**This publicly available curated dataset of almost 100 million photos and videos is free and legal for all.**

BY BART THOMEE, DAVID A. SHAMMA, GERALD FRIEDLAND, BENJAMIN ELIZALDE, KARL NI, DOUGLAS POLAND, DAMIAN BORTH, AND LI-JIA LI

# YFCC100M: The New Data in Multimedia Research

THE PHOTOGRAPH AND our understanding of photography transitioned from a world of unprocessed rolls of C-41 sitting in a refrigerator 50 years ago to sharing photos on the 1.5-inch screen of a point-and-shoot camera 10 years ago. Today, the photograph is again something different. The way we take photos has fundamentally changed from what it was. We can view, share, and interact with photos on the device that took them. We can edit, tag, or "filter" photos directly on the camera at the same time we take the photo. Photos can be automatically pushed to various online sharing services, and the distinction between photos and videos has lessened. Beyond this, and more important there are now lots of them. As of 2013, to Facebook alone more than 250 billion photos had been uploaded and on average received more than 350 million new photos each day,[6] while YouTube reported in July 2015 that 300 hours of video were uploaded every minute.[22] A back-of-the-envelope calculation estimated 10% of all photos in the world were taken in the last 12 months, as of more than four years ago.[8]

Today, a large number of shared digital media objects have been uploaded to services like Flickr and Instagram, which, along with their metadata and social ecosystem, form a vibrant environment for finding solutions to many research questions at scale. Photos and videos provide a wealth of information covering entertainment, travel, personal records, and various other aspects of life as they were when taken. Viewed collectively, they represent knowledge beyond what is captured in any individual snapshot and provide information on trends, evidence of phenomena or events, social context, and societal dynamics. Consequently, media collections are useful for qualitative and quantitative empirical research in many domains. However, scientific endeavors in fields like social computing and computer vision have generally relied on independently collected multimedia datasets, complicating research and synergy. Needed is a more substantial dataset for researchers, engineers, and scientists around the globe.

## » key insights

- **In the same way freely licensed works have advanced user-generated content rights, we propose a common dataset to advance scientific discovery and enhance entrepreneurship across academia and industry.**

- **We introduce a dataset with the volume and complexity to address questions across many fields—from computer vision to social computing to artificial intelligence and sensemaking.**

- **The core photos and videos in the dataset reveal a surprising amount of detail about how people experience and interact with the world and with each other; the multimedia commons extends the dataset annotations in a community-driven manner.**

To address the call for scale, openness, and diversity in academic datasets, we take the opportunity in this article to present a new multimedia dataset containing 100 million media objects we developed over the past two years and explain the rationale behind its creation. We discuss its implications for science, research, engineering, and development and demonstrate its usefulness toward tackling a range of problems in multiple domains. The release of the dataset represents an opportunity to advance research, giving rise to new challenges and addressing existing ones.

## Sharing Datasets

Datasets are critical for research and exploration,[16] as data is required to perform experiments, validate hypotheses, analyze designs, and build applications. Over the years, multimedia datasets have been put together for research and development; Table 1 summarizes the most popular multimedia datasets over time. However, most of them cannot truly be called multimedia, as they contain only a single type of media, rather than a mixture of modalities (such as photos, videos, and audio). Datasets range from one-off instances created exclusively to support

the work presented in a single paper or demo, or "short-term datasets," to those created with multiple related or separate endeavors in mind, or "long-term datasets." A notable problem is the collected data is often not made publicly available. While this restriction is sometimes out of necessity due to the proprietary or sensitive nature of the data, it is not always the case.

The topic of sharing data for replication and growth has arisen several times over the past 30 years alone[2,7,18] and has been brought into discussion through ACM's SIGCHI.[20] This "sharing discussion" reveals

many of the underlying complexities of sharing, with regard to both the data (such as what exactly is considered data) and the sharing point of view (such as incentives and disincentives for doing so); for example, one might be reluctant to share data freely, as it has a value from the often substantial amount of time, effort, and money invested in collecting it. Another barrier arises when data is harvested for research under a general umbrella of "academic fair use" without regard to its licensing terms. Beyond the corporate legal issues, such academic fair use may violate the copyright of the owner of the data that in many user-generated content sites like Flickr stays with the creator. The Creative Commons (CC), a nonprofit organization founded in 2001, seeks to build a rich public domain of "some rights reserved" media, sometimes referred to as the "copyleft movement." The licenses allow media owners to communicate how they would like their media to be rights reserved; for example, an owner can indicate a photo may be used for only noncommercial purposes or someone is allowed to remix it or turn it into a collage. Depending how the licensing options are chosen, CC licenses can be applied that are more restrictive (such as CC Attribution-Non-Commercial-NoDerivs/CC-BY-NC-ND license) or less restrictive (such as CC Attribution-ShareAlike/CC-BY-SA) in nature. A public dataset with clearly marked licenses that do not overly impose restrictions on how the data is used (such as those offered by CC) would therefore be suitable for use in both academia and industry.

We underscore the importance of sharing—perhaps even its principal argument—is it ensures data equality for research. While the availability of data alone may not necessarily be sufficient for the exact reproduction of scientific results (since the original experimental conditions would also have to be replicated as closely as possible, which may not always be possible), research should start with publicly sharable and legally usable data that is flexible and rich enough to promote advancement, rather than with data that serves only as a one-time collection for a specific task and that

cannot be shared. Shared datasets can play a singular role in achieving research growth and facilitating synergy within the research community otherwise difficult to achieve.

