify of the query and narrow down potential large quantities of retrieved images/videos with respect to particular properties such as user distribution, time distribution, and geo-location.
- **URL Downloads:** providing a list of URLs matching the query. This way only a subset of all images and videos must be retrieved to streamline research efforts.

In the future, the browser aims to serve as a experimental platform for further retrieval mechanism and additional annotations to augment the dataset.

2. RELATED WORK

This section describes the related work with respect to making large scale datasets accessible to the research community. It provides an overview of currently popular image datasets and their online community tools.

MIR Flickr.

One of the first Flickr datasets, is the MIR Flickr collection from 2008 / 2010 [7, 8]. It presents a curated dataset of Flickr CC images and comes with an online website accompanied with global statistics about the dataset. This includes an overview of top tags, EXIF information, and annotations with respect to relevance and abstraction level. Although an important step towards providing real world data for research purpose, the website does not offer any browsing or tag search interface and only displays 6 sample images of the dataset.

ImageNet.

Besides providing a large-scale image dataset with expert annotations, the *ImageNet* dataset [5] also provides a comprehensive online website to browse and visualize its taxonomy of concept labels. The website consists of a textual search and a tree browser allowing researchers to quickly navigate to the desired concept (i.e. synset) and see basic statistics, example images of the synset, the underlying subsynsets summarized by a image mosaic. Further, it allows the user to download the list of URL from the synset.

[1]http://www.yfcc100m.org

Table 1: Availability of titles, descriptions, and tags.

Titles		
empty	3,835,258	3%
generic [1]	25,971,801	26%
non-generic	70,192,941 (avg. 3.08 words / title)	70%
Descriptions		
w/o description	68,277,216	68.3%
with description	31,722,784 (avg. 22.5 words / desc.)	31.7%
Tags		
none	31,028,877	31%
at least 1 tag	68,971,123 (avg. 7.06 tags / item)	69%

[1] e.g. "IMG_012345" or "DSC_12061999"

In contrast to the ImageNet website and because of the nature of the YFCC100m datasets with its user generated metadata, the proposed browser is able to provide additional information associated with the images and videos such as in-depth statistics about co-occurring tags, user and geo distribution in real-time to allow for quick evaluation of the dataset for a particular concept.

MS COCO.

The recently released *Microsoft Common Object in Context* (MS COCO) dataset [10] follows a similar approach. It provides access to its dataset via an online website to browse its object vocabulary, annotations (including category labels, bounding boxes, object segmentation, instance counts). Access is established with generic icons depicting common objects such as "car" and via a search box. Since the list of objects is limited to 80 categories the MS COCO browser is focusing individual images with its object segmentation boundaries.

Because of the large-scale size of the YFCC100m dataset, the proposed browser focuses on real-time accessibility of 100 million images with respect to an undefined concept vocabulary. It provides overview statistics for a customized subset of the dataset as compared to segmentation boundaries of pre-defined objects for each image.

MIT Places.

Another specialized dataset focusing entirely on scenes and places is the MIT Places dataset [16] covering 205 scene categories. The dataset comes with a online website listing all categories with samples images. Browsing capabilities are limited to listing a mosaic of images per category and therefore are more of a static nature as compared to the proposed browser with query refinement capabilities.

Summarizing, the presented YFCC100m browser provides online access mechanism for researchers to quickly identify a subset of the dataset being relevant for their work by formulating unconstraint queries and query dependent statistics and reporting.

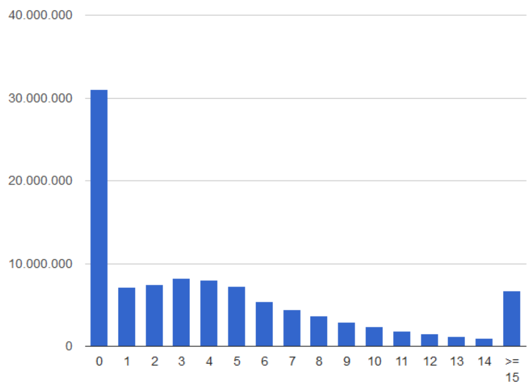

Figure 2: Distribution of tags given to images or videos in the dataset. Users either do not tag their images at all or, add between 1 to 10 tags for an item with exceptions of excessive taggging beyond 15 and more tags per items.

