New Trends and Ideas in Visual Concept Detection

The MIR Flickr Retrieval Evaluation Initiative

Mark J. Huiskes
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA Leiden
The Netherlands
markh@liacs.nl

Bart Thomee
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA Leiden
The Netherlands
bthomee@liacs.nl

Michael S. Lew
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA Leiden
The Netherlands
mlew@liacs.nl

ABSTRACT

The MIR Flickr collection consists of 25000 high-quality photographic images of thousands of Flickr users, made available under the Creative Commons license. The database includes all the original user tags and EXIF metadata. Additionally, detailed and accurate annotations are provided for topics corresponding to the most prominent visual concepts in the user tag data. The rich metadata allow for a wide variety of image retrieval benchmarking scenarios.

In this paper, we provide an overview of the various strategies that were devised for automatic visual concept detection using the MIR Flickr collection. In particular we discuss results from various experiments in combining social data and low-level content-based descriptors to improve the accuracy of visual concept classifiers. Additionally, we present retrieval results obtained by relevance feedback methods, demonstrating (i) how their performance can be enhanced using features based on visual concept classifiers, and (ii) how their performance, based on small samples, can be measured relative to their large sample classifier counterparts.

Additionally, we identify a number of promising trends and ideas in visual concept detection. Based on these trends and ideas, we formulate new directions for the MIR Flickr Retrieval Evaluation, resulting in two new initiatives to extend the original image collection. First, the collection will be extended to one million Creative Commons Flickr images. Second, a number of state-of-the-art content-based descriptors will be made available for the entire collection.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing Methods*

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search Process, Relevance Feedback, Query Formulation*.

General Terms

Algorithms, Experimentation, Human Factors, Measurement, Performance, Standardization.

Keywords

Visual concept detection, classification, trends, content-based image retrieval, image collections, benchmarking.

1. INTRODUCTION

The MIR Flickr collection [14] consists of 25000 high-quality photographic images of thousands of Flickr users, made available under the Creative Commons license. The database includes all the original user tags and EXIF metadata. Additionally, detailed and accurate annotations are provided for topics corresponding to the most prominent visual concepts in the user tag data. The rich metadata allow for a wide variety of image retrieval benchmarking scenarios.

In this paper we discuss new trends and ideas in visual concept detection as well as their relation to the MIR Flickr Retrieval Evaluation initiative. In Section 2, we first discuss a number of state-of-the-art algorithms and approaches to visual concept detection. This includes a discussion of the 2009 ImageCLEF large-scale visual concept detection task, which used the MIR Flickr image collection. Next, we present an overview of promising trends and ideas for improving the performance of visual concept classifiers.

In Section 3 we discuss results from a number of experiments in combining social data and low-level content-based descriptors to improve the accuracy of visual concept classifiers. In particular, we show how the use of Flickr user tags can improve performance, and analyze how classifier accuracy differs for different types of annotations. Next, in Section 4, we present results obtained by a number of relevance feedback methods, demonstrating (i) how their performance can be enhanced using features based on visual concept classifiers, and (ii) how their performance, based on small samples, can be measured relative to their large sample classifier counterparts.

Finally, in Section 5, we announce two extensions to the current MIR Flickr collection. First, the collection will be extended to one million Flickr images, again under Creative Commons licenses. Second, a number of state-of-the-art content-based descriptors will be made available for the entire collection.

2. NEW TRENDS AND IDEAS

2.1 State-of-the-Art

The MIR Flickr collection served as the main benchmark collection in the large-scale visual concept detection and annotation task (LS-VCDT) in ImageCLEF 2009. In total 18000 images of the collection were used (5000 for training, 13000 for testing). For this subset, annotations for 53 concepts were obtained. Altogether 19 research groups participated and submitted 73 runs [22].

The team with the best results (ISIS, University of Amsterdam) achieved an average AUC of 84% on their best run [25]. Their approach first determines a codebook of visual words corresponding to SIFT feature clusters in different color spaces. The feature sampling is based on a combination of a spatial pyramid approach and salient point detection, and concept classifiers are obtained by taking the word frequencies as input for an SVM classifier with χ^2 kernel.

Overall, results on the LS-VCDT task confirm that currently approaches based on visual words obtained by clustering of SIFT-like features achieve the best results for most visual topics. Combinations of local and global features can provide comparable performance results, see for instance [34].

The visual words approach (e.g. [28, 36]) has several possible variations: in feature sampling approaches (dense sampling schemes or interest points), in choice of descriptors (e.g. SIFT, SURF) and color spaces, in how to build the vocabulary (K-means, supervised and unsupervised tree-based methods, e.g. [20]), how to assign features to visual words (e.g. by soft assignment, [8]), in the choice of classification method, and in the case of SVMs, in the choice of kernel; [32] presents an interesting overview of trade-offs that can be made between accuracy and speed of computation.

In the following we discuss a number of trends and ideas with the potential to bring us beyond current level of performance.

