Performance Evaluation of Relevance Feedback Methods

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

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

ABSTRACT

In this paper we review the evaluation of relevance feedback methods for content-based image retrieval systems. We start out by presenting an overview of current common practice, and argue that the evaluation of relevance feedback methods differs from evaluating CBIR systems as a whole. Specifically, we identify the challenging issues that are particular to the evaluation of retrieval employing relevance feedback.

Next, we propose three guidelines to move toward more effective evaluation benchmarks. We focus particularly on assessing feedback methods more directly in terms of their goal of identifying the relevant target class with a small number of samples, and show how to compensate for query targets of varying difficulty by measuring efficiency at generalization.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance feedback Retrieval models Selection process Information filtering*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Algorithms; Experimentation; Human factors; Measurement; Performance

1. INTRODUCTION

Contrary to most well-known search engines to date, search systems that incorporate relevance feedback (RF) do not just present a ranking of results, but also let the user provide feedback on the relevance of these results. Using relevance feedback the user can indicate *by example* which items he finds relevant to his search task, thus helping the system to improve its suggestions iteration by iteration. Particularly in image retrieval, relevance assessment is truly at-a-glance: users can easily pick out the images relevant to them.

CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.

This paper provides a review on *performance evaluation* of relevance feedback methods used for image retrieval. Several reviews are already available that discuss *methods* of relevance feedback analysis in content-based image retrieval (CBIR): [38], [50],[6] and [23]. The first and last of these are reviews on the state-of-the-art of the entire CBIR field and contain short discussions on relevance feedback; [50] and [6] provide overviews of RF algorithms but do not address performance evaluation. The current paper thus differs from these earlier reviews by explicitly focusing on the performance evaluation of RF algorithms.

In [31] an introduction to the subject of CBIR system evaluation is provided, mainly from an information retrieval (IR) perspective. However, the paper does not explicitly address the evaluation of relevance feedback methods, and the authors observe that much research remains to be done on the evaluation of interactive systems and the inclusion of the user in the query process. Another, short, introduction to system evaluation is given in [27].

In section 2, we start out by reviewing current practice in relevance feedback evaluation. This includes a discussion of promising benchmarking initiatives that are currently underway.

In the remainder of the article we approach RF testing by focusing on its role in the retrieval process: to infer what a user is looking for *in a particular search.* Images generally have many qualities that are potentially relevant, and so their meaning to a user may differ from the context of one search task to another. It follows that the image representation process should, accurately, capture the *potentially* relevant qualities, while the RF algorithm should figure out which of these qualities is *actually* relevant in a particular search task.

This division of work implies that RF performance strongly depends on the quality of the image representations. In section 3 we discuss this issue, as well as a number of additional challenges to effective evaluation, such as the difficulty of achieving realistic testing conditions due to the high cost of obtaining consistent relevance feedback and ground truth.

In section 4, we propose a number of guidelines for setting up effective benchmarks for RF method evalation. The aim has been to bring benchmarking campaigns closer to the goal of testing under realistic conditions while also ensuring that the procedures are feasible in practice.

Next, in section 5, we work out in more detail how to implement these guidelines. We describe a testing approach that measures the efficiency of methods at generalization from a limited number of sample images. Most importantly,

this allows us to compensate measurements for query targets of varying difficulty.

2. CURRENT PRACTICE

To evaluate RF algorithms most authors have relied on a standard information retrieval approach, sometimes called the Cranfield paradigm ([45]). Two main characteristics of this approach are: (i) its laboratory-like setting: once the search problems have been specified, evaluation of a new method can proceed automatically, without need for further user interaction; (ii) systems are tested *as a whole*: the RF method is only one link in a chain of factors determining overall system performance.

We shortly discuss the main stages of this approach in the context of RF method evaluation.

Setup: Image Collection and Representation. Many authors have used the Corel Photo CD collection to compile their test image database (e.g. [47], [35], [2], [49], [5], [43], [44]). The total Corel collection consists of more than 800 Photo CDs, each containing 100 broadly similar images of a certain category (e.g. `africa`, `dogs`, `yosemite`, `castles`, `roses`).

In most cases, research groups have made their own selection from the available categories, usually amounting to a very small subset of the entire collection (e.g. 50, 20 or even fewer categories). This can result in a test collection consisting of dissimilar groups of images with a relatively high within-group similarity [31]. Even though it is clear that the choice of sets can greatly influence evaluation results, usually no motivation for a particular choice of categories is provided.

For representation of the images, there has been a bias towards low-level feature spaces. In the majority of articles proposing relevance feedback algorithms, images are primarily represented by low-level color and texture features. Typical choices are color histograms to represent color content, and Gabor filter-based features to represent texture patterns. Other popular choices are color moments, color layout and coherence features, and edge histogram and discrete wavelet transform-based texture features. In some cases also shape features are used. Image representation is discussed in more detail in section 3.

Search tasks and Ground Truth. Systems are tested on a set of search tasks, each with a pre-defined information need or *topic*. RF methods are compared by how many images relevant to the topic they help retrieve. Topic *ground truth* consists of a binary relevance relationship which indicates for each image in the collection if it is relevant to the topic or not. The relevant images are also known as the *target* set of the search task.

In most relevance feedback studies to date, ground truth is based on Corel category names: an image is relevant to a topic if and only if it has the specific category label. No further effort is expended to check if all images relevant to the topic have been found. This often leads to inaccuracies. For example, if the Corel `night` category defines ground truth, then many other "night"-images (e.g. in the `Paris`, or `volcano` categories) are missed and assigned as not relevant to the topic. As mentioned, it is common to construct test image collections that consist of only a limited number of categories. In [30] ("The truth about Corel - evaluation in image retrieval") a detailed analysis is presented of how this practice can inflate performance. Also various other practices are discussed that can lead to artificially high performance, e.g. the method by which query images are selected.