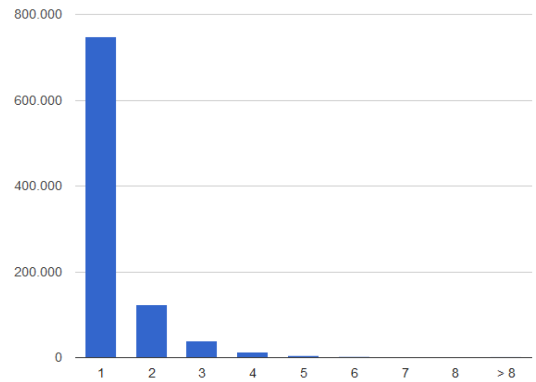

Figure 3: Flickr allowed over time to have tags consisting of multiple words like "New York". Although tags up to 5 terms can be found in the dataset, most of the tags consist of one term.

Table 2: Top occurring tags in the dataset and their counts

No.	**Tag**	**count**	No.	**Tag**	**count**
1	square	1,429,645	11	2011	1,063,045
2	iphoneography	1,369,398	12	2012	1,052,044
3	square format	1,321,876	13	2009	1,031,310
4	instagram app	1,313,837	14	london	996,166
5	california	1,226,796	15	2008	951,965
6	nikon	1,195,576	16	japan	932,294
7	travel	1,195,467	17	france	917,578
8	usa	1,188,344	18	nature	872,029
9	2010	1,109,926	19	art	854,669
10	canon	1,101,769	20	music	816,277

Table 3: Top occurring tag categories

Tag Categories	**count**
Dates	41,121,781
Locations	28,160,636
Camera descriptions	5,969,169
Flickr-specific[1]	3,711,140
App-Generated	3,438,022
URLs	2,522,917

[1] Tags like: 'square', 'black and white', 'project365', 'high-res', ...

3. ANALYSIS

The YFCC100m dataset is exclusively constructed from Flickr providing a rich repository of user generated images and videos with its associated metadata, including various information such as e.g. titles, descriptions, tags and others (please see [14] for more details). This section provides a global analysis of the dataset. Similarly analysis can be performed with the online browser.

3.1 Titles, Descriptions, Tags

The Flickr upload mechanism, among other things, allows users to upload images with or without title, add an optional free-text description to an image, and annotate the image with an arbitrary number of tags (or no tags at all). Due to this mechanics, such metadata should be considered incomplete and inhomogeneously distributed. This has multiple implications with respect to the usability of the dataset.

A global analysis of the dataset images and videos metadata as illustrated in Table 1, yields the following observation: While roughly 96% of all images have been given a title, a high proportion of those titles is machine-generated by the capture device or upload programs. A simple regular expression matching against all non-empty titles in the dataset found around 26% of all titles consisting of one to five capital letters, followed by an optional underscore and a number (e.g. "DSC_12061999"). This leaves only around 70% of items or less with a descriptive title.

With respect to descriptions, a different observation was made. Here, only 31.1% of all items in the dataset have a non-empty description. This might be caused by the more time consuming nature of creating a description for each image or video during the upload.

Luckily the proportion of images and videos having at least one tag is much larger. In this context 69% of all items in the dataset have a tag being linked to the video or image.

3.2 Tags

On average there are 7.06 individual tags for each item in the dataset. The great majority of images and videos has been tagged with only a few tags, usually between one and ten. (see Fig. 2). The entire dataset has a total of 486,435,393 tags, reducible to 7,940,039 distinct tags defining the broad vocabulary of the Flickr community.

