Real-time Analysis and Visualization of the YFCC100m Dataset

Firstname Lastname
Institute
City, Country
firstname.lastname@domain.tld

ABSTRACT

With the Yahoo Flickr Creative Common 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 possibilities 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 [14], 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 make progress in improving visual recognition performance.

One recently released dataset lining up in this context is the Yahoo Flickr Creative Commons 100 Million (YFCC100m) dataset [13]. 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 items and their associated meta data as available on Flickr. This includes also the URL for direct download of the photos and videos itself. Its vast size allows not only the analysis of how user generated media content is shared and annotated, but also provides a richness of content to serves as a large-scale source for training statistical methods [11] or the creation of specialized subsets [1, 4]. However, to make this dataset accessible for a broader audience (in particular of this size) it requires further mechanism than the released text file.

For example consider the case that a computer vision researcher wants to train a tree classifier (Fig. 1). Does the YFCC100m dataset contain a proper set of tree 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 come from the same user (a dataset of trees pictures from a single user might be biased). Further, the researcher might be interest in trees from a particular region such as Australia or be interested in a particular season in 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 particular query is of very helpful for a researcher to make the larger decision to download the dataset and all of the 100 Million images and video of the dataset.

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

- **Global Statistics**: a 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 how many images and videos does the dataset contain for a particular user defined query and browse these images on the fly.

- **Query Statistics**: providing general statistics for a user generated query allowing to get an idea about co-occurring tags, user distribution, time distribution, and geo-location.

- **Query Refinement**: to allow for modification of the query to 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.

### 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 popular 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 sub-synsets summarized by a image mosaic. Further, it allows the user to download the list of URL from the synset.

In contrast to ImageNet website and because of the nature of the YFCC100m datasets (user generated metadata), the proposed browser provides besides images associated with a concepts also an in-depth statistics about co-occurring tags, user and geo distribution in real-time to allow for quickly 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 on a online website to browser 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 online browser is focusing on the images and present object with its segmentation boundaries.

Because of the vast amount of images and videos, the proposed browser in contrast focuses on real-time accessibility of 100 million images with respect to 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.

In contract to the above mentioned online access mechanism, the proposed browser provides a query mechanism and query dependent statistics reporting to allow researchers to quickly identify a subset of the dataset relevant for their work, download it and start to work with the dataset and benefit from it. This query mechanism and reporting is in particular of importance because of the vast size of the dataset.

### 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 [13] for more details). This section provides a global analysis of the dataset. Similarly, user generated queries and their statistics can be performed on the online browser\(^1\).

\(^1\)http://yfcc100m.appspot.com/
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 tagging beyond 15 and more tags per items.

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

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

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 must 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 typing a description of each image or video during 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 if 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, reducable to 7,940,039 distinct tags demonstrating the broad vocabulary of the Flickr community.

However, popular tags can very often be grouped into one or more of the following categories: app-generated, camera descriptions, trends, activities, locations, dates and sometimes URLs. We found, that in most cases these categories of tags are highly descriptive concerning the image object. Especially tags falling into the camera- and app-type categories have a pre-eminent stop word character, for 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 implausible geolocations and timestamps, if needed. Table 2 shows the most popular tags and tag categories (Table 3). Although it is possible to insert a larger text as a single tag, 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 uploading 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 most active users (a number of only 9894 different users) are responsible for 50% of all uploaded items for the most active users). This observation indicates a strong bias towards these users making training data curation for classifier training sensitive to these user if mechanisms are not 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 the 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 users 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 89,038,561 images and videos provide a geo-position in form of a latitude and longitude value. Using this information we were able to map 86% of all items in the dataset to a single country. Around 3% 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. Obviously, seen on the linear scale, the majority of images is taken in 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. Congo, 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. 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 notable 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.

### 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 subsets of the YFCC100m dataset, we present the YFCC100m Browser, which can be found under http://yfcc100m.appspot.com. The browser is designed to filter and explore the entire dataset of 100 million images and
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, generated (see 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 dataset, ranked by their occurrence count. Clicking at a tag, or the small plus besides, allows to launch a refined search either only 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 defined 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\(^2\). The frontend is using in the Google AppEngine environment\(^3\), 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 back-end realizing search and query mechanism of the browser is running Google BigQuery\(^4\). This includes the retrieval, aggregation and temporarily 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.

\(^2\)https://cloud.google.com/compute/
\(^3\)https://cloud.google.com/appengine/
\(^4\)https://cloud.google.com/bigquery/
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 let's 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

With the YFCC100m dataset a great dataset for various computer vision tasks has been introduced. 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 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 in this paper, we provide a tool to the community, which provides quick access 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 dataset content. With its real-time performance, iterative refinement mechanism and adaption to subsets of the complete dataset researcher 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. REFERENCES