Despite the well-known findings on the subjectivity of testing with self-chosen categories from the Corel set, also many recent articles still take the same approach, see for instance [8], [12], [48], [14], [18], [17], [24] and [25]. Other authors have used different databases, e.g. [16], [3], [42], or [33], but also in these cases no shared benchmark databases are used.

Fortunately, the situation may be improving. Even though so far only few RF papers have taken advantage, there are now several promising benchmarking campaigns which do aim to provide shared databases and search topics. It must be noted though that in most cases relevance feedback is treated as one of more approaches to improving overall performance, and so far no targeted tasks have been developed specifically for comparison of relevance feedback algorithms[1]. We mention:

- **The Benchathlon Network** ([26]). Currently no further benchmarking activities are taking place, but a free image collection is still available.
- **TRECVid** (e.g. [37]). Supported by a very active community of researchers. It provides large test collections, uniform scoring procedures, and a forum for organizations interested in comparing their results. The benchmarks are aimed at general video retrieval and define a number of tasks such as shot boundary determination, high-level feature extraction and search. In the latter task, user interaction is encouraged so it provides an opportunity to demonstrate performance gains through the use of RF. However, RF is only one of the interaction strategies that may be employed here, next to, for instance, simply scanning the videos. The benchmark tests largely follow the steps described in the current section; a difference is that ground truth is often obtained by pooling (see section 3).
- **ImageCLEF** ([4]), directed at multi-language retrieval. It aims to test systems based on their ability to maintain retrieval performance independent of the language used to express the associated texts or textual queries. ImageCLEF uses a number image collections, e.g. the **IAPR TC-12 Benchmark** [22], which also serves as a benchmark by itself. ImageCLEF 2006 and 2007 offered an interactive retrieval evaluation using a database provided by FlickR (iCLEF2006/7).
- **ImagEVAL** (e.g. [11]). Aims to test retrieval systems under actual conditions of use. It is funded by the French government and is open to participation to all Europeans.
- **SHREC** (3D Shape Retrieval Contest), organized by the EU-funded network AIM@SHAPE. Its objective is to evaluate the effectiveness of 3D-shape retrieval algorithms. Its 2007 edition offers a dedicated relevance feedback track. It uses two collections of 3d shapes and a fixed set of geometrical shape features. The evaluation setup follows the steps outlined here.

[1]With the exception of the SHREC 3D 2007 benchmark, which includes a task specifically aimed at relevance feedback; see below.

Retrieval. In the retrieval phase, the system must find images relevant to the testing topics. First an initial ranking is produced; next, subsequent RF by selection of example images should produce better rankings.

In nearly all papers proposing RF algorithms, no actual online feedback is used for the example selection; instead, user behavior is simulated based on a "consistent-user" assumption (e.g. [35], [42]). This assumption states that the simulated user judges image relevance exactly according to the ground truth supplied for the topic, and that he will select feedback examples accordingly.

For reasons explained in section 3, we generally support this system-based approach and favor it over a more user-centric testing approaches; nevertheless, testing can benefit from more realistic simulations, e.g. by also modeling partial relevance of images (see section 3).

Simulations differ in how many feedback examples are selected per iteration. Sometimes only a single extra image per iteration is used. Often, more images are selected, e.g. in [15] three positives and three negatives are selected at each iteration when these are available. They also differ by *scope*, i.e. by how far down the ranking the user is simulated to look.

Many RF algorithms are essentially transformations of feedback example sets to new relevance rankings of the database. Input is either a positive example set, or, if also negative feedback is used, a pair of a positive and a negative example set. An algorithm can be understood this way if the example set (pair) of the most recent iteration is assumed to capture the information available to the search system and no "memory" of previous interactions is used to determine new rankings. Note this means that the user is expected to maintain the example sets such that they represent his wishes as well as possible, e.g. he may need to delete images from the sets that are no longer sufficiently relevant.

Algorithms of this type can be tested simply based on the quality of the rankings they produce based on single example set pairs. Instead of simulating iteration sequences, performance is tested on a large variety of possible example sets as if these had occurred somewhere during the process. This avoids the need for ad hoc assumptions on how users select relevant examples from rankings to go from iteration to iteration. See [21] and [2] for examples of this approach.

There are also various approaches to initiate queries. Initial rankings can be obtained by random sampling or produced through independent means, e.g. a keyword search. The search process can also be started by applying the RF method to a set of initial feedback examples. Sometimes a single query image from the target set is used (e.g. [48]); often, a small subset of the target category images is used to provide initial positive examples; similarly, negative example images may be sampled from the images not belonging to the target class.

Ranking Assessment. The Cranfield paradigm evaluates the performance of the retrieval system as a whole. RF algorithms are evaluated by comparing the quality of their rankings when all conditions (image collection, image features, topic set, (simulated) user behavior) are kept equal, except the RF algorithm itself. The most common way to evaluate ranking quality is by means of precision-recall graphs.

There has been considerable debate about how precision-recall graphs must be averaged to summarize results from multiple relevance rankings. This is important because methods are generally tested on a collection of topics. Also within a topic several precision-recall graphs may be produced corresponding to, for instance, results from different simulations, or different initial queries. In [34] several possible approaches to averaging precision-recall graphs are covered. In [19] the influence of the fraction of relevant items, or generality, on performance is discussed; the authors recommend adding the generality of the target class as an additional dimension to the performance graph, as well as a restriction to specific normalized scope values.

Several authors have recommended additional performance measures derived from the relevance rankings, e.g. [30] and [7]. In particular MAP, Mean Average Precision, is often used as a summarizing evaluation measure; it corresponds to the area under the (normalized) precision-recall graph. Other common performance graphs are precision versus scope and recall versus scope, where scope is the number of highest ranking images considered. All of these measurents can be computed after each iteration of the feedback process. Typically, authors consider anywhere between 3 and 10 rounds of feedback. To show the improvement of the rankings it is also common to graph precision at a specific scope versus iteration number. In [25] a distinction is made between *actual* recall and precision, *new* recall and precision and *cumulative* precision and recall. These quantities can be used to measure improvement between iterations.