Categorizing popular tags, they can be grouped into one or more of the following categories: "app-generated", "camera descriptions", "Flickr-specific", "App-generated", "activities", "locations", "dates", and "URLs". We found, that in most cases these categories of tags are highly indescriptive concerning the image object. Especially tags falling into the camera- and app-type categories have a pre-eminent stop word character, since they are used by a lot of different users for many items regardless of their content. Location- and date-type tags on the other hand could, despite of their general unreliability, serve as a fallback or verification basis for missing or unplausible geolocations and timestamps. Tables 2 and 3 show the most popular tags and tag categories respectively. Although it is possible to insert a larger text as a single tag, the number of N-grams used as tags in the dataset goes seldom beyond four (see Fig. 3).

3.3 User Distribution

Throughout the YFCC100m dataset the activity of different users varies vastly. Altogether a total of 578,262 different Flickr users contributed to the complete dataset by upload-

Table 4: Geo-information as provided in the dataset

provide data	within country borders	48,296,858	48.3%
	other	172,971	0.2%
	sum	48,469,829	48.5%
without data		51,530,171	51.5%

ing their images or videos under any form of the Creative Commons License. That makes an average of 173 items per user. The vast majority – however – uploaded only a low number of images while some users turned out to be very active. The dataset reveals that the top 1.7% of all users (a number of only 9894 different users) are responsible for 50% of all uploaded items). This observation indicates a strong bias towards these users, making training data curation for classifier training sensitive to these users, if no mechanisms are established to balance out user contribution. Looking at the long tail, 36% which is more than a third of all users, uploaded only five or less items each, while around 17% of all users even only contributed a single image or video. This phenomenon is especially leveraged on Flickr, since user accounts exist which are associated with applications, allowing groups of people to upload their images via one single account. One such example is the "friendly.flickr" account.

3.4 Geographic Information

As shown in Table 4 a number of 48,469,829 images and videos provide a geo-position in form of a latitude and longitude value. Using this information, we were able to map 48% of all items in the dataset to a single country. Around 0.4% of those items have a position pointing to oceans or polar regions, and could therefore not be mapped to any specific country. The absolute distribution of all mappable items over countries is shown in Fig. 4. As seen on the linear scale, the most prominent location is the USA including Alaska (32.89%). A map illustrating the same distribution in logarithmic scales (Fig. 5) yields a better visualization of the overall global activity. The distribution shows that also Brazil, Canada, India, China, Australia and Central Europe are relatively active in comparison to other regions. The least active region is Central Africa. According to the plot, there are five major countries (D.R. Kongo, Southern Sudan, Romania, Serbia, Montenegro and Kosovo) and some island states without a single image or video in the dataset.

3.5 Creation- and Upload-Times

Every item in the dataset provides a timestamp indicating the upload to the Flickr servers. For almost every item the creation time is also given. While the upload times are usually reliable, the capture timestamps unfortunately suffer from inaccurate date and time settings of the capture devices and software tools, used for upload [13]. Leaving the original capture times unfiltered, the corresponding years range from 1 to 9999. The proportion of items with a capture year before 2000 or beyond 2015 however is below 1% of all images and videos and therefore might be neglectable. Nevertheless this unreliability should be kept in mind when working with time information in the dataset.

3.6 Tags over time

Since time information is provided, an interesting further analysis is the distribution of tags over time. In particular: how does the distribution of popular tags change over time? Fig. 6 shows the popularity over all months of the year for selected exemplary tags (due to space constraints limited to 4 tags, more are available online). Obviously, the tags "winter" and "snow" have a strong correlation over the year. This kind of correlation over different timespans could e.g. be used as a basis for automated clustering of related terms. On the other hand it also mirrors the datasets enormous bias towards the northern hemisphere, where winter and snow are typically associated with the months of November to February. The tag "fireworks" is maybe surprisingly not very prominent in January around new years eve in comparison to July, where there is the 4th of July, which is often celebrated with fireworks in the USA only. As a control tag, the mostly app-generated tag "instagram" - as expected - shows a relatively homogenous distribution over the year, mirroring, that it is not related to seasonal changes.