## YFCC100M Dataset
We created the Yahoo Flickr Creative Commons 100 Million Dataset[a] (YFCC100M) in 2014 as part of the Yahoo Webscope program, which is a reference library of interesting and scientifically useful datasets. The YFCC100M is the largest public multimedia collection ever released, with a total of 100 million media objects, of which approximately 99.2 million are photos and 0.8 million are videos, all uploaded to Flickr between 2004 and 2014 and published under a CC commercial or noncommercial license. The dataset is distributed through Amazon Web Services as a 12.5GB compressed archive containing only metadata. However, as with many datasets, the YFCC100M is constantly evolving; over time, we have released and will continue to release various expansion packs containing data not yet in the collection; for instance, the actual photos and videos, as well as several visual and aural features extracted from the data, have already been uploaded to the cloud,[b] ensuring the dataset remains accessible and intact for years to come. The YFCC100M dataset overcomes many of the issues affecting existing multimedia datasets in terms of modalities, metadata, licensing, and, principally, volume.

**Metadata.** Each media object included in the dataset is represented by its Flickr identifier, the user who created it, the camera that took it, the time it was taken and uploaded, the location where it was taken (if available), and the CC license under which it was published. The title, description, and tags are also available, as are direct links to its page and content on Flickr. Social features, comments, favorites, and followers/following data are not included, as such metadata changes from day to day. This information is, however, easily obtained by querying the Flickr API.[c]

We are working toward the release of the Exif metadata of the photos and videos as an expansion pack.

*Tags.* There are 68,552,616 photos and 418,507 videos in the dataset users have annotated with tags, or keywords. The tags make for a rich, diverse set of entities related to people (`baby`, `family`), animals (`cat`, `dog`), locations (`park`, `beach`), travel (`nature`, `city`), and more. A total of 3,343,487 photos and 7,281 videos carry machine tags—labels automatically generated and added by camera, computer, application, or other automated system.

*Timespan.* Although the YFCC100M dataset contains media uploaded between the inception of Flickr in 2004 and creation of the dataset in 2014, the actual timespan during which they were captured is much longer. Some scans of books and newspapers have even been backdated to the early 19th century when originally published. However, note camera clocks are not always set to the correct date and time, and some photos and videos erroneously report they were captured in the distant past or future; Figure 1 plots the moments of capture and upload of photos and videos during the period 2000–2014, or 99.6% of the media objects in the dataset.

*Locations.* There are 48,366,323 photos and 103,506 videos in the dataset that have been annotated with a geographic coordinate, either manually by the user or automatically through GPS. The cities in which more than 10,000 unique users captured media are Hong Kong, London, New York, Paris, San Francisco, and Tokyo. Overall, the dataset spans 249 different territories (such as countries and islands) and includes photos and videos taken in international waters and international airspace (see Figure 2).

*Cameras.* Table 2 lists the top 25 cameras used to take the photos and videos in the dataset as overwhelmingly digital single lens reflex (DSLR) models, with the exception of the Apple iPhone. Considering the most popular cameras in the Flickr community are primarily various iPhone models[d] this bias in the data is likely due to CC licenses attracting a certain subcommu-

---

a   Dataset available at https://bit.ly/yfcc100md
b   Photos, videos, and features available at http://www.multimediacommons.org/
c   https://www.flickr.com/services/api/

---

d   https://www.flickr.com/cameras/

**Table 1. Popular multimedia datasets used by the research community. When various versions of a particular collection are available, we generally include only the most recent one. PASCAL, TRECVID, MediaEval, and ImageCLEF are recurring annual benchmarks that consist of one or more challenges, each with its own dataset; here, we report the total number of media objects aggregated over all datasets that are part of the most recent edition of each benchmark.**

| Year | Dataset | Type | Image | Video | Audio | License | Accessibility | Content |
|------|---------|------|-------|-------|-------|---------|---------------|---------|
| 1966 | Brodatz | texture | <1K | - | - | © | ⊕ | |
| 1996 | COIL-100 | object | 7K | - | - | ⓘ | ⊕ | ★ |
| 1996 | Corel | stock | 60K | - | - | © | 🛒🚚 | ★ |
| 2000 | FERET | face | 14K | - | - | © | ⊕✎ | ★ |
| 2005 | Yale Face B+ | face | 16K | - | - | © | ⊕ | ★ |
| 2005 | Ponce | texture | 1K | - | - | © | ⊕ | ★ |
| 2007 | Caltech-256 | object | 30K | - | - | ⓘ | ⊕ | ★ |
| 2007 | Oxford | buildings | 5K | - | - | © | ⊕ | ★ |
| 2008 | CMU Multi-PIE | face | 750K | - | - | © | 🚚✎ | ★ |
| 2008 | Tiny Images | web | 80M | - | - | © | ⊕ | ★ A ⊞ |
| 2008 | MIRFLICKR-25K | Flickr | 25K | - | - | ⓒⓒ | ⊕ | ★ 📷 ⊙ ⊞ 👤 🏷 |
| 2009 | NUS-WIDE | Flickr | 270K | - | - | © | ⊕ 🔗 | ★ 📷 ⊙ ⊞ 🌐 🏷 |
| 2009 | ImageNet | web | 14M | - | - | © | ⊕✎→ | ★ A ⊞ ▢ |
| 2010 | SUN | web | 131K | - | - | © | ⊕ | ★ A ⊞ |
| 2010 | MIRFLICKR-1M | Flickr | 1M | - | - | ⓒⓒ | ⊕ | ☆ 📷 ⊙ ⊞ 👤 🏷 |
| 2012 | PASCAL | Flickr | 23K | - | - | © | ⊕✎→🏆 * | ★ ▢ ⬭ |
| 2013 | MS Clickture | web | 40M | - | - | © | 🚚✎→🔗 ** | ☰ |
| 2014 | Sports-1M | sports | - | 1M | - | ⓒⓒ | ⊕🔗 *** | ★ |
| 2014 | MS COCO | Flickr | 330K | - | - | ⓒⓒ | ⊕ | ★ A ⊙ ⬭ >_ ⚒ |
| 2014 | YFCC100M | Flickr | 99M | 800K | - | ⓒⓒ | ⊕✎→🔗 **** | ☆ 📷 ⊙ ⊞ 👤 🌐 🏷 ⚒ |
| 2015 | TRECVID | mixed | - | 220K | - | © | ⊕✎→🏆 | ★ >_ ⋯ |
| 2015 | MediaEval | mixed | 6M | 51K | 1,4K | © | ⊕✎→🏆🔗 ***** | ☆ ⊞ ⬭ >_ ⋯ |
| 2015 | ImageCLEF | mixed | 500K | - | - | © | ⊕✎→🏆 | ☆ A ⊞ ⋯ |