2.2 Massive Data

Today's near-limitless availability of imagery and image metadata (see also Section 2.3) represents both a challenge and an opportunity. The challenge lies in keeping up with the ever-growing size of our multimedia collections, particularly in offering sufficiently accurate high-level metadata to make these collections searchable. Dealing with very large collections and large numbers of visual concepts has several computational implications, for instance with respect to feasible indexing structures to cluster and access the images; see [2] for a recent discussion. To be able to keep features of a large number of images in memory, we need effective small-size descriptors. In [31] a hierarchical representation scheme is presented that can greatly reduce the size of descriptors, while retaining good discrimination performance; [12] demonstrates the application of a similar size-reduction scheme for visual word-type of features. Part of the solution may also lie in various forms of cloud computing technology that is making it increasingly practical and feasible to scale up computations and is now quickly gaining ground. Interesting opportunities in this direction are offered by open-source projects such as Hadoop MapReduce [4] and HBase (modeled after Bigtable, [6]).

On the other hand, large-scale data can also directly *benefit* visual concept detection. The idea is that rather than designing more intelligent classification algorithms and image representations, we simply use more data. As outlined in for instance [30] the billions of images freely available on the internet provide us with a dense sampling of the visual world. Using this data even very simple classification algorithms can achieve decent classification accuracy.

For benchmarking purposes it is important that, although data and various types of metadata are now easy to come by (see again Section 2.3), *reliable* annotations are still hard and expensive to produce. For the MIR Flickr evaluation we aim to strike a balance in this situation by offering both a reasonable size core set that is annotated in great detail, and a much larger collection of additional imagery with automatically supplied metadata (tags, visual descriptors, EXIF etc.). The extension to the original core set is discussed in Section 5.

Yet another sense in which massive data hold great promise for visual concept detection is in the exploitation of *continuous* data streams. Using any video streams, it is now easy to feed learning methods with one or several continuous streams of visual data. The time-varying views on the visual concepts, offered by such streams, may turn out to be a crucial step forward compared to learning from static images. Some further possibilities in this direction are discussed in Section 2.4.

2.3 Social Data Analysis

On social networks (e.g. Facebook) and content-sharing websites (e.g. Flickr and YouTube), image and video content are often accompanied by various forms of metadata like tags, ratings, comments, EXIF data, as well as with information about the uploaders and their social network. These "social data" make that it is now much easier to amass training data for visual concept detection. Also, their strongly subjective nature has great potential to improve the performance of the classic concept detection approaches by complementing the manual concept annotations traditionally used for training.

As shown in for instance [16] and in the current special session ([34], and further on in this paper), the social data, in particular the user tags, can serve directly as image features for learning visual concepts. In particular for topics that are hard to learn with low-level visual descriptors alone, the improvement in classification accuracy is often considerable; see Section 3.

However, it is well known that image labels originating from user tags have a number of problems. Tags supplied for an image often provide an incomplete description of the visual content, focusing mainly on the interest of the user 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.

This sets an interesting challenge to (i) determine the visual relevance of the tags that are present, (ii) deciding if other tags apply to the image besides the ones that were provided by the user. A recent approach to the former problem is offered, in the context of video sequences, by [33]. In this paper a supervised learning approach is proposed to establish when the tags of a "weakly labeled" video, visually apply to a frame in the sequence. Another basic approach that can be used for both types of

problems is to consider the presence of tags in the visual neighbors of tagged images, e.g. [18]. Two papers that have used the MIR Flickr collection to analyze social data are [3], which studies tag set analysis to clarify the precise meaning of tags and the semantic relations between tags based on tag co-occurrence models, and [21], which uses tag networks to provide automatic tag translations.

We consider such analysis of social data as important goals of MIR Flickr: both their direct use in concept detection and their automatic clarification and enrichment. We hope that the extension of its collection with a large set of images, which are described both by automatic visual descriptors and by social data will be an important step in this direction; see Section 5.

Various other sources of data can also serve as training data for visual concept detectors. We mention for instance image labels obtained through annotation games [1], click-through data of search engines (and, similarly, logs of relevance feedback interactions), and finally automatically generated EXIF data. For the latter particularly geo-location data has great potential for clarifying the semantic content of photos.

2.4 Beyond Bags of Features

The successful bag-of-visual-words approach is limited in its descriptive power by not taking into account the spatial layout of the feature patterns. Nevertheless, it has turned out to be a challenging problem to build representations that can provide a robust improvement of performance. A first idea that consistently outperforms simple bag-of-feature histograms is the use of spatial pyramids of local features [17]. In this approach the image is first partitioned into increasingly fine sub-regions, and next the histograms of local features found inside each sub-region are computed.

Aiming towards more invariant feature representations, further improvement may lie in a hierarchical organization of the visual words themselves. For instance, spatial patterns of low-level visual words can be combined into new intermediate-level visual words. This brings us close to a number of cognition-inspired concept detection mechanisms, such as [27] and [9]. In the latter, using a mixture of Markov chains, also temporal sequences of local features are clustered, corresponding perhaps roughly to learning *sentences* (or at least phrases) of visual words.

Next to these bottom-up approaches to building a more precise and discriminative visual language, also top-down mechanisms might be exploited to further improve performance. Currently, visual concepts are detected largely independently of each other; usually, no attempt is made to combine the resulting labels into a coherent description of the image. An initial attempt to exploit a concept hierarchy and simple concept relations to improve detection performance was part the 2009 LS-VCDT task (see above). Several groups could indeed improve their performance by analyzing label co-occurrence; see [23] for a first analysis of the results.

