

What's with the Free Images?

A Study of Flickr's Creative Commons Attribution Images

Herkko Hietanen
Helsinki Institute for Information
Technology, HIIT
PO BOX 19215
00076, Aalto, Finland
herkko.hietanen@hiit.fi

Kumaripaba Athukorala
Helsinki Institute for Information
Technology, HIIT
PO BOX 19215
00076, Aalto, Finland
kumaripaba.athukorala@hiit.fi

Antti Salovaara
Helsinki Institute for Information
Technology, HIIT
PO BOX 19215
00076, Aalto, Finland
antti.salovaara@hiit.fi

ABSTRACT

Our survey of the Flickr photo site's Creative Commons attribution licensed images reveals that there is a wide variety of high quality of relevant stock images available. However, searching the images can be demanding since the image metadata is inconsistent. The main problem of finding open images is that the search tools are mostly based on user generated tags. The search results would benefit from human sorting and simple machine vision analysis. These steps might be able to close the gap between commercial stock photo and open image collections.

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Storage and Retrieval – *information search and retrieval.*

General Terms

Management & Human Factors.

Keywords

Flickr, Creative Commons, image search, ImageNet, stock photo, metadata, tagging.

1. INTRODUCTION

In the not-too-distant past, image searching was a time consuming and costly exercise. Now there are stock photos online on services such as Getty and Corbis Images. Online microstock photo companies have opened the stock photo markets to a new class of amateur and semiprofessional photographers. The increased supply of stock images has reduced their price considerably [8] [1]. The commercial services are no longer the only sources for quality stock images. There are tens of millions of images online which are licensed free of charge with open content licenses.

The falling price of the stock photos combined with the growth of open content images raises questions: Is the zero-price the result of the process of commodification of stock images? Can the open images compete with the commercial stock photo services?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MindTrek'11, September 28-30, 2011, Tampere, Finland.

Copyright 2011 ACM 978-1-4503-0816-8/11/09...\$10.00.

While the license price of open images is zero, there are transactional costs to finding the right images. These transaction costs could increase the opportunity costs of open image collections and eventually make the paid stock photo services the less expensive alternative. In this research we examine whether the Creative Commons (CC) images are accurately retrievable and how their quality compares to commercial stock images. We do this to determine how much work needs to be invested to improve the open image repositories to the level of professional stock photo sites.

2. THE STATE OF THE ART

There exist several commercial photo collections which sell professional quality images by professional photographers. The images are also screened by paid editors before publishing. However, the services have slightly different approaches to selecting the images they sell.

Magnum Photos is a photographic cooperative owned by its photographer-members. It has a long tradition of photo journalism. Magnum has an online archive of over 500,000 photos. Photographers retain their copyright to photographs which Magnum licenses. Magnum recently started using crowdsourcing methods to create image tags. Their crowdsourcing tests have shown that workers optimally assign between four and eight keywords per image --- any more resulted in duplication [10].

The online image trade is dominated by relatively new companies like Getty and Corbis Images. Getty has grown aggressively to become one of the biggest image services by acquiring traditional photo agencies and archives, by digitizing their collections and enabling easy online distribution. While much of Getty's income derives from professional photography, Getty bought one of the biggest microstock photo providers, *iStockphoto*, in 2007. Unlike Magnum, *iStockphoto* has very liberal entry requirements for its contributors. At *iStockphoto* any photographer can choose to upload their images to the Web site. The catalogue of over 6 million files is crowdsourced [4] mainly by amateurs and semiprofessionals. *iStockphoto* acts as an agency by selling the images to clients. Individual photographers make small profits per download, while *iStockphoto* takes a portion of the profits. Many of the contributors are motivated by money [1]. In 2011 *iStock* expects to pay over 2 million dollars in royalties per week to contributors [9].

Amateurs have uploaded nearly 6 billion images to online photo sharing service *Flickr*. Over 190 million of the images are licensed with one of the Creative Commons (CC) licenses. Each of the six CC licenses has a different set of restrictions. The most

restrictive license permits only unmodified use of images for non-commercial usage. The most permissive CC-Attribution (CC-BY) license requires licensees merely to attribute the author. Previously photographers have used the more restrictive CC-licenses, but Flickr’s data shows that recently users have started to prefer more permissive licenses [7]. In 2011, over 25 million of the CC images are licensed with CC-BY licenses.

