Boosting Image Retrieval through Aggregating Search Results based on Visual Annotations

Ximena Olivares
Universitat Pompeu Fabra
Passeig Circumval.lacio 8
08003 Barcelona, Spain
ximena.olivares@upf.edu

Massimiliano Ciaramita
Yahoo! Research
C/ Ocata 1
08003 Barcelona, Spain
massi@yahoo-inc.com

Roelof van Zwol
Yahoo! Research
C/ Ocata 1
08003 Barcelona, Spain
roelof@yahoo-inc.com

ABSTRACT

Online photo sharing systems, such as Flickr and Picasa, provide a valuable source of human-annotated photos.Textual annotations are used not only to describe the visual content of an image, but also subjective, spatial, temporal and social dimensions, complicating the task of keyword-based search. In this paper we propose a method that exploits visual annotations, e.g. notes in Flickr, to enhance keyword-based systems retrieval performance. For this purpose we adopt the bag-of-visual-words approach for content-based image retrieval as our baseline. We then propose to use rank aggregation over the top 25 results obtained with a set of visual annotations that match the keyword-based query.

The results on retrieval experiments show significant improvements in retrieval performance when comparing the aggregated approach with our baseline, which also slightly outperforms text-only search. When using a textual filter on the search space in combination with the aggregated approach an additional boost in retrieval performance is observed, which underlines the need for large scale content-based image retrieval techniques to complement the text-based search.

Categories and Subject Descriptors

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

Keywords

Image retrieval, visual annotations, rank aggregation

1. INTRODUCTION

The popularity of recent on-line photo sharing services, such as Flickr [11] and Picasa Web [23], has produced very large, continuosly growing, corpora of human-annotated digital images, where millions of photos are uploaded and annotated on a daily basis. The annotations provided by users are essential to make photos retrievable by search engines, as keyword-based search is the de-facto model for query formulation on the Web.

However, retrieval models that are generally effective for text retrieval do not work as well for text-based image retrieval. Three factors complicate the matters for text-based retrieval. First of all, textual annotations of images are rather sparse and short as most users use only a few keywords to annotate their photos. Furthermore, the annotations provided do not solely serve the purpose of describing the visual content of a photo. Annotations often include spatial, temporal, and social references, as well as subjective/personal descriptions. This further diffuses the results achieved with keyword-based search on images. Finally, the keyword-based query formulation is powerful, but lacks the expressiveness that is inherent in an image. It is difficult for a user to express the visual characteristics of the desired image only using textual clues.

The latter problem has been extensively studied in content-based image retrieval, where the objective is to include the visual characteristics of an image into the search process. Using the query by image content (QBIC) search paradigm similar images are retrieved for a given sample image by extracting visual features from all the images in the collection. The down-side of this approach is that the user need to begin the query process with a sample image. Alternatively, high level concepts are derived for the low level features that are extracted from the image content. The problem with this approach is often referred to as the semantic gap problem [13], where for each concept a special concept detector is needed to translate the user information need into low-level image features. The latter makes the approach less suitable for widespread application on the Internet, where no domain restrictions are in effect.

In this paper we propose a method that deploys visual annotations, e.g. notes in Flickr, to enhance the retrieval performance of key-word based queries. With a note, the user can highlight a certain region in the photo and associate a tag (label) with the region. To illustrate this, Figure 1 shows examples of notes with the tag "British telephone booth". Though people annotate notes in a similar fashion as they annotate photos, i.e. their intentions are diverse, the bounding box on the region makes the note a good candidate for a visual query.

For that purpose we adopt the bag-of-visual-words approach for content-based image retrieval introduced by Sivic et al. [24] as our baseline system. Generally speaking, the retrieval performance for content-based image retrieval is lower than the performance of keyword-based image retrieval.

Figure 1: Examples of visual annotations for a telephone booth.

In this paper, we therefore explore two complementary paths. On one hand, we are interested in combining visual and textual search to improve the precision of our search results. In addition we want to exploit the large set of visual annotations that form the collective knowledge in Flickr.

One advantage of using the visual annotations in Flickr is that there are many examples that can serve as input for the retrieval process. We therefore propose to use rank aggregation for merging the result sets obtained with the content-based image retrieval system that is fed with the visual annotations, which match a given keyword-based query. Rank aggregation is primarily used by meta-search engines, where the results from different search engines are merged into a new ranked list of results. A simple, and commonly used method is the Borda count model [3] that assigns a score to each element in the set of ranked lists, and then sums the scores for each individual element.

Based on a retrieval performance experiment we evaluate and compare the performance of tag-only search, content-based image retrieval using visual annotations, and content-based image retrieval using visual annotations in combination with rank aggregation. For that purpose we sampled a collection of annotated photos from Flickr, defined a set of topics based on frequent searches in Flickr, and created a set of 10 visual annotations for each topic.

The remainder of this paper is organised as follows. Section 2 presents the related work on media annotation, content-based image retrieval, and rank aggregation. In Section 3 we then describe the system for content-based image retrieval, and introduce the method for rank aggregation of the results for the visual annotation in Section 4. The set-up and results of the retrieval performance experiment are presented in Section 5 and the conclusions and future work is discussed in Section 6.

2. RELATED WORK

Photo sharing systems allow the user to search image collections by submitting a keyword-based query. The images are then retrieved using sophisticated text retrieval models. This type of search is based on the text that describes the images, such as their title, description, and tags. Photos in Flickr contain different types of meta-data, ranging from technical details to more subjective information. At low level, they contain information about the camera, shutter speed, rotation, etc. At a higher level, the user that uploaded the image can include a title and description, which are more likely to be used to describe the image as a whole. The use of tags permits the user to describe the what he thinks is relevant to the image using simple keyword combinations.