The icons represent the following:

- ©  Some or all content in dataset is copyrighted.
- ⓒⓒ  All content in dataset has a Creative Commons license.
- ⓘ  Content in dataset can be freely used on condition of citing the dataset paper.
- ⊕  Dataset has to be downloaded.
- 🛒  Dataset has to be purchased.
- 🚚  Dataset is delivered by mail.
- ✎  Dataset can only be obtained by accepting a license agreement.
- →  Dataset can only be obtained after creating an account.
- 🏆  Dataset can only be obtained by participating in a benchmark competition.
- 🔗  Dataset contains URLs to the content instead of the content itself.
- ★/☆  Dataset is fully/partially annotated with class labels.
- A  Dataset contains content found by querying search engines with dictionary words.
- ⊞  Dataset contains generated features.
- >_  Dataset contains subtitles, transcripts, or captions describing the content.
- ☰  Dataset contains search engine click log data.
- 👤  Dataset contains user information.
- 📷  Dataset contains camera information.
- ⊙  Dataset contains timestamps.
- 🌐  Dataset contains locations.
- 🏷  Dataset contains tags.
- ▢  Dataset contains object bounding boxes.
- ⬭  Dataset contains object segmentations.
- ⚒  Dataset is still evolving.
- ⋯  Dataset changes from year to year.

* The PASCAL training and development data can be freely downloaded, but the test data requires registration.

** Reduced-resolution images are included in the dataset, while full-resolution images must be downloaded separately.

*** The Sports-1M dataset has a CC license, though the videos it links to are hosted on YouTube and copyrighted.

**** The photos and videos have been uploaded to the cloud; like the metadata, the photo and video data can be mounted as a read-only network drive or downloaded.

***** Most MediaEval challenges include the media objects in their dataset, though some provide only URLs; in previous editions data had to be purchased and delivered by postal mail in order to participate in certain challenges.

nity of photographers that differs from the overall Flickr user base.

*Licenses.* The licenses themselves vary by CC type, with approximately 31.8% of the dataset marked appropriate for commercial use and 17.3% assigned the most liberal license re-

quiring attribution for only the photographer who took the photo (see Table 3).

**Content.** The YFCC100M dataset includes a diverse collection of complex real-world scenes, ranging from 200,000 street-life-blogged photos

by photographer Andy Nystrom (see Figure 3a) to snapshots of daily life, holidays, and events (see Figure 3b). To understand more about the visual content represented in the dataset, we used a deep-learning approach to detect a variety of concepts (such as

people, animals, objects, food, events, architecture, and scenery). Specifically, we applied an off-the-shelf deep convolutional neural network[13] with seven hidden layers, five convolutional layers, and two fully connected layers. We employed the penultimate layer of the convolutional neural network output as the image-feature representation for training the visual-concept classifiers. We used Caffe[11] to train 1,570 classifiers, each

a binary support vector machine, using 15 million photos we selected from the entire Flickr corpus; positive examples were crowd-labeled or handpicked by us based on targeted search/group results, while we drew negative examples from a general pool. We tuned the classifiers such that they achieved at least 90% precision on a held-out test set; Table 4 lists the top 25 detected concepts in both photos and videos (using the

first frame). We see a diverse collection of visual concepts being detected, from outdoor to indoor images, sports to art, and nature to architecture. As we view the detected visual concepts as valuable to the research community, we released them as one of our expansion packs in July 2015.

Flickr makes little distinction between photos and videos, though videos do play a role in Flickr and in the YFCC100M dataset. While photos encode their content primarily through visual means, videos also do so through audio and motion. Only 5% of the videos in the YFCC100M dataset lack an audio track. From a manual examination of more than 120 randomly selected geotagged videos with audio, we found most of the audio tracks to be diverse; 60% of the videos were home-video style with little am-