2.5 Benchmarking

Benchmarking of visual concept detection algorithms has greatly improved over the last decade, with initiatives such as TRECVid [29], ImageCLEF and Pascal VOC [7]. Also freely available annotated image collections such as IAPR TC12 [10] and MIR Flickr have contributed to testing under realistic conditions. Despite these favorable developments, [35] still raises the important concern of measuring how well concept classifiers generalize across domains. The MIR Flickr collection covers a very diverse domain of photography; nevertheless, in future additions to our annotation scheme we will take into account their potential for cross-domain evaluation.

One specific goal of the MIR Flickr annotations is to improve benchmarking of retrieval systems based on relevance feedback. Below we will demonstrate a number of interesting relations between benchmarking such systems and visual concept classification.

3. VISUAL CONCEPT DETECTION

In the following, we contribute a number of experiments in combining social data and low-level content-based descriptors to improve the performance of visual concept classifiers. First we describe the MIR Flickr tags and the special structure of its annotations.

3.1 MIR Flickr Tags and Annotations

The MIR Flickr collection 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).

In [14] we propose a hierarchical annotation procedure that makes it possible to generate consistent and realistic queries at greatly reduced cost. The decrease in annotation effort is mainly realized by reducing the size of the *annotation set*, i.e. the image set that needs to be considered for the annotation of a topic. The method reduces the annotation set while making sure to retain all potentially relevant images in the set.

Table 1. Most common Flickr tags in the MIR Flickr collection corresponding to visual concepts.

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

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 (at least) weakly relevant to strongly relevant

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 by definition, images relevant in this latter, stronger, sense are always also relevant in the weaker sense.

The MIR Flickr annotations were set up with the goal of evaluating the performance of retrieval systems based on *relevance feedback*. Concretely, this means the main annotations should represent so-called "full topic" annotations. These have two main aspects: (i) a topic label is added to an image only when the topic is relevant according to the subjective interpretation of the annotator, and (ii) a single annotator (effectively) annotates the entire collection this way, as opposed to the usual practice where several annotators collaborate on annotating a topic. It turns out that this latter, expensive-seeming, aspect does not mean that the annotator necessarily has to visit all images of the collection.

The top level topics used were chosen to cover many interesting topics as proper subtopics. They also have a large overlap with the most common Flickr tags in the collection. See Tables 1 and 2.

Concretely, the annotation process is divided into two main stages. First, we need a relatively costly stage, referred to as the pre-annotation stage, in which all concepts are interpreted in the wide sense described above. This stage proceeds down the semantic hierarchy. It is performed by an initial group of annotators who perform the hard work to realize a lighter subsequent, subjective, annotation stage. In this second stage follow the actual topic annotations, where many annotators can provide their personal interpretation of the topic.

In the pre-annotation stage all images are identified that annotators in the second stage reasonably might find relevant to the topic (or its subtopics). For this reason, we refer to these pre-annotations as *potential* labels. To receive a label the topic 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 the purpose of creating ground truth for testing queries, the pre-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 can take `tropical-sea`, `sea-at-sunset`, `sea-only` as additional subtopics. Especially the annotation of these latter specific annotations can typically be obtained with very little effort.

Table 2. Current MIR Flickr topics. The listed numbers indicate the fraction (in percents) of images in the collection that have been annotated with the corresponding topic. The first percentage represents the potential labels, the second percentage the regular annotations (see text). Missing numbers correspond to annotations in progress.

General Topic	Occurrence (%)		Subtopics	Occurrence (%)	
sky	32		clouds	15	5.4
water	13		sea/ocean	5.2	0.9
			river	3.6	0.6
			lake	3.6	
people	41	13	portrait	16	15
			male	24	15
			female	25	16
			baby	10	4.6
night	11	2.7			
plant life	35		tree	19	2.7
			flower	7.3	4.3
animals	13		dog	2.7	2.4
			bird	3.0	1.9
man-built structures	40		architecture city/urban building house bridge road/street		
sunset	8.5				
indoor	33				
transport	12		car	4.7	1.5
food	4.0				

The currently available annotations are summarized in Table 2, together with their occurrence rates. In particular the number of regular annotations will be extended according to the scheme described above.

Additional annotations are available from the ImageCLEF 2009 LS-VCDT task at http://imageclef.org/2009/PhotoAnnotation. Also these annotations will be extended, possibly by means of the Amazon Mechanical Turk, see [24].

3.2 Image Representation

For the image classification and relevance feedback tests we combined the following five sets of image features. Adding local features is left to future work.

1. **HMMD Color Histogram** descriptor. The non-uniform quantization of the HMMD color space is similar to that of the MPEG-7 color structure descriptor (CSD, [19]). Based on its 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, [19]). This descriptor captures the spatial distribution and orientation of edges by grouping of local edge direction histograms.
4. **MPEG-7 Homogeneous Texture** descriptor (HTD, [19]). 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**. See also below. 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 MIR Flickr collection.

Including the Flickr tags, this adds up to a total of 2341 features per image.