Ames and Naaman [1] present a qualitative study describing the motivations behind users tagging their pictures. They defined two main dimensions of the motivations: social and functional, and characterized the motivations whether they were used for themselves or their family (social), or as a way of complementing the context of the image (functional). Furthermore, Dubinko et al [8] show that tags not only describe the specific contents of the images, but also additional information. They observed recurring categories such as: events (e.g. "Valentine's days" or "Thanksgiving"), personalities (e.g. "Pope"), and social media tagging (e.g. "What's in your fridge"). Another important characteristic of tag-based systems is the way people use the tags. In Marlow et al [18] they analyzed the tags used to describe the images, and observed that most users have few distinct tags, while a small group of users have large sets of tags.

The variety in which users tag their photos, has an impact on retrieval. As a consequence, tag-only search can become very noisy. We therefore argue that it is important to include the intrinsic information of the image into the retrieval process. In this paper we are interested in using the visual content of an image. The literature in this area is extensive, but in the context of this paper we can limit ourselves to the state of the art in image object retrieval. Furthermore, we are interested in approaches that can be applied at large on the Internet without training a large set of concept detectors [25].

Sivic and Zisserman [24] introduced the bag-of-visual-words architecture, and successful results have been reported for object retrieval on a large image collection containing buildings in Oxford. For every image in the collection, affine regions are extracted, and described by a SIFT [16] descriptor. This set of vectors are quantized to build a visual vocabulary as proposed in [7]. This approach allows to represent every image as a set of visual words, hence making it possible to describe them with a weighted vector and use standard text retrieval techniques to determine the similarity between images. Since the spatial arrangement of the visual words is crucial, they add a simple constraint to the spatial distribution of words. We adopt this approach as the baseline for our research.

In [24] a region of a video frame is selected to obtain all the frames in the video where the selected object appears. Based on this work, Philbin et al [22], have built a large-scale object retrieval system using a combination of images extracted from Flickr and images from the Oxford building database. A query image is submitted and a set of ranked images is returned. They presented considerations for building a visual vocabulary and tested their results using building landmarks. In addition, Chum et al [5], presents a query expansion approach, where some of the retrieved images were used to reformulate the original image query improving the results obtained in [22].

This line of research uses an image query as input for the retrieval system. We envision that the information obtained from the tags, as well as the information contained in the image must be combined to obtain better results, in order successfully use a keyword-based query to search over the collection. In [15], an analysis of the patterns is presented that exist between the visual words of images that share a

common set of tags.

In our research we deploy visual annotations, e.g. notes in Flickr, that associate a label to a region of the image. An example of visual annotation is shown in Figure 1. This type of annotations are valuable, since the associated text is highly relevant to the highlighted area in the image, Given a keyword-base query, it is possible to obtain a set of visual annotations that can be used to search over the collection based on the content of the image. This will lead to partial list of results from several sources (each of the different visual annotations). It is then necessary to decide how to merge these results. This problem can be compared to the problem of determining the ranked list of winners in an election. A simple, and commonly used method is the Borda count model [3] that assigns a score to each element in the set of ranked lists, and then sums the scores for each individual element. Various methods have been proposed for rank aggregation on the Web, in the context of meta-search engines [10, 9, 2]. To the best of our knowledge it is the first time this approach is applied to keyword-based image retrieval using visual annotations. Aslam and Montague [2] investigate the problem of meta-search and compare different models. Their results show that Borda count is a simple and efficient algorithm with good performance. For this reason, we use this mechanism in the aggregation stage of our system

Our work is inspired by the state of the art methodologies for image retrieval which we aim to extend by combining textual and visual information to improve retrieval performance, and specifically precision.

3. CONTENT-BASED IMAGE RETRIEVAL

In this section, we describe the system for image (object) retrieval. As a baseline, we adopt the framework proposed by Sivic and Zisserman in [24, 22] to handle the retrieval of photos based on visual characteristics. They successfully applied this framework on a domain-restricted collection to detect the same object in different photos, i.e. in their experiments they focussed on detecting near-identical representations of buildings in Oxford. Their results are promising both in the dimension of scalability and retrieval performance.

In short, the framework consists in the following steps, for which a parallel with text retrieval can be made:

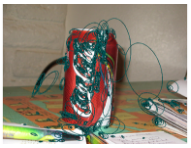
1. Extract visual features (salient regions) from the images in the collection, and describe them with a high-dimensional descriptor.
2. Build a visual vocabulary from the high-dimension descriptions by quantising and clustering them into a vocabulary of visual words. In this step, the high-dimensional descriptions are lemmatised into similar visual words. Each image can then be described as an histogram of visual words.
3. Using the bag-of-words approach, existing text-retrieval model can be invoked to build an index over the image collection, and similar image can be found using the query by image content paradigm.
4. Finally, a post-retrieval step is needed to re-rank the results to take the spatial structure of the image into account, which is vastly more dominant in image retrieval, than in text retrieval [24].

In the sections below, we provide a more detailed outline of this approach, complemented with some of the implementation specifics used in our experiments. After explaining the baseline system, we present our approach for aggregation of visual annotations in Section 4.

3.1 Feature Extraction

In the literature, many approaches to extract visual information (features) from images have been proposed [20]. A combination of these features is typically used to retrieve similar images. In our work, we limited ourselves to extracting high-dimensional region descriptors from images, based on Harris affine and Hessian affine regions, as introduced by [19], because of their invariance to rotation, translation and scale. Harris affine regions are based on the points obtained with the Harris detector, which are later processed obtaining affine viewpoint covariant regions that represent corner structures. On the other hand, Hessian affine regions are based on processing the points obtained by the Hessian detector, resulting in affine viewpoint covariant regions, which represents blob structures.

When processing the image collection we extracted on average 1,000 Harris regions and 1,066 Hessian regions per image. Each region is then described using a 128-dimension SIFT [16] descriptor. Figure 2.1 shows the extracted Harris regions for one of the images in our collection. When a visual annotation is drawn over the image to mark an object, we can select only the feature descriptors that are inside the bounding box of the annotation (see Figure 2.2), and ultimately, as shown in Figure 2.3, we only use those features to describe the object as input for searching.



