# Automatically Annotating the MIR Flickr Dataset

## Experimental Protocols, Openly Available Data and Semantic Spaces

Jonathon S. Hare
jsh2@ecs.soton.ac.uk

Paul H. Lewis
phl@ecs.soton.ac.uk

School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ
United Kingdom

## ABSTRACT

The availability of a large, freely redistributable set of high-quality annotated images is critical to allowing researchers in the area of automatic annotation, generic object recognition and concept detection to compare results. The recent introduction of the MIR Flickr dataset allows researchers such access. A dataset by itself is not enough, and a set of repeatable guidelines for performing evaluations that are comparable is required. In many cases it also is useful to compare the machine-learning components of different automatic annotation techniques using a common set of image features.

This paper seeks to provide a solid, repeatable methodology and protocol for performing evaluations of automatic annotation software using the MIR Flickr dataset together with freely available tools for measuring performance in a controlled manner. This protocol is demonstrated through a set of experiments using a "semantic space" auto-annotator previously developed by the authors, in combination with a set of visual term features for the images that has been made publicly available for download. The paper also discusses how much training data is required to train the semantic space annotator with the MIR Flickr dataset. It is the hope of the authors that researchers will adopt this methodology and produce results from their own annotators that can be directly compared to those presented in this work.

## Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance Evaluation*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.4.9 [**Artificial Intelligence**]: Applications


*MIR'10,* March 29–31, 2010, Philadelphia, Pennsylvania, USA.


## General Terms

Experimentation, Measurement, Performance, Algorithms

## Keywords

Evaluation, Automatic Annotation, Semantic Image Retrieval, Visual-terms, Semantic spaces, Image Content Analysis

## 1. INTRODUCTION

In the past, it has been difficult to accurately compare the results of automatic annotation and semantic retrieval techniques from different research groups because there has been no standard, freely available dataset with which to experiment. Many authors chose to use the Corel Stock Photography collection [19], following its first use for annotation [4]. The Corel dataset has been criticised in the past as both being "too easy" and too small for proper retrieval evaluation [25, 19]. The Corel dataset also had many other deficiencies, such as the lack of control over which images were used (Duygulu et al [4] used 5000 images, but other authors have used both more and less images), and issues of image quality (the original images were relatively high resolution, but most researchers have only been able to obtain much smaller versions with less than 200 pixels on the longest side). The issue of image quality is particularly important with modern image feature representations, which often fail to work with very small images; for example, attempts to use difference-of-Gaussian based SIFT features with the small Corel images have inevitably failed due to the sparsity of extracted interest regions.

A second problem in comparing automatic annotation systems is that they tend to be highly sensitive to the particular image features that have been selected. In many situations, it would be insightful to be able to compare the performance of the machine-learning component of different automatic annotation systems. However, for this to happen in practice, the machine-learning components need to be tested with the same training data.

The release of the MIR Flickr dataset in 2008 now affords researchers access to a collection of 25000 high quality annotated photographic images which are freely available for the comparative evaluation of automatic annotation, semantic retrieval and relevance feedback systems. This paper seeks to explore the use of the MIR Flickr dataset in the automatic annotation context through a number of contributions. In Section 2, we describe an experimental protocol that extends the protocol described in the original MIR Flickr work [8] by

considering different sized training sets and evaluation metrics. Together with the experimental protocol, we provide information on software tools that can be used for evaluating results. In Section 4, in addition to briefly describing our auto-annotation technique, we describe the creation a set of visual-term features which we have made available for download from our website[1] and have provided links to the tools that were used to create these features. In Section 4, we apply the experimental protocol to our existing Semantic Space auto-annotator and describe the results attained. Finally, in Section 5, we provide some concluding discussion of the results of our experiments in the form prescribed by our experimental protocol.

It is our hope that other researchers will be able to use the protocol, data and tools that we describe in order to produce results that can be objectively compared to our own in the future.

# 2. EXPERIMENTAL PROTOCOL

In the original MIR Flickr paper [8], a standardised "Visual Concept/Topic Recognition" task was proposed. The task proposed splitting the dataset into 15000 training images and reserving the remaining 10000 for testing, and evaluating using precision-recall measures. We propose that the task is extended in the following way:

- Create three training sets of 5000, 10000 and 15000 images, and a test set of 10000 images. To avoid bias, the data sets should be created by dividing every five images of the dataset and assigning the first image to the 5000 image training set, the first and second to the 10000 image training set, the first, second and third to the 15000 image training set, and the forth and 5th to the 10000 image test set.

- Train the automatic annotator system using images from each of the three training sets, together with the relevant ("REL") labels alone, the potential ("POT") labels alone, and the relevant and potential labels combined (here-after referred to as "ALL"). Researchers are free to choose their own subsets of training data for cross-validation and optimisation.

- Evaluate the system on the test set, and present results using the following measures:

  - Interpolated precision-recall graphs from a hypothetical retrieval experiment carried out by retrieving ranked lists of images for each of the annotation terms.
  - Average precision per annotation term from the above hypothetical retrieval experiment.
  - Graphs of precision versus number of documents retrieved (from 1 to 1000 documents) from the above hypothetical retrieval experiment.
  - Equal Error Rate (EER) and Area Under Curve (AUC) values calculated from the ROC curve for each annotation term and averaged over all terms, as per the ImageCLEF 2008 and 2009 visual concept detection tasks [3, 21].

[1] `http://users.ecs.soton.ac.uk/jsh2/mirflickr/`

In addition, when researchers present their results, we suggest that it is good practice to describe how computationally efficient their technique is (for example; how long feature extraction takes, how long training takes and how long it takes to annotate a new unannotated image). Also implementation details should be described (for example; programming languages or tools used and whether the code is single threaded, multithreaded, or runs on a cluster).

