Vocabulary-Supported Image Retrieval

Bernt Schiele and Julia Vogel
Perceptual Computing and Computer Vision Group
Department of Computer Science
ETH Zurich, Switzerland
{schiele,vogel}@inf.ethz.ch

Abstract

Today's content-based image retrieval systems (CBIR) mostly rely on a predefined set of low-level image features and incorporate user-interactions using techniques such as relevance feedback. These systems however do not take advantage of the fact that in many applications queries can be formulated using a vocabulary. In this paper we propose a general framework which allows to use vocabulary at several levels. The framework should be seen as an extension of today's CBIR systems enabling the use of vocabulary as well as online learning techniques such as relevance feedback. The image detectors supporting the vocabulary can be either implemented directly or learned offline from examples and user-interactions.

1. Introduction

During the last years content-based image retrieval (CBIR) systems gained more and more attention in both the computer vision and the database research communities. This research is triggered by the fact that digital libraries of images and video are rapidly growing in size and availability. In order to avoid the expense and limitations of text annotations, there is considerable interest in navigation by perceptual attributes which can be extracted automatically from images.

The declared goal of CBIR research is to enable the design and implementation of general purpose tools for image retrieval. This general formulation does not allow to make any assumption about the image content or the images themselves to be analyzed. The space of image representations explored traditionally has been therefore restricted to those of a generic nature. Consequently, many CBIR systems only use low-level image descriptors. The assumption these CBIR systems rely upon however is that an optimal combination of generic image descriptors is sufficient to satisfy the query and intention of the user. Unfortunately, it is far from obvious that the direct combination of low-level image descriptors will be powerful enough to find any object or express any concept, the user might be interested in.

On the other hand CBIR systems have access to feedback from their users that can be exploited to simplify the task of finding the desired images. That feedback is used to find the optimal combination strategy of low-level features. This is done either manually by an experienced user or via a relevance feedback mechanism. In both cases the user has to develop an understanding of the underlying data structures and algorithms. When the user has to choose the combination of features manually it is clear that the user has to understand their meaning in order to obtain an effective combination strategy. But even in the case of relevance feedback, the user has to develop a certain understanding of the system in order to give feedback the system can exploit. Thus, often too much is asked for of an average user.

Current CBIR research can be divided into three main categories. The first set of systems makes only use of low-level features such as color, texture, shape or spatial information for describing the images' content [1, 2, 3]. Generally, the user chooses which features and feature weights to utilize or to combine. In the second category of CBIR systems, the expertise of the user is integrated in the retrieval loop. These relevance feedback mechanisms allow the user to specify relevant and non-relevant images given by the system. The advantage is that the retrieval converges more directly to the desired target and that possible changes of the users' goal are detected. In [4, 5], queries are shifted due to relevance feedback, whereas the systems in [6, 7] formulate the feedback and retrieval procedure as a Bayesian inference problem. The PicHunter [8] system belongs to a third category since it incorporates another important aspect into the retrieval process: the user model. Here, the next set of results is not only based on the user feedback, but also on a probabilistic model of the users' behavior.

None of these systems however makes use of the following: Thinking about the humans' way to describe images and their content, it can be observed that in many applications the user can formulate the query using a set of words or, more general, by using a particular vocabulary. We therefore propose to extend today's retrieval systems by mechanisms which support the direct use of vocabulary. Since the use of vocabulary will be intuitive to most users we expect to improve the usability of today's systems substantially. Obviously, useful vocabulary will be - at least to a certain degree - application-dependent. Consequently a retrieval system should not only provide a basic set of vocabulary useful for many applications. Additionally, the system should be able to learn new vocabulary from the user or even allow the user to define new vocabulary. Learning of words and concepts should be simple and intuitive in the sense that an unexperienced user can train the system to support new, application specific vocabulary. In our opinion teaching of particular words and concepts will be more suitable than general relevance feedback mechanisms for two reasons. The first reason is that the user can give explicit examples of a particular word or concept off-line and independent of a particular retrieval session. By applying the learned concept immediately to new images the user can directly verify the quality of the new word or concept. The second reason is that relevance feedback mechanisms attempt to learn the intention of the user form a small number of interactions. The main difficulty of this task is that the intention of the user is often composed of several concepts and words at the same time.

It is expected that vocabulary-supported queries will improve the usability of CBIR systems. In general however, the vocabulary cannot be complete and will be limited. The aim of this papers is therefore to extend the capabilities of today's retrieval systems rather than replacing them. It is important to note that the general framework proposed here includes not only low-level features and standard relevance feedback mechanisms but also supports the direct use of vocabulary, vocabulary learning and long-term learning.

