

Hierarchical Annotation of Large Image Collections

Mark J. Huiskes

Leiden Institute of Advanced Computer Science

Leiden University

markh@liacs.nl

Michael S. Lew

Leiden Institute of Advanced Computer Science

Leiden University

mlew@liacs.nl

Abstract

This paper describes the procedure we have followed to manually annotate the MIRFLICKR collection. It can be characterized as a stepwise refinement of annotations along two, orthogonal, hierarchies. After a substantial initial investment, the hierarchical structure allows for fast generation of new queries for the evaluation of various tasks in image retrieval. The annotation structure is particularly geared to generating realistic queries for evaluation of relevance feedback (RF) systems. We present our first results obtained with the annotations, both for building classifiers for automatic annotation, and for testing RF-based retrieval methods.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Human Factors, Performance, Experimentation, Standardization

Keywords

Benchmarking, Annotation, Relevance Feedback, Image Collections

1 Introduction

In this paper we describe the procedure we have followed to manually annotate the MIRFLICKR-25k ([4]) collection. The hierarchical structure of the procedure allows for fast generation of new queries for the evaluation of various tasks in image retrieval. A specific goal has been to facilitate the generation of queries for the evaluation of retrieval systems based on relevance feedback. In the following, we first introduce the MIRFLICKR collection. Next, in section 2, we discuss the hierarchical annotation procedure. Finally, in section 3, we present our first testing results obtained with the new annotations.

The MIRFLICKR image collection consists of 25000 images downloaded from the social photography site Flickr.com, through its public API. All images are made available under Creative Commons licenses, allowing for image use as long as the photographer is credited for the original creation. In some cases, use is granted under additional restrictions, but none of these preclude the use of the images for benchmarking purposes.

The MIRFLICKR database supplies all original tag data supplied by the Flickr users; in the collection there are 1386 tags which occur in at least 20 images, with an average total number of 8.94 tags per image. Table 1 lists the most common tags corresponding to concrete visual concepts (colors, seasons and place names were left out).

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

Table 1: Most common Flickr tags corresponding to concrete visual concepts.

When available, also EXIF (exchangeable image file format) metadata are supplied: its fields represent a number of properties and settings of the digital camera at the time of taking a picture. This includes information on the camera (e.g. brand, manufacturer), camera settings (exposure, aperture, focal length, ISO speed etc), image settings (orientation, resolution, compression), time and data, and increasingly also, geolocation. EXIF data are available for about 85% of the MIRFLICKR images. Table 2 below lists a number of common fields.

EXIF Field	Field Possession (%)
Aperture	95
Exposure Time/Shutter Speed	96
Focal Length	95
ISO Speed	79
Flash	75

Table 2: Common EXIF fields. Field possession indicates the fraction of images that have a value for the EXIF field given that EXIF data is available for the image.

The collection and metadata(tags, EXIF, annotations) are described in detail in [4] and can be downloaded at <http://press.liacs.nl/mirflickr>.

2 Hierarchical Annotation

This paper describes the procedure we have followed to annotate the MIRFLICKR collection. The goal of this procedure has been to obtain annotations that correspond to ground truth for realistic search queries, in particular for testing image retrieval systems.

In [5] we have analyzed the requirements for ground truth annotations suitable for such testing, specifically for testing retrieval systems that employ relevance feedback. For testing such systems the ground truth is not only used to evaluate retrieval accuracy, but also for performing realistic simulations of the relevance feedback interactions. For this reason, one important requirement for the annotation process is that the relevance to a given query is decided by single assessors. However, given the effort required, this requirement is hard to fulfill for large image collections.

The procedure proposed in this paper makes it possible to generate consistent and realistic queries at greatly reduced cost. In the following, we will refer to the image set that needs to be considered for the annotation of a topic as the *annotation set*. The decrease in annotation effort is mainly realized by reducing the size of an annotation set, while making sure to retain all potentially relevant images in the set. This is achieved by building a hierarchical structure of annotation sets, refining the sets along two dimensions:

1. Abstraction level: from general to specific categories

The first hierarchy consists of a regular semantic concept hierarchy, branching from general into more specific categories. To reduce annotation cost for subtopics we use the parent topic as annotation set. This is made possible by the orthogonal, relevance, hierarchy.