*Training and test set word statistics.*

It is important that the proportions of each annotation word in the three training sets and the test set are approximately the same as each other to avoid bias. Figure 1 illustrates the number of occurrences of each annotation in the four sets. The plot shows that the four sets have roughly equivalent distributions of annotations, so first order bias is not a problem. Potential higher-order bias between sets of multiple terms (i.e. pairs of co-occurring words) is unlikely to be an issue as most current auto-annotation systems assume term independence or weak dependence. However, this factor should be considered in future work.

## 2.1 Tools and Settings

In order to generate the results (precision/recall and ROC statistics), we propose that a standard set of tools and settings should be used. The standard tool for generating retrieval statistics is the `trec_eval` tool[2], which originally arose from the TREC text retrieval conferences. `trec_eval` takes two files as input; a set of ground-truth relevances for a set of queries (known as a QRELS [Query Relevance] file), and results lists of retrieved documents for each of those queries. The output of the `trec_eval` tool is a detailed report of retrieval statistics for each query, and the average of these statistics over all queries. For evaluating MIR Flickr annotations, we suggest using the tool with only the `-q` option (this produces statistics for each query in addition to the summary). This usage differs slightly from standard TREC usage, which also specifies `-M 1000` and limits the evaluation to the top 1000 documents; in evaluating automatic annotation we believe that this is undesirable, and that all images should be considered. Pre-made QRELS files for the dataset are available for download from our website, together with a tool for creating QRELS for any subset of images.

In order to generate the ROC curve statistics, we propose that the `eval_tool` script[3] created for the ImageCLEF concept detection task is used. This script is implemented in GNU `octave`, but should also work in `matlab`. The script requires a ground-truth input in the form of a binary array, a list of classes (annotation terms), and a results file which contains a matrix of confidences for each image/annotation pair. More information can be found on our website.

# 3. AUTOMATIC ANNOTATION

As with many auto-annotation approaches, the methodology we have applied to demonstrate the protocol involves extracting feature vectors for each of the images, and then feeding the features of a training set, together with annotations to a machine learning system. Our machine learning system attempts to learn low-level relationships between all

[2] `http://trec.nist.gov/trec_eval/`

[3] `http://www.imageclef.org/system/files/evaltool.tgz`

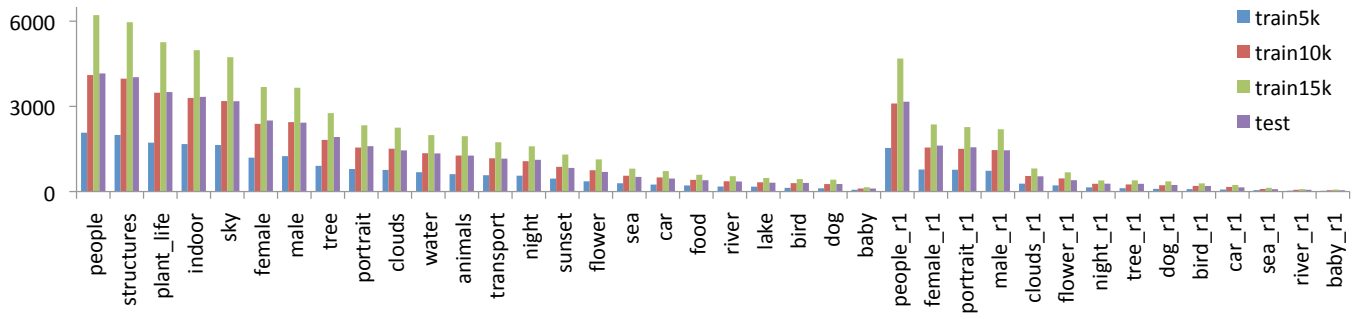



**Figure 1: Plot illustrating the number of occurrences of each annotation in each of the proposed subsets of the MIR Flickr dataset.**

of the features and annotations. Once the training phase is complete, features from unannotated images can be fed into the system to use the learnt relations to get predictions of annotations.

## 3.1 Image Features

Recently, it has become popular to transform image features into discrete elements or *terms*. These so-called "visual terms" are elegant because they enable image content to be described in much the same way as a text document. Typically, an image is represented by a histogram of the number of occurrences of each distinct visual term [23]. This kind of approach is often called a "bag of words" (or "bag of visual terms") model, as the terms are treated completely independently of each other, regardless of their relative or absolute positioning in the image.

In order to extract visual terms for the images in the dataset we propose to use the idea of detecting salient interest regions within the images from which descriptors can be calculated in order to create terms. We chose two feature morphologies and two region detectors for the task. These are described briefly below.

### *Region detectors.*

Lowe [13, 12] showed that by searching a difference-of-Gaussian pyramid for local peaks, both spatially and across scale, it is possible to select points robust to a range of projective transformations. The difference-of-Gaussian closely approximates the scale-normalised Laplacian-of-Gaussian [15, 13], $\sigma^2\nabla^2 G$. Mikolajczyk [17] showed that the minima and maxima of $\sigma^2\nabla^2 G$ produced the most stable interest points when compared to a range of other operators.

The Maximally Stable Extremal Region (MSER) detector [16] finds arbitrarily shaped regions in the form of connected components of an appropriately thresholded image. The regions are extremal because all of the surround pixels have either higher or lower intensity than the pixels within the region. The regions are *maximally stable* because of the optimal threshold selection process. The stability is measured as a function of how stable the local binarisation of the pixels is over a range of thresholds. As the threshold changes, the number of pixels within a connected region will likely change as well; if the number of pixels is fairly constant, then the region is stable. This definition of region stability based on relative area change is affine-invariant.