3.3 Tags As Features

Figure 1 compares the classification accuracy between classifiers based on low-level features only (Set 1-4 above), and classifiers that additionally use the Flickr tags (Set 5 above) as features.

The classifiers were trained on the 24 potential labels, and 14 regular (subjective) annotations. The mean average precision (MAP) is obtained using 5-fold cross-validation on training sets of 15000 images and test sets of 10000 images. Results are shown for two classification methods: a linear discriminant classifier (LDA) and a support vector machine classifier (SVM) with RBF kernel. The SVM classifications (C-SVM) were obtained using LIBSVM [5]. Values for the C (cost) and γ (RBF) parameters were selected using subsets of 1000 images of the 15000 training images.

As can be observed in Figure 1, the inclusion of the tags as features often greatly improves classification accuracy, in

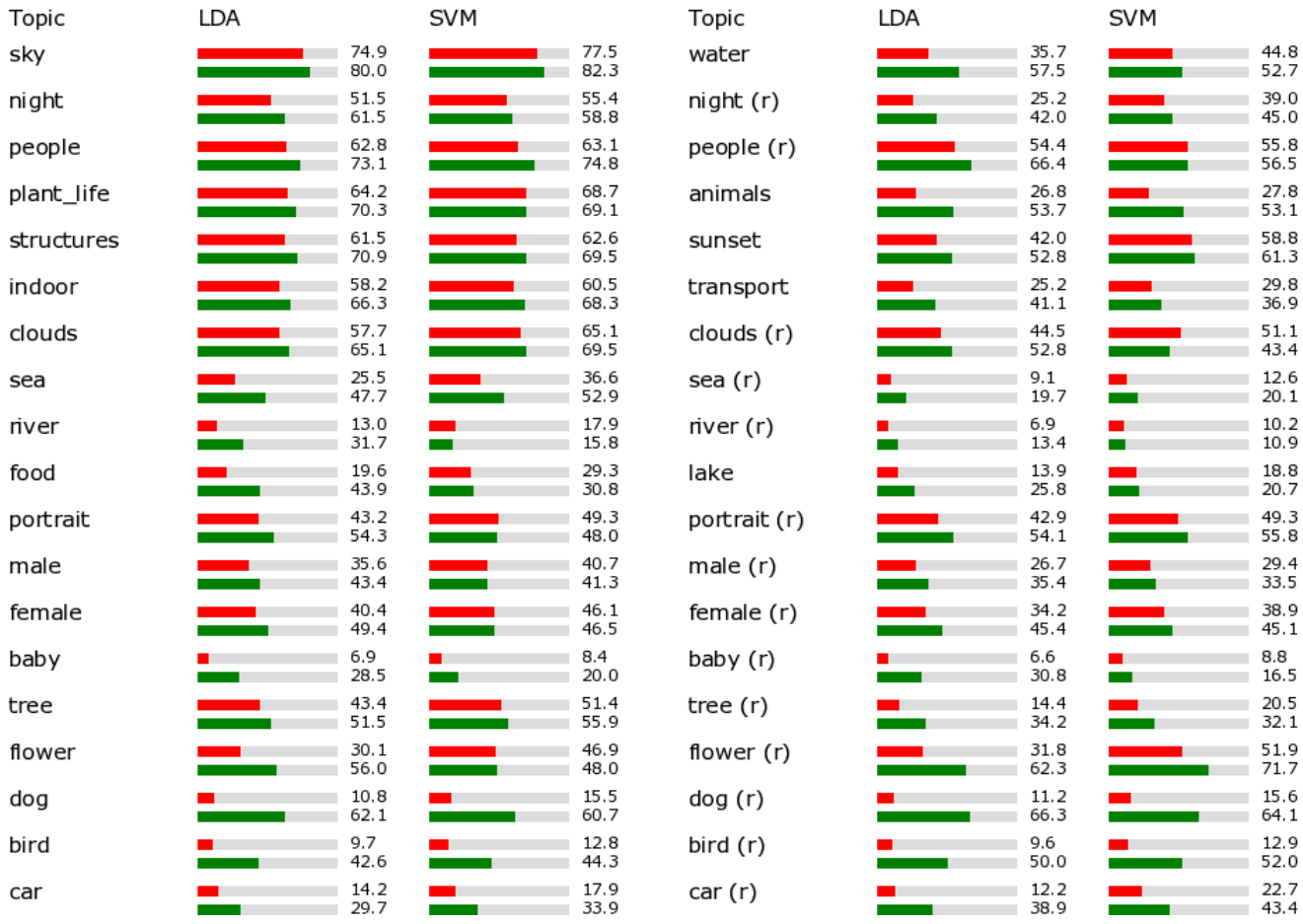

Figure 1. Comparison of mean average precision (MAP) for classification with and without Flickr user tags, for two classifiers (LDA and SVM). Topics are the 24 pre-annotations, and 14 regular annotation topics (with (r)), see text. The red bars display performance of classifiers trained and tested on low-level visual features only. The green bars display performance of classifiers trained and tested on the same features together with the Flickr tag features.

particular for topics that have relatively low accuracy when using the low-level features only.