Another important performance measure is response time (e.g. [11]), i.e. the time it takes for the system to present a new relevance ranking; since most users require real time performance, this put strict limits on feasible execution times for RF analysis (see also [30]). Also important in this respect is scalability with database size; [7] provides an extensive list of further performance measures, e.g. stability to query and error resilience.

3. CHALLENGES IN RF EVALUATION

Despite progress in general retrieval benchmarking, testing relevance feedback algorithms remains a challenge. Here we discuss a number of reasons why this is so. First we discuss consequences of the dominant effect of representation quality on retrieval performance. Next, we show that, despite strong consensus in the research community that testing should be performed under realistic conditions of use, this remains difficult to achieve. One reason is the cost and subjectivity of feedback by human subjects, another is the difficulty of providing ground truth for realistic and varied topics.

1. Image representation quality

To be able to search digital content it needs to be represented in a form accessible to the search engine. The richness and accuracy of description determine how well user needs can be met. We first discuss the importance of representation quality to retrieval performance in general, and next consider consequences particular to RF evaluation.

For text documents a very reasonable level of representation is reached by simply considering the words occurring in the texts. In some applications, images can also be represented by keywords, mainly by one of two ways: either by using the textual context of the image, or by manual annotation. The former is the method used in most current search engines for web images. Search engines use for instance the image title, filename, caption, or the text surrounding the

image on the web page, to obtain keywords describing the image. Because the content of the image itself is not taken into account, these keywords tend to be less accurate than for ordinary text documents. The other method is direct manual annotation of images. This comes with a number of relatively minor challenges of its own, but it has the distinct advantage of reaching a high semantic level very directly. It's single real drawback is of course that it is not automatic, and thus costly. For many applications neither of the two is feasible and we must turn to methods that are content-based and automatic (or semi-automatic).

Many low-level descriptors have been developed to characterize the color and texture content of images. Despite progress in these techniques, as well as in image segmentation, a strictly bottom-up approach to image understanding has so far not been feasible. A lot of effort is now expended in combining low-level image analysis with methods of machine learning and pattern recognition to train classifiers for visual concepts (e.g. [39]). The basic idea is to provide high level concept annotations for a small but representative part of a collection, and use these examples to learn to predict the occurrence of the concepts. The remaining, much larger, part of the collection is annotated automatically by means of the resulting classifier.

An effort to standardize annotations using a taxonomy on the order of 1000 concepts is underway as part of the Large Scale Concept Ontology for Multimedia-initiative (LSCOM, [32]). The concepts (e.g. `car`, `indoor`, `Pope`) are selected based on their utility for actual searches, frequency of occurrence, and feasibility to build classifiers for them with reasonable accuracy. In [40] classifiers for a subset of 101 semantic concepts are presented. For each concept an SVM is trained using a generic image representation; classifier performance is reported on data of the TRECVid benchmarking campaign ([37]). From this study and many others, e.g. within the TRECVid high-level feature extraction track, it has become clear that although useful concept classifiers can be learned, accuracy remains fairly limited for many categories. A similar situation exists for detectors that do not just classify concept presence, but also aim to detect the location of the concept in the image frame. An interesting recent benchmark in this respect is the PASCAL Visual Object Classes Challenge ([10]).

So, even though successful detection and recognition of concepts leads to very useful metadata, so far the quality of content-based representations is generally poor compared to even simple textual representations. For general domains, classifier error rates tend to be high, and the visual concepts do not provide sufficiently detailed coverage of the semantic interest space. As a result, it is expected that - for the time being - the quality of automatic or semi-automatic image representations will remain a serious bottleneck for the performance of CBIR systems. Since most relevance feedback algorithms are applied in content-based systems using such representations, this is also a key property to take into account when designing the evaluation environments for the RF algorithms.

The first, and most important, consequence is that for many search topics it is unreasonable to expect satisfactory retrieval performance: topics are simply not represented sufficiently well by the available metadata. Many concepts are hard to learn even when a lot of examples are supplied, let alone for the typical small sample size situation that RF methods have to work with. In section 5 we propose an evaluation approach where first the feasibility of the search task is assessed; subsequently, performance is measured relative to what can reasonably be expected given the available image representation quality.

A second consequence is that the performance of RF algorithms is usually only a second order factor in the overall performance of a retrieval system. This means that if we let users explore retrieval systems and measure their satisfaction, these measures will tend to be dominated by the effects of poor image representation. The same holds true if we compare retrieval systems by how well they assist the human in performing his task. RF can definitely improve satisfaction and utility but such subjective measures will be too indirect to reliably quantify the performance of different algorithms consistently (see for instance [28]) and [45]).

2. Cost and inconsistency of RF

For effective testing we need to compare performance of RF methods on a large number of concrete search topics. Given the subjectivity, i.e. user- and task dependence, of image relevance to the topics (e.g. [36]), ideally both the ground truth and the relevance feedback for a particular topic should be provided by single users. In practice, however, this is usually not feasible for the following reasons:

- The number of relevance judgments required from individual assessors is prohibitive.
- The assessors need to be consistent in their interpretation of a search task between iterations, and even more demanding, between using different methods and systems.
- Assessor efforts on providing relevance feedback cannot be re-used. Whenever a new approach is tested, also new relevance feedback is required.
- It is hard to make sure that assessors are unbiased to the different methods. For instance, fair testing requires that the users are equally familiar with each of the methods and their optimal use, since a good understanding of the inner workings of a system can have great benefit for performance (e.g. [5]).