1 Extracted regions.

2 Overlaying a visual annotation.

3 Features describing the object.

Figure 2: Example feature extraction.

3.2 Visual Vocabulary

Once features have been extracted from the images in the collection, a visual vocabulary needs to be build. The vocabulary can be generated by clustering the SIFT descriptors into k clusters. Based on a learned clustering model a *visual word*, a cluster label, is associated with all the elements contained in a cluster. Clustering large amounts of data, for large values of k, as in this case where k can be in the order of tens of thousands, is a challenging task. As shown in [17, 14] approximate k-means clustering can adequately scale up for this type of task. Similarly, we implemented an approximate k-means algorithm paired with a kd-tree on the cluster set. Search for the nearest neighbor in the tree is carried out using a priority queue for the nodes, which are ranked according to the distance of the nodes hyper-rectangle from the query point. Search terminates when the queue is empty, if the exact nearest neighbor has been identifies, or after reaching a maximum number of comparisons. In our clustering model we use only one kd-tree, rather than

several randomized ones, since we found limited benefit from using several trees, over one tree with a higher threshold for the maximum number of comparisons. The maximum number of comparisons was set to 1,200.

We learned the clustering model on a set of 1 million SIFT descriptors randomly selected from the image collection. We experimented with various sizes of the vocabulary, ranging from 1,500 to 10,000 clusters. For the experiment described here we settled upon a vocabulary of 10,000 words. The remaining descriptors are classified based on the learned k-means model. Outlier descriptors are removed from the set: an outlier is a datapoint whose distance to the nearest centroid is greater than the average distance in that cluster plus twice the standard deviation of these distances. Similarly to stop word filtering in text retrieval, we removed the top 2.5% of the clusters with the largest population. An independent vocabulary is created for each feature, e.g. Harris affine and Hessian affine. A third vocabulary of 20,000 words is created by merging the vocabularies generated using the two independent features representations.

3.3 Vector Space Model

Following the traditional bag-of-words approach for text retrieval, an image can be represented as a weighted-term vector in the vector space model. Using the analogy to text retrieval, we used the tf-idf weight of the visual words to create the corresponding vector. The similarity between the images can then be measured by calculating the cosine similarity of the weighted vectors, obtaining a normalized value ranging between 0 and 1. Alternatively, we can use one of the object annotations to search the vector space, to find images that are likely to contain the object.

3.4 Spatial coherence filter

A limitation of the bag-of-words approach is that all structural information contained in the image is lost. Although two images can have a high degree of cosine similarity, the relative spatial coherence of the visual words between these two images can be low, which indicates that they are visually not similar at all. Therefore an analysis of the spatial arrangement of the visual words between the query image and each of the retrieved images is needed, as also argued in more detail in [24, 22].

In the present work we have implemented a simple spatial coherence filter. For every common visual word between two images, we analyze the common visual word present in the surrounding area. This spatial constraint generates an additional similarity measure that is used to discriminate images that only have the visual words in common with the ones that also satisfy the spatial distribution of the elements.

4. AGGREGATED SEARCH WITH VISUAL ANNOTATIONS

In the introduction of the paper, we have discussed how users annotate images at large in on-line photo sharing services such as Flickr. In particular users can attach labelled notes to photos on Flickr. Though not as popular as the photo annotations, notes can be valuable to learn different visual representations of an object. This observation leads to the main contribution of this article, where we aim to improve the retrieval performance by aggregating the result sets for searches with visual object annotations in photos. Although the use of textual information in Web image retrieval systems has matured, we propose that it can be improved by complementing it with visual information, especially when the user's information need is specific, and can not easily be described by a combination of keywords.

The widespread availability of visual annotations in Flickr provides us with a base collection of annotated objects where for each high-level concept a set of visual annotations is available that can be used to aid the user in his search. The portion of the image enclosed by the visual annotations contains a set of visual words (as defined in Section 3), that are mapped to a particular concept defined by the text describing the annotation. When a user submits a keyword-based query, the system will use the visual annotations to obtain images that answer the user query. Each annotation is used to search for similar images, using the cosine similarity between the images that have a textual annotation that also matches the user query. As a result we obtain, for every annotation, a set of similar images which are re-ranked using the spatial coherence filter described in Section 3.4.

In Figure 3 the results for the query "apple logo" are shown. The top three rows show the top 10 search results using three different visual annotations. To limit the search space, we have used a filter on the image tags. Obviously, this already improves the results when searching with a single visual annotation. In the experiment of Section 5 we will present a comparison of tag-only, tag & visual, and visual search that illustrates how the retrieval performance is influenced for each of the different combinations. The bottom row of Figure 3 shows the aggregated results.

The results from each of the visual annotations can be seen as individual sources of information that need to be merged into a single set of results. This problem is similar to the one of a metasearch engine that needs to combine search results, or essentially, the combination of any set of ranked lists. Using the ranked position of the images, the results are merged using a voting mechanism [3]. Borda models this problem as a set of *voters* (in our case each visual annotation) that must sort a set of *candidates* (the set of results) by assigning points to each of them, and a final list of ranked candidates must be obtained. For this, every voter assigns points to each of the candidates, based on their position in their ranked list. The first element in the list is assigned with n points, the second element is given $(n-1)$ points, until the last element is assigned 1 point. To obtain the final list of results, all the candidates are sorted by their total number of points.

The aggregated ranking favours images that are ranked high in several of the partial rankings. Whereas outliers, e.g. those results only retrieved by one of the sample images, will be degraded in the aggregated ranking. The intuition is that even though they match the textual tag, their content might not match the concept behind the query. Figure 3 shows a diagram of the aggregation process. For each of the samples their ranked list of results is presented. Every result image is assigned points according to their position in the list. This is illustrated in the first three rows. Finally, the aggregated results corresponds to the list of candidates, sorted by their total number of points, as in the last row. We can observe that the rank of the image returned in first position is a combination of the partial ranks, and likewise for the subsequent results.