### *SIFT and Colour SIFT.*

There are a large number of different types of feature descriptors that have been suggested for describing the local image content within a salient region; for example colour moments and Gabor texture descriptors [22, 24, 14]. However, many of these descriptors are not robust to poor imaging conditions. It was shown in [18] that the Scale Invariant Feature Transform (SIFT) descriptor [13], was superior to other descriptors found in the literature, such as the response of steerable filters or orthogonal filters. The performance of the SIFT descriptor is enhanced because it was designed to be invariant to small shifts in the position of the sampling region, as might happen in the presence of imaging noise.

The SIFT descriptor is a three-dimensional histogram of gradient location and orientation. Lowe suggested that gradient location be quantised into a $4 \times 4$ location grid, and gradient angle be quantised into 8 orientation bins [13]. The resulting descriptor has 128 dimensions. Illumination invariance is obtained by normalising the descriptor by the square root of the sum of the squared components.

Recently the SIFT descriptor has been extended to work with colour gradients [1, 27]. The colour SIFT descriptor contains three vectors of 128 dimensions; the first is like regular SIFT and contains intensity gradient information, and the other two are colour based.

### 3.1.1 MIR Flickr Features

For our experiments with the MIR Flickr dataset, visual-terms were created by finding interest points within the images, extracting local feature descriptors, and then quantising to a pre-determined codebook of visual terms. We used a combination of multiscale difference-of-Gaussian interest regions with SIFT features [13], MSER regions [16] with SIFT features, and MSER regions with colour-SIFT features [1].

Each of the three region/feature combinations had its own 3125 term codebook created by applying hierarchical k-means [20] (5 levels with 5 clusters per node). The codebook size was not optimised in any way, and was chosen based on a best guess basis from previous experience with these feature morphologies and the machine learning technique described in the next subsection. The final image representation was created by appending the term-occurrence vectors from each of the region/feature representations to create a vector with 9375 dimensions.

Over the entire collection of 25000 images, there were over 20 million occurrences of visual terms from the difference-of-Gaussian detector, and over 5 million from each of the

MSER detector based features. The difference reflects the coverage attained with the two salient region detectors; the MSER detector picks large regions with stable characteristics (i.e. regions with little intensity gradient), whilst the difference-of-Gaussian technique locks on to regions with large, rapidly changing intensity gradients. The occurrences of visual terms across the entire dataset exhibit a ubiquitous characteristic seen in natural languages — that of Zipf's Law. Zipf's law states that the frequency of any term is approximately inversely proportional to its rank in a sorted frequency table. Figure 2 illustrates the Zipfian nature of the three visual vocabularies. The plot shows that the three vocabularies have approximately the same distribution (although the difference-of-Gaussian curve has higher values because there are about four times as many occurrences). In this work, we do not try and exploit the Zipfian nature of the visual data, however, it is possible to use the distribution to filter terms that occur very frequently (and hence are non-descriptive) [23, 5].

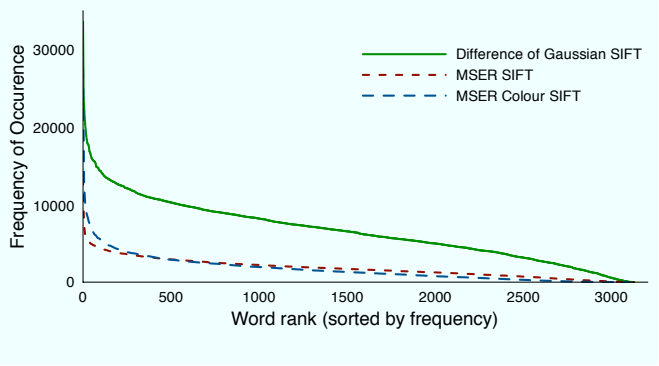



**Figure 2: Frequency of visual terms from all 25000 images plotted against their rank sorted by decreasing frequency. The curves are approximately Zipfian with the few dominant terms on the left and long tail towards the right.**

*Implementation.*

The difference-of-Gaussian SIFT features were extracted using David Lowe's `keypoints` executable[4]. The MSER regions were detected using Jiri Matas's detector[5]. The SIFT features for the MSER regions was extracted using Krystian Mikolajczyk's `compute_features` executable[6]. The chosen Colour SIFT feature is a chromatic descriptor called "InvC" (see [1]), and was extracted using the MSER regions with Jan-Mark Geusebroek's modified version of the `compute_features` program[7].

The codebooks and visual terms were created using our `ClusterQuantiser` software which uses the Hierarchical Integer K-Means implementation from the open-source `VLFeat` library[8] as the underlying computational engine.

[4]`http://people.cs.ubc.ca/~lowe/keypoints/`
[5]`http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html#binaries`
[6]`http://www.robots.ox.ac.uk/~vgg/research/affine/descriptors.html#binaries`
[7]`http://staff.science.uva.nl/~mark/downloads.html#colorsift`
[8]`http://vlfeat.org/`

The extracted features, visual-terms and word-occurrence information for each of the 25000 images, together with the `ClusterQuantiser` software are available for download from our website[9].

## 3.2 Semantic Spaces for Annotation

In this work we re-apply an existing automatic-annotation and semantic-retrieval technique, called a linear-algebraic semantic space, that we developed in previous work [6, 7]. The approach is based on the idea of creating a high dimensional vector space in which annotations are placed along with unannotated images. The placement of images in the space is such that they lie near to the words that describe them. In our case, we also place each visual term used for indexing the images into the space. Similar images and/or terms in this semantic-space share similar positions within the space. In order to build the space, we developed an approach that generalises a text indexing technique called Cross-Language Latent Semantic Indexing [10], which itself extends a technique known as Latent Semantic Analysis [2]. In the training stage, images, their respective annotations and visual term counts are used to learn a basis for the space that maps related items to similar locations. Once learned, the basis can be used to project unannotated images into the space using their respective visual term counts.