The fact that user consistency is so hard to achieve and the large and non-reusable efforts required for a more user-centric testing approach, are important reasons to favor the system-based approach. As discussed in section 2 this has led to the common practice of simulating feedback judgments based on pre-defined ground truth. The great advantage of this approach is that ground truth assessments are re-usable, so that, once the data have been collected, new or modified methods can be tested without requiring further user effort. However, despite the widespread use of feedback simulation, little research has been done to what extent the simulations correspond to actual user behavior. One particular interest in this respect is feedback example selection based on *partial relevance*. Real users will not only select images that are fully relevant to a topic, but also images that exhibit only one or a few properties they are searching for. In particular in the initial stages of a search when few or no fully relevant images are available, systems that can exploit partial relevance will be at an advantage. Unfortunately, in most feedback simulations partial relevance is not taken into account (notable exceptions are [33] and [20]).

Improving this situation is an interesting area for future research.

3. Topic ground truth

To test image retrieval systems we need ground truth for realistic search topics. As we have just seen, for RF evaluation the ground truth is not only needed for measurement of search effectiveness, but also to simulate the relevance feedback itself. Providing ground truth is a labor intensive and, for most, tedious process. In regular retrieval benchmarks, i.e. those not particularly aimed at RF, it is common practice to reduce this cost by *pooling* relevance judgments (see for instance [41]).

First, only images (or documents, shots etc.) are considered that appear in the top N of images ranked most relevant of at least one approach participating in the benchmark. This way only a small subset of the total collection of images needs to be assessed for a given search topic. Images that were not retrieved within any of the top N's are assumed not to be relevant.

Next, the resulting image pool is randomized, assuring that ranking and system information are no longer available, and the images are divided over the participating groups to further reduce the work of providing the relevance judgments.

Note that defining ground truth through pooling assumes the existence of a shared understanding of the topic, i.e. we are no longer considering individual interpretations of a search topic. Also, since the relevance of a large number of images remains unverified, the accuracy of the performance measures is affected (see [11]). Precision measurements are exact for the first top N images but generally deteriorate when the scope is extended. Recall measurements are also directly affected since the correct number of relevant images is not known. See also [46] for a discussion of re-usability of ground truth that has been obtained through pooling, e.g. on the effect for systems that did not take part in the original evaluation.

Since, in RF evaluation, ground truth is not only used for precision and recall measurements, but also for simulation of the relevance feedback itself, pooling becomes still more problematic. In particular, when RF approaches are judged based on several feedback iterations, where at each iteration the resulting ranking needs to be judged for relevance, pooling becomes progressively harder to use. For these reasons it is preferred to work with complete ground truth for RF evaluations. Moreover, since RF strategies are designed specifically to accommodate for the user- and task-dependence of topic relevance, the evaluation environment should allow for demonstrating the ability to adapt to individual differences. So, although we advocate that ground truth should be supplied by *several* users, ideally the ground truth of a *single* search task should be based on the relevance interpretation of a *single* assessor.

We have already discussed some of the problems resulting from ground truth based on pre-defined (Corel) database categories. Studies such as [1] and [29] indicate an additional problem: realistic query topics are usually more complex than those captured by single simple semantic label. It is therefore a challenge to provide diverse search tasks at varying levels of abstraction. One recommendation would be to define composite topics, e.g. not only "flowers", but also "red flowers against a blue sky background".

In [9] it is found that specific photo needs, e.g. representing concrete objects like named persons, buildings or places, can dominate the use of photo archives. Ideally we should thus include this type of search tasks in our testing setup. However, we must also make sure that the topics can be learned given the available image representations. We return to this issue in the next section.

4. EFFECTIVE RF EVALUATION

In light of the issues discussed so far, we provide three guidelines for more effective scientific reporting on RF method performance. The first two cover the choice and design of search topics for testing. The third guideline proposes a new testing framework that allows us to factor in the difficulty of the search topics and thereby allowing us to measure more directly if the goals of RF are being met.

1. Start with consistent ground truth.

Methods need to be tested using a collection of topics for which accurate and consistent ground truth is available. For reasons discussed earlier, we prefer ground truth for individual topics to be provided by *single* assessors. This leads to ground truth that is consistent with a realistic individual information need, providing exactly what is required given that RF methods are designed to accommodate the wishes of individual users.

In [45] it is shown that the ground truth targets that different users supply, even for generic topics, often differ by more than 40%. For the purposes of RF evaluation we recommend to gather a number of such targets for a given topic, and subsequently test the ability of the RF methods to adapt to the different interpretations.

Finally we reiterate that we should avoid defining ground truth based on database category labels without checking if other images, without the label, also apply to the topic (see section 2).

2. Consider only feasible search tasks.

The role of RF in the retrieval process can be summarized as identifying relevant regions in our feature space by generalizing from a small number of examples. Performance should thus be measured only on problems where such relevant regions indeed exist. If a search topic is not sufficiently well represented by the available features to learn it even for large samples, then it will also be impossible to generalize from small samples.

We must acknowledge that many search problems of practical interest are currently beyond the quality of available image representations. For illustration, try finding images of `Amsterdam` based solely on color features. It is an important challenge to collect search tasks that trade off practical interest and feasibility, and to keep topic collections up to date with progress in representation power.

We will work out the feasibility of search problems in more detail in section 5.

3. Measure relative to representation structure.

Performance on any given search task is to a large extent determined by the difficulty of that task. Factors that determine the difficulty are the number of images of the target class available in the database, the number of images in the database that are similar to the target images but do not belong to the target class, and most importantly, by how well the desired properties are captured by the image features. Search problems with target classes that can be isolated well

from the non-relevant images in representation space lead to good performance; target classes that are hard to separate from non-relevant images give rise to poor performance.

Together we can refer to such aspects as the topic *representation structure*, i.e. the structure of the target class in the space induced by the image representation and its embedding in the non-relevant images.