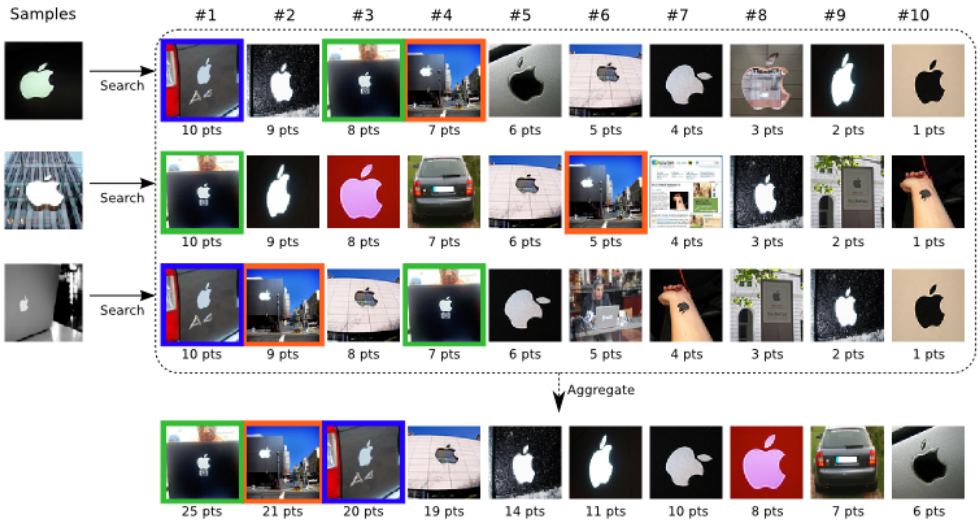


Figure 3: Aggregating the search results for the query "apple logo" using visual annotations.

5. EVALUATION

In this section we describe the set-up and outcome of the retrieval performance experiment that we performed to compare tag-based search, visual search based on sample object annotations, and aggregated visual search based on object annotations. First we address the hypotheses behind the evaluation. We then describe the set-up of the experiment and finally present the results.

5.1 Evaluation task

We formulate the following hypotheses, which we will investigate in the retrieval performance experiment:

H1: Rank aggregation over the results sets of content-based image retrieval with the visual annotations will significantly improve the retrieval performance in terms of precision. The agreement between the different result sets for the partial searches will lead to a more focussed result set for the aggregated result set with a higher precision at the top of the ranking.

H2: Tag-based search combined with content-based image retrieval, using visual annotations will improve the retrieval performance, in terms of precision. When performing a textual search over an image collection a rather diverse set of results will be retrieved, as the annotations are usually very sparse and the textual clues do not allow for visual disambiguation. When searching with visual annotations it is possible to discover the different aspects of an object, and in combination with a filter on the textual annotations we can retrieve more relevant results at the top of the ranking.

5.2 Experimental setup

For the experiment we have defined a set of topics, compared five different systems, collected relevance judgements on the results obtained by the different systems in a TREC-style fashion. Below, details on the different facets of the experiment are described.

5.2.1 Image collection

Different image collections have been used for object recognition, such as the CalTech collection [12], COIL collection [21], and the Corel collection [6]. They are widely used for object classification, recognition, and categorization tasks. The main characteristic of these collections is that they have well defined visual attributes for the objects represented in the images. Usually they contain images with uniform size, and low level of cluttering, which is not coherent with the scenario on the Web, where diversity is present on all possible dimensions. Although some of these collections were created by downloading images from Web pages, they have been manually selected to match a set of constraints. In our work we will be focusing on images with high variability, Web-extracted, and annotated by Web-users. For this reason, instead of using one of the previous collections, we used images that are collected from Flickr [11], without manually selecting them.

The collection contains 12,000 images that were crawled through the public Flickr API, based on a set of tags, that corresponds with the topics that are used for the experiment. As a result we obtained a set of images that at least had one of the tags, but we made no restriction on whether they were relevant to their surrounding tags, or whether the object actually appeared on the image. In addition we collected the title, tags, and description for each of the photos. The collection contains 59,693 unique tags (from a total of 229,672 tags). Photos in Flickr are made available in various resolutions, ranging from thumbnail size to the original size uploaded by the user. To leverage the number of features that can be extracted from the image and its corresponding processing time, we downloaded the medium size image, which have a resolution of at most 500x333 pixels.

5.2.2 Topics

We have pooled a set of 30 topics, which where derived from Flickr search logs. The queries were sorted by descending frequency, and we filtered them for objects. We can basically overlay the topics with four broad categories:

Topic	Description
American flag	Picture of a cloth-made American flag.
Big Ben clock tower	View of the clock tower.
Arc de Triomphe	Front view of the arc.
Clock	Round mechanical clock.
Coke can	Can of coke.
CN tower	View of the skypod.
Dice	Any view of a dice.
Eiffel tower	Picture of the tower, taken from the base.
Engagement ring	Upper view, containing a stone.
Guitar	Body of a classical or electric guitar.
Soccer ball	Picture of an official-size soccer ball.
Statue of Liberty	Top view of the Statue of liberty.
Apple logo	Logo from Apple brand.
Rose	Top view of a rose.
Parthenon	Front facade.
Strawberry	Picture where the skin of the fruit is clearly shown.
Daisy	Top view of a daisy.
Moai	At least one visible Moai statue.
Sunflower	Top view of a sunflower.
Sushi roll	Piece of a cut sushi roll.
Golden Gate bridge	View of at least one of the main pillars.
McDonald logo	Big "M" from the McDonald logo.
Taj Mahal	Taj Mahal front facade.
Hot air balloon	Fully inflated hot air balloon without the basket.
Petronas Twin Towers	View of both towers with the sky-bridge between them.
Telephone booth	Classic UK red telephone boxes.
Butterfly	Picture containing the butterfly's wings.
Converse	Converse sneakers.
Watermelon	Watermelon showing the skin.

