ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT

Pattern Recognition III (IIII) III–III

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Performance evaluation and optimization for content-based image retrieval

Julia Vogel[a,*], Bernt Schiele[b]

[a]Department of Computer Science, Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
[b]Department of Computer Science, Darmstadt University of Technology, Germany

Received 30 November 2004; received in revised form 28 October 2005; accepted 28 October 2005

Abstract

Performance evaluation of content-based image retrieval (CBIR) systems is an important but still unsolved problem. The reason for its importance is that only performance evaluation allows for comparison and integration of different CBIR systems. We propose an image retrieval system that splits the retrieval process into two stages. Users are querying the system through image description using a set of local semantic concepts and the size of the image area to be covered by the particular concept. In Stage I of the system, only small patches of the image are analyzed whereas in the second stage the patch information is processed and the relevant images are retrieved. In this two-stage retrieval system, the retrieval performance, that is precision and recall, can be modeled statistically. Based on the model, we develop closed-form expressions that allow for the prediction as well as the optimization of the retrieval performance. As shown through experiments, the retrieval precision can be increased by up to 55% and the retrieval recall by up to 25% depending on the user query.

Keywords: Content-based image retrieval; Performance prediction; Performance characterization; Computer vision; Image semantics

1. Introduction

Since long-time performance evaluation of computer vision algorithms has been recognized as being of utmost importance for the advancement of the field [1,2]. The goal of performance evaluation or characterization here refers to the analysis of the quality, not the speed, of a particular vision algorithm. Usually, this goal requires the generation of large benchmark sets with hand-labeled or synthetically produced ground truth, a very tedious process often leading to inconsistent annotations. However, this effort is undertaken regularly for some computer vision and pattern recognition applications such as tracking and surveillance with the PETS workshop started in 2000 [3], document analysis with the TREC seri1s carried out for the first time in 1992 [4], face recognition [5], and image flow, vehicle detection, symbol and shape recognition [6]. Only recently, as part of TREC 2001, a video track devoted to the research in automatic segmentation, indexing, and content-based retrieval of digital video was put together [7]. In addition, the problem of performance evaluation and validation was the topic of several dedicated workshops [8–10] and discussions [11–13].

Also in the context of content-base image retrieval (CBIR), performance evaluation of the proposed retrieval systems is essential as argued by Smith [14], or Müller et al. [15]. Performance evaluation is indispensable since it allows the comparison of different systems and the analysis of how those systems perform depending on the application. Knowing the performance of different algorithms and systems also permits their combination and their integration into larger and more powerful CBIR systems. Some systems, for example, are better to coarsely limit the search space and therefore might be used as the front-end of a larger system. Other systems, on the other hand, might work well on small, preprocessed subsets of the database.

* Corresponding author. Laboratory for Computational Intelligence, Department of Computer Science, University of British Columbia, Vancouver, Canada. Tel.: +1 604 822 6281; fax: +1 604 822 5485.
E-mail addresses: julia.vogel@alumni.ethz.ch (J. Vogel), schiele@informatik.tu-darmstadt.de (B. Schiele).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.10.024

However, evaluation of content-based image retrieval systems is particularly difficult. On a complete system level, the ultimate performance measure should be user-centric and related to user satisfaction. Obviously, such performance measures are not only difficult to define but will also vary greatly between individual users, tasks and applications, and even between sessions with the same user. Establishing generally accepted ground truth in that context seems to be tedious if not impossible due to the required amount of consistently annotated images. Images are very complex carriers of information. Thus, the hand-annotation of images and the definition of benchmark sets is not only application- and user-dependent but also often ambiguous. In addition, the semantic gap between the user's image understanding and the low-level image representation of the computer complicates the definition of benchmarks.

Recognizing the fact that performance evaluation is extremely important and at the same time very difficult, this paper introduces a two-stage retrieval system which makes performance characterization manageable. The key idea is that the user queries the database with high-level semantic concepts [16,17]. An immediate benefit is that for those local semantic concepts the acquisition of large amounts of consistent ground truth is feasible because the semantic content of the concepts is less complex than that of full images. In addition, the two-stage retrieval process allows to model the retrieval results statistically. Based on the statistical model, we develop closed-form equations for the prediction of retrieval precision and retrieval recall. In the two-stage retrieval system it is also possible to optimize the retrieval performance. The performance optimization uses the results of the performance prediction. Depending on the user query, the optimization step resets internal parameters in order to increase precision and/or recall of the retrieval.

In general, the intention of our system is to unburden the user from any parameter settings or complicated query modes. The query based on local semantic concepts is similar to the way humans describe images. Thus, with the two-stage retrieval system the semantic gap between the image understanding of the user and the computer decreases. Other systems propose relevance feedback to capture the user's high-level query and perception subjectivity (e.g. [18,19]). The downside of these approaches is that the user is required to undergo possibly several rounds of feedback, and that it is not possible to predict or optimize the retrieval quality beforehand. In our system, the performance optimization step is transparent to the user. The retrieval performance increases without requiring the user to set any parameters. In addition, many current retrieval systems follow the query-by-example paradigm in which the user is searching for images based on an example (for an overview of the current state of CBIR research refer to Ref. [20]). Our system is based on the idea that the user describes the desired image using a set of semantic concepts.

The rest of the paper is organized as follows: Section 2 introduces the two-stage retrieval system and the employed query mode. In Section 3, closed-form expressions for the retrieval performance in the two-stage system are derived. The performance of the system can be optimized in three ways: in Stage II of the system (Section 4), in Stage I of the system (Section 5) or jointly in both stages (Section 6). In Section 7, an additional, approximate optimization method is introduced that does not require specific database information. The proposed methods are summarized and discussed in Section 8.

2. Two-stage retrieval system

In the proposed retrieval system, users describe the images they are looking for by using a set of local semantic concepts (e.g. 'sky', 'water', 'building', 'rocks', etc.) and the size of the image area to be covered by the particular concept. Thus, an exemplary query might be: "Search images with 20–40% of 'sky' ". Fig. 5 depicts exemplary retrieval results for that query. Note that here, due to the semantic query mode, the concept 'sky' corresponds to very different occurrences of sky (e.g. clear sky, cloudy sky, overcast sky, etc.). The interval-based query mode might seem artificial at first sight. However, the user interval could also be mapped to descriptors such as "very little", "half of", "most of", etc. In addition, the combination of the search for several concepts in the same images leads to a powerful global image representation that can be used for scene categorization or retrieval as shown [21].

The technical realization of the retrieval is split into two stages (see Fig. 1). In order to enable the use of concepts for querying, the system provides a set of so-called concept detectors. In Stage I of the system, the database images are analyzed by these concept detectors. They return a binary decision whether a particular image region contains the concept (positive patch) or not (negative patch). In the current implementation, each image is subdivided into a regular grid of patches each comprising 1% of the image. However, the system can be extended to arbitrary patch sizes. In Stage II, the patch-wise information of the concept detectors is processed according to the user interval to actually retrieve a set of images. The performance *optimization* affects the selection of the appropriate concept detector in Stage I and the setting of an internal parameter, the so-called system interval $S = [S_{low}\%, S_{up}\%]$, in Stage II. Here, the main idea is to internally adapt the system interval in order to compensate for some of the concept detectors' errors and to thus optimize the system performance.