We propose to measure performance relative to the representation structure. This has the distinct advantage that we provide all measurements with a shared scale of reference. Currently, when methods are tested on collection of topics, the resulting precision-recall measurements are averaged without any reference to topic difficulty. In practice, this means that the aggregated precision-recall graphs are mainly determined by the mix of topic difficulties, making it very hard to compare outcomes between different studies. Assessing topic representation structure beforehand has the additional advantage that we can also detect topics that are too difficult and are not feasible to be learned for the available data (see previous guideline).

In the next section, we outline how to characterize the representation structure by estimating the optimal performance it allows for a classifier with a large part of the ground truth at its disposal.

5. STRUCTURAL NORMALIZATION

Instead of characterizing RF methods by their absolute performance, which is largely determined by the difficulty of the topic, we quantify how well the methods can approximate the optimal performance achievable for a topic. We first show how optimal performance can be characterized using statistical learning and decision theory. Next, we show concretely how to normalize performance with respect to representation structure, by providing three approaches to estimate the optimal classifier for RF method benchmarking. The normalized performance measure is called *generalization efficiency* because it measures how well we can approximate optimal performance with only a small number of example images. Finally, we show how the framework can be used to evaluate the consistency of RF methods by testing their ability to select features that are known to be relevant.

5.1 Statistical Decision Theory

To explore the relationship between relevance feedback and representation structure in more detail, we model the problem of finding relevant images by means of binary classification. Refinements to this model, e.g. allowance for partial relevance, are possible but will not be considered here.

Let D be a database of n images. Image i of the database is represented by feature vector x_i in feature space[2] $\mathcal{X}$.

Given a search problem and database D, associated ground truth $\mathcal{G}$ is denoted as

$$\mathcal{G} = \{(x_i, y_i) | \, i = 1, \ldots, n;\, y_i \in \{0, 1\}\}\,, \tag{1}$$

i.e. as a set of feature-label pairs where an image is labeled as $y_i = 1$ when it is relevant according to ground truth, and as $y_i = 0$ when it is not. The target class $\mathcal{T} \subset D$ of all relevant images is

$$\mathcal{T} = \{i \in \{1, \ldots, n\} | \, y_i = 1\}. \tag{2}$$

A classifier is a function $f : \mathcal{X} \to \mathcal{Y} = \{0, 1\}$ that assigns feature vectors $x \in \mathcal{X}$ as either relevant or not.

For the images in the ground truth set $\mathcal{G}$ a classifier can make two types of mistakes: *false positives* (images classified as relevant that are, according to ground truth, not relevant), and *false negatives* (images classified as not relevant that are, in fact, relevant). The two mistake types are closely related to precision and recall. If we take the scope of retrieved images equal to the number of images assigned as relevant by the classifier, precision is equal to one minus the fraction of false positives; recall is one minus the fraction of false negatives.

In the following we will assess the quality of classifiers using standard statistical decision theory (SDT, e.g. [13]). In this theory the aim is to find (or approximate) classifiers which minimize the expected, possibly weighted, loss, resulting from the two types of mistakes. This approach allows us to analyze the ability of RF methods to generalize from a small sets of example images. In particular, SDT allows us to define optimal classifiers that are realistic in the sense that they do not simply classify ground truth to full perfection[3], but rather provide the optimal generalization performance achievable *given the available image representation*. For instance, if a problem has a representation structure where the target class can not at all be separated from the non-relevant images (e.g. refer back to the problem of finding images of "Amsterdam" using only color features), even an optimal classifier will make many mistakes. In these cases, representation is often such that there are no, or very few, regions in feature space for which the target class is sufficiently probable to justify target class assignment.

Optimal classifiers minimize the loss associated with false positives and false negatives. A loss function $\mathcal{L}(y, f(x))$ defines a cost for each combination of classification outcome $f(x)$ and true value y. The often used zero-one loss function

$$\mathcal{L}(y, f(x)) = \begin{cases} 0 & \text{if} \quad f(x) = y \\ 1 & \text{if} \quad f(x) \neq y \end{cases}\,, \tag{3}$$

assigns zero cost to correctly classified instances, and equal (viz. unity) cost to both false positives and false negatives. More intricate loss functions usually further refine the second term by splitting based on the y-value.

The optimal classifier with best generalization performance minimizes expected loss for the "true" joint probability $p(X, Y)$ of feature vectors X and associated outcomes Y. Note that since $p(X, Y) = p(Y|X)p(X)$ this joint probability can be interpreted as a combination of the prior probability of image feature occurrence and the probability that such feature value leads to an image perceived as relevant (or not relevant) to the user. For the zero-one loss this classifier is called Bayes' classifier, which is given by

$$f(x) = \underset{k \in \mathcal{Y}}{\operatorname{argmax}}\, p(k|X = x), \tag{4}$$

i.e. it simply assigns to the class that is most likely given the

[2] $\mathcal{X}$ need not be a single metric space; often it is a composite product of various metric spaces

[3] In [19] a classifier that is perfect (no matter which data it has to work with) is called the Total Recall Ideal System (TRIS).

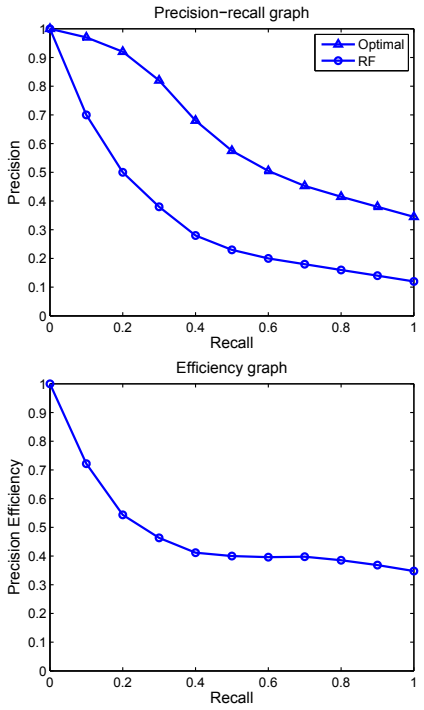

Figure 1: From precision-recall to efficiency graph

feature value. For unsymmetric loss functions, the optimal classifier will be biased toward the class with smaller loss.