Table 1: List of topics.

fruits & flowers, *monuments & buildings*, *brands & logos*, and *general objects*. Table 1 shows the list of selected topics. For each topic a short description is defined that details the visual requirements, which can not be easily expressed in keyword-based search. This additional information will be used to guide the assessors in their judgements.

In addition to the topic descriptions, we provide a visual example for each topic, as depicted in Figure 4. Finally, for each topic a set of 10 visual annotations is created that is used to feed the content-based image retrieval system with the visual examples. For example, see the annotations shown in Figure 1.

5.2.3 Systems

For the experiment we can differentiate five variants of our system (S1-5). Each system uses as input a keyword-based query, and returns a ranked list of image results.

S1: *Text-based retrieval.* The textual baseline for our experiment is based on the vector space model for text retrieval. Using the textual annotations (tags) of the images related images are retrieved for a given keyword-based query, by measuring the cosine-similarity between the query and the image annotations.

S2: *Content-based image retrieval using visual annotations.* This system uses the keyword-based query to select (at random) one of the ten visual annotations that matches the query. Based on the extracted visual features that are within the bounding box of the visual annotation related images are retrieved, as described in detail in Section 3. As we are selecting visual annotations at random for each topic, we constructed 25 random runs for which we report the average performance over the 25 repeated measurements in the results section.

S3: *Aggregated ranking over the results of content-based image retrieval using visual annotations.* For this system, we search with all 10 visual annotations and apply rank aggregation over the 10 partial result lists that are computed for each topic, as discussed in more detail in Section 4. The top 25 results of the 10 partial rankings is used as input for the aggregation step.

S4: *Content-based image retrieval using visual annotations and a tag filter.* The approach of this system is similar to system S2, with an additional filter over the image annotations, which requires that the tags match with all the query terms.

S5: *Aggregated ranking over the results of content-based image retrieval using visual annotations and a tag filter.* The approach of this system is similar to system S3, with an additional filter over the image annotations, which requires that the tags matches with all the query terms.

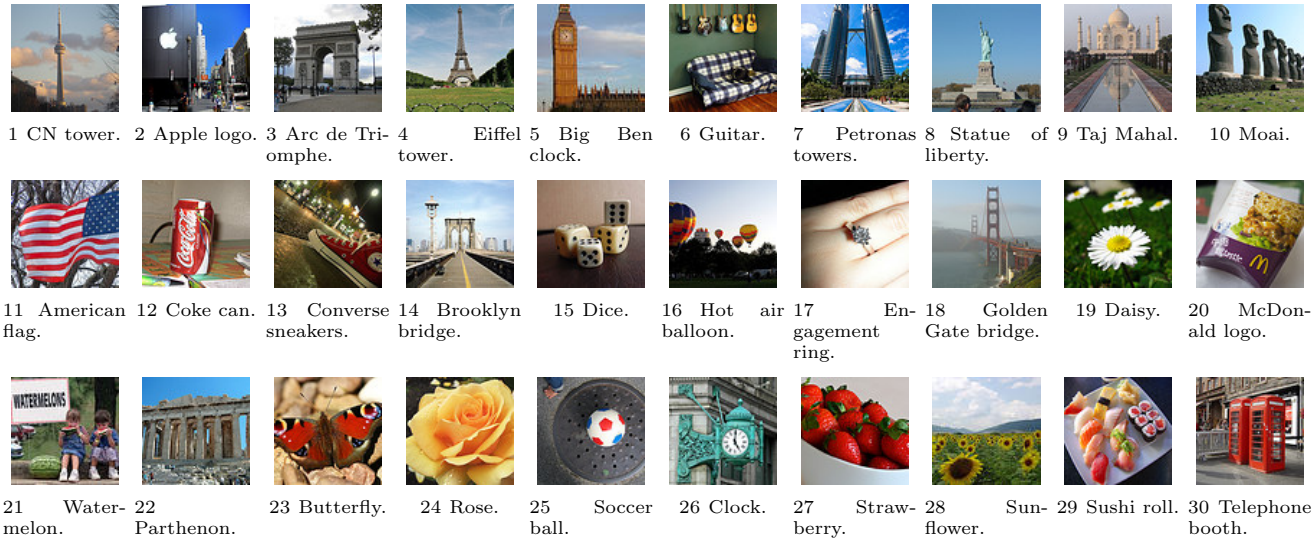
Comparison of system S2 versus S3 (or S4 versus S5) allows for testing hypothesis H1, which states that the retrieval performance benefits from the rank aggregation over the partial results obtained by the visual annotations. Likewise, the comparison of S1 with S4 and S5 allows us to test hypothesis H2, where we are interested in improving the retrieval performance by combining visual and textual search.

5.2.4 Pooling and assessments

We have implemented a blind review pooling method, as is commonly used in TREC [26]. The topic pools are based on the top 25 results for each topic retrieved by each of the systems. For the systems S2-5 we have pooled by selecting the top 25 results for each visual annotation, and we have included a separate run for each of the three features (Harris, Hessian, and combined). The assessors were asked to judge the relevance of the results for a given topic on a binary scale, and they were instructed to take the information provided by the topic description into account. The assessment interface provided the assessor with the image, title, tags and description.

5.2.5 Evaluation measures

In this experiment we were mainly interested in achieving a high precision at the top of the ranking and not so much in recall. In the results section we therefore focus on P@N, with N ranging from 1-25, which allows us to investigate the quality of the ranking at early cut-off. Furthermore, we will report mean average precision (MAP) and binary preference (BPREF), which is claimed to be more stable for incomplete judgements [4].



1 CN tower. 2 Apple logo. 3 Arc de Triomphe. 4 Eiffel tower. 5 Big Ben clock. 6 Guitar. 7 Petronas towers. 8 Statue of liberty. 9 Taj Mahal. 10 Moai.