3.4 Patterns in Topic Accuracy

Figure 2 shows the classifier accuracy in greater detail. It is interesting to observe that although the SVM classifiers generally give higher precision@50 values, their MAP values are quite similar to the LDA values. The LDA classifiers, however, have a great advantage in both ease of computation (e.g. 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.

Figure 2 also allows us to compare the performance differences between the pre-annotations (potential labels) and the regular annotations (with (r)). For some topics, often corresponding to somewhat "scenic" concepts (e.g. night, clouds, sea, river), classification performance degrades when we go from wide - weak relevance- interpretations to the more subjective - strong relevance - interpretations. For other categories, often corresponding mostly to foreground objects (e.g. portrait, baby, flower, bird, car) performance stays more or less the same or even improves for the more subjective interpretations.

A possible explanation that is consistent with the results presented here consists of an interaction of two competing effects. On the one hand the, wide-interpretation, pre-annotations are more visually consistent and thus easier to learn because their labeling does not depend on the personal interpretation and preference of the annotator. For instance, by his subjective interpretation an annotator might select mainly cloud images which for some reason or other appear "crisp" to him, while ignoring many images of the more plain-clouded-sky variety.

On the other hand, the subjective interpretation annotations may actually become more visually consistent due to a second effect that mostly images are selected where the concept is visually prominent. For example for the flower concept, the pre-annotations will contain many images where flowers are in the background, whereas the regular subjective annotations will consist to a large extent of images with a large flower at the center.

In general, the effect of a subjective concept interpretation on the visual consistency of the resulting topic can of course go either way.

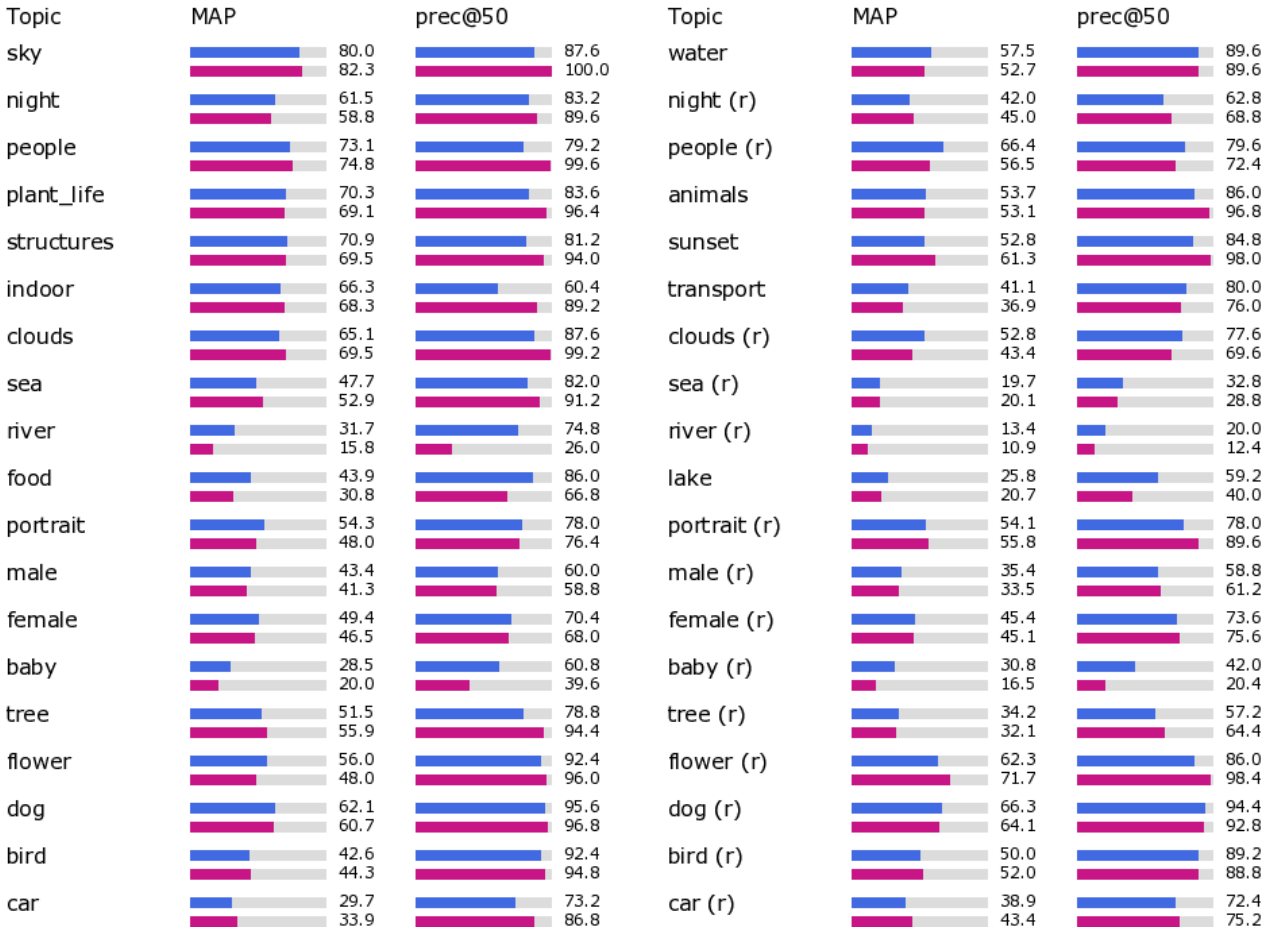

Figure 2. Mean average precision (MAP) and precision@50 for the current MIR Flickr topics. The top bars (blue) display performance for the LDA classifiers; the bottom bars display performance for the SVM classifiers.