Figure 1. Number of captured and uploaded media objects per month in the YFCC100M dataset, 2000–2014; the number of uploads closely follows the number of captures, with the number of more-recent uploads exceeding the number of captures as older media is uploaded.
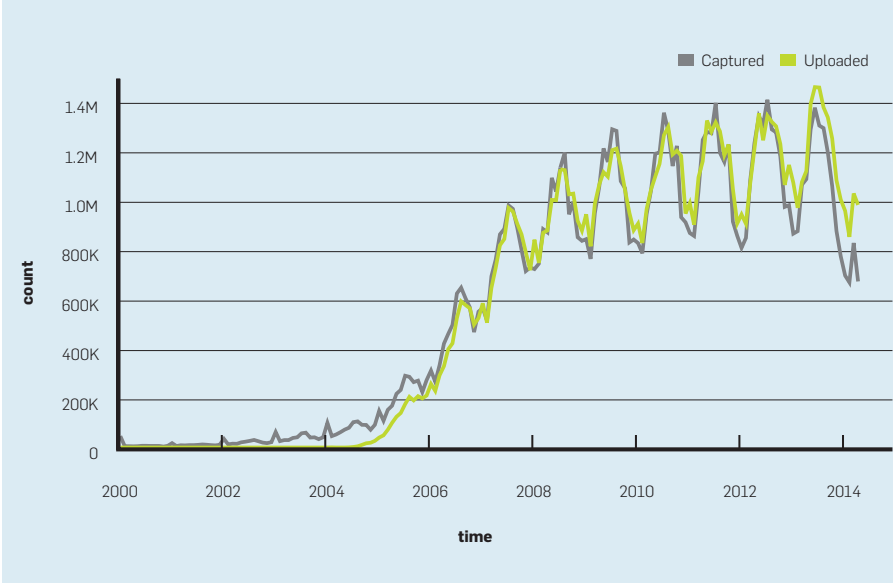


Figure 2. Global coverage of a sample of one million photos from the YFCC100M dataset; *One Million Creative Commons Geotagged Photos* by David A. Shamma ⓒ ⓘ ⊜ (https://flic.kr/p/o1Ao2o).
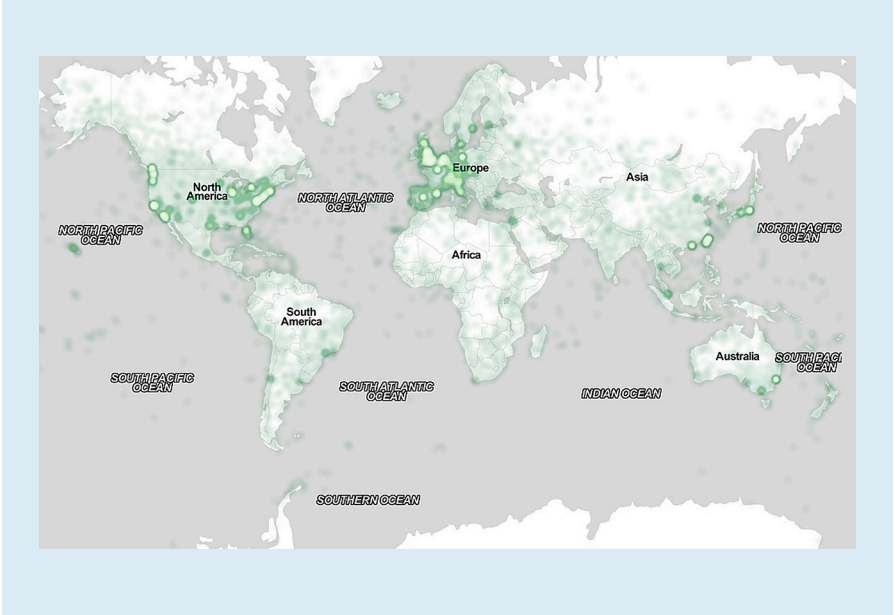


Table 2. Top 25 cameras and photo counts in the YFCC100M dataset; we merged the entries for the Canon models in the various markets, European (such as EOS 650D), American (such as EOS Rebel T4i), and Asian (such as EOS Kiss X6i).

| Make | Model | Photos |
|---|---|---|
| Canon | EOS 400D | 2,539,571 |
| Canon | EOS 350D | 2,140,722 |
| Nikon | D90 | 1,998,637 |
| Canon | EOS 5D Mark II | 1,896,219 |
| Nikon | D80 | 1,719,045 |
| Canon | EOS 7D | 1,526,158 |
| Canon | EOS 450D | 1,509,334 |
| Nikon | D40 | 1,358,791 |
| Canon | EOS 40D | 1,334,891 |
| Canon | EOS 550D | 1,175,229 |
| Nikon | D7000 | 1,068,591 |
| Nikon | D300 | 1,053,745 |
| Nikon | D50 | 1,032,019 |
| Canon | EOS 500D | 1,031,044 |
| Nikon | D700 | 942,806 |
| Apple | iPhone 4 | 922,675 |
| Nikon | D200 | 919,688 |
| Canon | EOS 20D | 843,133 |
| Canon | EOS 50D | 831,570 |
| Canon | EOS 30D | 820,838 |
| Canon | EOS 60D | 772,700 |
| Apple | iPhone 4S | 761,231 |
| Apple | iPhone | 743,735 |
| Nikon | D70 | 742,591 |
| Canon | EOS 5D | 699,381 |

bient noise; 47% had heavy ambient noise (such as people chatting in the background, traffic sounds, and wind blowing into the microphone); 25% of the sampled videos contained music, played in the background of the recorded scene or inserted during editing; 60% of the videos did not contain any human speech at all, while for the 40% that did contain human speech, 64% included multiple subjects and crowds in the background speaking to one another, often at the same time. The vocabulary of approximately 280,000 distinct user tags used as video annotations indeed shows tags describing audio content (music, concert, festival) and motion content (timelapse, dance, animation) were more frequently applied to videos than to photos. When comparing the videos in the dataset to those from YouTube, 2007–2012, we found YouTube videos are on average longer (Flickr: 39 seconds, YouTube: 214 seconds). This is likely due to the initial handling of videos on Flickr where their length until May 2013 was restricted to a maximum of 90 seconds; recent videos uploaded to Flickr tend to be longer.

