The MIR Flickr Retrieval Evaluation

Mark J. Huiskes Michael S. Lew
LIACS Media Lab, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
{markh, mlew}@liacs.nl

ABSTRACT

In most well known image retrieval test sets, the imagery typically cannot be freely distributed or is not representative of a large community of users. In this paper we present a collection for the MIR community comprising 25000 images from the Flickr website which are redistributable for research purposes and represent a real community of users both in the image content and image tags. We have extracted the tags and EXIF image metadata, and also make all of these publicly available. In addition we discuss several challenges for benchmarking retrieval and classification methods.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection, Dissemination, Standards*.
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query Formulation*.

General Terms

Experimentation, Human Factors, Measurement, Performance, Standardization

Keywords

Content-based image retrieval, relevance feedback, image collections, benchmarking

1. INTRODUCTION

Arguably, the most frequently used test set in content-based image retrieval [2] is the Corel Stock Photography collection [5]. The total collection consists of more than 800 Photo CDs, each containing 100 broadly similar images of a certain category. In most cases, research groups have made their own selection from the available categories, usually amounting to a varying subset consisting of 3,000 to 10,000 images.

New test sets [1-10] for the image retrieval community are important for benchmarking, finding weaknesses in systems, and determining the significance of novel algorithms. The ideal test set should have the following requirements.

First, the test set should be *representative* of an interesting image retrieval area. In the past, researchers often would come up with their own ad hoc test sets, perhaps even from their personal imagery collections, which then obviates representativeness in that we really require many different sample points, together covering the entire spectrum of the imagery sources. Ideally, thousands of individuals should contribute to the test set, and the set should be sufficiently large to be representative of the whole.

Second, *ground truth* should be available for the test set so that objective evaluations can be performed. Many recent benchmarking initiatives (e.g. in the high-level feature extraction and search tasks of TRECVid (e.g. [7]), and the IAPR TC12 ([8]) benchmark), rely on search topic ground truth that has been obtained through *pooling*. In this approach (e.g. [11]), only images are annotated that appear in the top N of most relevant images in the ranking of at least one approach participating in the benchmark. Pooling reduces the cost of annotation considerably, but it leaves large parts of collections unlabeled, thereby hindering accurate measurement of precision and recall. Also note that when descriptive keywords and text descriptions are available, these generally do not provide a sufficiently exhaustive content characterization to allow for direct use as ground truth; the Flickr tags discussed below are certainly a good example to this effect. For accurate evaluation, ground truth should be available for the entire test set. As argued in [10] this is also particularly important for the performance evaluation of retrieval systems employing relevance feedback methods, since there ground truth is also needed for realistic simulation of the relevance feedback.

Third, the test set really should be easily *accessible* and freely *redistributable*. No copyright forms should be required, and any researcher should be able to legally distribute the test set. In the case of the Corel collection, it is no longer sold so it certainly is not easily accessible. The MPEG7 test collection was available to the scientific world for a few years, but now it is nearly impossible to find and cannot be freely redistributed. It is important to note that the accessibility is also essential for reviewing: the reviewer should be able to access the collection if he feels it is relevant to his decision. A single paper may show 10 or 20 samples, but that often is insufficient to give a good overview of the collection. For example, the St. Andrews collection originally used in the ImageCLEF evaluation is only available legally for researchers who have officially registered with ImageCLEF. Other researchers or reviewers cannot view it. Another interesting test set is the Corbis database but it is also not legally redistributable.

Fourth, we think that it is important to have a set of *standardized tests* associated with the test database. In current literature, it

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-312-9/08/10...$5.00.

frequently happens that different researchers will perform different performance tests on the same database which can make it impossible to perform comparative benchmarking. The set of standardized tests should at least include a varied collection of challenging search topics, with accurate ground truth as discussed above, as well as detailed guidelines on uniform performance measurement and reporting.

In this paper, we present the MIR Flickr test set which is designed to address the four main requirements: representative of an area; accurate ground truth; freely redistributable; and standardized tests.

2. THE MIR FLICKR SET

This new image collection consists of 25000 images that were downloaded from the social photography site Flickr.com through its public API. The color images are representative of a generic domain and are of high quality. This is guaranteed by the high "interestingness"[1] of the images: this image score represents an evolving measure of quality determined by factors such as where clickthroughs on the image are coming from, who comments on it and when, or who marks the image as a favorite.

Figure 1 shows a number of example images from the collection.

Figure 1. Examples from the image collection. Also listed are Creative Common attribution license icons and the creators of the images.