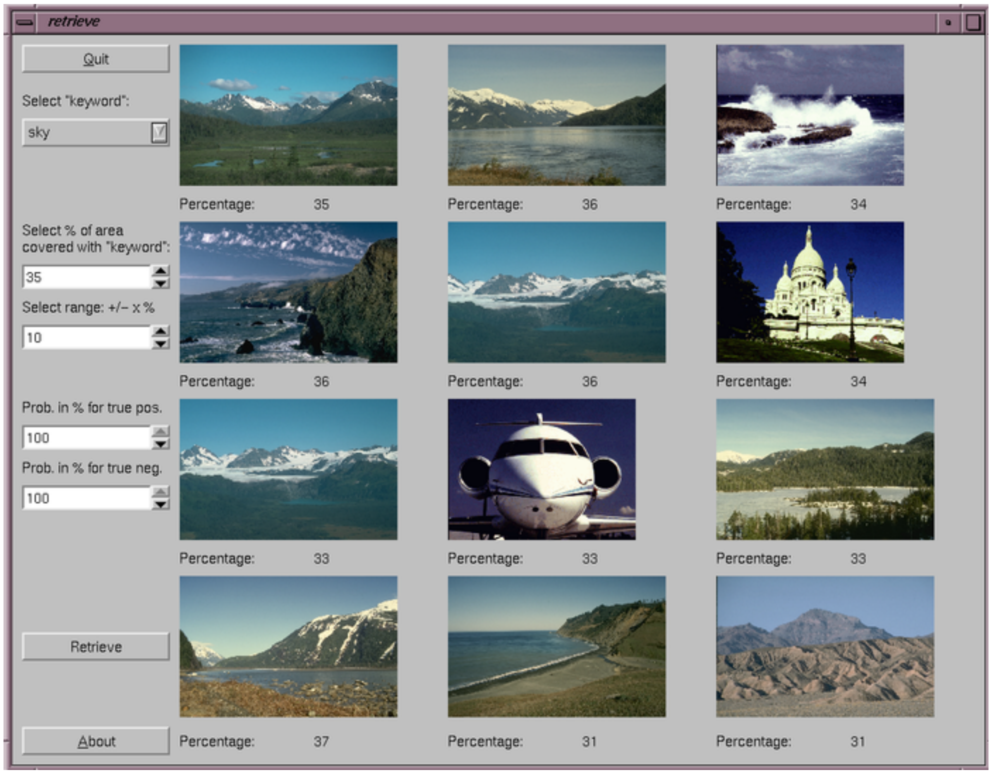
2. General Framework

The system of vocabulary-supported queries can be thought of as a front-end or addition to any standard image retrieval system. It enables the user to formulate queries in a more intuitive and easy-to-learn way. Also, the concept of using vocabulary to describe images provides a framework for learning new detectors and improving the ones currently available. In the following, we first describe the concept of vocabulary-supported queries in more detail. Subsequently, we discuss the embedding of vocabulary-supported queries into a general multi-stage image retrieval system.

2.1. Vocabulary-Supported Queries

A human trying to describe to another person what kind of image he or she is looking for may use the following or similar syntax: "I am looking for a picture that is outdoors with some water, sky and also some forest on it. But I don't want any mountains." The same syntax will be used in vocabulary-supported queries. The user specifies a set of vocabulary that has to appear in the image. In our current implementation the user can also specify the percentage of the image that should be covered by a particular concept. Additionally, the user may indicate concepts that may not appear in the image. In general, the combination of the vocabulary is based on the boolean operators AND, OR and NOT. Fig. 1 shows a screen-shot the of our current system. The interface of the system allows to specify keywords and the percentage range of the image which should be covered.

In order to enable the use of vocabulary, the system has to provide image descriptors and detectors supporting the vocabulary. These detectors may either be implemented specifically as it may be meaningful in the case of face detection. Alternatively, detectors can be learned or improved by means of learning. For several object classes such as faces or people there exist elaborate detectors with high detection rates [9, 10, 11]. Besides these detectors, several authors have shown that one can relatively easily learn detectors for concepts such as sky, grass, buildings, cars or indoor vs. outdoor scenes [12, 13]. Even though one can implement several detectors such as face and people detectors specifically, most detectors supporting vocabulary will be application-specific and therefore have to be learned. Addition-



Select "keyword": sky

Select % of area covered with "keyword": 35

Select range: +/− x % 10

Prob. in % for true pos. 100

Prob. in % for true neg. 100

Figure 1. Screen-shot of the current interface to the vocabulary-supported CBIR system. Images are returned according to the amount of "sky". Note the unusual diversity of the images all containing the concept "sky". The diversity of the images is due to the fact that the query is based on vocabulary rather than on low-level features directly.

ally, the system should enable the non-expert users to increase the number of detectors depending on their application and need.

