The New Data and New Challenges in Multimedia Research

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde,
Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li

***Abstract*—We present the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M), the largest public multimedia collection that has ever been released. The dataset contains a total of 100 million media objects, of which approximately 99.2 million are photos and 0.8 million are videos, all of which carry a Creative Commons license. Each media object in the dataset is represented by several pieces of metadata, e.g. Flickr identifier, owner name, camera, title, tags, geo, media source. The collection provides a comprehensive snapshot of how photos and videos were taken, described, and shared over the years, from the inception of Flickr in 2004 until early 2014. In this article we explain the rationale behind its creation, as well as the implications the dataset has for science, research, engineering, and development. We further present several new challenges in multimedia research that can now be expanded upon with our dataset.**

***Index Terms*—Dataset, Yahoo, Flickr, Creative Commons, 100 Million, YFCC100M.**

I. Introduction

The photograph and our understanding of photography is ever changing and has transitioned from a world of unprocessed rolls of C-41 sitting in a fridge 50 years ago to sharing photos on the 1.5” screen of a point and shoot camera 10 years back. And today the photograph is again something different. The way we take photos is fundamentally different. We can view, share, and interact with photos on the device they were taken on. We can edit, tag, or “filter” photos directly on the camera at the same time the photo is being taken. Photos can be automatically pushed to various online sharing services, and the distinction between photos and videos has lessened. Beyond this, and more importantly, there are now lots and lots of them. To Facebook alone more than 250 billion photos have been uploaded and on average it receives over 350 million new photos every day [1], while YouTube reports that 100 hours of video are uploaded every single minute [2]. A back of the envelope estimation reports 10% of all photos in the world were taken in the last 12 months, and that was calculated already more than three years ago [3].

Today, a large number of the digital media objects that are shared have been uploaded to services like Flickr or Instagram, which along with their metadata and their social ecosystem form a vibrant environment for finding solutions to many research questions at scale. However, scientific endeavors in social computing and computer vision alike have generally relied on various sized efforts of multimedia datasets that complicates research growth and synergy. There is the need for a more substantial dataset for researchers, engineers, and scientists around the globe.

To meet 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, and explain the rationale behind its creation, as well as the implications the dataset has for science, research, engineering, and development. With it comes the opportunity to advance current research in multimedia and computer vision, as well as to bring new challenges to social computing and digital preservation. Further, real world data does not have well formed annotations which presents the sense-making of the dataset itself as an area of investigation. We touch upon all these topics in this paper.

Bart Thomee and David A. Shamma are with Yahoo Labs, San Francisco, CA, USA. Contact: bthomee@yahoo-inc.com.
Gerald Friedland, Benjamin Elizalde and Damian Borth are with ICSI, University of California, Berkeley, USA.
Karl Ni and Douglas Poland are with Lawrence Livermore National Laboratory, Livermore, CA, USA.
Li-Jia Li is with Snapchat, Venice, CA, USA; this work was done while she was at Yahoo Labs.

II. Sharing Datasets

Datasets are critical for research and exploration [4] as, rather obviously, data is required for performing experiments, validating hypotheses, analyzing designs, and building applications. Over the years a plurality of multimedia datasets have been put together for research and development. These datasets range from the one-off instances that have been exclusively created for supporting the work presented in a single paper or demo (‘short-term datasets’) to those that have been created with multiple related or separate endeavors in mind (‘long-term datasets’). Most datasets cannot be truly called multimedia collections, due to only containing a single type of media. The photos and/or videos contained in these datasets have typically been collected from one or more sources, such as from personal collections, stock collections, web pages, search engines, news broadcasters, and social media. Here we briefly highlight the characteristics of the most well-known datasets currently used by the research community.

MIRFLICKR has had two releases of Flickr photos. The initial 2008 release [5] included 25,000 images with ground truth annotations for 29 concepts, while the 2010 release [6] contained one million images without these annotations, but including visual features. It was an early attempt of making a real and functional dataset from modern images with social tag annotation and Exchangeable Image File Format (EXIF) data.