2. Relevance level: from wide to narrow topic interpretation

Moving down the second hierarchy we proceed from a very wide interpretation of topic relevance, where images that are even weakly relevant to the topic are already assigned with the topic label, to a more narrow and subjective interpretation of relevance to the topic. Note that given these characterizations, images relevant in this latter, stronger, sense are by necessity also relevant in the weaker sense. The hierarchy resulting from different annotators providing their interpretations at different levels of relevance can then serve a dual purpose: to allow for annotation down the semantic hierarchy, and to further refine the annotation sets. This is discussed in further detail below.

The top level topics used for the annotation of MIRFLICKR, listed in Table 3, were chosen to cover many interesting topics as proper subtopics. They also have a large overlap with the most common Flickr tags (see Table 1).

General topic	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: First annotations supplied for the MIRFLICKR-25k database.

2.1 Staged Top-Down Annotation

The concrete annotation process is divided into two main stages. First, we need a relatively costly stage, in which all concepts are interpreted in the wide sense described above. This stage proceeds top-down the semantic hierarchy and is performed by an initial group of annotators.

Second, follows the stage of performing annotations corresponding to ground truth for actual topic queries, in which many individual assessors can provide their subjective interpretation of the topics.

In the first stage all images are identified which annotators in the second stage may reasonably find relevant to the topic, or its subtopics. For this reason we refer to these annotations as *potential* labels. The topic, usually a concept that can be visually identified, does not need to appear prominently: it is sufficient when it is visible or applicable at least to some extent. In this way the potential labels act as a greatest common denominator for the concept, allowing the resulting images to serve as annotation set for both the individual subjective interpretations of the topic, and for the annotation of concepts deeper in the hierarchy. For our purpose of creating ground truth for testing queries, this annotation stage is sufficiently objective to require only a single main annotation round. However, preferably one or more additional rounds would repeat the effort to correct for oversights and errors of the original annotators.

In the second stage we first proceed by letting individual annotators provide their interpretation of the main topics of the hierarchy, considering only images with the corresponding potential label. Interpretations may range from quite wide to very narrow and specific. A useful approach is to first interpret the topic in a general sense, selecting only images in which the topic is considered to be saliently present. Subsequently, ground truth for a large number of additional queries can be generated by choosing more specific subtopics, e.g. for the “sea” topic, we might consider “tropical sea”, “sea at sunset”, “sea only” as additional subtopics. Especially the annotation of these latter specific annotations can be obtained with only little effort.

2.2 Annotation Interface

Annotation at different levels in the hierarchical topic structure can benefit from interaction strategies that take into account the specific requirements at these levels. The annotation of the potential labels, particularly at the higher semantic levels when all images still need to be considered, requires a careful and detailed consideration of the image. To avoid doubling effort in repeatedly reaching a sufficient understanding of the images, at this stage images are best considered one at a time while assigning several labels simultaneously. This can be realized by various interface designs, e.g. by means of extensive annotation tools for annotating an entire hierarchy at once (e.g. [7]). Given the size of the MIRFLICKR collection, we have aimed to minimize the effort required for the tagging actions. Since our concept hierarchy is relatively small, we chose a configurable keyboard-based approach where a tag (or group of tags) is assigned by pressing the corresponding key.

At later stages, where context is more uniform as a result of smaller annotation sets, it is often faster to consider several images in parallel. This is particularly the case for annotations corresponding to very specific queries where the topic visually “pops out”. It may then be feasible to process a grid of small images at a time, clicking only the relevant images. An example of such case is given in the figure below, where images of “tropical sea” need to be labeled given an annotation set of “sea” images. In the intermediate stages, the interaction strategies can be tailored to the topic characteristics by combining these main modes of interaction in various ways, e.g. by using a grid of images and still assigning several labels at a time (using keys to switch active tags). It also pays off to adapt the size and number of images displayed to the detail of understanding required for accurate labeling.

Our tagging script is available at <http://github.com/huiskes/tag.pl>. Additional details on the various annotation mechanisms, e.g. on autoforwarding images, can be found in the manual.