3. Performance prediction

The goal of the performance prediction is to make a forecast on the performance of the retrieval depending on certain parameters. We define the retrieval performance by precision, which is the percentage of the retrieved images that are also relevant, and recall, which is the percentage of the relevant images that are retrieved. In the remainder of this

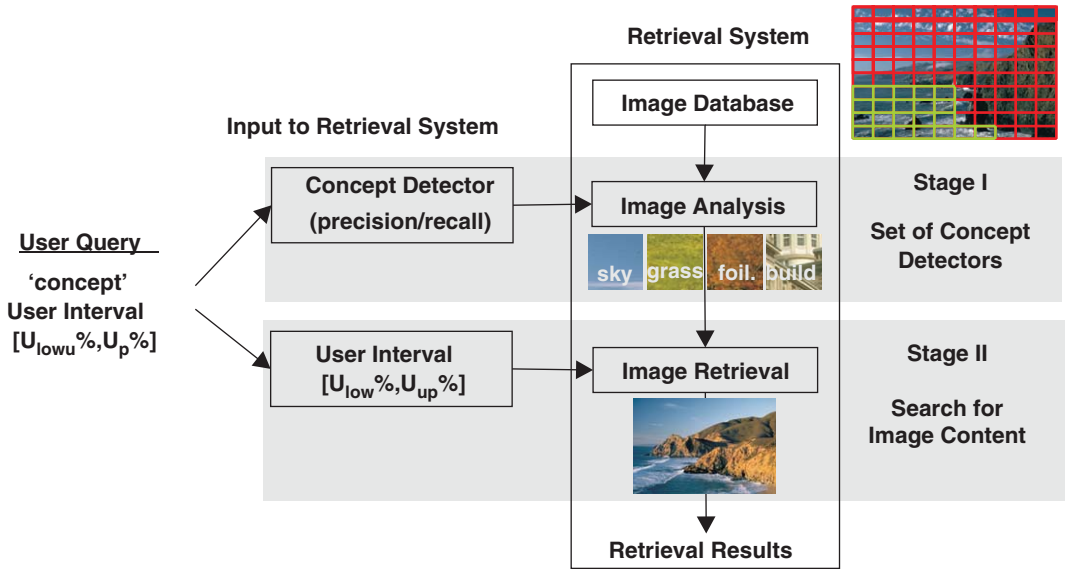

Fig. 1. Two-stage retrieval system.

section, the prerequisites and the derivations for the performance prediction are derived separately for each stage.

3.1. Performance of the concept detectors

A valid user query consists of the concept being searched for and a user interval $U = [U_{low}\%, U_{up}\%]$ specifying the amount of the image to be covered by the concept. Since we divide the image into a grid of 10×10 image patches, U_{low} and U_{up} also correspond to the *number* of image patches covered by the desired concept. In Stage I of the retrieval system, there exists a multitude of detectors for various concepts. This implies especially that there might be multiple concept detectors for one single concept with different performance characteristics. According to the user query, the appropriate concept detector is selected from those detectors, the image patches are analyzed and the classification results per patch are passed to Stage II. This analysis stage can be performed off-line. The performance characteristics of the concept detectors are modeled by the probability p for correctly detecting a positive patch (true positives) and the probability q for correctly detecting a negative image region (true negatives).

The concept detectors are usually trained off-line. In Section 5, we will discuss the learning of the concept detectors in more detail. The goal is to have multiple concept detectors with varying performance characteristics for each concept. This can be obtained by using one classifier with different confidence thresholds or by using different classifiers. In general, any classifier with known performance characteristics can be employed such as for example the semantic classifiers of Town and Sinclair [22] or the texture models for automatic annotation of Picard and Minka [23].

3.2. Mathematical framework

The prerequisite for the performance prediction is that the performance of the employed concept detector is known in the form of the detectors' parameters p and q. For now, we assume the concept distribution $P(N_P)$ to be known. This assumption will be relaxed in Section 4.3. Precision and recall without subscript always refer to the overall retrieval performance.

The derivations are based on the assumption that the concept detectors decide independently on each patch. Thus, the probability $p_{true}(k)$ and the probability $p_{false}(k)$ are binomially distributed. ($N =$ total number of patches, $N_P =$ number of positive patches per image.)

$$p_{true}(k) = \binom{N_P}{k} p^k (1-p)^{N_P-k}, \tag{1}$$

$$p_{false}(k) = \binom{N-N_P}{k} (1-q)^k q^{N-N_P-k}. \tag{2}$$

If a total of i positive patches is to be retrieved, both the true positives and the false positives add up to the total number of detected patches. Thus, the probability to retrieve i positive patches, given that a particular image has in fact N_P true positive patches, is

$$P(N_{retr} = i|N_P) = \sum_{j=0}^{i} p_{true}(i-j)\, p_{false}(j). \tag{3}$$

Similarly, if the interval $U = [U_{low}\%, U_{up}\%]$ of positive patches is searched for, Eq. (3) has to be summed over this interval to obtain the probability $P(N_{retr} \in U|N_P)$.[1]

$$P(N_{retr} \in U|N_P) = \sum_{i\in U} \sum_{j=0}^{i} p_{true}(i-j)\, p_{false}(j). \tag{4}$$

[1] Since $N = 100$ in our experiments, the percentages $U_{low}\%$ and $U_{up}\%$ can be treated as integers in the summations. Otherwise, a normalizing constant would be necessary.

If Eq. (4) is weighted with the concept distribution $P(N_P)$, we obtain the probability to retrieve images (not patches!) that satisfy the query U relative to the image database:

$$P_{retr}(U) = \sum_{N_P=0}^{N} P(N_{retr} \in U|N_P)P(N_P). \quad (5)$$

Precision and recall depend on the probabilities for relevant images $P_{relevant}(U)$ and for true-positive images $P_{true\_pos}(U)$. In order to obtain $P_{true\_pos}(U)$, $P(N_{retr} \in U|N_P)$ is only weighted with the part of the concept distribution $P(N_P)$ that lies inside the user interval $U = [U_{low}\%, U_{up}\%]$:

$$P_{true\_pos}(U) = \sum_{N_P \in U} P(N_{retr} \in U|N_P)P(N_P). \quad (6)$$

The probability that images satisfy the user query depends on the user interval U and the concept distribution:

$$P_{relevant}(U) = \sum_{N_P \in U} P(N_P). \quad (7)$$

Finally, Eqs. (5)–(7) lead to a closed-form expression for the probability of precision and the probability of recall:

$$P_{precision}(U) = \frac{P_{true\_pos}(U)}{P_{retrieved}(U)} = \frac{\sum_{N_P \in U} P(N_{retr} \in U|N_P)P(N_P)}{\sum_{N_P=0}^{N} P(N_{retr} \in U|N_P)P(N_P)}, \quad (8)$$

$$P_{recall}(U) = \frac{P_{true\_pos}(U)}{P_{relevant}(U)} = \frac{\sum_{N_P \in U} P(N_{retr} \in U|N_P)P(N_P)}{\sum_{N_P \in U} P(N_P)}. \quad (9)$$

Thus, with Eqs. (8) and (9), precision and recall of the retrieval can be predicted. The expressions for precision and recall have been validated on a database of 1073 images. The database images have been divided into a regular grid of 10×10 image patches, and the patches have been manually annotated with the concepts 'sky', 'water', 'grass', 'buildings', 'face', and 'car'. All simulations in the following are based on this ground truth. Depending on the selected detector parameter p and q, the annotations are randomly falsified. For the exemplary query $[20\%, 40\%]$ of 'sky' ($p = 0.9$, $q = 0.8$), the prediction is 22.7% precision and 54.5% recall. Simulation of 133 rounds leads to an average of 22.75% precision ($\sigma = 0.65\%$) and an average of 54.52% recall ($\sigma = 1.8\%$) which are close to the predicted value. The higher standard deviation of the recall is due to the smaller amount of images in the estimation of $P_{relevant}(U)$ than in the estimation of $P_{retrieved}(U)$ (compare Eqs. (8) and (9)).