**Representativeness.** In creating the dataset, we did not perform any special filtering other than to exclude photos and videos that had been marked as "screenshot" or "other" by the Flickr user. We did, however,

include as many videos as possible, as videos represent a small percentage of media uploaded to Flickr, and a random selection would have led to relatively few videos being selected. We further included as many photos as possible associated with a geographic coordinate to encourage spatiotemporal research. These photos and videos together form approximately half of the dataset; the rest is CC photos we randomly selected from the entire pool of photos on Flickr.

To investigate whether the YFCC-100M dataset includes a representative sample of real-world photography, we collected an additional random sample of 100 million public Flickr photos and videos, irrespective of their license, uploaded during the

Table 3. A breakdown of the 100 million photos and videos by their kind of ⓒⓒ Creative Commons license, ⓘ attribution, ⊖ no derivatives, ⓐ share alike, and Ⓢ noncommercial.

| License | Photos | Videos |
|---|---|---|
| ⓒ ⓘ | 17,210,144 | 137,503 |
| ⓒ ⓘ ⓐ | 9,408,154 | 72,116 |
| ⓒ ⓘ ⊖ | 4,910,766 | 37,542 |
| ⓒ ⓘ Ⓢ | 12,674,885 | 102,288 |
| ⓒ ⓘ Ⓢ ⓐ | 28,776,835 | 235,319 |
| ⓒ ⓘ Ⓢ ⊖ | 26,225,780 | 208,668 |
| Total | 99,206,564 | 793,436 |

Table 4. The top 25 of 1,570 visually detected concepts in the YFCC100M dataset; photos and videos are counted by how often they include visual concepts.

| Concept | Photos | Videos |
|---|---|---|
| Outdoor | 44,290,738 | 266,441 |
| Indoor | 14,013,888 | 127,387 |
| People | 11,326,711 | 56,664 |
| Nature | 9,905,587 | 47,703 |
| Architecture | 6,062,789 | 11,289 |
| Landscape | 5,121,604 | 28,222 |
| Monochrome | 4,477,368 | 18,243 |
| Sport | 4,354,325 | 25,129 |
| Building | 4,174,579 | 7,693 |
| Vehicle | 3,869,095 | 13,737 |
| Plant | 3,591,128 | 11,815 |
| Black and White | 2,585,474 | 10,351 |
| Animal | 2,317,462 | 9,236 |
| Groupshot | 2,271,390 | 4,392 |
| Sky | 2,232,121 | 11,488 |
| Water | 2,089,110 | 15,426 |
| Text | 2,074,831 | 5,623 |
| Road | 1,796,742 | 12,808 |
| Blue | 1,658,929 | 10,273 |
| Tree | 1,641,696 | 6,808 |
| Hill | 1,448,925 | 6,075 |
| Shore | 1,439,950 | 8,602 |
| Car | 1,441,876 | 4,067 |
| Head | 1,386,667 | 8,984 |
| Art | 1,391,386 | 2,248 |

Figure 3. Two photos of real-world scenes from photographers in the YFCC100M dataset: (a) *IMG_9793: Streetcar (Toronto Transit)* by Andy Nystrom ⓒ ⓘ Ⓢ ⊖ (https://flic.kr/p/jciMdz) and (b) *Celebrating our 6th wedding anniversary in Villa Mary* by Rita and Tomek ⓒ ⓘ Ⓢ ⓐ (https://flic.kr/p/fCXEJi).



(a)

(b)

same time period as those included in the dataset. We then compared the relative frequency with which content and metadata are present in the YFCC100M dataset and in the random sample. We found the average difference in relative frequencies between two corresponding visual concepts, cameras, timestamps (year and month), and locations (countries) was only 0.02%, with an overall standard deviation of 0.1%. The maximum difference we observed was 3.5%, due to more videos in the YFCC100M having been captured in the U.S. than in the random sample (46.2% vs. 42.7%). While we found little individual difference between the relative frequency of use of any two corresponding cameras in the YFCC100M dataset and in the random sample, at most 0.5%, we did find the earlier mentioned tendency toward more professional DSLR cameras in the dataset rather than regular point-and-shoot cameras. This tendency notwithstanding, the dataset appears to exhibit similar characteristics as photos and videos in the entire Flickr corpus.

**Features and annotations.** Computing features for 100 million media objects is time consuming and computationally expensive. Not everyone has access to a distributed computing cluster, and performing even light processing of all the photos and videos on a single desktop machine could take several days. From our experience organizing benchmarks on image annotation and location estimation we noted accompanying the datasets with pre-computed features reduced the burden on the participating teams, allowing them to focus on solving the task at hand rather than on processing the data. As mentioned earlier, we are currently computing a variety of visual, aural, textual, and motion features for the dataset and have already released several of them. The visual features span the gamut of global (such as Gist), local (such as SIFT), and texture (such as Gabor) descriptors; the aural features include power spectrum (such as MFCC) and frequency (such as Kaldi) descriptors; the textual features refer to closed captions extracted from the videos; and the motion features include dense trajectories and shot boundaries. These features, as computed descriptors of the photos

and videos, will be licensed without restriction under the CC0 (©⓪) license. Real-world data lacks well-formed annotations, thus highlighting the sense-making of the dataset itself as an area for investigation. Annotations (such as bounding boxes, segmentations of objects and faces, and image captions) are not yet available for the YFCC100M, though generating and releasing them is on our roadmap.