Once the unannotated images have been projected into the space, semantic retrieval and automatic annotation are straightforward. For retrieval, the query term is located in the space, and the images are ranked based on decreasing cosine similarity between their respective position, the origin and the query term's position. For annotation the process is reversed, and potential annotation terms are ranked based on their respective cosine similarity with the image vector. It should be emphasised that this process won't actually give you the exact annotations for the images, but rather a list of all the annotation terms in order of decreasing likelihood of them applying to the image in question.

# 4. DISCUSSION OF EXPERIMENTAL RESULTS

Using the automatic annotator and image features described in Section 3, we have performed the experimental workflow described in Section 2. In this section we describe the results attained.

## 4.1 Semantic space parameter setting

The semantic space auto-annotation technique has a single parameter, the dimensionality of the space, that needs to be optimised before annotation can proceed. Following the methodology used in previous work, we removed a number of images from each of the training sets to create validation sets. A space was then created for each of the reduced training sets, and the validation sets were projected in. We then calculated the optimal dimensionality by choosing the number of dimensions that gave the highest mean average precision. Figure 3 illustrates how the MAP changes as the number of dimensions increases. Once the optimal dimensionality had been determined, new spaces were created using the complete sets of training data (i.e. including the respective validation sets). The selected dimensionality for each of the nine spaces being evaluated is shown in Table 1.

[9]`http://users.ecs.soton.ac.uk/jsh2/mirflickr/`

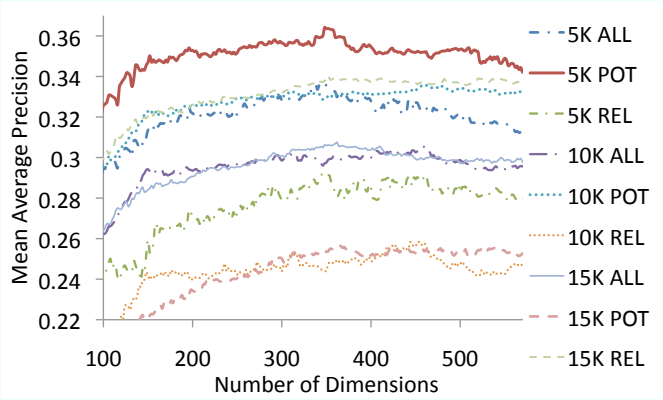



**Figure 3: Variation in performance (measured by mean average precision) as the number of dimensions of the semantic space changes when using the validation data sets.**

## 4.2 Retrieval Experiments

Table 2 shows the mean average precision for each of the different training sets and annotation sets. Figure 4 shows the results of using the REL annotations over the different training set sizes in the form of interpolated precision-recall curves. The table and graph show a number of interesting features. Most notable is the lack of effect from increasing the training set size; the larger training sets do produce a very slight increase in precision, but it would be difficult to justify this increase given the extra work involved in processing more images. It would be very interesting to see whether this effect only applies to our particular annotation technology, or whether other machine learning techniques also exhibit the same behaviour with the same training data. One possible hypothesis for this effect is that our semantic space is particularly good at learning course-grained relationships in the data, but is less good at learning finer-granularity relationships that have relatively little data support. These fine-granularity relationships would potentially be lost as noise in the dimensionality reduction stage of the technique. Adding more data would not help much because in addition to bolstering the finer-grained relationships, the signals from the course relationships would also get bolstered, cancelling out the potential improvement.

Table 2 also shows that the potential "POT" annotations were learned better than the relevant "REL" annotations. The combined "ALL" annotations exhibit a tradeoff in precision between the potential and relevant annotations. It is not clear (although it currently seems unlikely) that the slightly higher precision of the "POT" annotations would be enough to aid in the filtering of of images to reduce the workload of a human annotator marking relevant images in the same way the potential annotations were originally used to create the relevant annotations [8].

**Table 1: Selected number of dimensions for each of the semantic space annotators**

| Annotation Set | Training set size | | |
|---|---|---|---|
| | 5000 | 10000 | 15000 |
| ALL | 344 | 458 | 358 |
| POT | 348 | 458 | 352 |
| REL | 344 | 452 | 362 |

**Table 2: Mean average precision of all queries, separated by training set size and annotation set.**

| Training set size | Annotation Set | | |
|---|---|---|---|
| | ALL | POT | REL |
| 5000 | 0.276 | 0.313 | 0.216 |
| 10000 | 0.287 | 0.324 | 0.237 |
| 15000 | 0.292 | 0.327 | 0.240 |

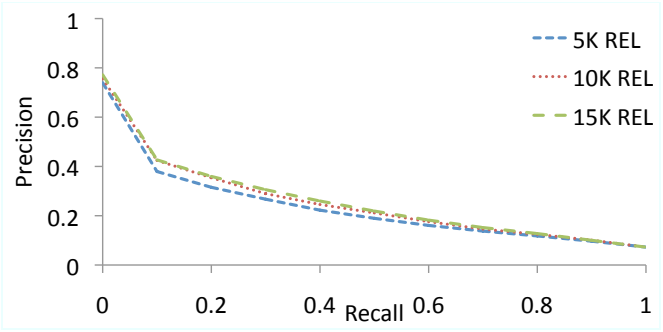



**Figure 4: Interpolated precision-recall curves for each training set size using the REL annotations.**

The precision-recall curves in Figure 4 display a distinct shape. The curves suggest that the first few (in terms of percentage recall) retrieved documents are on average about 80% relevant, but there is then a large drop such that after the first 10% of relevant retrieved documents, the precision is only around 40%. Between 10% and 100% recall the precision drops almost linearly. Figure 5 shows interpolated precision-recall curves for a selection of queries from the REL annotation set (training used 15000 images). These curves indicate that not all concepts are able to be learned equally; this is an expected result seen in all annotation systems. In particular, the graph shows that the "river_r1"