2.1 Copyright and Licenses

Most images on Flickr are not copyright-free and are published with all right reserved. However, a considerable number of images have been offered under a Creative Commons license[2]. The image collection presented here consists only of images with one of the Creative Commons *attribution* licenses. Licenses of this type allow for image use as long as the photographer is credited for the original creation. Possibly, the use is granted under additional restrictions, which may include: only non-commercial use allowed, not allowed to create derivative works (i.e. the image can only be distributed in its original state), and distribution of derivative works allowed only under a "Share Alike" condition (i.e. with a license identical to the original). None of these restrictions precludes the use of these images for benchmarking purposes.

When downloading the images care was taken to collect as much information as possible on the creator of the image. The creator information together with the exact license type and image title is collected in a license metafile associated with each image. Additionally, the creators will be acknowledged on the MIR Flickr website.

2.2 Collection Process

The selective downloading of images by license type is made possible through the Flickr API[3]. The 25000 images were collected for original upload dates from March 21, 2007 to June 30, 2008, i.e. covering a time period of approximately 15 months. The 25000 images of the collection were supplied by a total of 9862 Flickr users; 5566 users are represented in the set by a single image. The user[4] with the highest number of images in the set has a contribution of 41 images.

As mentioned, the images were downloaded by querying for images with high interestingness: for each date in the period given above, the 500 most interesting images uploaded at that date were requested, from which then only the images with an attribution license were actually downloaded. The Flickr API calls were made from our Perl script (slurpr.pl), which additionally allows us to collect owner information, the image tags and EXIF metadata.

The Flickr tags and EXIF metadata are discussed in more detail in Sections 3 and 4, respectively.

The images and associated metadata can be downloaded at http://press.liacs.nl/mirflickr/. The first version of the image collection is available as a single zip file (mirflickr08.zip, approx. 3GB, MD5: A23D0A8564EE84CDA5622A6C2F94778500).

3. FLICKR TAGS

One of the great attractions of Flickr is the platform it offers its users to search and share their pictures based on image tags. We offer the tags of the images in two forms: the raw form in which they are obtained from the users and a processed form where the raw tags have been cleaned by Flickr. This process includes for instance removing capitalization, spaces and various special characters. The average number of tags per image is 8.94. In the collection there are 1386 tags which occur in at least 20 images.

[1] http://flickr.com/explore/interesting/

[2] http://creativecommons.org/

[3] http://flickr.com/services/api/

[4] Trey Ratcliff: thanks!

Although most tags are in English, some foreign terms occur as well.

Image Tag	Frequency
sky	845
water	641
portrait	623
night	621
nature	596
sunset	585
clouds	558
flower/flowers	510/351
beach	407
landscape	385
street	383
dog	372
architecture	354
graffiti/streetart	335/184
tree/trees	331/245
people	330
city/urban	308/247
sea	301
sun	290
girl	262
snow	256
food	225
bird	218
sign	214
car	212
lake	199
building	188
river	175
baby	167
animal	164

Table 1: Frequency of tags in the MIR Flickr set

The tags can be subdivided in various categories. The most useful tags for research purposes are most likely those that clearly describe the images, preferably with a direct relation to the visual content of the image (e.g. snow, sunset, building, party).

Table 1 shows the most common content-based tags of this type. In this table we have left out simple colors (by decreasing frequency: black-and-white, blue, red, green, white, yellow, black, pink and orange), seasons, and locations. The latter may or may not have a clear visual relation to the content, but they offer proof that the image collection contains images taken at many interesting locations around the world. The most common tags in this category are California, New York, London, Japan, Italy, USA and Canada.

Sometimes tags refer to more abstract concepts (e.g. love, travel) or adjectives (e.g. old, cute, vintage), but these are rare among the most common tags. Less useful but very common categories of tags refer to Flickr-related terminology (e.g. explore, interestingness, abigfave, anawesomeshot, naturesfinest, diamondclassphotographer) and camera brands and types (nikon, canon, d40).

4. EXIF METADATA

Most modern digital cameras embed metadata on camera type, camera settings, time and date and, in some cases, geolocation information in the image by means of EXIF[5] metadata tags. An overview of commonly used tags is given in Table 2. When the owners of pictures have allowed it (which is mostly the case), we have collected these tags, and any other tags that were supplied.

A number of recent papers have investigated the usefulness of this metadata for image classification and retrieval, e.g. [12] and [13]. These show that there is a definite promise of improved performance by taking into account the "optical context" in which a picture was taken.