Flickr provides three easy ways to filter search the Creative Commons licensed images: by using the options in the Advanced Search, by browsing through recent images uploaded under each type of licence or through the Flickr API which allows other applications to access the images and their metadata freely.

Flickr provides a way to sort the search results according to their relevance and interestingness. Flickr determines the interestingness by determining where the clickthroughs are coming from, who comments on the images and who marks them as favorites and by examining image tags, description, title and several other factors which are constantly changing [2].

It is well known that image labels originating from user tags have a number of problems. Previous research shows that amateur photographers typically optimize their image sharing for browsing and not for searching [6]. Their tags often provide an incomplete description of the visual content, focusing mainly on the interest of the photographer while leaving out many “plain”, yet visible, objects. At the same time, and for the same reason, a large proportion of tags may refer to information not directly visible in the image.

There are several initiatives to improve search results which are based on textual user created links. ImageNet [3] is a repository of over 3 million images built upon the backbone of the WordNet ontology. The images are quality controlled and human annotated through crowdsourcing methods. ImageNet is designed for the purpose of supporting object recognition, image classification and automatic object clustering applications. This design goal makes the collection less interesting for regular users for two reasons: 1) Although, ImageNet has clean and full resolution images, most of them are “all rights reserved” images and thus not freely usable. 2) While ImageNet has over 80,000 noun word sets, it does not contain verbs or adjectives.

MIRFLICKR is another image retrieval test data set. It is a subset containing 25,000 of Flickr’s CC-BY licensed images with their original tags and EXIF image metadata. MIRFLICKR has also been developed for training of content-based image retrieval systems [5]. Unlike ImageNet, MIRFLICKR’s collection is not built on any ontology and it has only ten basic categories tagged and verified by experts.

3. METHOD AND RESULTS

Our two tests collected information about the quality of open images. Our automated script made nearly 150,000 searches to the Flickr’s CC-BY licensed images. The goal of this experiment was to collect information about the metadata that these open images contain. In the second experiment we conducted a study where participants rated the quality of the search results of different collections.

3.1 Retrieval test

In the first step of this research we built a dataset of images and matching metadata. The approach we followed was similar to that

of ImageNet. Unlike ImageNet, we did not restrict our searches to nouns only. We used WordNet’s 147,306 term ontology as a list of search terms to retrieve CC-BY licensed images from Flickr. Our importer script retrieved hundred images that Flickr determined to be the most relevant [1] for each search term and their metadata. We gave the retrieved images points according to the order in which they showed up in the search results. We gave the first image 99 points and the second 98 points. This way the first 99 images received a *relevancy score*. The importer also retrieved all the tag words that the photographer had associated with the image. Those tags were not issued any relevancy score unless they were among the first hundred search results for that given word. Our database is a collection of images with both weighted WordNet terms and additional Flickr tags.

147,306 searches retrieved the total of 5,671,643 images. Thirty parallel importer instances retrieved one million images per week. The searches returned an average of 39 images for each Wordnet keyword searched. Several images showed up multiple times in the search results. Because of this, only 2,866,612 of the retrieved images were unique. The retrieved images had 28,200,124 tags and there were 1,170,349 different terms. The average image had 9.8 tags while the median was 7. Only a few images had more than a thousand tags each. The highest number of tags used in an image was 1886. Originally the image had only 29 tags, but it had an 8894-word essay attached into the image description field. Because Flickr’s searches are not limited only to the tags but analyzed also the image name and description fields, this image showed up in several searches.

Figure 1 shows the number of images against the number of tags for images. Many of the images were poorly tagged, with only 30.6% of the images having more than 10 tags. Given that a fairly high percentage of these tags are uninformative (e.g., meaningful only to the photographers), the number of truly useful tags is low. For example the most used tag was “2010” which was associated with 53,806 images. There were 6,386 tags that had more than 500 images associated to them.