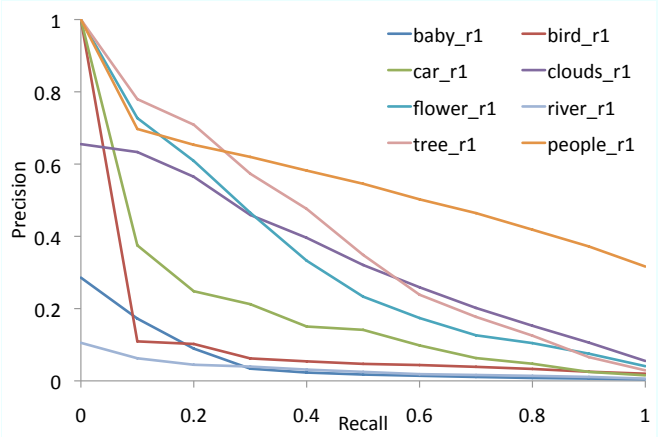



**Figure 5: Interpolated precision-recall curves for selected REL annotations using the 15000 image training set.**

term is not learnt particularly well, whereas the "people_r1" concept was learned much more successfully. The "bird_r1" concept is an interesting case; the graph indicates that concept was learnt partially — there were a few good results at the beginning of the ranked search results list, but the remainder were spread out (the graph doesn't tell us anything about the number of images you would have to look at to get a given recall, but looking at the `trec_eval` statistics we can see that after 1000 images had been retrieved, only 58 of the 196 relevant "bird_r1"images had been seen). There are a number of reasons for this, but there are two major ones. Firstly the visual feature representation may be insufficient to accurately model the concept; for example, it would be difficult to learn the concept of a particular colour using intensity gradient features alone. Secondly, the concept may be visually diverse or biased and not accurately captured by the training data. An example of this would be if the training data contained a number of images of birds, most of which are flying in a against a clear blue sky. In this case, the annotator is more likely to associate birds with a "blue sky" visual feature, and would probably fail at annotating images of birds sitting in their nests.

Figure 6 illustrates the precision of our annotator in a different way — the figure shows plots of precision against the number of retrieved documents. The curves for the different annotations are all quite different, although three of them ("sea_r1", "people_r1" and "portrait_r1") show an initial increase in precision as more documents are retrieved, followed by a peak and gradual decrease. The "dog_r1" curve shows a fairly constant drop in precision as more images are retrieved.

Figure 7 graphically shows the top five retrieved images for three different REL queries (corresponding to a range of average precision), using both the 15000 image and 5000 image training sets. Whilst the order of retrieved images changes with the different training set sizes, the overall average precision per term is about the same. Figure 8 shows the relative R-Precision histogram between results from the 15000 and 5000 image training sets with the REL annotations. The histogram shows that with the exception of the "baby_r1" annotation, all the annotations get a minor precision improvement with the increased training set size, although the improvement is not equally spread across the annotations.

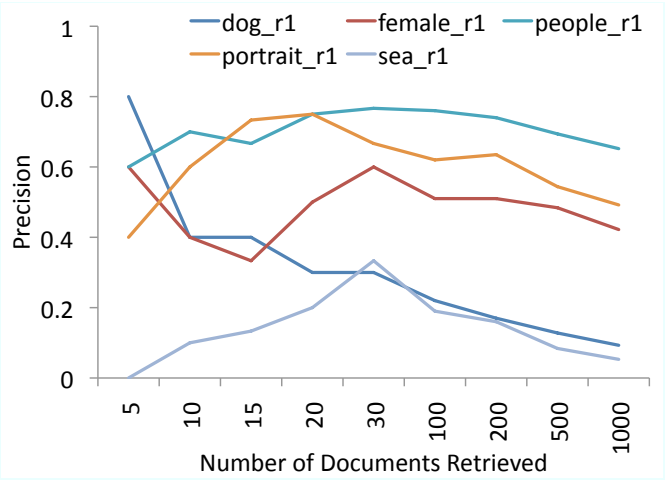



**Figure 6: Plot of precision versus number of retrieved documents for selected REL annotations using the 15000 image training set.**

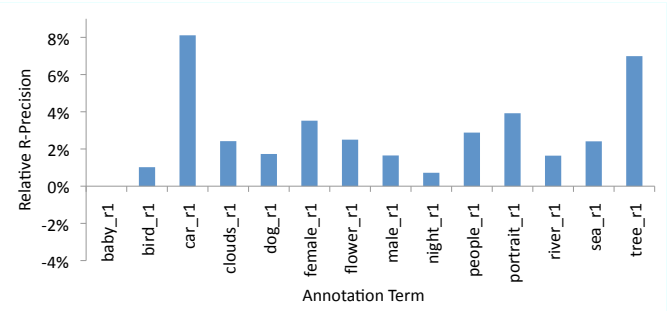



**Figure 8: Relative R-Precision of semantic spaces trained using 5000 and 15000 images and the REL annotations. The fact that all the bars are positive indicates that increasing the size of the training set only increases precision in this case.**

## 4.3 Annotation (ROC) Experiments

The averaged Equal Error Rate (EER) and Area Under Curve (AUC) results extracted from the analysis of the annotator using ROC curves are shown in Table 3. These results mirror the MAP results presented in Table 2 in that increased training set size gives a slight performance boost (increased AUC, decreased EER). However, they suggest that using the REL annotations outperforms the POT annotations (with the combined ALL annotations in between the two). This is a complete reversal of the results from analysing the MAP!