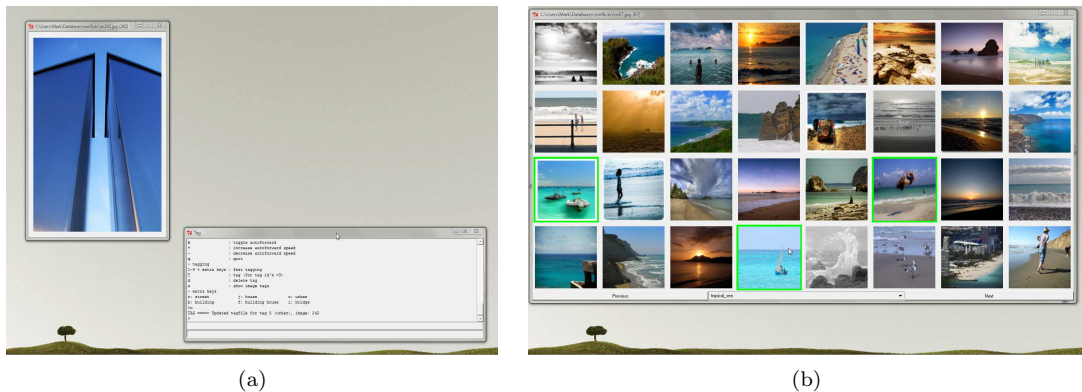


Figure 1: tag.pl (a) single image view mode, (b) grid view mode

3 Classification and Relevance Feedback

The MIRFLICKR collection was developed with the specific goal of offering a better evaluation benchmark for RF systems, following [5]. We present our first results of building classifiers for the annotation categories and testing different RF methods. We also discuss how features based on classifiers trained on the potential, wide interpretation, labels can greatly benefit relevance feedback results for queries consisting of more narrow concept interpretations.

3.1 Image Representation

For the image classification and relevance feedback tests we combined the following five sets of image features:

1. **HMMD Color Histogram descriptor.** The nonuniform quantization of the HMMD color space is similar to that of the MPEG-7 color structure descriptor (CSD, [6]). Based on the difference variable the color space is divided into five subspaces; for each subspace a customized number of hue and sum levels are selected. The hue quantization was tailored to make the hue bins correspond closely to main color names. The histogram is extended by grouping elementary bins by hue, difference (similar to saturation) and sum (similar to intensity).
2. **Spatial Color Mode descriptor.** Based on the same HMMD color space quantization as the previous descriptor, this histogram describes the spatial occurrence of dominant colors. The dominant colors are determined by counting only pixels which colors occupy at least 30% of 15x15 pixel structuring elements. The spatial occurrence is measured by splitting each bin based on 3 horizontal and 3 vertical image sections. Again the histogram is extended by lumping bins over the color dimensions and image sections.
3. **MPEG-7 Edge Histogram descriptor (EHD, [6]).** This descriptor captures the spatial distribution and orientation of edges by grouping of local edge direction histograms.
4. **MPEG-7 Homogeneous Texture descriptor (HTD, [6]).** The features are obtained by first filtering images with a bank of orientation and scale sensitive filters, and computing the mean and standard deviation of the filtered outputs in the frequency domain. An extended histogram is obtained by summing over orientations and scales. In our experiments the standard deviation features for the individual outputs were not used.
5. **Flickr tags.** A set consisting of 293 binary features indicating Flickr tags of visual concepts. Each selected tag is associated with at least 50 images in the MIRFLICKR collection.

This adds up to a total of 2341 features per image.

3.2 Classification

We have trained classifiers based on the annotations of 23 potential labels, and 17 subjective queries. Table 4 lists the mean average precision (MAP) and precision@50 of the classifiers for the potential labels based on two classification methods. Table 4 (a) corresponds to classification by linear discriminant analysis. The results are obtained for a test set of 10000 images after determining discriminant directions based on a training set of 15000 images. Table 4 (b) lists results obtained by SVM classification (C-SVM) using LIBSVM ([1]). Values for the C (cost) and γ (RBF) parameters were selected using subsets of 1000 images of the 15000 image training set. The testing results are again for a test set of 10000 images, disjoint with the training and validation sets.