Some of the Flickr image tags refer to similar information, e.g. the already mentioned tags for camera brands and types, as well as tags such as macro, bokeh (i.e. the appearance of out-of-focus areas), photoshop and geotagged.

Camera Settings and Sensor Readings		
Aperture/Fnumber	Shutterspeed/ExposureTime	FocalLength
ExposureProgram	ExposureBias	MaxApertureValue
MeteringMode		
Camera Image Settings		
Orientation	Compression	x-Resolution
y-Resolution	ResolutionUnit	PixelXDimension
PixelYDimension		
Camera/Software Information		
Manufacturer	Model	Software

Table 2: Common EXIF tags

5. GROUND TRUTH

In Section 6 we will propose a number of standardized challenges for the MIR Flickr collection. These will include both topic classification tasks as well as challenges to propagate tags from the small group of images already labeled with a tag to the entire image collection. For both types of challenges we can consider

[5] http://exif.org

the tag as a search topic, and must thus assess the relevance of tags to images. As indicated in the introduction, we aim to facilitate accurate performance evaluation by providing ground truth *for the entire test set*. Following the guidelines proposed in [10], we additionally prefer to collect several complete interpretations on a topic, rather than one pooled interpretation. This means that annotators always provide their relevance assessment of a topic on the entire collection, making the resulting ground truth more consistent in the sense that it does not mix different interpretations. This is particularly important for the evaluation of image retrieval systems employing relevance feedback, but also for classic recognition tasks such repeated full set annotations provide useful statistics on the reliability of the annotations and can also serve to gain more insight in the robustness of the classifications.

We have also asked the annotators to annotate each tag-based topic in two ways. First, by interpreting relevance to the topic in a *wide sense*: as soon as a tag is at least somewhat relevant to the content of the image, he should label the image as relevant in this sense. Second, by interpreting the topic in a *narrow sense*: now images are tagged only if the tag concept is, according to the annotator's own subjective interpretation, saliently present in, or applicable to, the content of the image. In practice the difference between the two types of relevant image sets may correspond quite closely to the images assessed as "partially relevant" according to the scheme used in [1]: "consisting of images that are in some way relevant, but for which the annotator is not confident enough to label them as fully relevant".

Given these goals, choosing a particular order for the tag annotations can considerably relieve the total required effort. We start by annotating the most general topics, and do so first in the wide sense described above. The general topics were chosen in such a way that (i) they mostly correspond to common Flickr tags themselves, and (ii) they either contain some additional common tags as subtopics, or they are expected to be useful to this end in the future. The sunset tag may be an exception, which we mainly we found useful as a topic supplementary to the night tag. The general topics and corresponding subtopics selected for our first annotation effort are listed in Table 2. Only the general topics marked by an asterisk are not common tags themselves, but they are deemed sufficiently useful for current and future subtopic annotation to justify their inclusion.

By restricting subsequent annotation of subtopics and narrow sense interpretations to the labeled sets resulting from the wide sense interpretations, we can greatly reduce the number of images that need to be considered. With a sufficient number of annotators for the initial wide sense interpretations of the general topics, we expect that the union of their relevant sets will effectively contain all potential candidates for the subsequent more narrow interpretations. Restricting the search for relevant images to this set thus significantly reduces cost while still allowing us to meet all the objectives above. Finally we note that even though the listed general topics represent only an initial choice to achieve a feasible annotation effort, they already cover a wide range of interest and can contribute to reducing annotation cost for many future subtopic annotations.

General topics	Subtopics
sky	clouds
water	sea/ocean, river, lake
people	portrait, boy/man, girl/woman, baby
night	
plant life*	tree, flower
animals	dog, bird
man-built structures*	architecture, building, house, city/urban, bridge, road/street
sunset	
indoor	
transport*	car

Table 3: Topics and subtopic selected for full annotation

6. STANDARDIZED CHALLENGES

We define these standardized challenges so that researchers can compare results. In the first two challenges, the task is essentially the traditional pattern recognition goal of visual concept detection where the assumption is that the image tags are highly correlated with the constituent visual concepts within the image. In the third challenge, we investigate a very wide range of tags and focus on ranked tagging, where the researcher builds a system which recommends tags with a ranked order. The goal is to maximize the correlation between the ranked tags from the automatic system and the tags from the Flickr users.