NUS-WIDE [7] released 269,648 images along with semi-automatically obtained ground truth annotations for 81 concepts, such as events, scenes and people, organized in a multi-level hierarchy. Three smaller versions of the dataset were also released to accommodate those without sufficient computational machinery at their disposal to handle the entire collection.

PASCAL [8] dataset contains images with ground truth annotations for 20 concepts, including their bounding boxes and pixel-level segmentation. For the person concept additional body part annotations are also supplied, as well as one of 10 different actions the person may be performing. The dataset featured in a yearly challenge held between 2006 and 2012.

ImageNet [9] dataset is constructed by using the hierarchical structure of WordNet; concepts are grouped into sets of synonyms called *synsets*. Concepts vary from animals such as 'sea lion' to man-made objects like 'aircraft'. Manually labeled images were collected via crowdsourcing. In total, there are over 100,000 synsets in WordNet. The dataset, currently containing 14,197,122 images, is the largest dataset in terms of the number of classes, and it is used as a benchmark in various vision challenges. About 7% of the images have bounding box annotations, and 9% have visual features.

Microsoft COCO [10] dataset contains 91 object categories, where 82 of them are associated with more than 5,000 labeled instances. In total the dataset has 2.5 million object annotations and segmentations available for the 328,000 images. In contrast to the ImageNet dataset, COCO has fewer categories but more instances per category.

SUN [11] focused on building a large dataset of scenes. The dataset contains 131,072 images for 908 scenes that were collected from the Web and manually filtered to ensure relevance to the scene category. Using an online annotation tool people have been invited to draw the outlines of objects visible in the images, so far producing over 325,000 annotations for 5,650 object categories.

TRECVID [12] is a yearly recurring video retrieval benchmark that currently offers several datasets, which include approximately 18,500 Creative Commons licensed videos from the Internet Archive, 150 hours of airport surveillance videos, several seasons of a British soap opera, and 231,000 user-generated videos crawled from the Web. Each dataset requires its own license agreement to sign, some of which restrict the data to be exclusively used in the context of the benchmark.

While these datasets support a broad spectrum research topics in the field of multimedia, for instance ranging across object detection, image segmentation and action recognition, a substantial number of studies can be found in the literature that have collected their own data rather than using an existing dataset. One particular problem is that in those instances the collected data is typically not made publicly available. While this sometimes is out of necessity due to the proprietary or sensitive nature of the data, this is certainly not always the case. The question of sharing data for replication and growth has arisen several times in the past 30 years alone [13–15] and has been brought into discussion in ACM's SIGCHI [16]. This "sharing discussion" reveals many of the underlying complexities with sharing, both with regards to the data (e.g. what exactly is considered data) and to the sharing point of view (e.g. 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 that was invested in collecting it. Another barrier to sharing arises when data is harvested for research under a general umbrella of "academic fair use" without regards towards its licensing terms. Beyond the legal corporate issues, this clearly may violate the copyright of the owner of the data that, in many User-Generated Content (UGC) sites like Flickr, remains with the creator. For example, all aforementioned collections except MIRFLICKR include copyrighted photos and videos. Yet, not all data is covered by a license that prevents it from being shared and/or used for non-commercial or even commercial purposes.

The Creative Commons (CC), a nonprofit organization that was founded in 2001, seeks to build a rich public domain of *"some rights reserved"* media, sometimes referred to as the copyleft movement. Their licenses allow media owners to communicate how they would like their media to be rights reserved. For example, an owner can indicate that a photo may be used only for non-commercial purposes or whether someone is allowed to remix it or turn it into a collage. Depending on how the licensing options are chosen, CC licenses can be applied that are more restrictive (e.g. `CC Attribution-NonCommercial-NoDerivs` or `CC-BY-NC-ND` license) or less restrictive (e.g. `CC Attribution-ShareAlike` or `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, will therefore be suitable for use by both academia and industry.