4. Performance optimization in Stage II

In the following three sections, we will introduce several methods for performance optimization in our two-stage retrieval system. The general goal of the performance optimization is to increase precision and recall of the image retrieval. As will be shown, it is possible to optimize precision and recall separately as well as jointly depending on the user's request. In this section, we will only introduce methods that concern Stage II of the retrieval system. In Section 5, we discuss the optimization potential of the concept detectors in Stage I. In Section 6, it will be shown that a joint optimization of both stages is most beneficial. For a schematic view of the retrieval system with performance optimization see Fig. 2.

The derivation of performance optimization for Stage II of the retrieval system requires the introduction of an internal parameter: the system interval $S = [S_{low}\%, S_{up}\%]$. Since the detectors' decisions are only correct with a certain probability, the retrieval performance will vary if the system is queried internally with a query $S = [S_{low}\%, S_{up}\%]$ that differs from the user interval $U = [U_{low}\%, U_{up}\%]$. Intuitively, if the probability is high that the detector makes a false positive decision, it is necessary/sensible to raise the lower limit of the user interval U_{low} to $S_{low} = U_{low} + X$, $X > 0$. The following will formalize this intuition and determine a system interval $S \neq U$ for internal use that optimizes the retrieval performance.

First, Eqs. (4)–(9) have to be extended with the internal parameter S. From now on, the probability P_{retr} to retrieve images depends only on the system query S instead of the user interval U because the actual retrieval of the images in the database is governed only by S.

$$P_{retr}(S) = \sum_{N_P=0}^{N} P(N_{retr} \in S|N_P)P(N_P), \quad (10)$$

where

$$P(N_{retr} \in S|N_P) = \sum_{i \in S} \sum_{j=0}^{i} p_{true}(i-j)p_{false}(j). \quad (11)$$

The probability for true-positive images P_{true_pos} depends on both S and U. The retrieval is *performed* according to the system interval S, but is *evaluated* according to the user interval U:

$$P_{true\_pos}(U, S) = \sum_{N_P \in U} P(N_{retr} \in S|N_P)P(N_P). \quad (12)$$

Eq. (7) remains valid because only the user interval U decides whether an image is relevant for the retrieval. In summary, the probabilities for retrieval precision and recall become (compare Eqs. (8) and (9)):

$$P_{precision}(U, S) = \frac{\sum_{N_P \in U} P(N_{retr} \in S|N_P)P(N_P)}{\sum_{N_P=0}^{N} P(N_{retr} \in S|N_P)P(N_P)}, \quad (13)$$

$$P_{recall}(U, S) = \frac{\sum_{N_P \in U} P(N_{retr} \in S|N_P)P(N_P)}{\sum_{N_P=U_{low}}^{U_{up}} P(N_P)}. \quad (14)$$

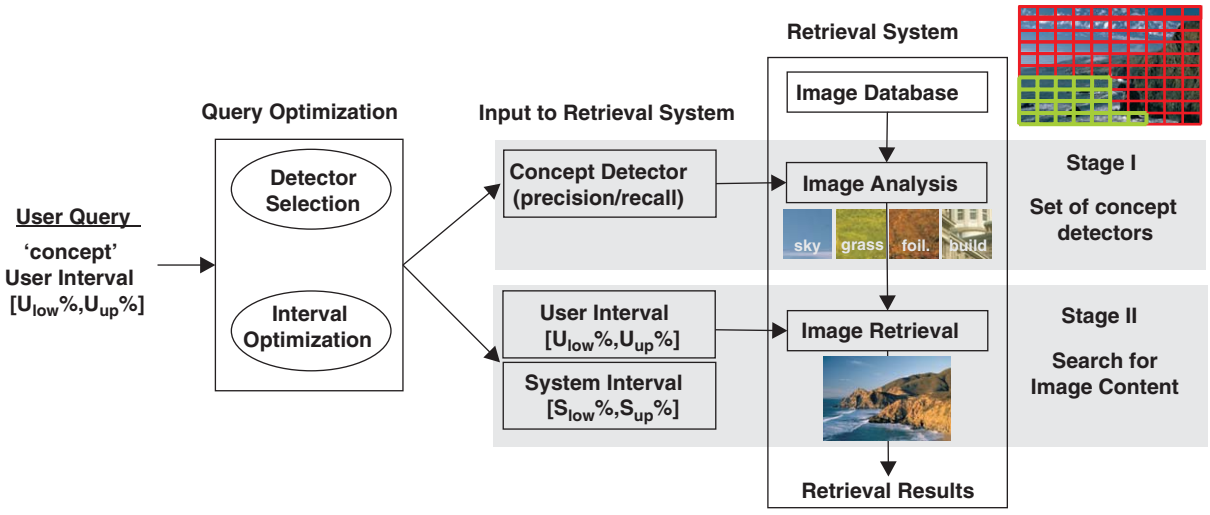

Fig. 2. Two-stage retrieval system with query optimization.

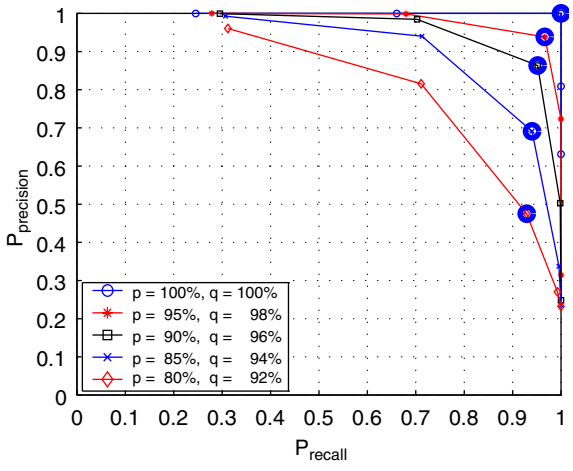

Fig. 3. Prediction of precision and recall for "10–30% *sky*" when varying the detectors' performance p and q and the system interval S.

These two equations are a closed-form expression for retrieval precision and recall when a user interval is given and the system is queried internally using S. In Section 4.1, an algorithm is presented that maximizes Eqs. (13) and (14) recursively and returns the optimal system interval S.

Fig. 3 illustrates the influence of the system interval S. The tested query is "Find images with 10–30% 'sky' ". The five curves correspond to five different sets for p and q as indicated in the legend of the figure. As before the manual annotations were randomly falsified depending on p and q. From left to right S_{low} and S_{up} are varied in the following way: $S = [S_{low}\%, S_{up}\%] \in \{[18\%, 22\%], [14\%, 26\%], [10\%, 30\%], [6\%, 34\%], [2\%, 38\%]\}$ while the user interval is $U = [10\%, 30\%]$ in all cases. As expected, the precision is very high when the system interval is narrow whereas the recall is low. By increasing the width of the system interval, the recall can be increased at the cost of the precision. The decrease of the precision is much faster for smaller values of p and q. This behavior is due to the fact that the user interval covers only about 20% of the image. Thus, the probability for the detection of false positives is much higher than the probability for the detection of false negatives. As a result, many of the retrieved images are not relevant and the precision drops.

4.1. Optimization algorithm

Eqs. (13) and (14) are closed-form expressions for precision and recall depending on the user interval and the system interval. This implies that the equations can be evaluated prior to retrieval making it possible to optimize the expected performance prior to retrieval. Because the equations do not allow us to find a closed-form expression for the system interval $S = [S_{low}\%, S_{up}\%]$ as a function of user interval and desired performance, we use a recursive algorithm for obtaining the system interval that optimizes the retrieval performance. The algorithm allows to choose an optimization constraint: maximum recall, maximum precision, or joint maximization of precision and recall. It is also possible to indicate a minimum value for precision and recall.