2.2. Multiple-Stage Image Retrieval System

The before mentioned tasks of vocabulary-supported queries, learning and improving of image detectors and, in addition, relevance feedback can be mastered in a multiple-stage image retrieval system.

The first stage consists of a purely vocabulary-based query where the user attempts to specify his or her query using the available descriptive vocabulary. The result of the first stage is therefore a reduction of the overall number of images the user is potentially interested in. From an abstract point of view this first stage can be seen as a pre-filtering step. We expect that already the most basic vocabulary consisting of only a small number of words will enable to reduce the search space considerably. The expected benefit of this first stage may compensate for the added effort required for the the vocabulary.

The second stage consists of one or several refinement steps where the user can make extensive use of the vocabulary. In the classical relevance feedback scenario the user can give positive as well as negative feedback depending on his or her intention. This feedback is given most often globally for entire images. In order to refine the query the task of the underlying learning algorithm is to find out which parts of the image caused the positive or negative feedback. In the proposed system the user is not restricted to this global feedback. The user may also specify which of the words contained in an image are relevant and which are not. In that sense, vocabulary enables the user to specifically state which local parts of an image are important and which are not. Besides enabling more specific feedback we expect that most users will find the proposed feedback supported by vocabulary more intuitive.

The third and final stage of the system builds upon standard relevance feedback mechanisms. This stage is necessary since the vocabulary will most often not suffice to specify the intention of the user

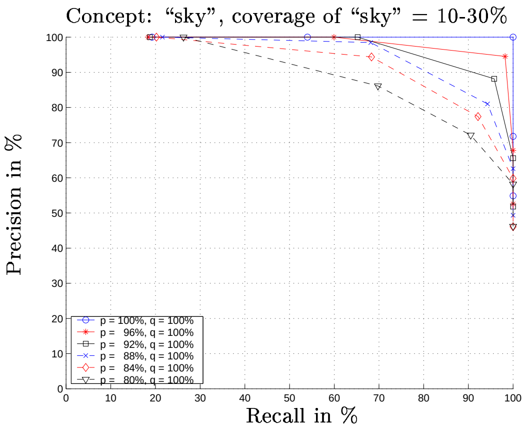


Figure 2. Precision vs. Recall with varying probability p (prob. for true pos.)

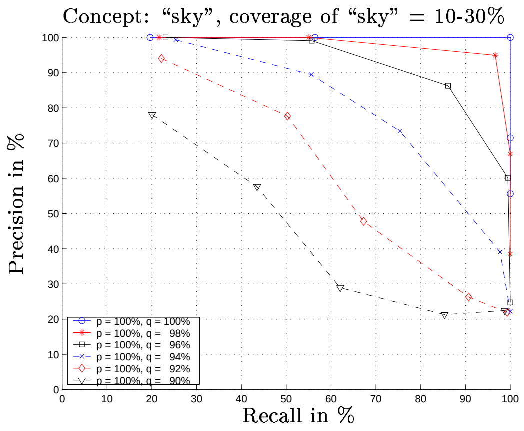


Figure 3. Precision vs. Recall with varying probability q (prob. for true neg.)

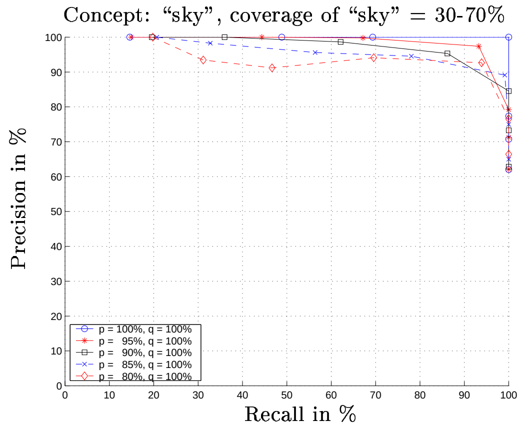


Figure 4. Precision vs. Recall with varying probability p (prob. for true pos.)