We underscore the importance of sharing—perhaps even its principal argument—is that 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 also would need 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 new advancements, rather than with data that only serves 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 in facilitating synergy within the research community that is otherwise difficult to attain.

III. The YFCC100M Dataset

We created the Yahoo Flickr Creative Commons 100 Million Dataset[1] (YFCC100M) as part of the Yahoo Webscope program. This dataset is the largest public multimedia collection that has ever been released, containing a total of 100 million media objects, of which approximately 99.2 million are photos

[1]The dataset is available for general (academic or commercial) use at http://bit.ly/yfcc100md.

and 0.8 million are videos, all of which have been uploaded to Flickr between 2004 and 2014 and published under a CC commercial or non-commercial license. Our dataset overcomes many of the issues affecting datasets such as the aforementioned popular ones, for instance in terms of *modalities*, *metadata*, *timespan*, *licensing* and principally *volume*; we discuss the strengths and limitations of our collection with respect to existing datasets in more detail below. As is the case with many datasets, the YFCC100M is constantly evolving, and over time we will release various expansion packs[2] containing data not yet present in the collection; 100 million can broadly address many challenges. For instance, several visual and aural features extracted from the videos have already been made available, and we are working towards releasing the EXIF metadata of the photos and videos as an additional expansion pack. The actual photos and videos are further currently being uploaded to the cloud to ensure the dataset will remain accessible and intact for years to come.

A. Metadata

Each media object included in the dataset is represented by its metadata in the form its Flickr identifier, the user that created it, the camera that took it, the time at which it was taken, the location where it was taken (if available), and the CC license it was published under. In addition, the title, user tags, machine tags, and description are also available, as well as direct links to its page and its content on Flickr. Social features, comments, favorites, and followers/following data are not included in the dataset as, by their nature, these change on a day to day basis. They can, however, be easily obtained by querying the Flickr API[3], and over time we will make snapshots of such social features available for download.

Tags. There are 68,552,616 photos and 418,507 videos in the set that have been annotated with user tags (or keywords). Figure 1 shows the frequency distribution of the top 100 user tags for photos and for videos. A total of 3,343,487 photos and 7,281 videos carry machine tags (labels that have been automatically added by a camera, computer, application, or some other automated system).

Cameras. Table I shows that the top 25 cameras used in the dataset are overwhelmingly digital single lens reflex (DSLR) models with the exception of the Apple iPhone. Considering that the most popular cameras in the Flickr community at the moment primarily consist of various iPhone models[4], this bias in our data is likely due to CC licenses attracting a certain subcommunity 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% has been assigned the most liberal license that only requires the photographer that took the photo to be attributed, see Table II.

[2]The dataset expansion packs will be made available at http://bit.ly/yfcc100mf
[3]https://www.flickr.com/services/api/
[4]https://www.flickr.com/cameras/

TABLE I
Top 25 cameras and photo counts in the YFCC100M dataset. We have merged the entries for the Canon models that have different names in the European (e.g. EOS 650D), American (e.g. EOS Rebel T4i) and Asian (e.g. EOS Kiss X6i) markets.