The algorithm proceeds in two steps. In the first step, a set of system intervals is generated that are most probably of interest to the user. Starting from the user interval $U = [U_{low}\%, U_{up}\%]$, precision and recall of that point and its four neighbors $[U_{low}\% \pm 1\%, U_{up}\% \pm 1\%]$ are calculated and stored in a hash table. Recursively, those of the four neighbors that improve the current performance are used as a starting point and the hash table is updated. Fig. 4 depicts the complete search space, that is the precision-recall pairs for all possible system intervals, for the query $U = [20\%, 40\%]$ of 'sky' and the detector parameters $p = 0.9$ and $q = 0.8$. Each point in the graph corresponds to a different set of system queries $S = [S_{low}\%, S_{up}\%]$. Note that two points that

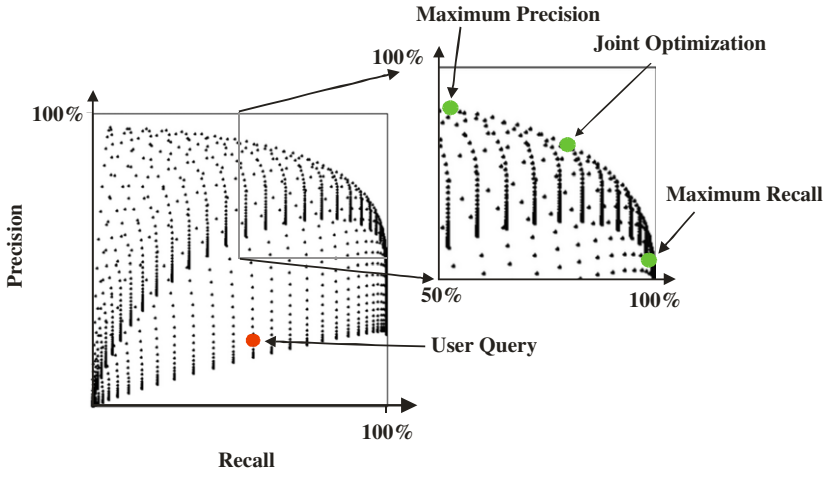

Fig. 4. Predicted search space for [20%, 40%] of 'sky', $p = 0.9$, $q = 0.8$.

are close to each other in the plot do not necessarily have similar system queries.

In the second step, the algorithm selects the point in the search space that meets the users' constraints. The two gray lines in Fig. 4 identify the desired minimum performance of 50%. The predicted performance of the user interval is marked by a black circle while the possible solutions are marked by gray circles. From left to right these are: "Maximum Precision", "Joint Optimization (of Precision and Recall)" and "Maximum Recall".

4.2. Results: Stage II performance optimization

Fig. 5 shows the retrieval results corresponding to the query [20%, 40%] of 'sky' ($p = 0.9$, $q = 0.8$). The user selected the joint optimization of precision and recall. Note the difference to most other retrieval systems. Since here the concept 'sky' is searched for, the retrieved images are very diverse but perfectly satisfy the user query. Only the first 12 retrieved images are displayed. On top of the display some statistics are summarized: the precision was predicted to increase from 22.7% to 80.9%. The actual retrieval resulted in a precision of 80.2%. The recall was predicted to increase from 54.5% to 80.8%. In the actual retrieval, the recall reached 83.2%. Thus, for this particular query, the precision could be improved by 58% and the recall by 25%. The bars visualize the relationship between non-relevant (dark gray), relevant (medium gray, left of the non-relevants) and retrieved (light gray) images. The length of the bars correspond to the amount of images.

Fig. 6 visualizes some optimization results for the optimization constraint "maximize recall with precision $> 50\%$" for the queries [10%, 30%] 'grass' ($p = q = 0.9$), [30%, 50%] 'grass' ($p = q = 0.8$), and [10%, 30%] 'buildings' ($p = q = 0.9$). The retrieval using the user interval is marked in black whereas optimized retrieval is marked in gray with the arrows pointing from the non-optimized to the optimized case. The optimized system intervals are $S = [17\%, 58\%]$ for query 1, $S = [37\%, 56\%]$ for query 2 and $S = [15\%, 42\%]$ for query 3. They are clearly different from the user intervals. Fig. 6 shows that the optimization constraints have been met. The precision increased significantly at all three queries.

In summary, the experiments of the Stage II performance optimization have two main results. First, depending on the user query, a gain of up to 60% in precision and up to 25% in recall can be reached. These results are obtained by probabilistic analysis and the resetting of an internal parameter, the system interval. The performance gain did not require the use of better classifiers. Second, the experiments show that the predicted value of the performance is closely met by the true retrieval performance.

4.3. Approximate performance optimization

Up to now, the assumption was that the concept distribution $P(N_P)$ is known. Thus, the results of the previous sections were obtained with the complete knowledge about the concept distribution used in Eqs. (13) and (14). Fig. 7a shows the concept distribution $P(N_P)$ of the concept 'sky'. However, it is not realistic to have the entire distribution at hand. So the dependency of the performance prediction and the performance optimization was tested using two approximate distributions. In the first test, the actual distribution of the concepts was completely neglected. Instead, it was assumed that the number of patches per image containing the particular concept, that is the positive patches, are uniformly distributed: $P_{uniform}(N_P) = 1/(N+1)$, where N is the maximum number of positive patches. The distribution is depicted in Fig. 7b.

In the second test, it was assumed that the a priori probability is available if a particular concept is present in an image or not. That leads to a two-class frequency distribution. Class A is the number of images that do not contain the