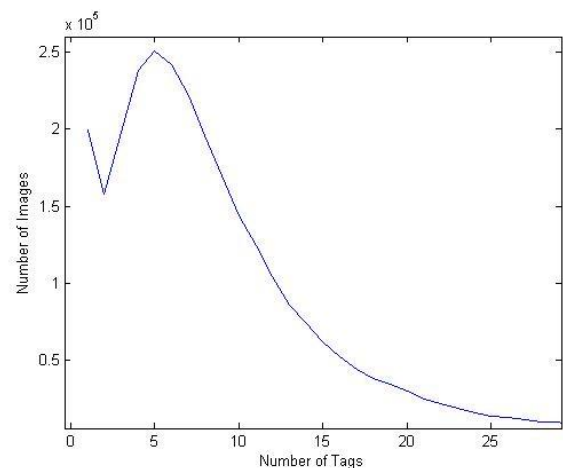


Figure 1. Number of Tags per Image

3.2 User study

In order to evaluate the retrieval results from Flickr, we carried out a small-scale user study. Ten participants were asked to review results from three image services. We asked the

participants to mark suitable and mismatching images, compare the search results and sort them in order.

Our user study ($N = 10$) examine the quality of search results in 3 image services: iStockphoto, Flickr, and Flickr-CC-By. We inspected Flickr's collection of the week's most interesting images, and used them to determine a list of 10 keywords (sunset, woman, flower, sea, bird, child, city, tree, car, and landscape). For these keywords, we retrieved the first 100 search results from each of the three services. This resulted in a dataset comprising $3 \times 10 \times 100 = 3000$ images which we printed in full color on thirty A3-sized papers, each printout containing the search results from one service and a keyword. These papers showed only the images and the search term – we had removed all the extraneous graphics and texts from the search results. To hide the origins of the images from the participants, we also made sure that the papers did not reveal the name of the service used.

We asked our participants to imagine a situation in which they needed to prepare a PowerPoint presentation with images that match our list of keywords. We asked them to inspect the printouts and to perform two tasks: 1) select five images that they would use in a power point presentation and 2) strike out the images that did not match the keyword. We advised the participants to strike out the images that had poor image quality, or poor match with the given keyword.

In the first part of the user study, with each participant having rated the retrieved images for all 3 services and 10 keywords, we could compare the answers given by each participant individually and then combine these analyses into the overall result. Therefore, we ran a repeated-measures ANOVA with 3×10 within-subjects design, with the number of strike-outs as the dependent variable. The results showed a difference between the results for different keywords ($p < .05$). This meant that there was a statistically significant difference between search results for different keywords: the retrieved images were better rated on some keywords than others. When comparing the results on the keyword level, we found no general factor that all good-quality or bad-quality search results would have shared in common. For instance, "landscape" and "sunset" images got generally speaking few strike-outs, but "sea", which is another keyword producing scenery, received high numbers of strike-outs.

A statistically stronger and more interesting result was that there was also a difference in retrieval quality between the services ($p < .01$). Of the 100 images, the numbers of images struck out for iStockphoto, Flickr and Flickr-CC-By were 10.4, 16.4, and 18.3, respectively. Only the difference between iStockphoto and the other two services was statistically significant; no difference was found between Flickr and Flickr-CC-By. In other words, iStockphoto was found the best of the three services, and Flickr and Flickr-CC-By were found equally good.

The second part of the study contained a more general rating task. For each keyword, we gave the participants the three clean 100-image printouts (with no strike-outs drawn over images), one for each service, and asked them to sort the printouts in order of increasing retrieval quality. Participants could rate the printouts as "equally good", having "slight difference" or having "clear difference". These differences were turned into scores for each service. The lowest-rated service always scored 0, the next service was scored either 1 or 2, depending on the level of difference, and similarly with the third service, when compared to the second one. Each service was therefore given a score between 0 and 4. For

each keyword, each service was finally given another score that told the difference of its 0...4 score to the average of the scores for that keyword. We used the same analysis as in the first part of the study. The difference scores were used as dependent variables in a 3×10 within-subjects repeated-measures ANOVA. The result from this overall analysis was the same as from the first part of the study: iStockphoto was significantly better than the Flickr services, and again there was no statistical difference between the Flickr services.

While evaluating the results in the second part of the study, we asked the participants to tell the reasons for their ordering. Common reasons for low ratings were that search results were too homogenous. For instance, 99/100 of the images retrieved by iStockphoto with the keyword "woman" portrayed women of a Caucasian origin, and a big percentage of the "city" images were skylines typical of American cities. Flickr-CC-By suffered from results coming from the same content provider and the same image set, with similar results. Another commonly stated reason was the keyword being in a secondary role in the retrieved results. iStockphoto's "flower" pictures mostly pictured flowers in female subjects' hands or hair, for instance. The third common reason for poor rating was the presence of unnecessary content in the images. Flickr and Flickr-CC BY suffered often from this, because their images were mostly unpolished photographs. There were also several duplicate images in both Flickr and Flickr-CC-By image sets.