(a)								
	sky	water	night	people	plant life	animals	structures	sunset
MAP	80.8	56.7	63.0	73.4	69.8	52.3	71.5	52.3
prec@50	98	90	92	78	92	90	86	81
	indoor	transport	clouds	sea	river	lake	portrait	male
MAP	66.8	40.8	65.2	46.6	33.4	26.1	54.7	43.0
prec@50	66	72	90	84	80	60	80	61
	female	baby	tree	flower	dog	bird	car	
MAP	50.3	32.1	52.0	55.8	63.0	39.2	27.8	
prec@50	68	70	84	94	98	92	68	
(b)								
	sky	water	night	people	plant life	animals	structures	sunset
MAP	82.6	58.5	63.4	71.9	76.2	35.2	69.0	63.8
prec@50	100	100	98	100	100	80	90	96
	indoor	transport	clouds	sea	river	lake	portrait	male
MAP	61.8	40.0	67.3	37.9	17.0	17.9	53.3	40.2
prec@50	84	90	100	68	32	24	90	54
	female	baby	tree	flower	dog	bird	car	
MAP	52.2	21.9	55.1	55.9	57.5	16.6	28.2	
prec@50	96	42	92	96	94	40	62	

Table 4: Classification results (a) LDA (b) SVM

As can be observed, the two classification methods often provide quite similar results. The linear discriminant classifiers, however, have a great advantage in both ease of computation, with no parameter selection required, and time and storage required for classification. Classification by LDA amounts to projection on a single discriminant direction vector, whereas the SVMs require a much more expensive combination of, generally many, support vectors.

3.3 Relevance Feedback

Figure 2 shows precision-recall (PR) curves obtained by two RF methods for 4 annotated topics, representing ground truth corresponding to subjective interpretation of the topic by a single annotator (see section 2). The methods shown are aspect-based relevance learning (ARL, [2, 3]) and SVM-based RF (using an RBF kernel function). The precision-recall curves are averages over 50 runs per method per class. In each run, 5 random positive examples were taken from the target class. For the SVM method also 10 negative examples were randomly selected. Results are shown

for two feature sets. The first set consists of the features described above in section 3.1. The results for this set are, despite the Flickr tag features, labeled as “low-level”. Using a second set, we consider how the performance of the RF approach may be improved if more high-level information is available. The information is obtained by using the classification results of the previous section. Specifically, we use the LDA classifiers obtained for the annotations resulting from a wide interpretation of the corresponding topics. The LDA classifiers were chosen over their SVM counterparts as, given their very simple underlying model, they are expected to suffer less from memorizing the training data. The added features consisted of classifier output for thresholds corresponding to a number of recall levels.

It is interesting to observe that the availability of classifiers for wide topic interpretations greatly enhances the performance of RF methods on more specific topic interpretations.

Also shown are the PR-curves corresponding to topic classifiers with best known performance (MAP) when trained on a full training set instead of a small set of examples. The curve is labeled as *best known classifier* (BKC).