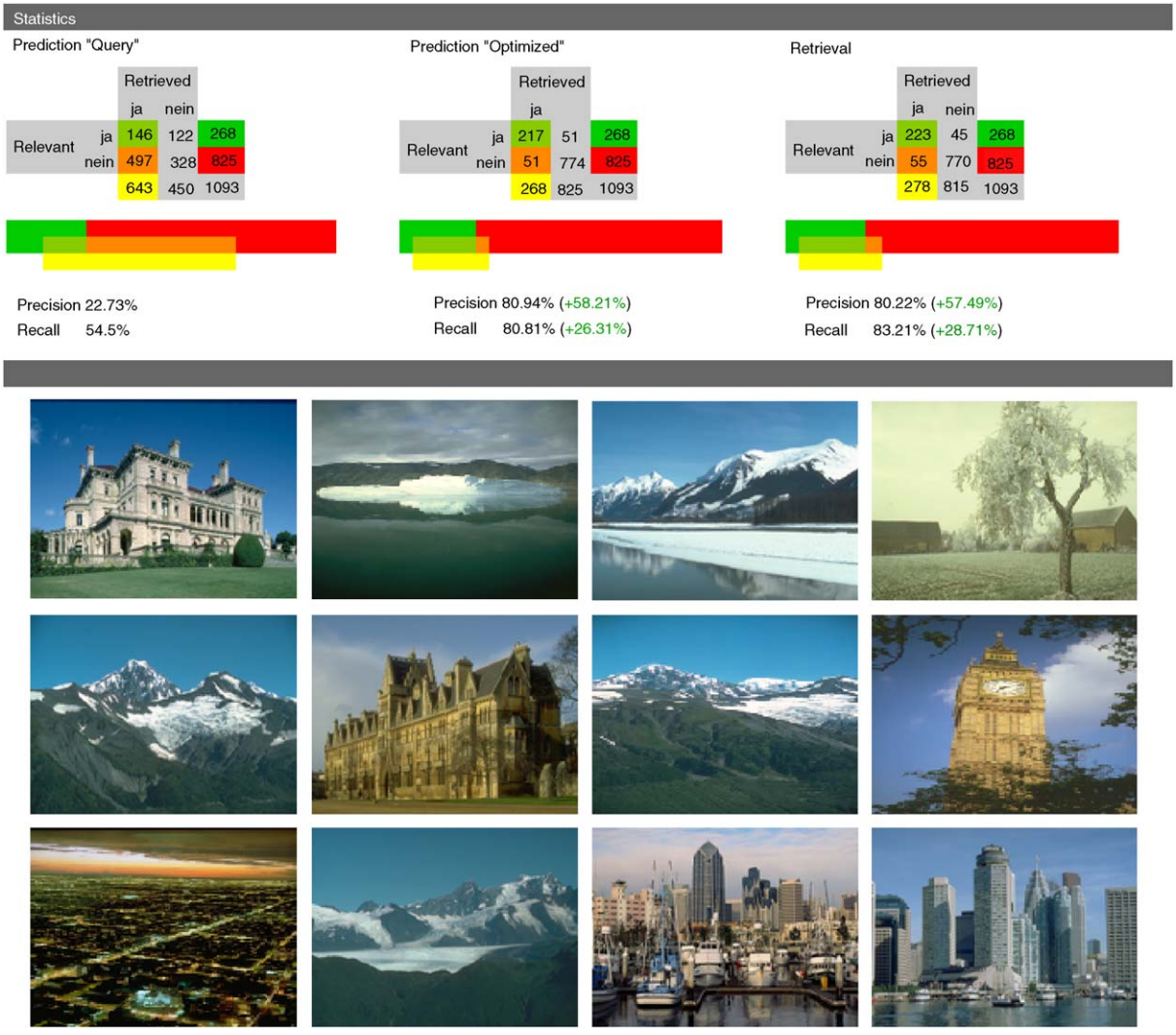

Fig. 5. Retrieval results for [20%, 40%] of 'sky', $p = 0.9$, $q = 0.8$.

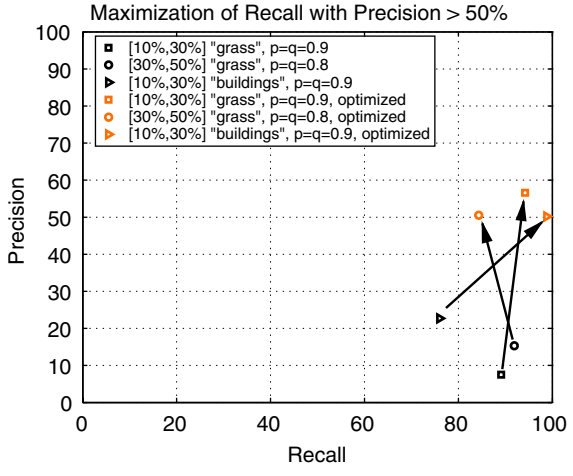

Fig. 6. Retrieval results of three queries: with/without optimization.

concept at all. Class B is the number of images that do contain one or more patches of the desired concept. The uniform distribution of the previous paragraph has been weighted with the two-class distribution. That is, $P_{two\_class}(0)$ contains the information of class A and the information of class B has been divided equally to $P_{two\_class}(N_P)$, with $N_P = 1, \ldots, 100$ (see Fig. 7c).

Both precision and recall are functions of the desired concept, of the detectors' performance specified by p and q and of the user interval $U = [U_{low}\%, U_{up}\%]$. In the experiments, the parameters were varied as follows: concept $\in$ {'grass', 'sky', 'buildings', 'water'}, $p = q \in \{0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$, $U \in \{[10\%, 30\%], [20\%, 40\%], [30\%, 50\%], [40\%, 60\%], [50\%, 90\%]\}$. Table 1 shows the results of some exemplary queries. The performance optimization based on the two approximate distributions is compared to the benchmark optimization results that have been generated with the complete distribution. The goal is to jointly maximize precision and recall. Table 1 shows

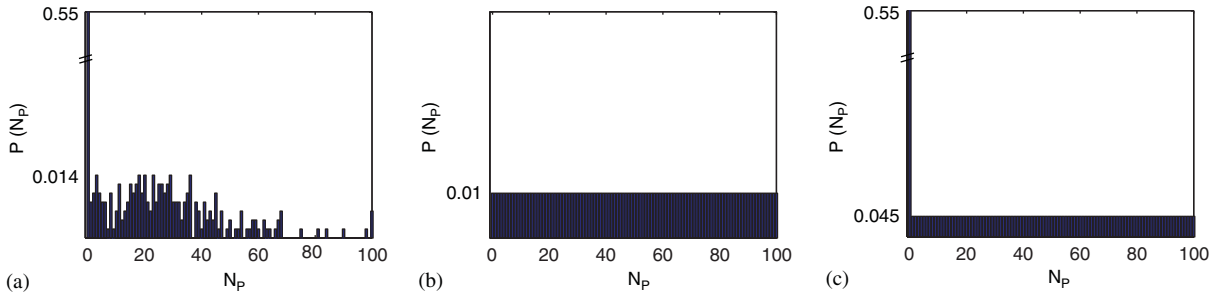

Fig. 7. Complete concept distribution for 'sky' and its approximations. (a) Complete concept distribution "sky"; (b) uniform approximation of (a); (c) twoclass approximation of (a).

Table 1
Uniform/two-class vs. complete distribution: joint optimization of precision and recall

User interval	Employed distribution	System interval (%)	Prediction		Retrieval mean	
			Precision (%)	Recall (%)	Precision (%)	Recall (%)
'Sky'	Complete	[18%, 35%]	85.4	88.9	85.2	88.9
[10%, 30%]	Uniform	[18%, 34%]	85.5	86.4	86.9	86.3
$p = q = 0.90$	Two-class	[18%, 34%]	82.8	86.4	86.7	86.6
'Water'	Complete	[29%, 44%]	78.6	80.8	78.7	81.2
[20%, 40%]	Uniform	[28%, 43%]	77.8	84.6	76.6	83.6
$p = q = 0.85$	Two-class	[29%, 43%]	79.6	81.7	79.8	79.1
'Sky'	Complete	[31%, 47%]	62.9	74.6	62.9	74.6
[10%, 30%]	Uniform	[28%, 43%]	64.0	81.0	53.6	81.3
$p = q = 0.75$	Two-class	[31%, 46%]	49.4	71.9	64.1	71.8
'Grass'	Complete	[41%, 53%]	50.0	67.1	51.2	67.4
[30%, 50%]	Uniform	[39%, 51%]	62.8	77.7	38.4	73.4
$p = q = 0.75$	Two-class	[40%, 51%]	63.1	73.7	45.1	66.9

that the system intervals are always close to the reference result based on the full distribution. That is S_{low} and S_{up} differ by only 1% or 2% from the reference. Accordingly, the results of the actual retrieval are similarly close to the reference retrieval. The performance *prediction* based on the approximate distributions differs from the actual retrieval, especially when the detectors' performance specifiers p and q are small. The prediction of the precision is more sensitive to approximations in the concept distribution than the prediction of the recall. Partly, the difference between prediction and actual retrieval exceeds 20%. Although the performance prediction based on the approximate distribution is not always correct, the results of the actual retrieval are close to the reference results. The reason is that the optimized system intervals are very close to the reference. The correctly estimated system intervals thus lead to a certain robustness with respect to the prediction. Over all 120 experiments, the optimized system interval, and thus the actual retrieval, are slightly better for the two-class distribution than for the uniform distribution.

It can be concluded that the optimized system intervals are so close to the benchmark that the actual retrieval results nearly reach the reference values. This is the case even though the performance prediction based on the approximate concept distributions is worse than in the reference cases. Also, often the difference between the reference values and the retrieval based on the approximate distributions is smaller than the standard deviation of the retrieval. In the case that the true distribution is sparse, the two-class distribution produces better system intervals. The outcome of experiments for other user constraints, such as the maximization of only the recall, is comparable.