11 American flag. 12 Coke can. 13 Converse sneakers. 14 Brooklyn bridge. 15 Dice. 16 Hot air balloon. 17 Engagement ring. 18 Golden Gate bridge. 19 Daisy. 20 McDonald logo.

21 Watermelon. 22 Parthenon. 23 Butterfly. 24 Rose. 25 Soccer ball. 26 Clock. 27 Strawberry. 28 Sunflower. 29 Sushi roll. 30 Telephone booth.

Figure 4: Topic image examples.

5.3 Results

5.3.1 *Feature selection*

Before addressing the main research questions, we have to analyze the retrieval performance when varying the feature selection. In Section 3 we have identified two features, Harris affine (HAR) and Hessian affine (HES), and a linear combination (COM) of the two features as our feature space. The feature selection affects all systems that use the visual search (S2-5). In table 2 we present the performance of each of the four systems with the different features.

The values in bold indicate the best performing variant per system for each of the three measures (MAP, BPREF, and P@10). Though the differences are not significant, the combined (COM) approach, where the two feature spaces are concatenated, clearly is the preferred method according to all the measures for each system. For the discussion of the results, we will therefore limit ourselves to the combined variant.

5.3.2 *Summary statistics*

Table 3 presents the summary statistics of the retrieval performance experiment of the five systems. Each of the systems returned the top 25 results for the 30 topics, except system S4 and S5, where the filtering had a small impact on the number of results retrieved. From the pool of nearly 9,000 images, 2,187 images were judged relevant. This indicates that there is a large diversity in the results returned by the different systems. Based on all four measures presented in the table, i.e. MAP, BPREF, P@5 and P@10 respectively, we can conclude that S5, the system that is based on aggregated ranking over the results of content-based image retrieval with a tag filter, clearly outperforms the other systems.

System	S1	S2	S3	S4	S5
Number of Topics	30	30	30	30	30
Images Retrieved	750	750	750	742	748
Relevant	2187	2187	2187	2187	2187
Relevant Retrieved	393	149	301	494	562
MAP	0.12	0.05	0.12	0.2	**0.24**
BPREF	0.2	0.07	0.15	0.26	**0.3**
P@5	0.53	0.34	0.55	0.72	**0.82**
P@10	0.49	0.31	0.48	0.71	**0.8**

Table 3: Summary statistics.

5.4 Precision at early cut-off

Figure 5 plots the graphs for precision at various cut-off points (PN). The graphs allow for a more detailed analysis of the systems and their ability to rank relevant results near the top of the ranking. For S1, the tag-only run, we find that the performance slightly decays from 0.57 to 0.49. As expected, the performance for S2, the system that uses content-based image retrieval with visual annotations, is lower than for S1 and ranges from 0.36 to 0.20. The results for system S3 show that the results can be significantly improved by performing rank aggregation of the results obtained for S2. With the precision ranging from 0.63 to 0.40, the precision is almost twice as high. In fact, the relevancy of the top 5 results is even higher than for the tag-only run.

The system variants S4 and S5 combine visual search with a textual filter. As shown in the figure, this leads to another significant increase in retrieval performance over the tag-only system S1 and the systems S2 and S3 that only use the visual features. We find that the precision over the top 25 ranges from 0.74 to 0.66 for S4, and that for S5 the precision is always higher than 0.75. We can therefore conclude that in all cases rank aggregation over the result sets for content-based image retrieval with visual annotations leads to a significant increase in retrieval performance as posed in hypothesis H1. Furthermore, the combined visual and textual approach shows significant improvements over the tag-only system, therefore we can validate hypothesis H2.

5.5 Topic analysis

In the final part of the evaluation, we put forward a topic analysis to detect whether the observations of the previous two sections are caused by abnormalities in the performance

Table 2: Retrieval performance for different features.

System	S2			S3			S4			S5		
Feature	COM	HAR	HES	COM	HAR	HES	COM	HAR	HES	COM	HAR	HES
MAP	**0,05**	0,04	0,04	**0,11**	0,10	0,10	**0,2**	0,18	0,19	**0,24**	0,23	**0,24**
BPREF	**0,07**	0,05	0,05	**0,14**	0,12	0,13	**0,26**	0,24	0,25	**0,30**	0,29	0,29
P@10	**0,31**	0,26	0,27	0,48	**0,49**	0,48	0,71	0,69	**0,72**	**0,80**	0,77	0,79