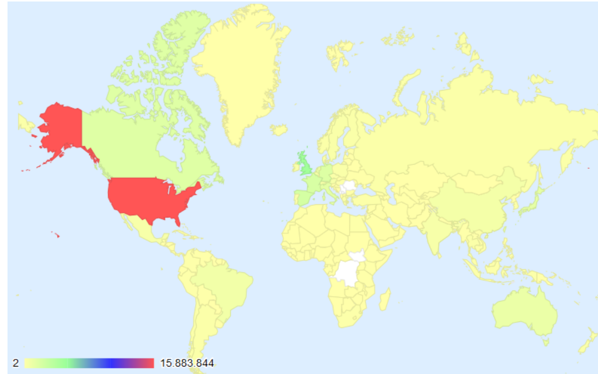

Figure 4: Geographic distribution on a linear scale

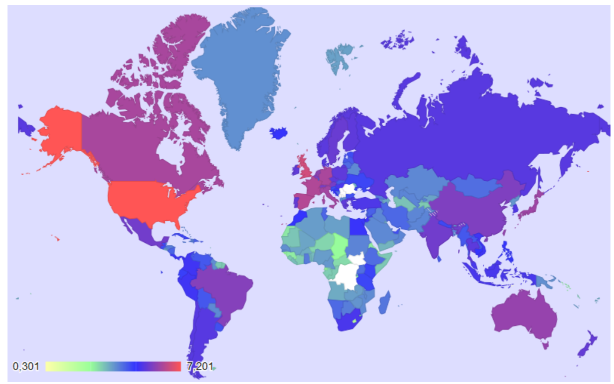

Figure 5: Geographic distribution on a logarithmic scale

4. BROWSER & VISUALIZATIONS

The previous section was describing global statistics about the YFCC100m dataset. However, often researchers are interested in a particular subset of a dataset, like for example the events in [1] or images and videos with geo-location information [4].

To enable easy and quick access for this type of queries which define a specific subset of the YFCC100m dataset, we present the YFCC100m Browser, which can be found under http://www.yfcc100m.org[2]. The browser is designed to filter and explore the entire dataset of 100 million images and videos in realtime. Subsets of the dataset can be retrieved by a straightforward keyword search.

[2]forwarding to http://yfcc100m.appspot.com

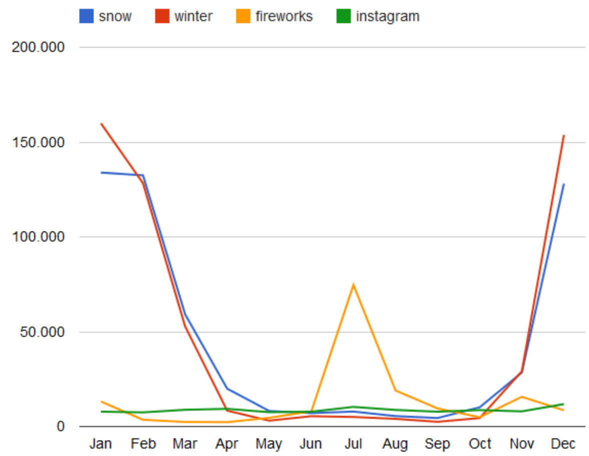

Figure 6: Popularity of selected tags with their items capture times mapped on months over the year..

4.1 Search & Browsing

Given a user query it retrieves the subset of images and videos matching the query and provides previews of images in form of thumbnails (see Fig. 7, **right**, for sample images of the query "trees"). Each item is linked to its associated Flickr page, where further information such as comments can be found. In addition, a set of statistics for the retrieved subset is generated dynamically. (seel Fig. 7, **left**, for statistics of the query "trees"). These statistics include a tag cloud visualizing the top 100 most common tags associated with the retrieved subset of the dataset, ranked by their occurrence count. Clicking at a tag, or the small plus besides, allows to launch a refined search either only for the selected tag or restricting the results to both tags, adding an explorative component to the browser. Further, a ranking of the most active users for the retrieved query is displayed. The user distribution is aiming to indicate a possible user bias. Finally, the distribution of the capture times filtered over the most relevant years (from 2002 up to 2015 inclusive) is visualized, as well as the global distribution of geo-locations is depicted on a world map. Images and videos for which either no geolocation in form of latitude and longitude was given or the given coordinates were not mappable to any country land mass (e.g. oceans, or polar regions) are excluded from the world map.