5. Performance optimization in Stage I

The performance optimization can be extended to Stage I of the retrieval system. Since this corresponds to the concept detectors, the first part of this section covers the training of the concept detectors and the second part the optimization depending on these detectors. It is desirable to have multiple detectors *per concept* with varying detector performance. Having several detectors per concept, during performance optimization the optimal one of this set can be selected. For training concept detectors with varying performance characteristics, we use AutoClass [24], an unsupervised Bayesian classification system that includes the search for the optimal number of classes.

5.1. Training of the concept detectors

The training of the concept detectors is performed offline. For this purpose, 4000 patches hand-labeled with 'sky', 'water', 'grass' and 'buildings' are used. Therefore, the classes can be very diverse. For example, a 'sky'-patch might comprise cloudy, rainy or sunny sky regions as well as sky regions during sunset. The patches are represented by 4^3-bin RGB-color histograms (col64), 4^3-bin histograms of third-order MR-SAR texture features (tex64) [25] and (2×4^3)-bin histograms (coltex128) that are combined of the 4^3-bin RGB-color histogram and the 4^3-bin texture histograms.

Depending on the feature set, AutoClass finds between 100 and 130 clusters in the data. In a supervised manner it can be determined which of the concept classes are represented by which cluster. Each cluster contains multiple classes resulting in different class probabilities for each cluster. Depending on the feature set, the *highest* class probability in each cluster ranges from 0.25 to 1. The availability of the class probabilities for each cluster provides us with three methods to obtain multiple classifiers. Firstly, in order to improve the precision of the concept detectors, only clusters with a class probability higher than a certain threshold are accepted. Obviously, this leads to a loss in recall. However, precision and recall *of the concept detectors* can thus be precisely controlled. Secondly, the classification using one feature set often performs much better for one class than for another. Thus, it is advantageous to use several feature sets. Thirdly, the classifications of two feature sets can be combined by means of the cluster precision: all cases are classified twice and the vote of the cluster with the higher precision counts.

The performance of various 'sky'- and 'grass'-detectors for different feature sets and feature combinations is shown in Fig. 8. As expected, the feature sets and combinations perform differently for different classes. For the 'sky'-detector, the color feature is not discriminant which lies in the fact, that the 'sky' class is very diverse in color. For the 'grass'-detector, the texture feature fails. This indicates that the employed texture feature catches primarily the structure on small scale and not the larger scale structure that exists in grass patches. The combination of two classifications as described above leads to an improvement in performance. In summary, the "tex64 + coltex128"-detector performs best for 'sky'-patches, whereas 'grass'-patches are detected best with the "col64 + coltex128"-detector.

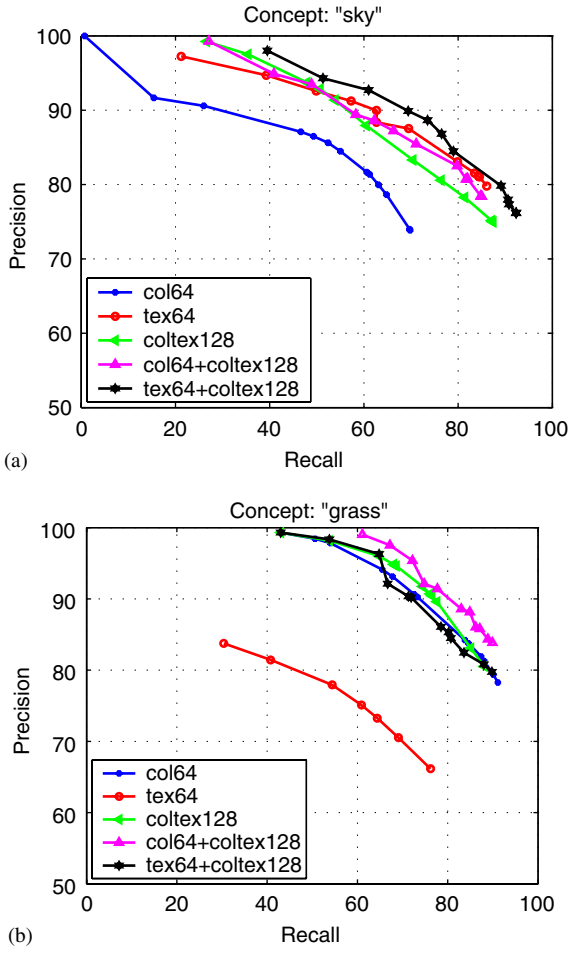

Fig. 8. $Precision_{det}$ vs. $Recall_{det}$ of various detectors. (a) Various 'sky' detectors; (b) various 'grass' detectors.

5.2. Results: Stage I performance optimization

Using the envelopes of the curves in Fig. 8, 13 discrete $\{Precision_{det}, Recall_{det}\}$ pairs that correspond to different concept detectors can be obtained. In order to identify the optimal concept detector for a given user query, Eqs. (8) and (9) are evaluated for each of these detector $\{Precision_{det}, Recall_{det}\}$ pairs.

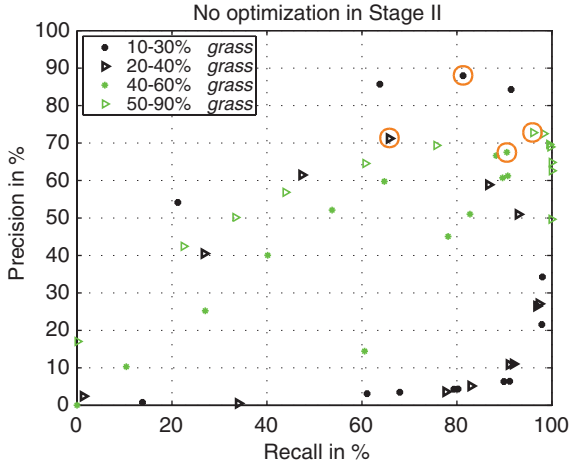

Fig. 9. Retrieval optimization in Stage I: predicted retrieval precision and recall with various 'grass'-detectors.

In Fig. 9 and in Table 2 (middle column), the result of the Stage I performance optimization is summarized. The diagram shows the influence of the detectors on the retrieval

Table 2
Best concept detectors for various 'grass' queries after optimization in Stage I (middle column) and joint interleaved optimization (right column)

Query 'grass'	Optimization only in Stage I		Joint interleaved optimization	
	$Precision_{det}$ (%)	$Recall_{det}$ (%)	$Precision_{det}$ (%)	$Recall_{det}$ (%)
[10%, 30%]	94	72	98	61
[20%, 40%]	94	72	94	72
[40%, 60%]	88	85	94	72
[50%, 90%]	88	83	85	89

performance of a set of 'grass' queries. Fig. 9 shows the predicted overall retrieval performance for each of the four queries and for each of the 13 'grass'-detectors. The points that belong to the same query but different detectors form an ellipsoidal curve. Points in the lower left-hand corner correspond to the *detector* with the highest precision, whereas the points in the lower right-hand corner correspond to *detectors* with low recall. The circles mark the best overall retrieval performance for each query. The corresponding detectors' performances are listed in the middle column of Table 2. Note that the best detector is different for each query: for example, the query [10%, 30%] 'grass' will be executed best with the $\{Precision_{det}=94\%, Recall_{det}=72\%\}$-'grass'-detector and the query [50%, 90%] 'grass' with the $\{Precision_{det}=88\%, Recall_{det}=83\%\}$-'grass'-detector. This supports our intuition that the retrieval performance can be improved by providing multiple detectors for the same concept.