### *4.3.1 Comparing EER, AUC and MAP*

We have already seen that the EER and MAP measures do not necessarily concur with each other. This is emphasised by Figure 9. This graph shows the EER, AUC and AP values for the REL and POT annotations estimated using a semantic space trained on the 15000 image set. The

**Table 3: EER and AUC of all queries, separated by training set size and annotation set.**

| Training set size | Annotation Set | | | | | |
|---|---|---|---|---|---|---|
| | ALL | | POT | | REL | |
| | EER | AUC | EER | AUC | EER | AUC |
| 5000 | 0.319 | 0.742 | 0.331 | 0.727 | 0.296 | 0.772 |
| 10000 | 0.315 | 0.748 | 0.326 | 0.733 | 0.283 | 0.789 |
| 15000 | 0.303 | 0.761 | 0.318 | 0.743 | 0.272 | 0.797 |

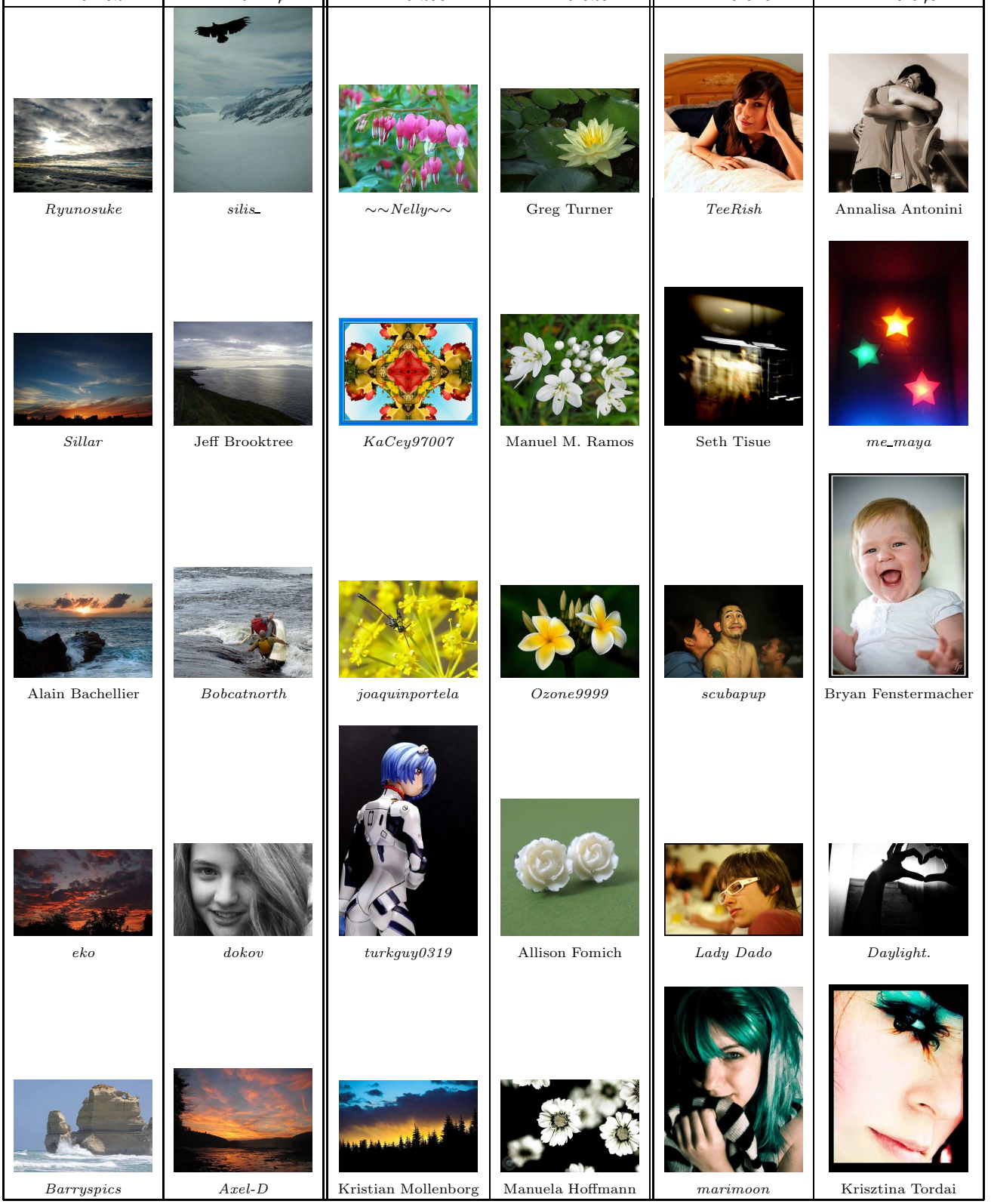
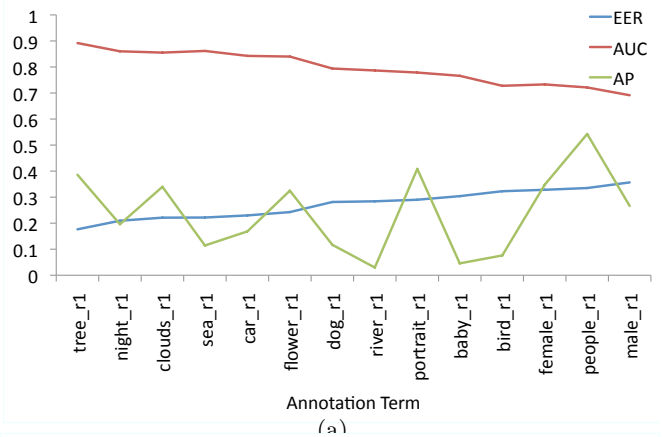
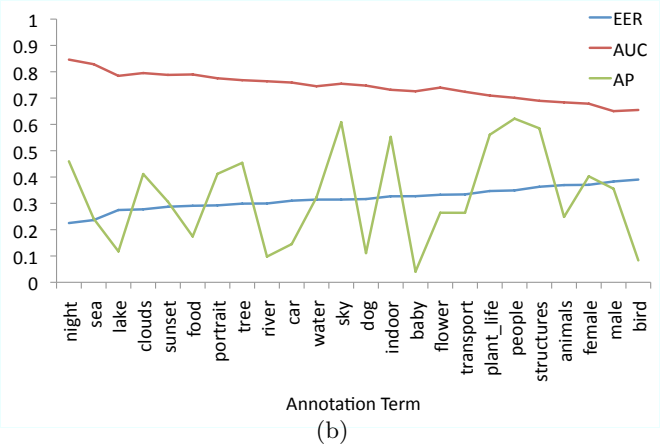