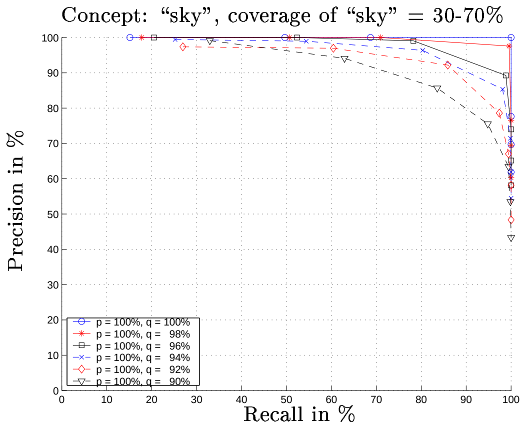


Figure 5. Precision vs. Recall with varying probability q (prob. for true neg.)

in a detailed manner. It is expected however, that the first two stages already reduced the number of potential images considerably. This will facilitate a much faster convergence of standard relevance feedback mechanisms.

Learning occurs at least during stage two and three. In the case the system does not know the vocabulary the user employed, an online-learning or in-session learning takes place. Via the relevance feedback given by the user it is possible to gain a representation of the desired concept. This representation can be used in future retrieval session. At the same time, the information provided by the feedback of the user can be employed in a long-term learning or between-session learning. Thus, concepts already known to the system and often asked for by users can be improved by between-session Bayesian inference (see [14]).

3. Experimental Results

The heart of vocabulary-supported image retrieval are the detectors for the objects or concepts the user might be looking for in an image. As mentioned before, these detectors can be implemented directly, e.g. for often used, common concepts, or be learned during an on- or offline stage. It is important to note that most vocabulary detectors are local by nature. This implies that the decision whether a certain concept exists in a given image is made locally. In our current implementation this decision is made for each image patch or segment. Thus, it is possible for the user to decide which area of the image should be filled by or correspond to a concept.

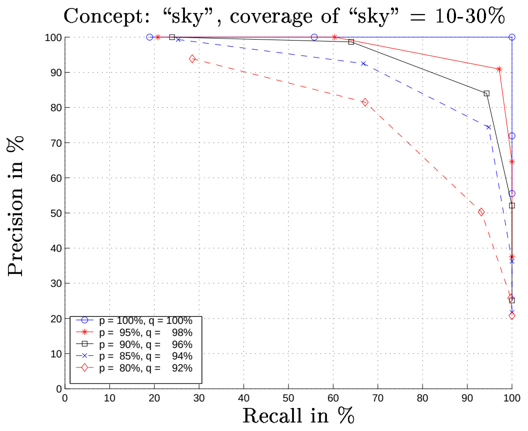


Figure 6. Precision vs. Recall with varying probabilities p (prob. for true pos.) and q (prob. for true neg.)

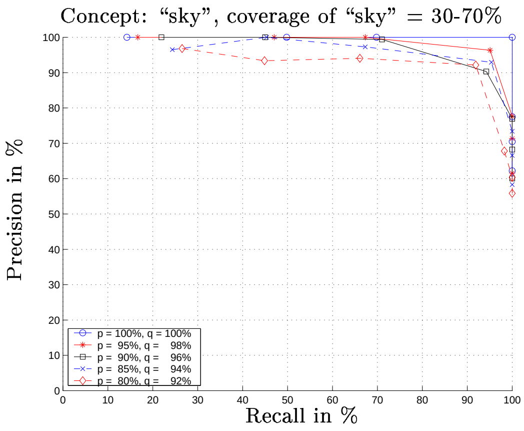


Figure 7. Precision vs. Recall with varying probabilities p (prob. for true pos.) and q (prob. for true neg.)

As introduced in the previous section vocabulary mainly used for stage one (query-formulation) and stage two (query-refinement). As pointed one of the potential benefits of vocabulary-supported queries is that they may reduce the search space significantly. The ultimate goal of any detector should be not only to obtain high detection rates but also to enable to find all images which are possibly interesting to the user. That means that our primary goal will be detectors with close to 100% recall rate and simultaneously keeping the precision of the detectors in a reasonable range.

The goal of the following experiment is to find out how good the performance of the vocabulary detectors needs to be in order to satisfy the retrieval requirements. Thereby, the retrieval requirements are measured by precision (percentage of the retrieved images that are relevant to the query) and recall (percentage of the relevant images that are retrieved). In order to have ground truth available, a set of 160 random images was annotated. The images were segmented into 100 patches and manually labeled whether the segments contained the concepts of “sky”, “water”, “grass”, “buildings”, “cars” or “faces”. Based on these annotations, the first stage of vocabulary-supported image retrieval with the user interface of Fig. 1 was implemented. Here, the user selects a concept and a percentage range that specifies the area of the image that has to be covered by the concept.

In order to simulate the influence of the detection probability on precision and recall, the annotations were randomized by two probabilities: p denotes the probability for the correct detection of true positives. If $p = 100\%$, the detector labels all and only these segments that contain the particular concept. q denotes the probability for the correct detection of true negatives. Or, in other words, $1 - q$ is the probability for the detection of false positives. Figs. 2 to 7 show the precision vs. recall-curves for several values of p and q and for different percentage ranges. The results for the concept “water” are similar to the shown results of “sky” since for both concepts the amount of relevant images in the annotated database is large enough to have statistically relevant results.