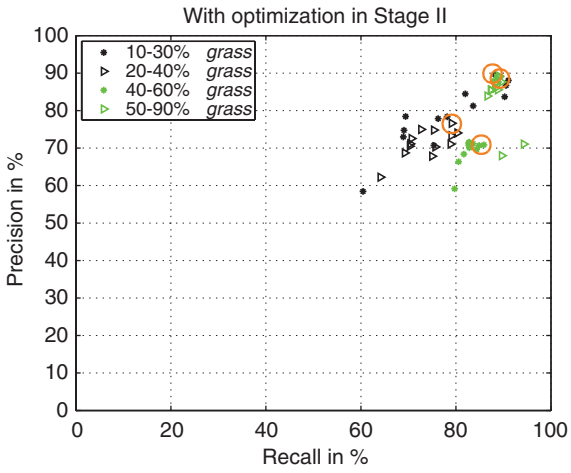

Fig. 10. Joint retrieval optimization in Stages I and II: predicted retrieval precision and recall with various 'grass'-detectors.

Table 3
Comparison of serial vs. interleaved combination of optimization stages

Query 'grass'	Serial combination		Interleaved combination	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
[10%, 30%]	86	90	88	91
[20%, 40%]	77	79	77	79
[40%, 60%]	70	85	71	85
[50%, 90%]	85	87	89	89

6. Joint two-stage performance optimization

There are two methods to combine the optimization stages in the two-stage retrieval system:

- *Serial combination* determine the best concept detector in Stage I as done in the middle column of Table 2. With the performance characteristics of that detector carry out the Stage II optimization in order to find the optimum system interval.
- *Interleaved combination* carry out the Stage II optimization for all detectors that are available for the requested concept in Stage I. Depending on the results, select the optimum system interval *S and* the optimum concept detector for the retrieval.

Fig. 10 corresponds to Fig. 9 *after* the second-stage performance optimization was carried out. The exemplary queries are the same as in Section 5.2. The optimization constraint is "joint optimization of precision and recall". The retrieval performance in Fig. 10 has improved substantially compared to Fig. 9. The circles mark the best overall retrieval performance for each query. The concept detectors corresponding to these best retrieval performances are listed in the right column of Table 2. These results correspond to an interleaved combination of the optimization stages.

For the serial combination of the optimization stages, the detectors of Table 2 (middle column) are used and the Stage II optimization is carried out. The first observation is that for most queries the detector in the middle column of Table 2 is not the same detector as in the right column. Thus, the performance of the overall retrieval will also differ. The overall retrieval performance for serial and for interleaved combination is analyzed in Table 3. As anticipated, the table shows that in all cases the interleaved combination of the optimization stages results in a better retrieval performance of 1–4% increase in precision and 1–3% increase in recall. Obviously, the interleaved combination is computationally more demanding than the serial combination because in the interleaved combination, the optimization algorithm in Stage II has to be evaluated for each detector present for a particular concept. In the case that the application is time critical, it might thus be advantageous to decide for the serial combination despite the lower performance gain.

7. Performance optimization by query mapping

Up to now, the user interval was mapped to an internal system interval in order to compensate for the wrong decisions of the concept detectors depending on the user interval, the detectors' parameters, and the concept distribution. However, the concept distribution is usually not fully available. Therefore, it needs to be estimated or approximated as shown in Section 4.3. Another approach is to only compensate for the probabilistic errors of the concept detectors.

Generally, the decision of a concept detector on a particular patch is only correct with the probabilities p and q. For that reason, the decision on the complete image is also influenced by those two parameters. The influence of p on the decision per image is larger when the user is looking for images covered with the concept by more than 50% and vice versa. In Section 3.2, the behavior of the concept detectors was modeled binomially.

The expected value of a binomially-distributed random variable with parameters n and p is $E(X_n) = pn$. Consequently, the expected values for retrieving true-positive patches (Eq. (1)) and for retrieving false-positive patches (Eq. (2)) are

$$E\{X_{true,retrieved}\} = pN_P, \tag{15}$$

$$E\{X_{false,retrieved}\} = (1-q)(N-N_P) \tag{16}$$

and the expected amount of positive patches that are retrieved out of N_P indeed positive ones is

$$E\{X_{retrieved}|N_p\} = pN_P + (1-q)(N-N_P). \tag{17}$$

We can use Eq. (17) to obtain a mapping from a user interval $U=[U_{low}\%, U_{up}\%]$ to a system interval $S=[S_{low}\%, S_{up}\%]$. Assuming that there are $N_P = U_{low}\%$ of a concept in an image, Eq. (17) returns the percentage of image area that is expected to be retrieved if the detector performs with the parameters p and q. This expected value can be used as new lower limit for the system interval S_{low} because it compensates for the errors of the concept detector. The new S_{low} takes into account that, on average and independent of the concept distribution, the detectors make wrong decisions. The reasoning for S_{up} is analogous.

$$S_{low} = pU_{low} + (1-q)(N-U_{low}), \tag{18}$$

$$S_{up} = pU_{up} + (1-q)(N-U_{up}). \tag{19}$$

Implicitly, Eqs. (18) and (19) are based on the assumption that the concepts are uniformly distributed. Nonetheless, even with this strong assumption, the performance gain is immense (see Fig. 11). For the exemplary query [10%, 30%] 'sky' and $p = q = 0.90$, the mapped system interval is $S = [18\%, 34\%]$ and on average the precision is increased from 41% to 87% and the recall from 77% to 87%. The "optimal" system interval is $S = [18\%, 35\%]$ and the one obtained with the uniform distribution is $S = [18\%, 34\%]$. This shows that the system queries are very similar. It also

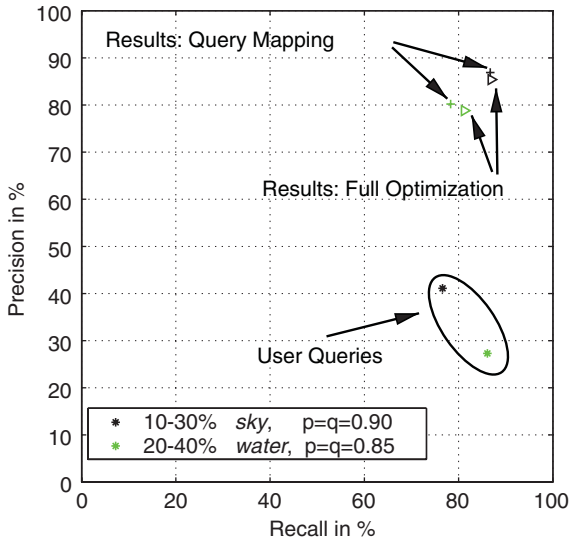

Fig. 11. Optimization by query mapping: comparison to full optimization.

demonstrates the above mentioned assumption. In Fig. 11, the retrieval results of the optimization by query mapping and by using the full concept distribution as a reference are plotted.

The query [20%, 40%] 'water' and $p=q=0.85$ leads to a mapped system interval $S = [29\%, 43\%]$ and to an average increase in precision from 27% to 80%. With the complete distribution the optimized query is $S = [29\%, 44\%]$ and with the uniform distribution $S = [28\%, 43\%]$. The recall decreases in this case on average from 86% to 78%. This example demonstrates the limitations of the query-mapping approach. With the mapping of the user interval to a system interval, precision and recall can be maximized only jointly. This can also lead to a decrease of one of the values. With the algorithms that were presented in Sections 4.1 and 4.3, precision and recall can also be optimized separately which is in many situations more desirable.

8. Discussion and conclusion

In this paper, we introduced a two-stage image retrieval system that allows us to predict as well as to optimize the retrieval performance. In particular, we developed closed-form equations for retrieval precision and retrieval recall that are based on a statistical model of the retrieval process. With the closed-form expressions for precision and recall, the performance of the system can be predicted as well as optimized by adapting an internal parameter according to a set of optimization constraints.