In practice, we cannot compute the exact optimal classifier, because we do not know $p(X, Y)$. However, following standard approaches from learning theory we can use the ground truth data to approximate the optimal classifier. The main concern is to avoid overfitting to the data. We take the following general procedure:

1. Choose a loss function. The zero-one loss function will often be the default choice, unless we want to further tune the tradeoff between precision and recall.
2. Compare classifier performances by training on the ground truth set. To avoid overfitting, use cross-validation.
3. Select the classifier that minimizes average loss over the cross-validation runs as estimate for the optimal classifier.

5.2 Generalization Efficiency

As mentioned, our aim is to use optimal classifier performance to normalize RF method performance. The general idea is illustrated in Figure 1.

The figure shows the precision-recall graph corresponding to the optimal classifier, as well as the precision-recall graph achieved by a certain RF method, both for a given search problem. The optimal precision-recall is the best we can do given the representation structure and available image representation. The details of moving back and forth between classifiers and rankers will be further discussed below. The second graph plots how well the RF method performance approximates (estimated) optimal performance.

The same idea can be used for all the usual evaluation graphs, e.g. precision-recall, precision-scope and recall-scope. Instead of using the absolute values as the dependent quantity we use values normalized by the optimal value. We refer to the normalized values as generalization *efficiency rates* and the resulting graphs as generalization *efficiency graphs*. In analogy to mean average precision (MAP), we reserve mean average efficiency (MAE) for the average precision efficiency rate integrated over the recall, i.e. the area under the precision efficiency-recall graph. Finally, we refer to the complementary rates (i.e. $1 -$ the efficiency rates) as *deviation rates*.

We discuss three methods to implement the general learning strategy sketched above to approximate the optimal learners in the context of RF method evaluation.

Approach I: Intra-RF method estimation. In this first approach a baseline estimate of optimal behavior is obtained using only the RF method itself. This is achieved by considering the relevance ranking obtained by applying the RF method under evaluation to a significant part of the ground truth data. If necessary, the method may be tuned to this new situation of having such a large training set at its disposal. Different parameter settings can be compared by evaluating the chosen loss function (step 1). Each ranking can naturally be associated with a classifier[4] by taking the retrieval scope equal to the size of the target class[5]. The images with rank lower or equal to this scope are classified as relevant, the images of higher rank as non-relevant.

The procedure can be summarized as follows

1. Choose a loss function $\mathcal{L}(y, f(x))$.
2. Determine a candidate set of variations of the original RF method, giving rankers $R_1, \ldots, R_M$. If there is only a single ranker ($M = 1$), steps 5, 7 and 8 can be omitted.
3. Divide the ground truth into training sets $\mathcal{G}_k^{\text{train}}$ and test sets $\mathcal{G}_k^{\text{test}}$, $k = 1, \ldots, K$ following a K-fold cross-validation scheme.
4. Apply the RF method variations R_j to the training set $\mathcal{G}_k^{\text{train}}$ giving rankings $r_{jk} = R_j(\mathcal{G}_k^{\text{train}})$. Note that the training set provides positive and negative example images as if they had been provided through feedback. The ranking r_{jk} assigns a rank to all database images (including the test images!).
5. Evaluate the loss function for the resulting rankings r_{jk} using the images in the test set $\mathcal{G}_k^{\text{test}}$. The estimate of expected loss L_{jk} is given by

$$L_{jk} = \frac{1}{|\mathcal{G}_k^{\text{test}}|} \sum_{(x_i, y_i) \in \mathcal{G}_k^{\text{test}}} \mathcal{L}(y_i, \hat{y}_i), \qquad (5)$$

   where $\hat{y}_i = 1$ if image i has rank $r_{jk}(i) \leq |\mathcal{T}|$ and $\hat{y}_i = 0$ otherwise; $|\mathcal{T}|$ is the size of the target class in the ground truth.
6. Repeat this procedure for each cross-validation division of the ground truth.

[4]Note that is not a classifier in the strict sense as its domain is restricted to feature values that occur in the ground truth set; however, for the normalization purposes described here this is sufficient.

[5]An alternative would be to minimize loss over all possible scopes, but this may be prone to overfitting.

7. Average the loss values L_{jk} over the different cross-validation runs k, giving an average loss $\bar{L}_j$ for each of the candidate rankers R_j.

8. Select the candidate ranker with lowest average loss as optimal ranker $R_{j_{\text{opt}}} : j_{\text{opt}} = \text{argmin}\, \bar{L}_j$.

9. Use its associated average precision-recall graph as ranking baseline as in Figure 1.

For some methods, e.g. SVM-based RF methods, this approach can be expected to give a good estimate of optimal performance. For other methods that were not designed for general classification, e.g. Rocchio's method, results can obviously be rather poor, in particular in the common case where the relevant class does not consist of single compact cluster. For those cases using one of the methods discussed below is preferable.

Approach II: Inter-RF method estimation. This approach can be applied when several RF methods are compared, e.g. in a benchmark situation. For each of the RF-methods an optimal ranking estimate is obtained using approach I above. The ranker with smallest loss over all the different methods is selected to provide the baseline optimal ranking. This baseline is then also used to normalize the measurements for all other RF methods.

Approach III: Direct estimation. In this final approach optimal baseline performance may be estimated directly by any suitable classification method. Not only the RF methods themselves can be used, but any learning method that is expected to perform well for the representation structure of the search problem.