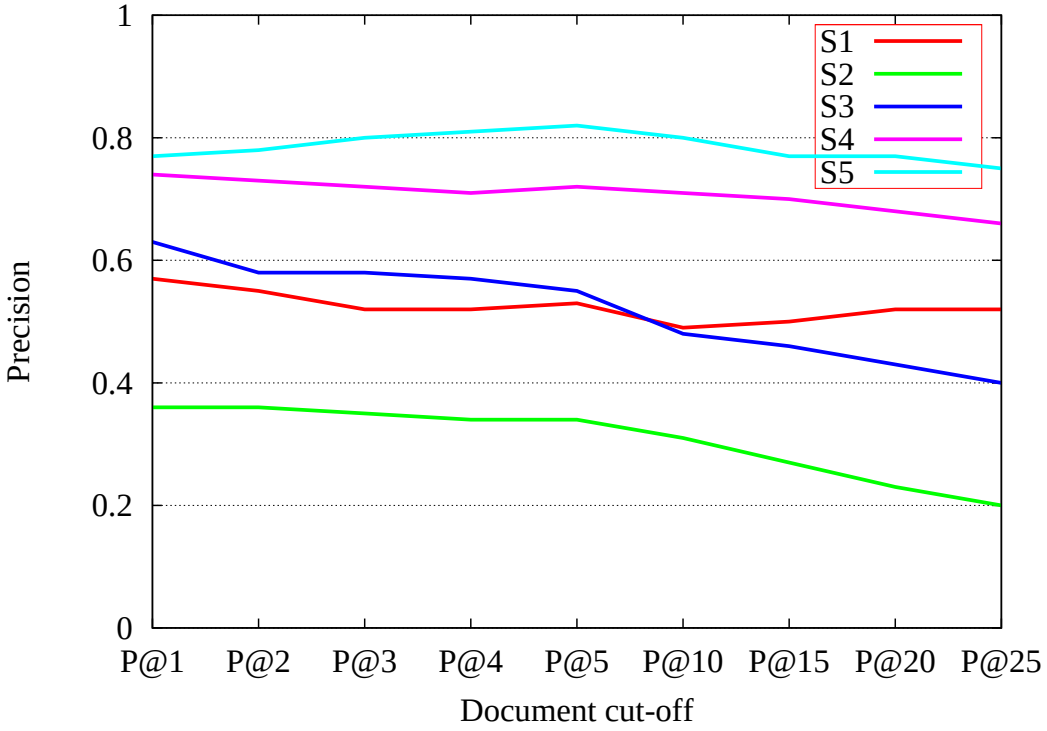


Figure 5: Precision at early cut off; systems overview.

for a subset of the topics. Figure 6 provides a topic histogram for the P@10. On the x-axis the P@10 (0.0 -10.0) is projected, while the y-axis projects the number of topics with the same P@10 rounded to one decimal precision.

For system S1 the average P@10 is 0.49, with a standard deviation of 0.24, while the average P@10 for system S5 is 0.8 with a standard deviation of 0.19. This indicates that there is a significant and uniform increase in retrieval performance for all topics.

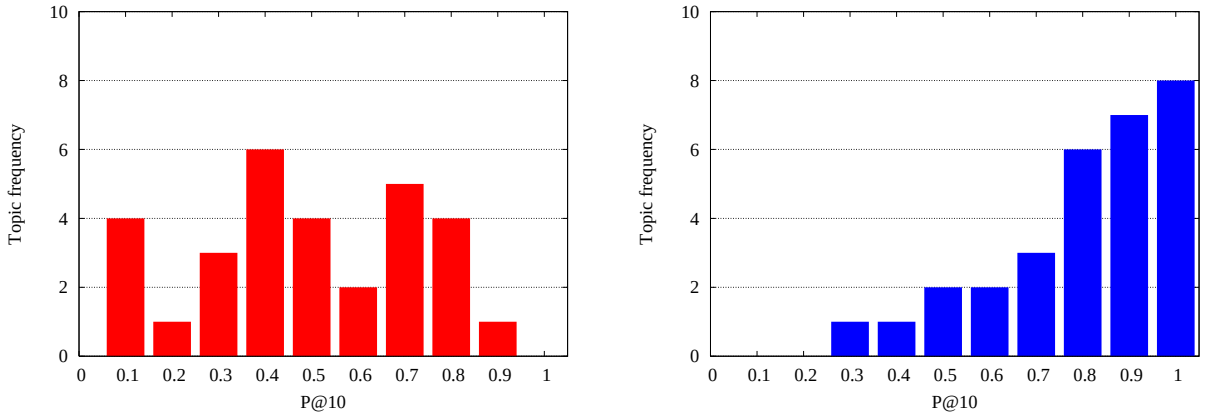
Finally, Figure 7 plots the MAP in a histogram for each individual topic per system. It allows for a per-topic comparison. A number of observations can be made. First of all, it reveals that S5 and S4 are consistently better than S1. However, the performance on a number of topics is weaker when no textual information is present to limit the search space, see for instance the performance of the topics: "butterfly", and "watermelon" with systems S2 and S3. An explanation is that those images (or visual annotations) contain many small non-characteristic visual words, which can easily be mistaken.

6. CONCLUSIONS

In this paper we have studied the problem of key-word based image retrieval on a diverse image collection, such as typically found in on-line photo sharing services. The available human annotations allow for existing text retrieval models to work on such large corpora, but due to the sparsity of the information provided with the photos these models are not optimal.

Central in our research was the question: "How can we deploy the visual annotations, also known as "notes" in Flickr, to enhance the retrieval performance?". In more detail, we have proposed to use rank aggregation to combine the result sets of a content-based image retrieval system that uses the visual annotations to retrieve similar images. The results of the retrieval performance experiment clearly showed that the quality of the results significantly improves when applying the rank aggregation on the results obtained with the content-based image retrieval system. Moreover, the results of our aggregated visual search show a marginal improvement when compared with the tags-only run. When extending the visual search with a textual filter on the tags we can further limit our search space, and show another significant boost in retrieval performance in terms of precision.

For future work, we plan to deploy alternative aggregation strategies that can be applied in a pre-retrieval fashion, rather than post-retrieval. The advantage would be a speedup in the retrieval process. In addition, we are in-



(a) P@10 for system S1 (per topic)

(b) P@10 for system S5 (per topic)

Figure 6: P@10: Precision after having seen the first ten results for systems S1 and S5.

terested in detecting different senses of the same keyword using a visual analysis. It would allow for decision-based diversity in the search results. Typical examples are of course "jaguar" and "apple". Last but certainly not least, we plan to investigate scalability issues with the content-based images retrieval techniques as presented in this paper.

7. ACKNOWLEDGMENTS

The images used in this paper were downloaded from Flickr and were posted by cindy47452, SolGrundy, wallyg, Atelier Teee, Bludgeoner86, ivanx, matsuyuki, borkurdotnet, dan.blanachard, riebschlager, Ctd 2005, engelcox, triciaward, Essjay in NZ, wallyg, Anushruti R, p!o, navonod, davesag, Zeetz Jones, marymactavish, selva, Blacknell, Wysz, Hearlover1717, el_en_houston, nutmeg66, kaneda99, foreversouls, xolivare, alexi1982, Fleur-Design, bolti22, triciaward, John Gevers, powerbooktrance, Steve Rhodes, Neil101, theparadigmshifter, larsomat, mundocuadro, xgravity23, Heavenbound, neiljmh, gspidermac.net, morebouncetotheounce, jthorstad, flex, richevenhouse, Jesman, Felix63, Platform 3, Mickeleh under Creative Commons (CC) license.

8. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human Factors in computing systems*, New York, NY, USA, 2007. ACM Press.

[2] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284, New York, NY, USA, 2001. ACM.

[3] J. C. Borda. Memoire sur les elections au scrutin. In *Histoire de l'Academie Royale des Sciences*, 1781.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA, 2004. ACM.

[5] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*, 2007.

[6] Corel clipart & photos. http://www.corel.com/products/clipartandphotos/, 1999.

[7] G. Csurka, C. Dance, J. Willamowski, L. Fan, and C. Bray. Categorization in multiple category systems. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 745–752, New York, NY, USA, 2006. ACM Press.

[8] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *ACM Trans. Web*, 1(2):7, 2007.

[9] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[10] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the workplace web. In *WWW*, pages 366–375, 2003.

[11] Flickr. http://www.flickr.com.

[12] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[13] A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 627–634, New York, NY, USA, 2007. ACM.

[14] V. Lepetit, P. Lagger, and P.Fua. Randomized trees for real-time keypoint recognition. In *Proceedings of Computer Vision and pattern Recognition (CVPR2005)*, San Diego, USA, June 2005.

[15] R. Lienhart and M. Slaney. Plsa on large scale image databases. In *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, 2007.

[16] D. Lowe. Distinctive image features from

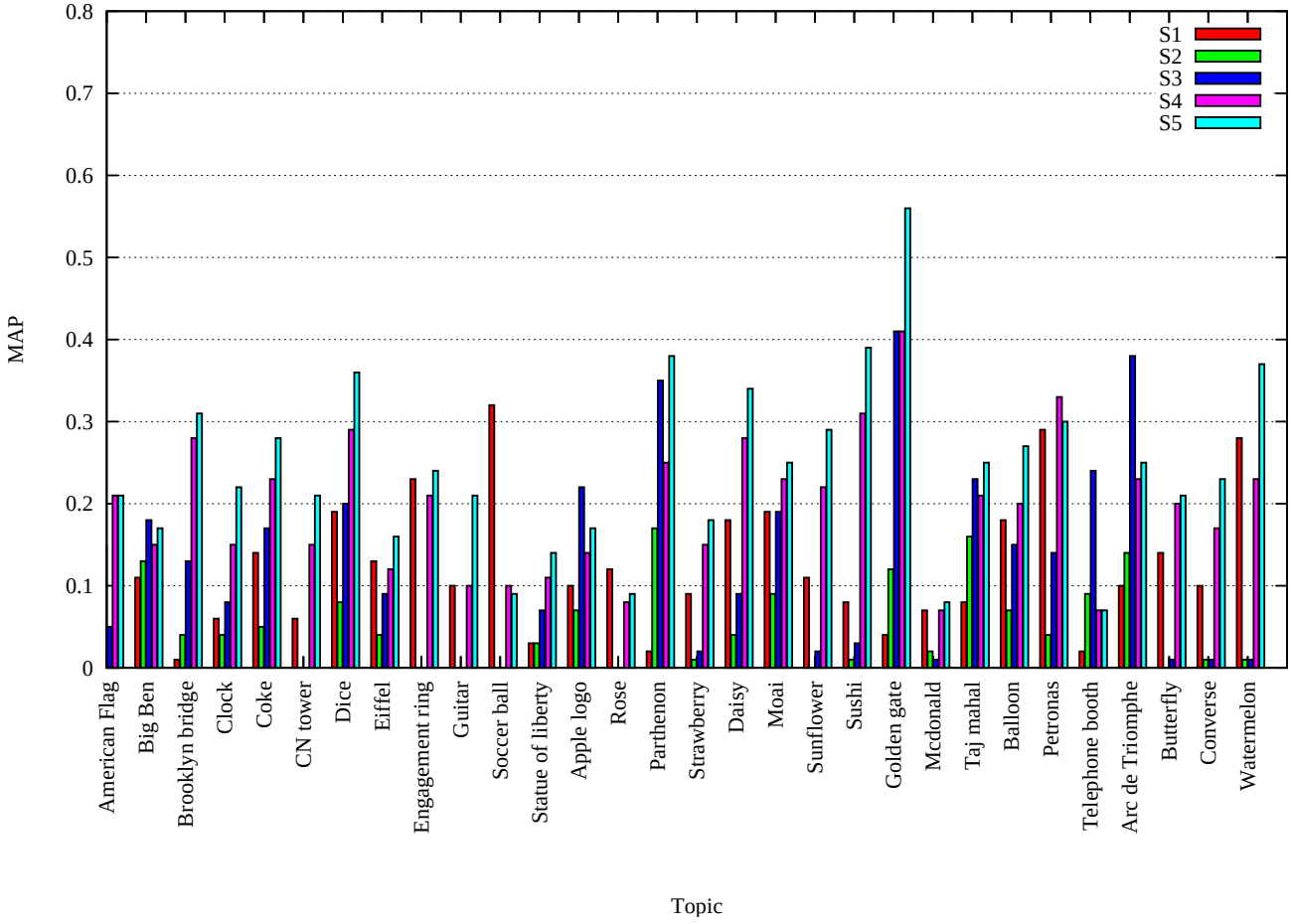


Figure 7: MAP histogram, per topic.

scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

[19] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, October 2004.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005.

[21] S. Nene, S. Nayar, and H. Murase. Columbia object image library: Coil, 1996.

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[23] Picasa. http://picasaweb.google.com.

[24] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 127–144. Springer, 2006.

[25] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9(5):975–986, 2007.

[26] Text retrieval conference homepage. http://trec.nist.gov/.