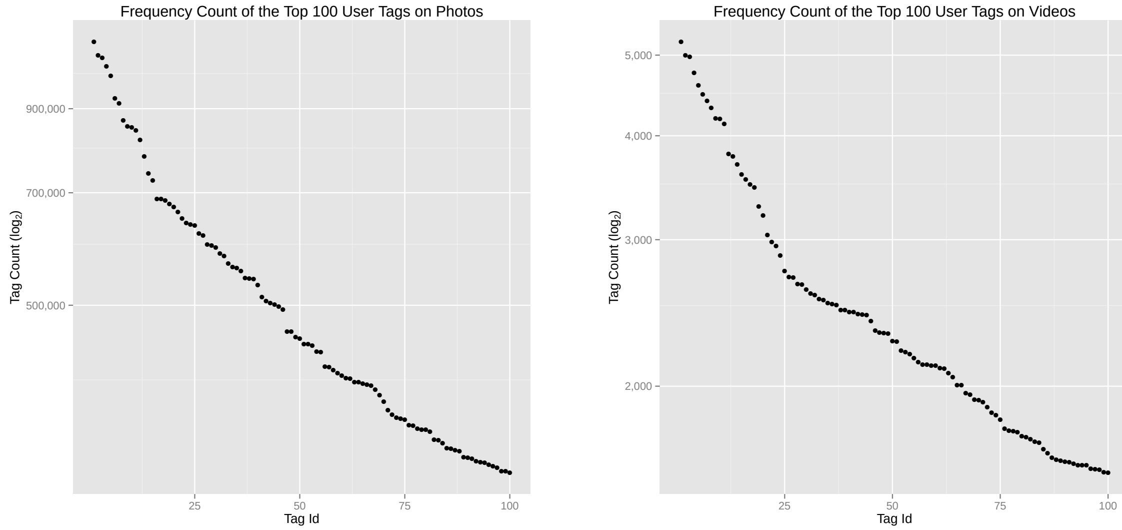
Make	*Camera*	*Count*
Canon	EOS 400D	2,539,571
Canon	EOS 350D	2,140,722
Nikon	D90	1,998,637
Canon	EOS 5D Mark II	1,896,219
Nikon	D80	1,719,045
Canon	EOS 7D	1,526,158
Canon	EOS 450D	1,509,334
Nikon	D40	1,358,791
Canon	EOS 40D	1,334,891
Canon	EOS 550D	1,175,229
Nikon	D7000	1,068,591
Nikon	D300	1,053,745
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

TABLE II
A breakdown of the 100 million photos and videos by their kind of 🅭 CreativeCommons license, as 🅯 By Attribution, ⊜ No Derivatives, 🄎 Share Alike, and 🄏 Non-Commercial.

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

B. Photos

To understand more about the visual content of the photos in the dataset, we used a deep learning approach to find the presence of a variety of concepts, such as people, animals, objects, food, events, architecture, and scenery. Specifically, we apply a deep convolutional neural network [17] with 7 hidden layers, 5 convolutional layers and 2 fully connected ones. The penultimate layer of the convolutional neural network output is employed as the image feature representation to train the concept classifiers. We trained 1000 visual concept classifiers and show the top 25 detected concepts in Table III. Here we see a diverse collection of visual concepts from indoor to outdoor images, faces to food, nature to automobiles. Beyond these concepts, the YFCC100M has a collection of complex real world images: from over 200,000 street life-blogged photos by photographer Andy Nystrom, Figure 2(a), to José-María Moreno García's photographs from life in central Spain, Figure 2(b).



(a) Distribution of user tags on photos.

(b) Distribution of user tags on videos.

Fig. 1. The frequency count of the top 100 user tags in the YFCC100M dataset on photos (a) and videos (b). The similar distributions indicate the user tag annotations of video content (0.8% of the overall corpus) match that of the photographic content.

(a) *IMG_9793: Streetcar (Toronto Transit)* by Andy Nystrom https://flic.kr/p/jciMdz

(b) *Celebrando El Tricampeonato De Europa De Futbol 2012 En Madridejos* by José-María Moreno García https://flic.kr/p/cpiC83

Fig. 2. Two photos of real world scenes from photographers in the YFCC100M dataset.

C. Videos

Flickr makes little distinction between a photo and a video; however, videos do play a role both on Flickr and in this dataset. While photos encode their content primarily through visual means, videos also convey this through audio and motion. The average length of a video in our dataset is 39 seconds with a tendency for recent video clips to be longer. This is likely due to the initial handling of videos on Flickr, where their length was initially restricted to a maximum of 90 seconds. Most of the videos in the dataset, 95.6% in total, have titles. This high percentage includes a large number of automatically generated ones such as *IMG_0520* or *MVI_0001*, which are responsible for 21% of the video titles. Beside titles, the vocabulary of approximately 280,000 distinct user tags used for the video annotations make for a rich and diverse set of entities related to humans (baby, family), animals (cat, dog), locations (park, beach), travel/vacation (nature, city), audio (concert, festival), and motion (timelapse, dance), to name just the top frequent few. When we compare the metadata of videos in the dataset to videos from YouTube taken in the time period 2007–2012, we observe that they have a similar distribution. We do see that YouTube clips have a higher video length and on average twice as many tags as Flickr videos.