4. RELEVANCE FEEDBACK

In the following we demonstrate two interesting relations between visual concept classifiers and image retrieval by relevance feedback. First, we show that features based on the output of classifiers trained on the pre-annotations, can greatly benefit relevance feedback accuracy for queries for more subjective concept interpretations. Second, we show how relevance feedback accuracy on a given topic can be measured relative to the performance of a visual concept classifier for the same topic, following the guidelines of [15].

Figure 3 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, [13]) 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 3.2. The results for this set are (despite the inclusion of the Flickr tag features) labeled as "low-level". Using a second set, we consider how the performance of the RF methods improves 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 pre-annotations. 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. Binary features were obtained from the classifier output by thresholding the discriminant projection values corresponding to different recall levels. We can observe that the inclusion of the features based on the classifiers for the wide concept interpretations of the pre-annotations greatly enhances the performance of RF methods on the more specific interpretations of the regular topic annotations.

Also shown are the PR-curves corresponding to topic classifiers with best known performance (MAP) based on cross-validation on the full collection training on 15000 images instead of the small set of feedback examples. The curve is labeled as best known classifier (BKC). These results are illustrative only, in an upcoming paper we present data demonstrating the increased robustness of measuring RF precision relative to classifier precision. See also [15].

5. EXTENSIONS TO MIR FLICKR

5.1 Image Collection and Descriptors

Given the new trends and ideas sketched in Section 2, MIR Flickr will be extended in two main ways. The collection will be extended to one million images and will be made available together with pre-computed content-based visual descriptors.

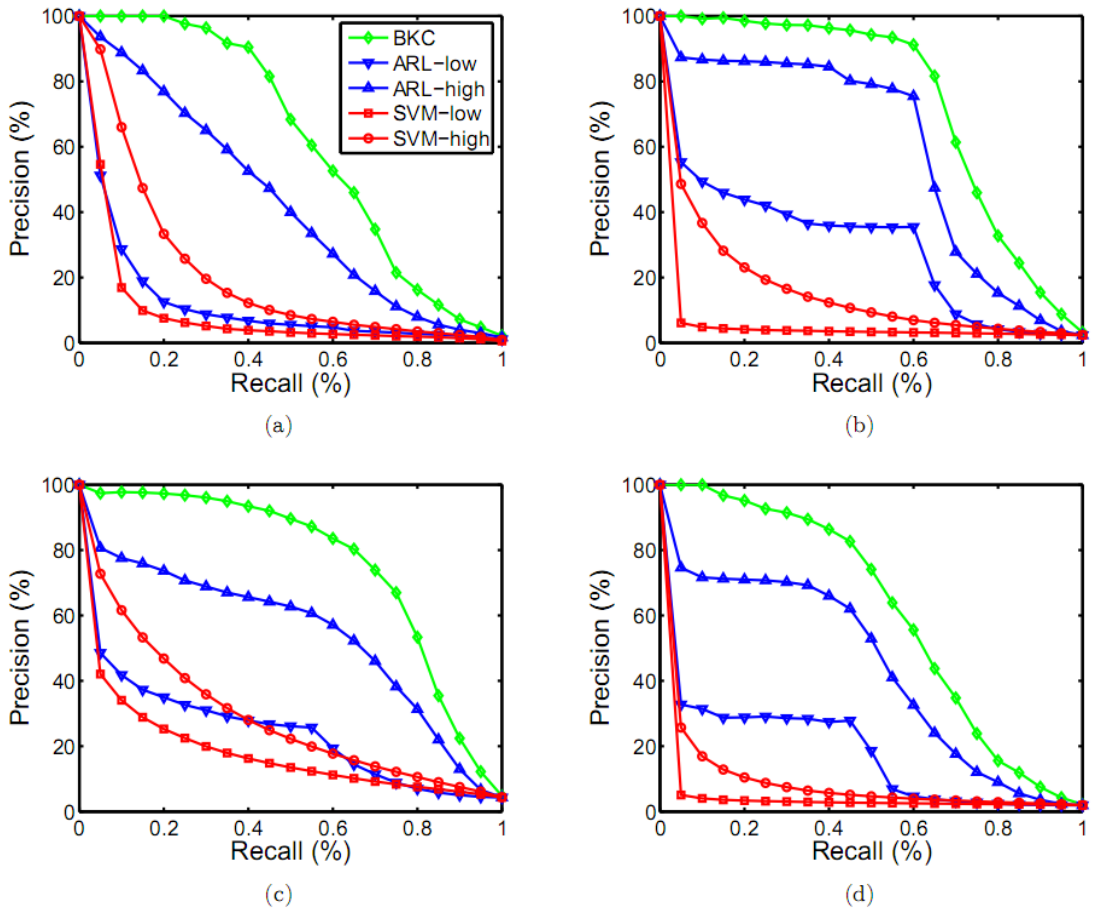

Figure 3. RF performance for 4 topics (a) dog on grass, (b) dog, (c) bird, (d) flower. See topic (a) for the legend. Topic (a) is not yet included in the standard MIR Flickr topics. The blue graphs correspond to RF by aspect-based relevance learning; the red graphs correspond to RF by support vector machines; the green graph shows the mean precision-recall for the best known classifier, see text.