The four prediction and optimization methods are compared in Table 4. The best results for both performance prediction *and* optimization are achieved if the complete concept distribution is available. In that case, the performance can be optimized for maximum precision as well as for

Table 4
Comparison of the methods for performance prediction and optimization

	Optimization precision and recall		Optimization recall	
	Estimated system interval	Prediction	Estimated system interval	Prediction
Complete distribution	++	++	++	++
Two-class distribution	+	Precision: − Recall: +	+	Precision: − Recall: +
Uniform distribution	+	Precision: − Recall: +	∅	Precision: − Recall: +
Query mapping	+	NA	NA	NA

maximum recall and for joint maximization of precision and recall. The predicted performance is always close to the actual one and the determined system interval is indeed optimal. Note that the system interval is the performance measure for the quality of the optimization. Since the complete concept distribution may not be available, the two-class and the uniform distribution have been evaluated for the performance prediction and optimization. Here, again the optimization is possible for all goals: maximum precision, maximum recall, or joint maximization of precision and recall. In nearly all cases the optimized system intervals are so close to the benchmark that the actual retrieval results are similar to the reference values. The performance prediction, however, is not as good as before, because, in particular, the precision prediction degrades. The prediction is slightly more reliable for the two-class distribution than for the uniform distribution since more information is available. In the fourth method, the system interval is obtained through a mapping that depends solely on the detectors' performance values. For that reason, a performance *prediction* is not possible, hindering for example the optimization for maximum recall. Even though absolutely no information about the concept distribution is used, the optimized system interval for joint optimization of precision and recall is as good as with the uniform distribution.

Being able to predict the retrieval performance opens up the possibility of combining our system with other retrieval systems. In particular, the vocabulary-based retrieval is suited as a pre-filtering system to reduce the retrieval search space. For these kinds of applications, a high recall is desirable. The proposed performance optimization method in combination with the "maximum recall"-optimization constraint ensures high recall even for a required minimum precision.

Acknowledgements

This work was supported in part by the CogVis Project, funded by the Commission of the European Union under Grant IST-2000-29375, and the Swiss Federal Office for Education and Science (Grant BBW 00.0617).

References

[1] R. Haralick, Computer vision theory: the lack thereof, in: Proceedings of the Third Workshop on Computer Vision: Representation and Control, Bellaire, MI, USA, 1985, pp. 113–121.
[2] K. Price, I've seen your demo: so what?, in: Proceedings of the Third Workshop on Computer Vision: Representation and Control, Bellaire, MI, USA, 1985, pp. 122–124.
[3] J. Ferryman, J. Crowley (Eds.), Sixth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance PETS-ECCV, Prague, Czech Republic, 2004.
[4] E. Voorhees, L. Buckland (Eds.), The Thirteenth Text REtrieval Conference (TREC 2004), NIST Special Publication 500-261, Department of Commerce, National Institute of Standards and Technology, Gaithersburg, ML, USA, 2004.
[5] P. Phillips, H. Moon, P. Rauss, S. Rizvi, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[6] S. Aksoy, M. Ye, M. Schauf, M. Song, Y. Wang, R. Haralick, J. Parker, J. Pivovarov, D. Royko, C. Sun, G. Farnebäck, Algorithm performance contest, in conjunction with ICPR 2000, 2000.
[7] TREC, Video Track, The Tenth Text REtrieval Conference, Video Track, 2001.
[8] A. Clark, P. Courtney (Eds.), Workshop on Performance Characterisation and Benchmarking of Vision Systems, Las Palmas, Spain, 1999.
[9] R. Haralick, R. Klette, S. Stiehl, M. Viergever (Eds.), Evaluation and Validation of Computer Vision Algorithms, no. 98111 in Dagstuhl Seminar, Dagstuhl, Germany, 1998. URL: http://www.dagstuhl.de/data/seminars/98/
[10] H. Christensen, W. Förstner, C. Madsen (Eds.), Workshop on Performance Characteristics of Vision Algorithms, Cambridge, United Kingdom, 1996.
[11] P. Courtney, N. Thacker, Performance characterisation in computer vision: the role of statistics in testing and design, in: J. Blanc-Talon, D. Popescu (Eds.), Imaging and Vision Systems: Theory, Assessment and Applications, NOVA Science Books, 2001.
[12] K. Bowyer, P. Phillips, Empirical Evaluation Techniques in Computer Vision, Wiley, IEEE Computer Society Press, New York, 1998.
[13] Y. Zhang, A survey on evaluation methods for images segmentation, Pattern Recognition 29 (1996) 1335–1346.
[14] J.R. Smith, Image retrieval evaluation, in: Workshop on Content-based Access of Image and Video Libraries CAIVL'98, Santa Barbara, California, 1998, pp. 112–113.
[15] H. Müller, W. Müller, D. Squire, S. Marchand-Maillet, T. Pun, Performance evaluation in content-based image retrieval: overview and proposals, Pattern Recognition Lett. 22 (2001) 593–601.
[16] J. Vogel, B. Schiele, On performance categorization and optimization for image retrieval, in: European Conference on Computer Vision ECCV'02, vol. IV, Copenhagen, Denmark, 2002, pp. 49–63.

[17] J. Vogel, B. Schiele, Query-dependent performance optimization for vocabulary-supported image retrieval, in: German Pattern Recognition Symposium DAGM'02, Zurich, Switzerland, 2002, pp. 600–608.
[18] Y. Rui, T. Huang, M. Ortega, S. Mehrotra, Relevance feedback: a power tool for interactive content-based image retrieval, IEEE Trans. Circuits Video Technol. 8 (5) (1998) 644–655.
[19] I. Cox, M. Miller, T. Minka, T. Papathomas, P. Yianilos, The bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments, IEEE Trans. Image Process. 9 (1) (2000) 20–37.
[20] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1349–1380.
[21] J. Vogel, B. Schiele, A semantic typicality measure for natural scene categorization, in: German Pattern Recognition Symposium DAGM'04, Tübingen, Germany, 2004, pp. 207–215.
[22] C. Town, D. Sinclair, Content based image retrieval using semantic visual categories, Tech. Rep. 2000.14, AT&T Laboratories, Cambridge, 2000.
[23] R. Picard, T. Minka, Vision texture for annotation, ACM J. Multimedia Syst. 3 (1) (1995) 3–14.
[24] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, D. Freeman, AutoClass: a bayesian classification system, in: International Conference on Machine Learning ICML'88, Ann Arbor, MI, USA, 1988, pp. 54–64.
[25] J. Mao, A. Jain, Texture classification and segmentation using multi-resolution simultaneous autoregressive models, Pattern Recognition 25 (2) (1992) 173–188.

About the author—JULIA VOGEL holds a M.Sc. degree in Electrical and Computer Engineering from Oregon State University, USA (1998), and a M.Sc. (Dipl. Ing.) degree in Electrical Engineering from the Technical University Karlsruhe, Germany (2000). In December 2004, she received a Ph.D. from ETH Zurich, Switzerland for her thesis "Semantic scene modeling and retrieval". Her research interests include semantic scene understanding, combined scene and object recognition, content-based image retrieval, and human scene perception. Funded through a research fellowship from the German Research Foundation, Julia is currently a postdoctoral fellow at the Laboratory for Computational Intelligence, University of British Columbia, Vancouver, Canada.

About the author—BERNT SCHIELE is a full professor of Computer Science at Darmstadt University of Technology, Germany. Prior to this appointment, he had been with ETH Zurich, Switzerland and the MIT Media Laboratory, Cambridge, MA, USA. His research interests include computer vision, perceptual computing, statistical learning methods, wearable computers, and integration of multimodal sensor data. Schiele received a Ph.D. in computer vision from the INP Grenoble, France.