D. Features and Annotations

Computing features for 100 million media items is a time consuming and computationally expensive task. From our experience with organizing benchmarks on image annotation and location estimation we noticed that accompanying the datasets with precomputed features reduced the burden on the participants, allowing them to focus on solving the task at hand rather than processing the data. We are currently computing

TABLE III
THE TOP 25 OF 1000 VISUALLY DETECTED CONCEPTS IN THE YFCC100M DATASET.

Concept	*Count*
outdoor	32,968,167
indoor	12,522,140
face	8,462,783
people	8,462,783
building	4,714,916
animal	3,515,971
nature	3,281,513
landscape	3,080,696
tree	2,885,045
sports	2,817,425
architecture	2,539,511
plant	2,533,575
house	2,258,396
groupshot	2,249,707
vehicle	2,064,329
water	2,040,048
mountain	2,017,749
automobile	1,351,444
car	1,340,751
food	1,218,207
concert	1,174,346
flower	1,164,607
game	1,110,219
text	1,105,763
night	1,105,296

a wide variety of visual, aural, textual and motion features for the dataset. These features, being computed descriptors of the images, will be licensed without any restrictions under the CC0 (🅭🄍) license. The visual features span the gamut of global (e.g. Gist), local (e.g. SIFT) and texture (e.g. Gabor) descriptors, the aural features include power spectrum (e.g. MFCC) and frequency (e.g. Kaldi) descriptors, the textual features refer to closed captions extracted from the videos, and the motion features include dense trajectories and motion boundaries. Certainly, the utility of the dataset will grow as more features and annotations are produced and shared, whether by ourselves or by others.

E. Strengths and Limitations

The YFCC100M dataset differs in its design from previous multimedia collections. The collection of photos, videos, and metadata in our dataset has been curated to be comprehensive and representative of real world photography, expansive and expandable in coverage, free and legal to use, and as such intends to consolidate and supplant many of the existing datasets currently in use. The vast majority of available datasets include media whose licenses do not allow them to be used without explicit permission from the rights holder. While fair use exceptions may be invoked depending on the nature of use, these will generally not be 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 will not only violate the ImageNet license agreement, but also very likely copyright law. Our dataset overcomes this issue providing rules on how the dataset should be used to comply with licensing, attribution and copyright.

Our 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 community. For example, the MediaEval Placing Task [27], 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. Other venues in which our dataset features are the ACM Multimedia 2015 Grand Challenge on Event Detection and Summarization, as well as the ACM Multimedia 2015 MMCommons workshop; the latter kicks off the development of a research community around annotating all 100 million photos and videos. This notwithstanding, our collection at present 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 noisy annotations, although through our efforts this will improve over time. We see this as more of a challenge than a limitation. Real world datasets require sense-making through structured annotation. At the scale of 100 million there are enough metadata labels for training and prediction of some features, i.e. geographic coordinates, and there exists opportunities to create new methods for labeling and annotation through explicit crowd-sourced work or through more modern techniques using social community behaviors. In addition, given the plurality of existing Flickr datasets and the size of our dataset, some overlaps are to be expected, where any existing annotations directly transfer over to our dataset. Of particular note is the Microsoft COCO [10] of which a third is present (about 100,000 images) in YFCC100M. Over time we will release the intersections with known datasets as expansion packs. We envision that the intersection with existing datasets allows researchers to expand upon what is known and actively researched.

IV. NEW CHALLENGES

Our dataset provides opportunities for large scale unsupervised learning, semi-supervised learning and learning with noisy data. Beyond this, YFCC100M offers new research questions to be explored in social computing, geography, and digital culture.