With this very vital information it is possible to get a first overview of the subsets as defiend by a user query and identify biases or get a quick impression of the quality of the associated images and videos.

4.2 Technical Specification

To allow high accessibility to the YFCC100m dataset and scalability with respect to multiple users simultaneously querying the browser, the online browser is build upon Google Compute Engine[3]. The frontend is using the Google AppEngine environment[4], a framework allowing to setup scalable web applications on Google's infrastructure i.e. after deployment, the application is spreaded across multiple servers and running instances are automatically spawned on demand to scale up with the application load. The backend, realizing search and query mechanism of the browser, is running Google BigQuery[5]. This includes the retrieval, aggregation and temporary storage of the search results. The main advantage of BigQuery is its database-like query languages and database schemas allowing to process large quantities of data including repeated and nested fields in a distributed way. Statistics of search and retrieval results are dynamically gathered and computed on the server side, while visualizations in form of charts are rendered clientside with Javascript.

Although BigQuery datasets can be easily accessed in an SQL-like language, some of its distributed computing characteristics lets it perform differently than regular database technologies. Queries on single static datasets, including grouping, sorting and selection are most often highly performant, while especially joins take comparatively more time. This is the reason for the statistics to sometimes take slightly longer to aggregate than the simple result preview, although altogether less data has to be processed for that. Still performance is high enough to view results in matters of seconds, enabling a fluid browsing experience.

5. CONCLUSION

The YFCC100m dataset introduces a great dataset for various computer vision tasks. However, the dataset's potential can only be fully utilized, if it is made easily accessible to the research community.

Considering the results from the global analysis, researchers using the dataset should also be aware of potential bias with respect to different dimensions of the dataset. Especially the highly inhomogenous distribution of ownership among users in the dataset and the uneven global geo distribution of images and videos must be considered. This – however – can only be taken into account when working with the dataset. Unfortunately, the huge number of 100 million items in the dataset makes processing this dataset challenging.

With the presented YFCC100m browser we provide a tool to the community, which provides quick access to the entire set of 100 million items and additionally provides overview statistics for user generated queries in real-time. The on-the-fly generated statistics offer first insights into the distribution of metadata annotations, visualize important biases and allow an evaluation of the quality of the dataset content. With its real-time performance, iterative refinement mechanism and adaption to subsets of the complete dataset, researchers from the computer vision and multimedia community can grasp the content (visual and metadata) of the dataset swiftly without having to download the whole dataset and therefore supporting its usage and increasing its visibility.

6. ACKNOWLEDGMENTS

This work was partially funded by the BMBF project Multimedia Opinion Mining (MOM: 01WI15002).

7. REFERENCES

[1] J. Bernd, D. Borth, B. Elizalde, G. Friedland, H. Gallagher, L. Gottlieb, A. Janin, S. Karabashlieva,

[3] https://cloud.google.com/compute/
[4] https://cloud.google.com/appengine/
[5] https://cloud.google.com/bigquery/