**Ecosystem**. The YFCC100M dataset has already given rise to an ecosystem of diverse challenges and benchmarks, similar to how ImageNet, PASCAL, and TRECVID have been used by the multimedia research community; for example, the MediaEval Placing Task,[3] an annual benchmark in which participants develop algorithms for estimating the geographic location where a photo or video was taken, is currently based on our dataset. To support research in multimedia event detection the YLI-MED corpus[1] was introduced September 2014 and consists of 50,000 handpicked videos from the YFCC100M that belong to events similar to those defined in the TRECVID MED challenge. Approximately 2,000 videos were categorized as depicting one of 10 target events and 48,000 as belonging to none of these events. Each video was further annotated with additional attributes like language spoken and whether it includes a musical score. The annotations also include degree of annotator agreement and average annotator confidence scores for the event categorization of each video. The authors said the main motivation for the creation of the YLI-MED corpus was to provide an open set without the license restrictions imposed on the original TRECVID MED dataset, while possibly also serving as, say, additional annotated data to improve the performance of current event detectors. Other venues incorporating the YFCC100M dataset are the ACM Multimedia 2015 Grand Challenge on Event Detection and Summarization and the ACM Multimedia 2015 MMCommons Workshop; the latter aims to establish a research community around annotating all 100 million photos and videos in the YFCC100M. The utility of the dataset is expected to grow as more features and annotations are produced and shared, whether by us or by others.

**Strengths and limitations.** Note the following strengths (⊕) and limitations (⊖) of the YFCC100M dataset.

⊕ *Design.* The YFCC100M dataset differs in design from most other multimedia collections. Its photos, videos, and metadata have been curated by us to be comprehensive and representative of real-world photography, expansive and expandable in coverage, free and legal to use, and intended to consolidate and supplant many of the existing datasets currently in use. We emphasize it does not challenge collections that are different and unique (such as PASCAL, TRECVID, ImageNet, and COCO); we instead aspire to make it the preferred choice for researchers, developers, and engineers with small and large multimedia needs that may be readily satisfied by the dataset, rather than having them needlessly collect their own data.

⊕ *Equality.* The YFCC100M dataset ensures data equality for research to facilitate reproduction, verification, and extension of scientific experiments and results.

⊕ *Volume.* Spanning 100 million media objects, the YFCC100M dataset is the largest public multimedia collection ever released.

⊕ *Modalities.* Unlike most existing collections, the YFCC100M dataset includes both photos and videos, making it a truly multimodal multimedia collection.

⊕ *Metadata.* Each media item is represented by a substantial amount of metadata, including some (such as machine tags, geotags, timestamps, and cameras) often absent from existing datasets. While social metadata is not included due to its ever-changing nature, it is readily obtained by querying the Flickr API.

⊕ *Licensing.* The vast majority of available datasets includes media for which licenses do not allow their use without explicit permission from the rightsholder. While fair-use exceptions may be invoked, they are, depending on the nature of use, generally not applicable to research and development performed by industry and/or for commercial gain; for example, a university spin-off offering a mobile product-recognition application that displays matching ImageNet images for each detected product

would violate not only the ImageNet license agreement but also very likely copyright law. The YFCC100M dataset prevents potential violations by providing rules on how the dataset should be used to comply with licensing, attribution, and copyright.

⊖ *Annotations.* The YFCC100M dataset reflects the data as it is in the wild; there are lots of photos and videos, but they are currently associated with limited metadata and annotations. Note the dataset may not and/or cannot offer every kind of content, metadata, and annotation in existing collections (such as object segmentations, as in COCO, and broadcast videos, as in TRECVID), although our efforts and those that spring from the ecosystem being constructed around it will offer more depth and richness to the existing content, as well as new material, making it more complete and useful over time. While a lack of annotations represents a limitation of the dataset, it is also a challenge. With 100 million media objects, there are enough metadata labels for training and prediction of some attributes (such as geographic coordinates) and opportunities to create new methods for labeling and annotation through explicit crowdsourced work or through more modern techniques involving social community behaviors. In addition, given the plurality of existing Flickr datasets and the size of our dataset, some overlap is to be expected, such that existing annotations directly transfer over to the dataset. Of special note is COCO, of which one-third (approximately 100,000 images) is included in the YFCC100M. Over time we will also release the intersections with known datasets as expansion packs. We envision the intersection with existing datasets will allow researchers to expand on what is known and actively researched.

**Guidelines and recommendations.** Although we consider volume a strength of the YFCC100M dataset, it can also pose a weakness when insufficient computational power, memory, and/or storage is available. The compressed metadata of all 100 million media objects requires 12.5GB hard disk space and, at the default pixel resolution used on Flickr, the

**Although we consider volume a strength of the YFCC100M dataset, it can also pose a weakness when insufficient computational power, memory, and/or storage is available.**

photos take up approximately 13.5TB and the videos 3.0TB. While the entire dataset—metadata and/or content—can be processed in minutes to hours on a distributed computing cluster, it might take a few hours to days on a single machine. It can still be used for experiments by focusing on only a subset of the data. Also, different fields of research, engineering, and science have different data requirements and evaluation needs, and all 100 million media objects in the YFCC100M dataset are not likely to be needed for each and every study. Note it is uncommon in the computer science literature for a paper to describe in enough detail how the dataset the authors used in their evaluations was created, effectively preventing others from fully replicating or comparing against the achieved results. One clear future challenge is how to ensure subsets of the dataset used in experiments can be reproduced accurately. To this end, we suggest researchers forego arbitrary selections from the YFCC100M dataset when forming a subset for use in their evaluations but rather use a principled approach that can be described succinctly. Such selection logic should examine one or both of two aspects of the dataset: the photos and videos in it are already randomized, and it consists of 10 consecutively numbered files. As such, a selection logic could be as simple as "We used the videos in the first four metadata files for training, those in the following three files for development, and those in the last three for testing" or in more complicated form as "From all photos taken in the U.S., we selected the first five million and performed tenfold cross-validation." Alternatively, the created subset can be made available for download described in terms of a set of object identifiers that index into the dataset. As an example, the organizers of the MediaEval Placing Task made the visual and aural features they extracted from the content available for download, in addition to the training and test sets. We envision the research community as likewise following this way of using and sharing the dataset.