4. CONCLUSIONS

The fact that our 147,306 searches to over 25 million CC-By images returned only under 3 million images suggests that most of the images cannot be easily found through common searches. It means that only 12 percent of the CC-By licensed images are easily findable.

One thing that could affect the results is the language. We used only English words. Flickr is currently available in ten languages and many of its users may have used non-English tags, rendering the images inaccessible to us. The other more likely explanation is that even though people choose to share their images with permissive licenses they don't bother to tag the images well enough for them to be found.

Our tests did show that the professional iStockphoto service provides considerably better search results in nearly all of our test searches. However, the difference between 3 million CC-By licensed images and the 6 billion general Flickr images suggest that the problem is not that there are not enough good quality open images available, but that the image search needs refinement. CC-By licensed pictures did not do worse than the All rights reserved pictures in our user evaluation. Therefore, a free image service is a realistic goal. On the other hand, there was a clear difference in search results between Flickr image sets and that of iStockphoto. Efforts should therefore be directed to improve search results and filtering in Flickr image searches. Simple selection of the images either by experts or by employing crowdsourcing methods could help to close this gap.

The feedback from our participants also suggests that prioritizing lighter colored images in image retrieval could improve the appeal of the search results. We believe that future work should therefore concentrate on creating models and incentives to harness crowds and machine vision to improve the quality of search results. Over time such improvements could help to attract a critical mass of

people to open image collection users. Such crowds could lead to emergent network effects that would unlock the value of open photos.

It is still too early to say whether the open content image collections are the final stage of the commodification of stock photos. However, the race to the bottom is fierce, and stock photo companies need to innovate in order to compete with the open image sources.

5. REFERENCES

- [1] Brabham, D., Moving the crowd at iStockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application, *First Monday* [Online], Volume 13 Number 6 (21 May 2008), available at <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2159/1969>
- [2] Butterfield, D., Fake, C., Henderson-Begg, J., Mourachov, S., (inventors) Interestingness ranking of media objects, Assigned to Yahoo! Inc. US Patent Application 20060242139 Published October 26, 2006.
- [3] Deng, J., Dong, W., Socher, R., Li L.-J., Li, K., Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition, 2009*, 248–255.
- [4] Howe, J., The rise of crowdsourcing, *Wired*, volume 14, number 6 (June), available at <http://www.wired.com/wired/archive/14.06/crowds.html>.
- [5] Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceeding of the 1st ACM international conference on Multimedia information retrieval (MIR '08)*. ACM, New York, NY, USA, 39-43. DOI=10.1145/1460096.1460104 <http://doi.acm.org/10.1145/1460096.1460104>
- [6] David Kirk, Abigail Sellen, Carsten Rother, and Ken Wood. 2006. Understanding photowork. In *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI '06)*, Rebecca Grinter, Thomas Rodden, Paul Aoki, Ed Cutrell, Robin Jeffries, and Gary Olson (Eds.). ACM, New York, NY, USA, 761-770. DOI=10.1145/1124772.1124885 <http://doi.acm.org/10.1145/1124772.1124885>
- [7] Linksvayer, M., Creative Commons licenses on Flickr: many more images, slightly more freedom. Creative Commons news post (2010), available at www.creativecommons.org/weblog/entry/20870.
- [8] Taub, E., When Are Photos Like Penny Stocks? When They Sell. *New York Times*, June 5, 2007, available at <http://www.nytimes.com/2007/06/05/technology/circuits/05syndicate.html?em&ex=1181188800&en=687225a44f80273c&ei=5087%0A>.
- [9] Thompson, K., Royalty Change Follow up, iStockPhoto forum, Sep 8, 2010, http://www.istockphoto.com/forum_messages.php?threadid=252322.
- [10] Wolmurth, P., Magnum Photo's tagging game, *British Journal of Photography*, 17 Feb 2011, available at <http://www.bjp-online.com/british-journal-of-photography/report/2027044/magnum-photos-tagging-game>.