**Figure 7: The five top ranked images for three different queries from semantic spaces trained with the 5000 and 15000 image training sets and the REL annotations.**

annotations have been sorted into order of increasing EER. The graph clearly indicates the relationship between EER and AUC (increasing EER is coupled with decreasing AUC), however, it is the lack of any relationship between EER and MAP that warrants further discussion.




(a)




(b)

**Figure 9: Relationship between EER, AUC and AP statistics per annotation. Data from the REL (a) and POT (b) semantic space annotators trained with 15000 training image instances.**

Equal Error Rate and Average Precision measure rather different things. The Average Precision is a measure that rewards a system's ability to retrieve relevant documents quickly (i.e. highly ranked). If in the list of ranked images, all the relevant images come first, then the Average Precision will be 1. The Equal Error Rate on the other hand describes the rate at which accept and reject errors are equal given a binary classifier created by thresholding the confidence values calculated by the annotation system. Taking the term "people_r1" which has an EER of 0.335 (15000 training images, REL annotations) as an example, this EER implies that for a given threshold a binary classifier can be created such that 33.5% of the images containing the "people_r1" would and labelled as **not** containing the annotation, and 33.5% of the images **not** containing "people_r1" would be labelled as containing the "people_r1" annotation. In the 10000 image MIR Flickr test set, this means that the annotator would misclassify 1060 of the 1364 "people_r1" images as not containing people and misclassify 2290 of the remaining images as containing people. Whilst the "people_r1" annotation has a relatively poor EER score, it has the best score in terms of Average Precision, which implies that the annotator does a reasonable job of ranking relevant images near the beginning of the list. The precision of the "people_r1" annotation after 1000 images have been retrieved is 0.652, implying that 652 of the first thousand images were relevant, but those relevant images must have been spread out amongst those 1000 images.

Whilst it is possible to have a (relatively) poor EER score with a (relatively) good AP score, the reverse is not true; As the average precision approaches 1, the EER must approach 0 because the confidences will be ranked such that a threshold exists that separates them into the correct classes.

It is not clear which of the evaluation metrics is the best; it really depends on the task for which the automatic annotations are required. In a ranked retrieval scenario, where a human is going to look at the first few images, it makes sense to try and maximise Average Precision so that more relevant images appear in the results. The ROC curve, and EER measure on the other hand gives useful information about how to set a threshold on the confidence values in order to build a completely automated binary classifier with a given performance. Therefore, maximising the EER may improve completely automated scenarios but its usage for a retrieval scenario will not necessarily give the human looking at the images satisfactory results. Conversely, maximising average precision simultaneously improves the user experience and EER.

## 4.4 Comparison to the photo annotation task in ImageCLEF 2009

The 2009 ImageCLEF photo annotation task [21] used a subset of the MIR Flickr dataset for the evaluation (a training set of 5000 images and a test set of 13000 images). Rather than using the MIR Flickr annotations, a different set of 53 visual concepts was provided. In our entry to the task, we used exactly the same feature representation and annotation system as described in this work. We also experimented with the use of the EXIF data, but were unable to get any satisfactory results using it. Overall, in the ImageCLEF evaluation, our annotator performed better than average compared to all the submitted runs, however, our EER/AUC scores were still a way off from the best runs. The best runs (by AUC/EER) came from the ISLA group at the University of Amsterdam. The ISLA approach combined multiple Colour SIFT sampling strategies and quantisers to created visual terms, and then applied Support Vector Machines (one per concept) using a $\chi^2$-kernel for classification [26]. It would be interesting to explore how well the ISLA SVM approach would work with the features we provide with this work.

## 4.5 Computational Performance and Annotator Implementation Details

The feature extraction phase was performed in parallel (4 images being processed at once) on a quad core machine (Intel Core 2 Quad @ 2.66Ghz, 8G ram, Redhat Enterprise 5.3). The time taken for image processing varied depending on both the size of the image, and the image content. Timings for a typical image from the training set are shown in Table 4.

Training a semantic space with a maximum of 500 dimensions takes about 5 minutes using the 5000 image training

**Table 4: Approximate timings for feature extraction on a typical image from the training set.**

| Feature | Time |
|---|---|
| Difference-of-Gaussian detection + SIFT extraction | $\approx$ 1.8s/image |
| MSER detection | $\approx$ 0.1s/image |
| SIFT extraction on MSER | $\approx$ 2.7s/image |
| Colour-SIFT extraction on MSER | $\approx$ 1.0s/image |
| Vector quantisation | <0.1s per set of extracted features |
| *Estimated total* | $\approx$ *5.9s/image* |

set, less than 8 minutes for the 10000 image set, and just over 10 minutes for the 15000 image set on a dual quad core 2.8GHz Xeon workstation running Mac OS X (the semantic space code is single threaded, so only uses a single core). We would estimate that no more than 1G of ram was used during the semantic space training phase. Projecting all the test image in bulk takes under 2 minutes, and it takes about 5 minutes to generate annotations or retrieve all the 10000 test images; so, in general, it take less than .05s to get from a list of visual terms to the suggested annotations for a single image.