**Future directions.** The YFCC100M dataset enables large-scale unsupervised learning, semi-supervised learn-

ing, and learning with noisy data. Beyond this, the dataset offers the opportunity to advance research, give rise to new challenges, and address existing problems; note the following challenges in which the YFCC100M dataset might play a leading role.

*Artificial intelligence and vision.* Large datasets have played a critical role in advancing computer vision research. ImageNet[5] has led the way toward advanced machine learning techniques like deep learning.[13] Computer vision as a field now seeks to do visual recognition by learning and benchmarking from large-scale data. The YFCC100M dataset provides new opportunities in this direction by developing new approaches that harness more than just pixels; for instance, new semantic connections can be made by inferring them through relating groups of visually co-occurring tags in images depicting similar scenes, where such efforts hitherto were hampered by lack of a sufficiently large dataset. Our expansion pack containing the detected visual concepts in photos and videos can help. However, visual recognition goes beyond image classification toward obtaining a deeper understanding of an image. Object localization, object interaction, face recognition, and image annotation are all important cornerstone challenges that will lead the way to retelling the story of an image—what is happening in the image and why it was taken. With a rich, diverse collection of image types, Flickr provides the groundwork for total scene understanding[14] in computer vision and artificial intelligence, a crucial task that can be expanded through the YFCC100M dataset, and even more once additional annotations are released.

*Spatiotemporal computing.* Beyond pixels, we emphasize time and location information as key components for research that aims to understand and summarize events. Starting with location, the geotagged photos and videos (approximately half of the dataset) provide a comprehensive snapshot of photographic activity in space and time. In the past, geotagged media was used to address a variety of research endeavors (such as location estimation,[9] event detection,[15] finding canonical views

**While the entire dataset—metadata and/or content—can be processed in minutes to hours on a distributed computing cluster, it might take a few hours to days to do it on a single machine.**

of places,[4] and visual reconstruction of the world[17]). Even styles and habits of understanding have been used to reverse-lookup authors online.[10] Geotagged media in the YFCC100M dataset can help push the boundaries in these research areas.

More data brings new discoveries and insights, even as it makes searching, organizing, and presenting the data and findings more difficult. The cameraphone has enabled people to capture more photos and videos than they can effectively organize. One challenge for the future is thus devising algorithms able to automatically and dynamically create albums for use on personal computers, cloud storage, or mobile devices, where desired media and moments of importance are easily surfaced based on simple queries. Harnessing the spatiotemporal context at capture time and query time will thus take a central role.

The challenge of automatically creating albums speaks toward social computing efforts aimed at understanding events in unstructured data through the reification of photos with space and time. While GPS-enabled devices are capable of embedding the precise time, location, and orientation of capture in a photo's metadata, this information (including seconds, hours, and sometimes even months) is often unavailable or out of sync. In addition, people frequently forget to adjust the camera clock to the correct time zone when traveling. Such issues pose problems for the accuracy of any kind of spatiotemporal data analysis, and new challenges in computational photography thus include devising algorithms that either fix or are resilient against erroneous information.

*Digital culture and preservation.* What we know to be user-generated content has grown from simple video uploads and bulletin board systems; life online has come to reflect culture. These large online collections tell a larger story about the world around us, from consumer reviews[21] on how people engage with the spaces around them to 500 years of scanned book photos and illustrations[12] that describe how concepts and objects have been visually depicted over time. Beyond archived collections, the pho-

tostreams of individuals represent multiple facets of recorded visual information, from remembering moments and storytelling to social communication and self-identity.[19] How to preserve digital culture is a grand challenge of sensemaking and understanding digital archives from nonhomogeneous sources. Photographers and curators alike have contributed to the larger collection of Creative Commons images, yet little is know of how such archives will be navigated and retrieved or how new information can be discovered therein. The YFCC100M dataset offers avenues of computer science research in multimedia, information retrieval, and data visualization, in addition to the larger questions of how to preserve digital libraries.

## Conclusion

Data is a core component of research and development in all scientific fields. In the field of multimedia, datasets are usually created for a single purpose and, as such, lack reusability. Moreover, datasets generally are not or may not be freely shared with others and, as such, also lack reproducibility, transparency, and accountability. That is why we released one of the largest datasets ever created, with 100 million media objects, published under a Creative Commons license. We curated the YFCC100M dataset to be comprehensive and representative of real-world photography, expansive and expandable in coverage, free and legal to use, and intentionally consolidate and supplant many existing collections. The YFCC100M dataset encourages improvement and validation of research methods, reduces the effort to acquire data, and stimulates innovation and potential new data uses. We have further provided rules on how the dataset should be used to comply with licensing, attribution, and copyright and offered guidelines on how to maximize compatibility and promote reproducibility of experiments with existing and future work.