The comparison of Figs. 2 to 5 exemplifies the different influence of p and q. In the case of a small desired coverage such as $20 \pm 10\%$ in Figs. 2 and 3, the influence of q is much larger than the influence of p. The reason is that for a small desired coverage the absolute amount of segments that might be influenced by q is up to 10 times larger. The influence of p and q is more similar for the symmetric desired coverage of $50 \pm 20\%$ in Figs. 4 and 5. Here, the small amount of relevant images in the database and their non-equal distribution over the accepted coverage range leads to the unexpected behavior in Fig. 4.

The first stage of vocabulary-supported image retrieval includes a restriction of the search space for future retrieval iterations. For that reason, it is important to know which limitations in precision have to be tolerated in order to have a maximum recall. In Figs. 6 and 7 both p and q are varied. Here, it can be seen that the behavior of precision and recall is again very dependent on the desired coverage. In Fig. 6, a recall of more than 90% leads already to the low precision of 50%. In contrast, for the desired coverage $50 \pm 20\%$, more than 98% of recall can be reached with a precision of nearly 70%. On the other hand, knowing these differences, the retrieval algorithm can react according to the coverage desired by the user. If the user is looking for only a small coverage (e.g. $20 \pm 10\%$) , the system might increase the search range in order to obtain a satisfying recall (e.g. to $20 \pm 14\%$ for recall rates larger than 90% in Figs. 2 and 3). The same might be reasonable if it is know that the detection probabilities p and q for certain detectors are below average.

4. Conclusion

This paper develops a general framework for vocabulary-supported image retrieval. After discussing the current state of the art the paper identifies the utility of vocabulary for content based image retrieval systems. The proposed framework consists of three main stages: the first stage enables the user to formulate his or her query directly using vocabulary, the second stage allows to refine the query again using vocabulary and the final stage consists of a standard relevance feedback mechanism. The main advantage of the proposed framework is that the use of vocabulary is intuitive for most users. Furthermore, vocabulary enables to reduce the search space considerably which may result ins faster convergence of the system. The experimental section gives quantitative data about the desired quality of detectors in order to obtain an appropriate reduction of the search space.

References

[1] M. Flickner et al., "Query by image and video content: the QBIC system", *IEEE Computer Magazine*, vol. 28, no. 9, pp. 23–32, 1995.

[2] A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases", *SPIE Storage and Retrieval of Images and Video Databases II*, February 1995.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying", submitted to PAMI.

[4] Y. Rui, T.S. Huang, S.Mehrotra, and M. Ortega, "A relevance feedback architecture for content-based multimedia information retrieval systems", in *IEEE Workshop on Content-based Access of Image and Video Libraries*, 1997.

[5] J. Assfalg, A. Del Bimbo, and P. Pala, "Image retrieval by positive and negative examples", in *International Conference on Pattern Recognition (ICPR'00)*, Barcelona, Spain, 2000.

[6] N. Vasconcelos and A. Lippman, "A bayesian framework for content-based indexing and retrieval", in *IEEE Data Compression Conference*, 1998.

[7] Ch. Meilhac and Ch. Nastar, "Relevance feedback and category search in image databases", in *IEEE Intl. Conference on Mulitmedia Computing and Systems*, June 1999.

[8] I.J. Cox, M.L. Miller, St.M. Omohundro, and P.N. Yianilos, "PicHunter: Bayesian relevance feedback for image retrieval", in *IEEE Intl. Conference on Pattern Recognition*, 1996.

[9] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation", *Pattern Recognition and Machine Intelligence*, vol. 19, no. 7, 1997.

[10] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection", in *International Conference on Computer Vision (ICCV'98)*, Bombay, India, January 1998, pp. 555–562.

[11] C. Papageorgiou and T. Poggio, "Trainable pedestrian detection", in *International Conference on Image Processing (ICIP'99)*, Kobe, Japan, October 1999.

[12] R.W. Picard and T.P. Minka, "Vision texture for annotation", *ACM Journal of Multimedia Systems*, 1995.

[13] M. Szummer and R.W. Picard, "Indoor-outdoor image classification", in *IEEE International Workshop on Content-based Access of Image and Video Databases*, Bombay, India, January 1998.

[14] N. Vasconcelos and A. Lippman, "Learning over multiple temporal scales in images databases", in *European Conference on Computer Vision*, 2000.