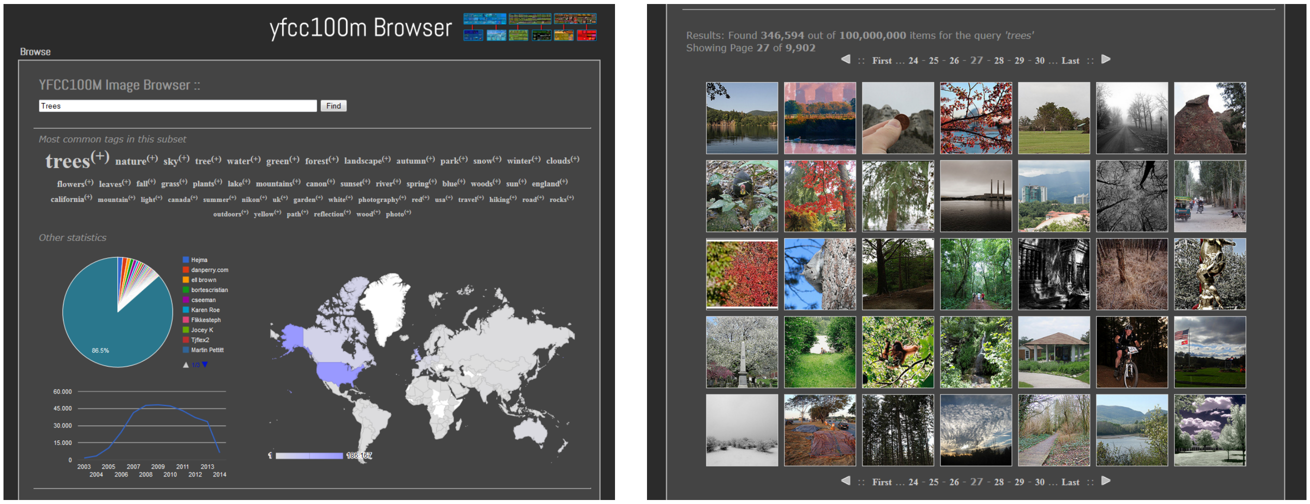

Figure 7: Screenshots of the YFCC100m Browser. For the query "trees", the browser was able to retrieve 346,594 items in the dataset. Results are illustrated as **Left:** Subset statistics including tag co-occurrence, user distribution, capturing time, and geo-location of the items. **Right:** Thumbnails of the images and videos found for this subset. Thumbnails can be paged through all 9,902 pages depicting all retrieved items.

J. Takahashi, and J. Won. The yli-med corpus: Characteristics, procedures, and plans. *arXiv preprint arXiv:1503.04250*, 2015.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 223–232, October 2013.

[3] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, Y. Mu amd A. Natsev, and J. Smith. IBM Research and Columbia University TRECVID-2011 Multimedia Event Detection (MED) System. In *Proc. NIST TRECVID Workshop (unreviewed workshop paper)*, December 2011.

[4] 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 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*, pages 27–31. ACM, 2014.

[5] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, July 2009.

[6] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Int. Journal of Computer Vision*, 88(2):303–338, June 2010.

[7] M. Huiskes and M. Lew. The mir flickr retrieval evaluation. In *Proc. ACM Int. Conf. Multimedia Information Retrieval (ACM MIR)*, October 2008.

[8] M. Huiskes, B. Thomee, and M. Lew. New Trends and Ideas in Visual Concept Detection: the MIR Flickr Retrieval Evaluation Initiative. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)*, pages 527–536, October 2010.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 1106–1114, December 2012.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[11] K. Ni, R. Pearce, K. Boakye, B. Van Essen, D. Borth, B. Chen, and E. Wang. Large-scale deep learning on the yfcc100m dataset. *arXiv preprint arXiv:1502.03409*, 2015.

[12] A. Smeaton, P. Over, and W. Kraaij. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer, 2009.

[13] B. Thomee, J. Moreno, and D. A Shamma. Who's time is it anyway?: Investigating the accuracy of camera timestamps. In *Proc. of the ACM Int. Conf. on Multimedia (ACM MM)*, pages 909–912. ACM, 2014.

[14] B. Thomee, D. A Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[15] V. Yanulevskaya, J. van Gemert, K. Roth, A. Herbold, N. Sebe, and J.M. Geusebroek. Emotional Valence Categorization using Holistic Image Features. In *Proc. IEEE Int Conf on Image Processing (ICIP)*, pages 101–104, October 2008.

[16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.