## Acknowledgments

Ⓒ

### References

1. Bernd, J., Borth, D., Elizalde, B., Friedland, G., Gallagher, H., Gottlieb, L.R., Janin, A., Karabashlieva, S., Takahashi, J., and Won, J. The YLI-MED corpus: Characteristics, procedures, and plans. Computing Research Repository Division of arXiv abs/1503.04250 (Mar. 2015).
2. Borgman, C.L. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology 63*, 6 (Apr. 2012), 1059–1078.
3. Choi, J., Thomee, B., Friedland, G., Cao, L., Ni, K., Borth, D., Elizalde, B., Gottlieb, L., Carrano, C., Pearce, R., and Poland, D. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the Third ACM International Workshop on Geotagging and Its Applications in Multimedia* (Orlando, FL, Nov. 3–7). ACM Press, New York, 2014, 27–31.
4. Crandall, D.J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. Mapping the world's photos. In *Proceedings of the 18th IW3C2 International Conference on the World Wide Web* (Madrid, Spain, Apr. 20–24). ACM Press, New York, 2009, 761–770.
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL, June 20–25). IEEE Press, New York, 2009. 248–255.
6. Facebook, Ericsson, and Qualcomm. *A Focus on Efficiency.* Technical Report, Internet.org, 2013; https://web.archive.org/web/20150402101302/http://internet.org/efficiencypaper
7. Fienberg, S.E., Martin, M.E., and Straf, M.L. Eds. (National Research Council). *Sharing Research Data.* National Academy Press, Washington, D.C., 1985; http://www.nap.edu/catalog/2033/sharing-research-data
8. Good, J. How many photos have ever been taken?. *Internet Archive Wayback Machine*, Sept. 2011; https://web.archive.org/web/20150203215607/http://blog.1000memories.com/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox
9. Hays, J. and Efros, A.A. IM2GPS: Estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Anchorage, AK, June 23–28). IEEE Press, New York, 2008.
10. Hecht, B., Hong, L., Suh, B., and Chi, E. H. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, Canada, May 7–12). ACM Press, New York, 2011, 237–246.
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, FL, Nov. 3–7). ACM Press, New York, 2014, 675–678.
12. Kremerskothen, K. Welcome the Internet archive to the commons. Flickr, San Francisco, CA, Aug. 2014; https://blog.flickr.net/2014/08/29/welcome-the-internet-archive-to-the-commons
13. Krizhevsky, A., Sutskever, I., and Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems* (Lake Tahoe, CA, Dec 3–8). Curran Associates, Red Hook, NY, 2012, 1097–1105.
14. Li, L., Socher, R., and Fei-Fei, L. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL, June 20–25). IEEE Press, New York, 2009, 2036–2043.
15. Rattenbury, T., Good, N., and Naaman, M. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the 30th ACM International Conference on Research and Development in Information Retrieval* (Amsterdam, the Netherlands, July 23–27). ACM Press, New York, 2007, 103–110.
16. Renear, A.H., Sacchi, S., and Wickett, K.M. Definitions of dataset in the scientific and technical literature. In *Proceedings of the 73rd Annual Meeting of the American Society for Information Science and Technology* (Pittsburgh, PA, Oct. 22–27). Association for Information Science and Technology, Silver Spring, MD, 2010, article 81.
17. Snavely, N., Seitz, S., and Szeliski, R. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics 25*, 3 (July 2006), 835–846.
18. Swan, A. and Brown, S. *To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs.* Technical Report. Research Information Network, London, U.K., 2008.
19. Van Dijck, J. Digital photography: Communication, identity, memory. *Visual Communication 7*, 1 (Feb. 2008), 57–76.
20. Wilson, M.L., Chi, E.H., Reeves, S., and Coyle, D. RepliCHI: The workshop II. In *Proceedings of the International Conference on Human Factors in Computing Systems, Extended Abstracts* (Toronto, Canada, Apr. 26–May 1). ACM Press, New York, 2014, 33–36.
21. Yelp. Yelp Dataset Challenge. Yelp, San Francisco, CA; http://yelp.com/dataset_challenge
22. YouTube. YouTube press statistics. YouTube, San Bruno, CA; http://youtube.com/yt/press/statistics.html

**Bart Thomee** (bthomee@yahoo-inc.com) is a senior research scientist in the HCI Research Group at Yahoo Labs and Flickr in San Francisco, CA.

**David A. Shamma** (aymans@acm.org) is director of the HCI Research Group at Yahoo Labs and Flickr in San Francisco, CA.

**Gerald Friedland** (fractor@icsi.berkeley.edu) is director of the Audio and Multimedia Lab at the International Computer Science Institute in Berkeley, CA.

**Benjamin Elizalde** (bmartin1@andrew.cmu.edu) is a Ph.D. student at Carnegie Mellon University in Mountain View, CA; this work was done while he was at the International Computer Science Institute in Berkeley, CA.

**Karl Ni** (kni@iqt.org) is a program lead and senior data scientist at In-Q-Tel's Lab41 in Menlo Park, CA; this work was done while he was at Lawrence Livermore National Laboratory in Livermore, CA.

**Douglas Poland** (poland1@llnl.gov) is a principal investigator at the Lawrence Livermore National Laboratory in Livermore, CA.

**Damian Borth** (damian.borth@dfki.de) is head of the Multimedia Analysis & Data Mining Group at the German Research Center for Artificial Intelligence in Kaiserslautern, Germany; this work was done while he was at the International Computer Science Institute in Berkeley, CA.

**Li-Jia Li** (lijiali.vision@gmail.com) is head of research at Snapchat, Venice, CA; this work was done while she was at Yahoo Labs, San Francisco, CA.

Watch the authors discuss their work in this exclusive *Communications* video. http://cacm.acm.org/videos/yfcc100m-the-new-data-in-multimedia-research