Standardized Challenge #1: Visual Concept/Topic Recognition

Based on the supplied ground truth annotations, train classifiers for each of the general topics and subtopics listed in Table 3. Report performance on the test set by means of precision-recall graphs and average precision for each topic. To this end the total image collection of 25000 images is split into a collection of 15000 training images and 10000 test images. To avoid bias, we divide every five images: the first three are assigned as training images, the last two as test images. Depending on the goal of the classification approach, researchers may focus on performance for the wide sense or narrow sense annotations.

Standardized Challenge #2: Tag Propagation

This constitutes a very challenging task of extending tag annotation to the entire collection of 25,000 images by using only the Flickr tags as a training set. One of the main questions to answer is whether tag propagation based on such small training sets is feasible. Using the supplied ground truth for the tags listed in Table 3, show the results in precision-recall and accuracy vs. rank graphs.

Standardized Challenge #3. Tag Suggestion

Consider images that possess at least one of the tags listed in Table 1. For each image, supply a relevance ranking of these tags. Given the current absence of complete ground truth for this task, use half of the images containing a tag as a training set and the other half as a test set (by taking every second image as a test image). Again the main question is if tag suggestion of this kind is feasible based on the limited number of samples. Measure the frequency at which at least one of the top 5 tags as recommended by the automatic system matches at least one of the Flickr tags.

Further details on the challenges will be made available on http://press.liacs.nl/mirflickr.

7. CONCLUSION AND PROSPECTS

There are a variety of content-based image retrieval benchmarking initiatives worldwide. We have not yet discovered one which addresses all four main requirements: representative of an area; accurate and complete ground truth; freely redistributable; and standardized tests. In this project, we hope to provide a new test set which will overcome the limitations of previous test sets such as the ubiquitous Corel set, and be timely for researchers in the MIR community.

In particular we offer a collection that can be downloaded and redistributed free of charge and without any registration. All that is asked is to respect the Creative Commons licenses and to keep the creator information alongside the images. In future work we may provide additional data sets following a similar collection strategy, e.g. to obtain additional training images for specific tags.

Additionally, we aim to extend the list of tags for which ground truth annotations are provided. A final prospect is to set up a search task specifically for benchmarking relevance feedback methods.

8. ACKNOWLEDGEMENTS

Leiden University and NWO BSIK/BRICKS supported this research under grant #642.066.603

9. REFERENCES

[1] P. CLOUGH, H. MUELLER, AND M. SANDERSON. 2004. The CLEF cross-language image retrieval track (imageCLEF) 2004. In *Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, LNCS, New York, NY, USA, ACM Press.

[2] M. S. LEW, N. SEBE, C. DJERABA, AND R. JAIN. 2006. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19.

[3] S. MARCHAND-MAILLET. 2005. The Benchathlon Network. http://www.benchathlon.net.

[4] S. MARCHAND-MAILLET AND M. WORRING. 2006. Benchmarking image and video retrieval: an overview. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 297–300, New York, NY, USA. ACM Press.

[5] H. MÜLLER, S. MARCHAND-MAILLET, AND T. PUN. 2002. The truth about Corel - evaluation in image retrieval. In: *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 38–49, London, UK. Springer-Verlag.

[6] X. ZHOU AND T. HUANG. 2003. Relevance feedback in image retrieval: a comprehensive review. *ACM Multimedia Systems Journal,* 8(6):536–544, April.

[7] A. F. SMEATON, P. OVER, AND W. KRAAIJ. 2006. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval,* pages 321–330, New York, NY, USA. ACM Press.

[8] C. H. C. LEUNG AND H. H.-S. IP. 2000. Benchmarking for content-based visual information search. In *VISUAL'00: Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pages 442–456, London, UK. Springer-Verlag.

[9] C. FLUHR, P.-A. MOELLIC, AND P. HEDE. 2006. Usage oriented multimedia information retrieval technological evaluation. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 301–306, New York, NY, USA. ACM Press.

[10] M. J. HUISKES AND M. S. LEW (2008). ``Performance Evaluation of Relevance Feedback Methods", ACM International Conference on Content-Based Image and Video Retrieval (CIVR'08), pages 239-248, Niagara Falls, Canada.

[11] K. SPARCK JONES AND C. VAN RIJSBERGEN (1975). Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.

[12] P. SINHA AND R. JAIN (2008). Classification and annotation of digital photos using optical context data. ACM International Conference on Content-Based Image and Video Retrieval (CIVR '08), pages 309-318, Niagara Falls, Canada.

[13] J. YEN, P. WU, AND D. TRETTER (2007) Knowledge discovery for better photographs, Proc. SPIE 6506, 65060B.