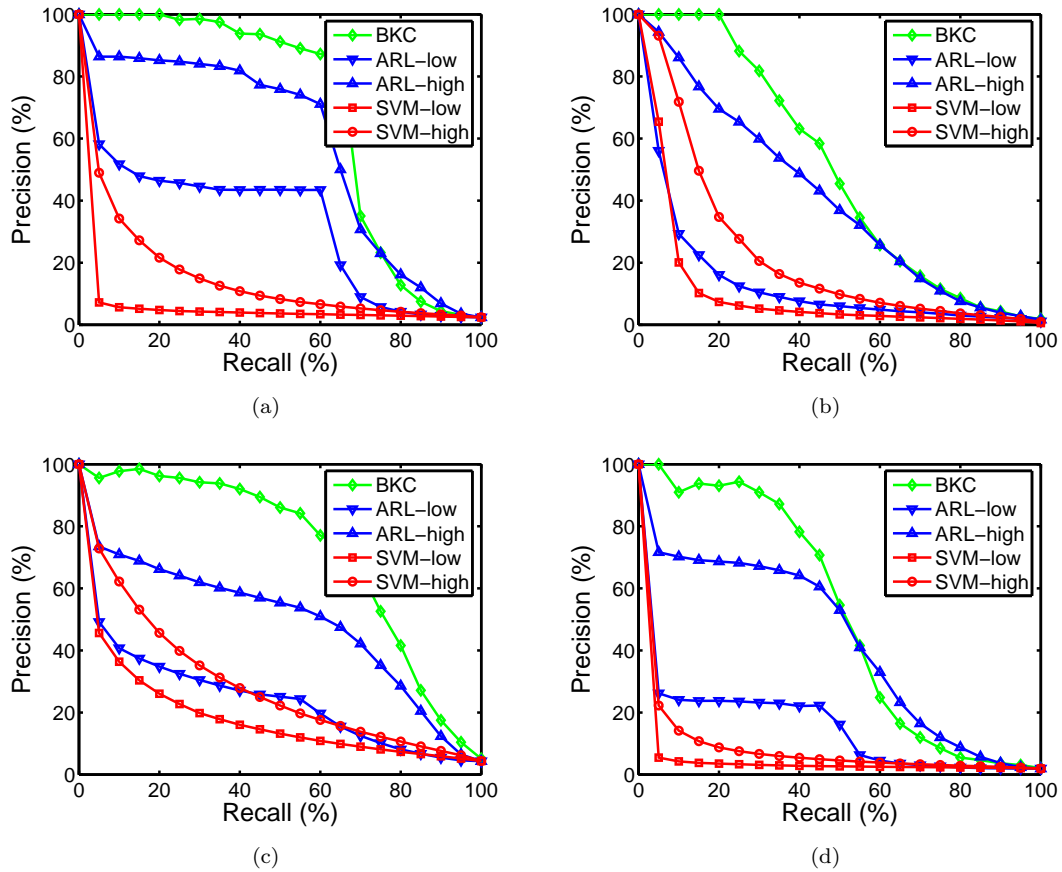


Figure 2: RF performance for 4 topics (a) dog, (b) dog on grass, (c) bird, (d) flower

In [5] it is recommended to normalize PR-curves by the best known classifier curves before averaging results over different topics. This factors out the main effect of image representation quality on the performance measures, providing a more meaningful average for cross-study comparisons. In future work we will provide a detailed comparison of RF methods following this approach.

4 Conclusion

In this paper we have outlined the procedure we have used to annotate the MIRFLICKR-25k collection. The main advantage of the approach is that, although a substantial investment by the initial annotators is required, annotation cost for subsequent topics and annotators is greatly reduced. This makes it feasible to generate consistent ground truth for a large number of realistic queries.

We presented a number of preliminary results obtained with the new annotations. In building topic classifiers, we found that linear discriminant analysis often strikes a better balance in accuracy and ease of computation than classification by SVMs. We also demonstrated the approach of [5] for factoring out representation quality in the evaluation of RF-based retrieval systems. From the evaluation results, it is interesting to observe that the availability of classifiers for wide topic interpretations greatly enhances the performance of RF methods on more specific topic interpretations.

Acknowledgements

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

References

- [1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [2] Mark J. Huiskes. Aspect-based relevance learning for image retrieval. In W.K. Leow, editor, *Proceedings of CIVR05, LNCS 3568*, pages 639–649. Springer, 2005.
- [3] Mark J. Huiskes. Image searching and browsing by active aspect-based relevance learning. In *Proceedings of CIVR06, LNCS 4071*, pages 211–220. Springer, 2006.
- [4] Mark J. Huiskes and Michael S. Lew. The MIR FLICKR retrieval evaluation. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, New York, NY, USA, 2008. ACM.
- [5] Mark J. Huiskes and Michael S. Lew. Performance evaluation of relevance feedback methods. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 239–248, New York, NY, USA, 2008. ACM.
- [6] B. S. Manjunath, Jens-Rainer Ohm, Vinod V. Vasudevan, and Akio Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11:703–715, 1998.
- [7] A. Th. (Guus) Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74, 2001.