The original MIR Flickr collection consists of 25000 fully annotated images. This paper announces the extension of the original core collection to one million images. The new images are obtained in the same way as the original images, and all images are made available under a Creative Commons Attribution Licence[1]. To obtain high quality photography, the images are also selected based on their Flickr interestingness score, see [14]. Note that the new images are not manually annotated like the core set of 25000 images, but all original Flickr user tag data, as well as the EXIF metadata, are made available.

In order to make it convenient to exploit the tagging data for visual concept detection, we additionally supply a number of content-based image features for the entire set of one million images. The following content-based descriptors are made available: the MPEG-7 Edge Histogram and Homogeneous Texture descriptors [19], and the color descriptor described in [26]. The latter descriptor, discussed above in Section 2.1, has also proven very effective in the TRECVid 2008 video retrieval benchmark, and the PASCAL VOC 2008 object classification competition. More details can be found at http://colordescriptors.com.

5.2 Distribution

The extended image collection and image representation data are made available in a number of ways. All original images are made available through BitTorrent. Since for many the full collection may prove too large to download, we also provide 64x64 pixel jpeg-thumbnails. Separate downloads are provided for the various types of metadata: the Flickr user tags, EXIF fields, and content-based visual features.

Further details, e.g. on the settings of the feature computations, and download instructions are on http://mirflickr.liacs.nl.

6. CONCLUSION

In this paper we have reviewed the recent developments in the MIR Flickr evaluation initiative. We have shown that the current collection offers a rich ground for experimentation on combining automatic content-based image descriptors and social metadata to improve the accuracy of visual concept classifiers.

We have also identified a number of promising trends and ideas in visual concept detection. Based on these trends and ideas, we have formulated new directions for the MIR Flickr Retrieval Evaluation, resulting in two new initiatives to extend the original image collection. First, the collection will be extended to one million Creative Commons Flickr images. The additional images will be made available with all original user tags. Second, a number of state-of-the-art content-based descriptors will be made available for the entire collection.

7. ACKNOWLEDGMENTS

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

[1] http://creativecommons.org/

8. REFERENCES

[1] von Ahn, L. and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria, April 24 - 29, 2004). CHI '04. ACM, New York, NY, 319-326. DOI= http://doi.acm.org/10.1145/985692.985733

[2] Aly, M., Welinder P., Munich, M. and Perona, P. 2009. Scaling Object Recognition: Benchmark of Current State of the Art Techniques. In *First IEEE Workshop on Emergent Issues in Large Amounts of Visual Data, IEEE International Conference on Computer Vision (ICCV) 2009*, Kyoto, Japan.

[3] Angeletou, S., Sabou, M., and Motta, E. 2009. Improving Folksonomies using Formal Knowledge: A Case Study on Search, *4th Asian Semantic Web Conference*, Shanghai, China.

[4] Borthaku, D. 2007. The Hadoop distributed file system: Architecture and design. From: lucene.apache.org/hadoop.

[5] Chang, C.-C. and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines*.

[6] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A.,Burrows, M., Chandra, T., Fikes, A., and Gruber, R. E. 2006. Bigtable: A distributed storage system for structured data. *Seventh Symposium on Operating System Design and Implementation (OSDI)*.

[7] Everingham, M., Zisserman, A., Williams, C. K. I., & Van Gool, L. 2006. The Pascal Visual Object Classes Challenge 2006 (VOC 2006) results (Technical report).

[8] van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J-.M. 2010. Visual Word Ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (in press).

[9] George, D. and Hawkins, J. 2009. Towards a mathematical theory of cortical micro-circuits. *PLoS computational biology*, Vol. 5, No. 10, e1000532.

[10] Grubinger, M., Clough, P.D., Müller, H. and Deselaers, T. 2006. The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems, *International Conference on Language Resources and Evaluation*, Genoa, Italy.

[11] Hare, J.S and Lewis, P.H. 2010. Automatically Annotating the MIR Flickr Dataset. In *Proceedings of the 2nd ACM international Conference on Multimedia information Retrieval* (MIR '10).

[12] Hörster, E. and Lienhart, R. 2008. Deep networks for image retrieval on large-scale databases. In *Proceeding of the 16th ACM international Conference on Multimedia* (Vancouver, British Columbia, Canada, October 26 - 31, 2008). MM '08. ACM, New York, NY, 643-646. DOI= http://doi.acm.org/10.1145/1459359.1459449

[13] Huiskes, M.J. 2006. Image Searching and Browsing by active aspect-based relevance learning. In *Proceedings of CIVR06*, LNCS 4071, 211-220. Springer.