Again a cross-validation scheme is used exactly as in approach I. Since in this case we are working with classifiers right from the start, the loss function can be evaluated directly, i.e. the classifier does not have to be derived from a ranking. A caveat here is that this advantage makes the final step of the algorithm of finding a baseline performance graph more involved. However, for most classification algorithms a natural method of ranking the samples is available (e.g. for SVMs a natural choice would be a ranking based on the sample distance to the separating hyperplane). In case no such ranking is available, any ranking that provides lower rank to relevant images than to non-relevant images will do to compute the baseline graphs.

This approach gives the greatest freedom in determining optimal performance achievable for a given search problem and is thus recommended for benchmarks that aim to become standards for the field.

Re-calibration of RF evaluations.

For each search problem, we store the baseline performance that was used to calibrate the evaluation measures.

In practice finding optimal baseline performance is often a gradual process and occasionally a classifier may be found that performs better than the baseline used thus far. In such case, previous measurements can easily be updated for a new optimal baseline.

We work this out for the case of precision efficiency-recall graphs. Let $\tilde{p}_r, r = 1, \ldots, N_r$ be the optimal precision values used thus far, at N_r recall levels. Similarly let $\tilde{p}_r^{\text{new}}$ be the updated optimal precision values. Let $p_r, r = 1, \ldots, N_r$ be the average precision efficiency values (e.g. aggregated over N_e experiments). Then the re-calibrated precision efficiency values are given by

$$p_r^{\text{new}} = \frac{\tilde{p}_r}{\tilde{p}_r^{\text{new}}} p_r. \tag{6}$$

This follows directly from the observation that $\tilde{p}_r p_r$ are the regular precision-recall values.

5.3 Consistency Testing

The framework described above can be used for a variety of consistency tests. For example, we can analyze the convergence of the efficiency with increasing size of the feedback sets. However, more interestingly, we want to focus here on a consistency test that assesses the ability of RF methods to select relevant features.

First note that good generalization already requires the, explicit or implicit, identification of features that genuinely contribute to the prediction of relevance. Similarly, the influence of features that do not contribute, i.e. merely serve as "noise", needs to be diminished and if possible eliminated altogether. Since feature selection is so integral to generalization success, the efficiency of RF methods at feature selection is already measured, be it indirectly, by the generalization efficiency just introduced.

In this section we focus on the special case of measuring the ability of RF methods to select relevant features when the target class of the search problem is included in the image representation, i.e. for the case where the image representation includes a feature that indicates exactly what is relevant to the user. This means we test performance in a situation where the usual lack of adequate representation is not the problem. Despite the explicit availability of the "answer", many RF methods still fail to perform satisfactorily for this case. The most common reason is that all the other features describing the images are not sufficiently weighted down and influence the resulting rankings too much. A classic example is Rocchio's method for which the relevance ranking is based on the distance to a query point. Both in the computation of the query point and the distance to the query point, the other features can completely dominate the effects of the correct feature. Many, more advanced, methods also suffer from similar problems, see [21].

In the following, we again use generalization efficiency to measure performance, except that now optimal generalization baseline performance is trivially available. Since the true answer is part of the representation, it is reasonable to measure performance relative to the Total Recall Ideal System (TRIS, [19]). More formally, we propose the following testing approach:

1. Add a target class indicator variable to the image representation: $\tilde{\mathcal{X}} = \mathcal{X} \times [0, 1]$, such that for image i: $\tilde{x}_i = (x_i, y_i)$, where $y_i = 1$ if $i \in \mathcal{T}$ and $y_i = 0$ otherwise.

2. Determine the TRIS baseline for the desired measurement. This is the performance corresponding to a perfect ranking where the first $|\mathcal{T}|$ highest ranked images are all relevant. For precision-recall this is simply $p(r) = 1$ for all normalized recall levels; for precision-scope we have $p(s) = 1$ for $n \leq |\mathcal{T}|$ and $p(s) = |\mathcal{T}|/s$ for $|\mathcal{T}| < s \leq n$. Recall-scope has $r(s) = |\mathcal{T}|/s$ for $s \leq |\mathcal{T}|$ and $r(s) = 1$ for $|\mathcal{T}| < s \leq n$.

3. Based on the extended image representation, compute the evaluation measure relative to the baseline; since $p(r) = 1$ for all r in the precision-recall case, this is simply the regular precision-recall graph for the extended problem.

This approach tests consistency with respect to feature selection in its most basic form. Naturally, more elaborate variations of this approach can be devised. For instance, ground truth on various elementary properties can be supplied, followed by tests on how well composite properties, defined in terms of the given elementary properties, are retrieved.

6. CONCLUSION

We are currently working towards the evaluation of a number of state-of-the art RF methods using the recommendations described in this paper. This also requires image feature sets that are sufficiently rich and heterogeneous to allow interesting topics to be learned. Despite the large effort required, we hope the resulting collection of search tasks will provide a benchmark that will offer more clarity on the strengths and weaknesses of the different RF methods.

7. ACKNOWLEDGMENT

We thank Bricks/NWO for funding this research.

8. REFERENCES

[1] L. Armitage and P. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997.

[2] Y. Chen, X. Zhou, and T. Huang. One-class SVM for learning in image retrieval. *Proc. IEEE ICIP 2001, Thessaloniki, Greece*, 1:34–37, 2001.

[3] G. Ciocca and R. Schettini. A relevance feedback mechanism for content-based image retrieval. *Information Processing and Management*, 35(5):605–632, 1999.

[4] P. Clough, H. Mueller, and M. Sanderson. The clef cross language image retrieval track (imageclef) 2004. In *Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004), LNCS*, New York, NY, USA, 2004. ACM Press.

[5] I. Cox, M. Miller, T. Minka, and T. Papathomas. The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Trans. Image Processing*, 9(1):20–37, 2000.

[6] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 253–262, New York, NY, USA, 2005. ACM Press.

[7] A. Dimai. Assessment of effectiveness of content based image retrieval systems. In *VISUAL '99: Proceedings of the Third International Conference on Visual Information and Information Systems*, pages 525–532, London, UK, 1999. Springer-Verlag.