AI and Vision

Large scale datasets have played a critical role in advancing computer vision research. ImageNet [9] itself led the way into the development of advanced machine learning techniques like deep learning [17]. The 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 for this effort by developing new approaches that harness more than just the pixels, such as geolocation and user tags, to advance the visual understanding. 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 have been hampered by the lack of a sufficiently large dataset.

However, visual recognition goes beyond image classification and involves obtaining a deeper understanding of images. Object localization, object interaction, face recognition, 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 was this image taken. Flickr, having a rich diverse collection of image types, provides the groundwork for total scene understanding [18] in computer vision and artificial intelligence, a crucial task that can now be expanded upon with our dataset.

Spatiotemporal Computing

Beyond the pixels, there is metadata, both automatically populated and UGC. Of this, we emphasize time and location as key components for research that aims at understanding and summarizing events. Starting with location, the geotagged photos and videos, about half of the dataset, provide a comprehensive snapshot of photographic activity over space and time. In the past, geotagged media has been used to address a variety of research endeavors, such as location estimation [19], event detection [20], finding canonical views of places [21], and visual reconstruction of the world [22]. Even understanding styles and habits have been used to reverse lookup authors online [23]. Geotagged media in the YFCC100M dataset can help push the boundaries in these research areas even further.

More data brings new discoveries and insights, yet it simultaneously also 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 therefore devising algorithms that are able to automatically and dynamically create albums for use on personal computers, cloud storage or mobile devices, where desired media and moments of importance can be easily surfaced based on simple queries. Harnessing the spatiotemporal context at capture time and at query time will take a leading role in successfully addressing this challenge.

The previous challenge speaks towards social computing efforts aimed at understanding events in unstructured data through the reification of the photo with space and time. While GPS-enabled devices are capable of embedding the precise time, location and orientation of capture in the metadata of a photo, in many instances this information is unavailable or out of date: seconds, hours or sometimes even months. In addition, people frequently forget to adjust the camera clock to the correct timezone when traveling. These kinds of issues pose problems for the accuracy of any kind of spatiotemporal data analysis, and new challenges in computational photography therefore include devising algorithms that can either fix or are resilient against the erroneous.

Digital Culture and Preservation

What we know to be UGC 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 [24] that speak to how people engage with the spaces around them to 500 years of scanned book photos and illustrations [25] that describe how concepts and objects have been visually depicted over time. Beyond archived collections, the photostreams of individuals represents many facets of recorded visual information, from remembering moments and storytelling to social communication and self-identity [26].

This presents a grand challenge of sensemaking and understanding digital archives from non-homogeneous sources. Photographers and curators alike have contributed to the larger collection of Creative Commons images, yet little is known on how such archives will be navigated and retrieved, or how new information can be discovered therein. This offers avenues of computer science research in multimedia, information retrieval, and data visualization in addition to the larger questions of digital libraries and preservation.

Reproducing and Comparing Results

The availability of data is a necessary condition for the exact reproduction of scientific results. However, in the computer science literature it is uncommon for a paper to describe in enough detail how the dataset they used in their evaluations was created, effectively preventing others from faithfully replicating or comparing against the achieved results. One clear challenge for the future is ensuring that (subsets of) datasets used in experiments can be accurately reproduced. Addressing this challenge, we note that 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. We suggest researchers to forego arbitrary selections from our dataset when forming a subset for use in their evaluations, but rather to use a principled approach that can be succinctly described. Such selection logic should examine one or both of the following two aspects of the dataset, namely that (i) the photos and videos in the dataset are already randomized, and (ii) the dataset consists of 10 consecutively numbered files. As such, a selection logic could be as simple as

> "We used the videos in the first four files for training, those in the following four files for development, and those in the last three for testing."

or more complicated as

> "From all photos taken in the United States, we selected the first 5 million and performed 10-fold cross-validation."

Alternatively, the created subset can be made available for download described as a set of object identifiers that index into the YFCC100M dataset. As an example, the organizers of the MediaEval 2014 Placing Task made, in addition to the training and test sets, the visual and aural features they extracted from the content available for download. We envision the research community to also follow this way of using and sharing the dataset.