[14] Huiskes, M.J. and Lew, M. S. 2008. The MIR Flickr retrieval evaluation. In *Proceeding of the 1st ACM international Conference on Multimedia information Retrieval* (Vancouver, British Columbia, Canada, October 30 - 31,

2008). MIR '08. ACM, New York, NY, 39-43. DOI= http://doi.acm.org/10.1145/1460096.1460104

[15] Huiskes, M. J. and Lew, M. S. 2008. Performance evaluation of relevance feedback methods. In *Proceedings of the 2008 international Conference on Content-Based Image and Video Retrieval* (Niagara Falls, Canada, July 07 - 09, 2008). CIVR '08. ACM, New York, NY, 239-248. DOI= http://doi.acm.org/10.1145/1386352.1386387

[16] Huiskes, M.J. and Lew, M.S. 2009. Hierarchical Annotation for Large Image Collections, *Theseus/ImageCLEF workshop on visual information retrieval evaluation*, Corfu, Greece.

[17] Lazebnik, S., Schmid, C., and Ponce, J. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2* (June 17 - 22, 2006). CVPR. IEEE Computer Society, Washington, DC, 2169-2178. DOI= http://dx.doi.org/10.1109/CVPR.2006.68

[18] Li, X., Snoek, C. G., and Worring, M. 2008. Learning tag relevance by neighbor voting for social image retrieval. In *Proceeding of the 1st ACM international Conference on Multimedia information Retrieval* (Vancouver, British Columbia, Canada, October 30 - 31, 2008). MIR '08. ACM, New York, NY, 180-187. DOI= http://doi.acm.org/10.1145/1460096.1460126

[19] Manjunath, B.S., Ohm, J., Vasudevan, V.V. and Yamada, A. 1998. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11:703–715.

[20] Moosmann, F., Nowak, E. and Jurie, F. 2008. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 9:1632–1646.

[21] Noh, T., Park, S., Yoon, H., Lee, S., and Park, S. 2009. An automatic translation of tags for multimedia contents using folksonomy networks. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in information Retrieval* (Boston, MA, USA, July 19 - 23, 2009). SIGIR '09. ACM, New York, NY, 492-499. DOI= http://doi.acm.org/10.1145/1571941.1572026

[22] Nowak, S. and Dunker, P. 2009. Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task. In *CLEF working notes 2009*, Corfu, Greece.

[23] Nowak, S. and Dunker, P. 2010. Performance measures for multilabel evaluation. In *Proceedings of the 2nd ACM international Conference on Multimedia information Retrieval* (MIR '10).

[24] Nowak, S. 2010. Reliable Annotations via Crowdsourcing? In *Proceedings of the 2nd ACM international Conference on Multimedia information Retrieval* (MIR '10).

[25] van de Sande, K.E.A., Gevers, T. and Smeulders, A.W.M. 2009. The University of Amsterdam's Concept Detection System at ImageCLEF 2009. *CLEF working notes 2009*, Corfu, Greece.

[26] van de Sande, K.E.A., Gevers, T. and Snoek, C.G.M. 2010. Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (in press).

[27] Serre, T., Wolf, L. Bileschi, S., Riesenhuber, M. and Poggio, T. 2007. Robust Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(3), 411–426.

[28] Sivic, J. and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 1470–1477.

[29] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international Workshop on Multimedia information Retrieval* (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM, New York, NY, 321-330. DOI= http://doi.acm.org/10.1145/1178677.1178722

[30] Torralba, A., Fergus, R., and Freeman, W. T. 2008. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 11 (Nov. 2008), 1958-1970. DOI= http://dx.doi.org/10.1109/TPAMI.2008.128

[31] Torralba, A. Fergus, R. Weiss, Y. 2008. Small codes and large image databases for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*, 1-8, Anchorage, AK.

[32] Uijlings, J. R., Smeulders, A. W., and Scha, R. J. 2009. Real-time bag of words, approximately. In *Proceedings of the ACM international Conference on Image and Video Retrieval* (Santorini, Fira, Greece, July 08 - 10, 2009). CIVR '09. ACM, New York, NY, 1-8. DOI= http://doi.acm.org/10.1145/1646396.1646405

[33] Ulges, A., Schulze, C., Keysers, D., and Breuel, T. 2008. Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proceedings of the 2008 international Conference on Content-Based Image and Video Retrieval* (Niagara Falls, Canada, July 07 - 09, 2008). CIVR '08. ACM, New York, NY, 9-16. DOI= http://doi.acm.org/10.1145/1386352.1386358

[34] Verbeek, J., Guillaumin, M., Mensink, T. and Schmid, C. 2010. Image Annotation with TagProp on the MIRFLICKR set. In *Proceedings of the 2nd ACM international Conference on Multimedia information Retrieval* (MIR '10).

[35] Yang, J. and Hauptmann, A. G. 2008. (Un)Reliability of video concept detection. In *Proceedings of the 2008 international Conference on Content-Based Image and Video Retrieval* (Niagara Falls, Canada, July 07 - 09, 2008). CIVR '08. ACM, New York, NY, 85-94. DOI= http://doi.acm.org/10.1145/1386352.1386367

[36] Zhang, J., Marszałek, M., Lazebnik, S. and Schmid, C. 2007. Local Features and Kernels for Classification of Texture and Objecs. Categories: A Comprehensive Study. *IJCV*, 73(2): 213–238.