[8] A. Dong and B. Bhanu. A new semi-supervised em algorithm for image retrieval. *cvpr*, 02:662, 2003.

[9] P. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 01(1):662, 1993.

[10] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

[11] C. Fluhr, P.-A. Moéllic, and P. Hede. Usage-oriented multimedia information retrieval technological evaluation. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 301–306, New York, NY, USA, 2006. ACM Press.

[12] P. H. Gosselin and M. Cord. Semantic kernel updating for content-based image retrieval. In *ISMSE '04: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering (ISMSE'04)*, pages 537–544, Washington, DC, USA, 2004. IEEE Computer Society.

[13] Y. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning – Datamining, Inference, and Prediction* . Springer, 2001.

[14] J. He, H. Tong, M. Li, H.-J. Zhang, and C. Zhang. Mean version space: a new active learning method for content-based image retrieval. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 15–22, New York, NY, USA, 2004. ACM Press.

[15] X. He, W.-Y. Ma, O. King, M. Li, and H. Zhang. Learning and inferring a semantic space from user's relevance feedback for image retrieval. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 343–346, New York, NY, USA, 2002. ACM.

[16] D. R. Heisterkamp and J. Peng. Kernel vector approximation files for relevance feedback retrieval in large image databases. *Multimedia Tools Appl.*, 26(2):175–189, 2005.

[17] C. Hoi and M. Lyu. Biased support vector machine for relevance feedback in image retrieval. In *Proc. Intl. Joint Conf. on Neural Networks, Budapest, Hungary*, pages 3189–3194, 2004.

[18] C.-H. Hoi and M. R. Lyu. A novel log-based relevance feedback technique in content-based image retrieval. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 24–31, New York, NY, USA, 2004. ACM Press.

[19] D. P. Huijsmans and N. Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):245–251, 2005.

[20] M. J. Huiskes. Aspect-based relevance learning for image retrieval. In W. Leow, editor, *Proceedings of CIVR05, LNCS 3568*, pages 639–649. Springer, 2005.

[21] M. J. Huiskes. Image searching and browsing by active aspect-based relevance learning. In *Proceedings of CIVR06, LNCS 4071*, pages 211–220. Springer, 2006.

[22] C. H. C. Leung and H. H.-S. Ip. Benchmarking for content-based visual information search. In *VISUAL '00: Proceedings of the 4th International Conference*

on Advances in Visual Information Systems, pages 442–456, London, UK, 2000. Springer-Verlag.

[23] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[24] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen. Semantic manifold learning for image retrieval. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 249–258, New York, NY, USA, 2005. ACM Press.

[25] J. Luo and M. A. Nascimento. Content-based sub-image retrieval using relevance feedback. In *MMDB '04: Proceedings of the 2nd ACM international workshop on Multimedia databases*, pages 2–9, New York, NY, USA, 2004. ACM Press.

[26] S. Marchand-Maillet. The Benchathlon Network. `http://www.benchathlon.net`, 2005.

[27] S. Marchand-Maillet and M. Worring. Benchmarking image and video retrieval: an overview. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 297–300, New York, NY, USA, 2006. ACM Press.

[28] G. Marchionini. Human performance measures for video retrieval. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 307–312, New York, NY, USA, 2006. ACM Press.

[29] M. Markkula and E. Sormunen. Searching for photos - journalists' practices in pictorial ir, 1998.

[30] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel - evaluation in image retrieval. In *CIVR '02: Proceedings of the International Conference on Image and Video Retrieval*, pages 38–49, London, UK, 2002. Springer-Verlag.

[31] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recogn. Lett.*, 22(5):593–601, 2001.

[32] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[33] J. Peng. Multi-class relevance feedback content-based image retrieval. *Comput. Vis. Image Underst.*, 90(1):42–67, 2003.

[34] V. Raghavan, P. Bollmann, and G. S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.*, 7(3):205–229, 1989.

[35] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[36] L. Schamber, M. Eisenberg, and M. S. Nilan. A re-examination of relevance: toward a dynamic, situational definition. *Inf. Process. Manage.*, 26(6):755–776, 1990.

[37] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[38] A. Smeulders, M. Worring, S. Santini, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):20–37, 2000.

[39] J. Smith, M. Naphade, A. Natsev, and J. Tesic. Multimedia research challenges for industry. *Proceedings of the International Conference on Image and Video Retrieval (CIVR 2005), Singapore.*, Lecture Notes in Computer Science(3568):28–37, 2005.

[40] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, New York, NY, USA, 2006. ACM Press.

[41] K. Sparck Jones and C. van Rijsbergen. Report on the need for and provision of an ideal information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

[42] D. M. Squire, W. Müller, H. Müller, and T. Pun. Content-based query of image databases: inspirations from text retrieval. *Pattern Recogn. Lett.*, 21(13-14):1193–1198, 2000.

[43] K. Tieu and P. Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1/2):17–36, 2004.

[44] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, New York, NY, USA, 2001. ACM Press.

[45] E. Voorhees. The philosophy of information retrieval evaluation. In *Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, 2001.

[46] T. Westerveld. Trecvid as a re-usable test-collection for video retrieval. In *Proceedings of the Multimedia Information Retrieval Workshop 2005*, Salvador, Brazil, August 2005.

[47] Y. Wu and A. Zhang. A feature re-weighting approach for relevance feedback in image retrieval, 2002.

[48] T. Yoshizawa and H. Schweitzer. Long-term learning of semantic grouping from relevance-feedback. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 165–172, New York, NY, USA, 2004. ACM Press.

[49] H. Zhang, C. Zheng, M. Li, and Z. Su. Relevance feedback and learning in content-based image search. *WWW: Internet and web information systems*, 6:131–155, 2003.

[50] X. Zhou and T. Huang. Relevance feedback in image retrieval: a comprehensive review. *ACM Multimedia Systems Journal*, 8(6):536–544, April 2003.