V. Conclusions

Data is one of the core components of research and development. In the field of multimedia, datasets are usually created with a single purpose in mind, and as such lack reusability. Additionally, datasets generally are not or may not be freely shared with others, and as such lack reproducibility, transparency and accountability. To overcome this conundrum, we have released one of the largest datasets ever created, consisting of 100 million media objects published under a Creative Commons license. To summarize, our dataset has been curated to be comprehensive and representative of real world photography, expansive and expandable in coverage, free and legal to use, and intends to consolidate and supplant existing collections. We have highlighted its use for the next generation of computer vision, human mobility, and social computing research. Our dataset encourages the improvement and validation of research methods, reduces the effort of acquiring 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.

Acknowledgements

We would like to thank Jordan Gimbel and Kim Capps-Tanaka at Yahoo, Pierre Garrigues, Simon Osindero, and rest of the Flickr Vision & Search team, Carmen Carrano and Roger Pearce at LLNL, as well as Julia Bernd, Jaeyoung Choi, Luke Gottlieb and Adam Janin at ICSI. Portions of this work were performed under the auspices of the U.S. Department of Energy by LLNL under Contract DE-AC52-07NA27344, and supported by the National Science Foundation by ICSI under Award Number 1251276.

References

[1] Facebook, Ericsson, and Qualcomm, "A focus on efficiency," Facebook Inc., Tech. Rep., 2013.

[2] "Youtube press statistics," http://youtube.com/yt/press/statistics.html, 2014, accessed March, 2015.

[3] J. Good, "How many photos have ever been taken?" https://web.archive.org/web/20150203215607/http://blog.1000memories.com/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox, September 2011, accessed March, 2015.

[4] A. H. Renear, S. Sacchi, and K. M. Wickett, "Definitions of dataset in the scientific and technical literature," in *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*, 2010, pp. 1–4.

[5] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 39–43.

[6] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, 2010, pp. 527–536.

[7] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes challenge – a retrospective," *International Journal of Computer Vision*, 2014.

[9] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] T. Lin, M. Maire, S. Belongie, J. Hays, P. Pietro, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," *arXiv.org*, 2014.

[11] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva, "SUN database: exploring a large collection of scene categories," *International Journal of Computer Vision*, pp. 1573–1405, 2014.

[12] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, 2006, pp. 321–330.

[13] S. E. Fienberg, M. E. Martin, M. L. Straf *et al.*, *Sharing research data*. National Academy Press, 1985.

[14] A. Swan and S. Brown, "To share or not to share: publication and quality assurance of research data outputs," Research Information Network, Tech. Rep., 2008.

[15] C. L. Borgman, "The conundrum of sharing research data," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 6, pp. 1059–1078, 2012.

[16] M. L. Wilson, E. H. Chi, S. Reeves, and D. Coyle, "RepliCHI: The Workshop II," in *Proceedings of the International Conference on Human Factors in Computing Systems, Extended Abstracts*, 2014, pp. 33–36.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[18] L. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: classification, annotation and segmentation in an automatic framework," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2036–2043.

[19] J. Hays and A. A. Efros, "IM2GPS: estimating geographic information from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[20] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," in *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, 2007, pp. 103–110.

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

[22] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3D," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 835–846, 2006.

[23] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "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*, 2011, pp. 237–246.

[24] "Yelp dataset challenge," http://yelp.com/dataset_challenge/, August 2014, accessed March, 2015.

[25] K. Kremerskothen, "Welcome the internet archive to the commons," https://blog.flickr.net/2014/08/29/welcome-the-internet-archive-to-the-commons/, August 2014, accessed March, 2015.

[26] J. Van Dijck, "Digital photography: communication, identity, memory," *Visual Communication*, vol. 7, no. 1, pp. 57–76, 2008.

[27] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland, "The placing task: A large-scale geo-estimation challenge for social-media videos and images," in *Proceedings of the ACM International Workshop on Geotagging and Its Applications in Multimedia*, 2014.