### *Implementation.*

The semantic-space software is written in C and makes use of Doug Rohde's SVDLIBC[10] for efficiently performing the large sparse SVD. The feature detector and descriptor software is written in C and C++. The image processing components were driven by standard UNIX make files[11], which enabled easy parallelisation using the `make` command's `-j` argument.

## 5. CONCLUSIONS AND FUTURE WORK

This paper has presented a methodology for performing automatic annotation, and visual concept detection tasks using the MIR Flickr dataset. The methodology emphasises the idea of making results from different systems comparable, and suggests freely available software tools for generating results. The second part of the paper was concerned with applying the methodology to our own annotation system, and presenting the results. For the evaluation we used quantised SIFT and Colour SIFT visual term features, which we have made available publicly. The results of our evaluation highlight two interesting features. Firstly, at least for our automatic annotator, the effect of changing the amount of training data between 5000 and 15000 images was surprisingly small. Secondly, the evaluation of automatic annotators using a retrieval-based framework as opposed to a classification-based framework (using ROC curves) can lead to remarkably different results.

There are a number of interesting avenues for further exploration of this work. Firstly, it would be interesting to see how far we can reduce the training set size before the annotator begins to break. Secondly, it would be interesting to further explore the effect of different visual features; in particular we are interested in the use of densely sampled SIFT features [27, 9], and spatial pyramids [11]. The incorporation of more traditional global features would also be interesting to study. Thirdly, we know from previous evaluations such as ImageCLEF, that our annotation technique is not the most powerful currently available (although this is difficult to quantify as the other annotation systems have been trained on differing feature morphologies); we are currently exploring alternative approaches such as the use of multiple Support Vector Machine annotators (e.g. [28]). Finally, the MIR Flickr dataset is provided with a number of user-generated "tags". The use of these tags for training annotators is another area open to exploration.

[10] `http://tedlab.mit.edu/~dr/SVDLIBC/`

[11] See `http://users.ecs.soton.ac.uk/jsh2/mirflickr/`

## 6. ACKNOWLEDGMENTS

This work was supported by the European Union under the Seventh Framework project LivingKnowledge (IST-FP7-231126), and the LiveMemories project, graciously funded by the Autonomous Province of Trento (Italy). The authors are also grateful to the creators of the original Flickr images in the MIR Flickr dataset for making them available for use in scientific evaluations.

## 7. REFERENCES

[1] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48 – 62, 2009.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[3] T. Deselaers and A. Hanbury. The visual concept detection task in imageclef 2008. In *CLEF Workshop 2008 / Evaluating Systems for Multilingual and Multimodal Information Access*, LNCS, Aarhus, Denmark, 17/09/2008 2009. Springer, Springer.

[4] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02*, pages 97–112, London, UK, 2002. Springer-Verlag.

[5] J. S. Hare and P. H. Lewis. On image retrieval using salient regions with vector-spaces and latent semantics. In W. K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, editors, *CIVR*, volume 3568 of *LNCS*, pages 540–549, Singapore, 2005. Springer.

[6] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom. A Linear-Algebraic Technique with an Application in Semantic Image Retrieval. In H. Sundaram, M. Naphade, J. R. Smith, and Y. Rui, editors, *CIVR 2006*, volume 4071 of *LNCS*, pages 31–40, Tempe, Arizona, 2006. Springer.

[7] J. S. Hare, S. Samangooei, P. H. Lewis, and M. S. Nixon. Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In *ACM CIVR '08*, pages 359–368. ACM, July 2008.

[8] M. J. Huiskes and M. S. Lew. The MIR Flickr Retrieval Evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[9] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 604–610, Washington, DC, USA, 2005. IEEE Computer Society.

[10] T. K. Landauer and M. L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38, UW Centre for the New OED and Text Research, Waterloo, Ontario, Canada, October 1990.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:2169–2178, 2006.

[12] D. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV*, pages 1150–1157, Corfu, 1999.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, January 2004.

[14] W. Y. Ma and B. S. Manjunath. A comparison of wavelet transform features for texture image annotation. In *ICIP '95: Proceedings of the 1995 International Conference on Image Processing (Vol.2)-Volume 2*, page 2256, Washington, DC, USA, 1995. IEEE Computer Society.

[15] D. Marr. *VISION: A computational Investigation into Human Representation and Processing of Visual Information*. W. H. Freeman and Company, 1982.

[16] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In P. L. Rosin and A. D. Marshall, editors, *BMVC*. British Machine Vision Association, 2002.

[17] K. Mikolajczyk. *Detection of local features invariant to affine transformations*. PhD thesis, Institut National Polytechnique de Grenoble, France, 2002.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 257–263, June 2003.

[19] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. In M. S. Lew, N. Sebe, and J. P. Eakins, editors, *CIVR*, volume 2383 of *LNCS*, pages 38–49. Springer, 2002.

[20] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *In CVPR*, pages 2161–2168, 2006.

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

[22] N. Sebe, Q. Tian, E. Loupias, M. Lew, and T. Huang. Evaluation of salient point techniques. *Image and Vision Computing*, 21:1087–1095, 2003.

[23] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, October 2003.

[24] M. A. Stricker and M. Orengo. Similarity of color images. In *SPIE Storage and Retrieval for Image and Video Databases*, pages 381–392, 1995.

[25] J. Tang and P. H. Lewis. A study of quality issues for image auto-annotation with the corel dataset. *IEEE Trans. Circuits Syst. Video Techn.*, 17(3):384–389, 2007.

[26] K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. The university of amsterdam's concept detection system at imageclef 2009. In *CLEF working notes 2009*, Corfu, Greece, 2009.

[27] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (in press), 2010.

[28] J. Zhang, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73:2007, 2007.