

The Perceived Similarity of Photos - A Test-Collection Based Evaluation Framework for the Content-Based Image Retrieval Algorithms¹

Eero Sormunen, Marjo Markkula & Kalervo Järvelin

Department of Information Studies, University of Tampere
Tampere, Finland

Abstract

Content-based image retrieval (CBIR) algorithms have been seen as a promising access method for digital photo collections, sooner or later replacing the traditional text-based methods. Unfortunately, we have very little evidence of the usefulness of these algorithms in real user needs and contexts. One problem is that appropriately designed test collections are not available even for the basic performance testing of the CBIR algorithms.

This paper proposes a task-oriented evaluation framework and an efficient procedure for constructing test collections for CBIR algorithms. First, the paper defines a plausible function for these algorithms in general purpose photo retrieval systems. We believe that the CBIR algorithms could be applied effectively in conjunction with text-based photo retrieval. Text-based methods are powerful in retrieving topically related items but do not support browsing. The CBIR algorithms could help in identifying visually similar photos within (often large) result sets of textual queries.

The proposed evaluation framework is based on the concept of perceived similarity and emphasises the role of expertise and realistic illustration tasks as a premise of similarity assessments. A major innovation of the proposed test collection is that it consists of an array of small test sets each built up of a tiny database, a query photo, and respective similarity assessments. The approach supports testing of prototype CBIR algorithms in short development cycles.

The empirical part of the paper reports how journalists were judging the similarity of photos while searching in the course of simulated, but realistic illustration tasks. The goal of the study was to exercise the construction process of the test collection. The results show that the task-oriented evaluation framework and the proposed procedures for constructing the test collection can be successfully applied. The lessons learned from the simulated illustration tasks, collection of similarity assessments and construction of the test collection are discussed.

1 Introduction

Traditional text retrieval methods dominate in operational photo archives. Access is based on manually assigned descriptions (captions, index terms, cataloguing data, etc.) and on standard text retrieval systems. Display and browsing support for thumbnail images is the only major enhancement in the access functions in the digital photo archives when compared with text retrieval systems [2, 5, 22].

Content-based image retrieval (CBIR) algorithms have been suggested to offer advantages over textual methods in photo retrieval, and are a subject of vast research efforts. Indexing is based on automatic identification and abstraction of indexable visual features within an image using image-processing transformations. One obvious bottleneck of the emerging technologies is that they operate at a very low level of visual abstraction. The most common features used are colour, shape and texture. [4, 9]. Conventional object recognition techniques cannot recognise very general objects in unrestricted contexts, for instance, to classify people and cows into different sets. Instead, approaches based on rich image descriptions of special object types (e.g. specific material surfaces, trees, naked people) have been successfully applied in identifying objects in the image [6].

A query is typically made by an example image (e.g. photo, drawing, sketch) and applying partial-match methods to rank retrieved photos into some calculated similarity order [7-9]. Picard and Minka [21] have crystallised the basic "paradigm" of the commercially available CBIR systems in the following way:

¹ This work was funded by the Academy of Finland. We also thank the newspaper *Aamulehti* and the photo agency *Lehtikuva* for their co-operation. Special thanks also to the journalists taking part in the exercise.

1. Pre-compute features or model parameters for each image.
2. Have the user specify which models or ranges of parameters are most important.
3. Have the user select sample images to initiate a query.

The authors state also that "The average person looking for images does not know how to choose model parameters as required in step 2. Moreover, ... the choice of parameters is non-intuitive even for the expert image processing researcher." These comments indicate that even the experts developing the algorithms have difficulties in understanding how different visual parameters should be applied in a particular context. The authors were focusing on the segmentation of images, a fairly low level abstraction problem. The choice of parameters might be even more non-intuitive with retrieval on higher abstraction levels.

Some CBIR products are currently being marketed for photo archive applications (e.g. QBIC by IBM, Visual RetrievalWare by Excalibur) including demos in Web, but we know very little about their potential usefulness in real life use. Rasmussen [22] concludes in her comprehensive review that the performance evaluation of the CBIR systems is nearly an untouched area of IR research in general purpose photo collections. Holt & Hartwick [13] is one exception. They tested QBIC on a tiny database of 650 art images, and obtained some promising but variable, and by no means conclusive results.

From the evaluation viewpoint, content-based image retrieval methods are at quite an early stage of development. Text retrieval systems have been exhaustively studied for over 40 years and standard test collections and evaluation methods are available for testing the matching algorithms. The performance characteristics of text matching algorithms are quite well understood [10, 28]. For CBIR algorithms, there are no standard test collections and evaluation frameworks available [22].

Our general goal is to develop an evaluation framework and a test collection for CBIR algorithms. As the first step, taken in this paper, we outline the basic ideas for the evaluation framework and report the lessons learned from an experiment on building such a test collection. Similarity is a core concept in the evaluation of CBIR algorithms. The algorithms derive estimates for similarity between photos by comparing their feature vectors. In well performing photo retrieval systems these similarity values should correlate with human perceived similarities. If they do not then it is hard to see how the algorithms could help the users in finding photos similar to those they find interesting. We used journalists as an example to examine the ways in which the similarity of photos is perceived by real users searching and selecting photos as a routine task.

The paper is organised as follows. In Section 2 we discuss the CBIR in the light of user needs, introduce the idea of a test collection and its use, and discuss the concept of photo similarity. In Section 3, we describe the exercise conducted on building a test collection. The results on how journalists perceived the similarity of photos are presented in Section 4. In Section 5, we draw conclusions and discuss the lessons learned from the exercise. We do not have the CBIR test results yet. Instead, in this paper, we introduce the evaluation framework.

2 Outline for an Evaluation Framework of CBIR Algorithms

In outlining an evaluation framework, the first task is to define the *function* of a system that is to be evaluated. The framework has to include a description of potential users, their needs and the performance criteria relevant to users. Based on these specifications and considering appropriate reliability and validity issues, one can design an experimental setting and data collection procedures to run experiments. [25, 28].

2.1 The Function of CBIR Algorithms in Photo Retrieval

The analysis of user needs in the photo archives embracing a variety of subject areas (e.g. museums, advertising, mass communications) suggests that text-based methods will remain the basic access method in the foreseeable future. Proper name and object type searches are the most commonly identified categories of photo needs. Users define their needs very often using contextual criteria that cannot be derived directly from the photos but rather from the assigned textual descriptions, e.g., news events or general themes [1, 16, 18]. Further, textual descriptions are assigned to photos by most suppliers (e.g. news photo agencies) regardless of the access methods applied. For instance, journalists need captions to check the background of a photo before selecting it for publishing [19].

CBIR algorithms could make a contribution in photo retrieval systems. Basically, as pointed out by Svenonius [27] and Enser [5], text based retrieval of images is a vicarious access method and visual access methods have a high potential to enhance retrieval capabilities. However, we want to emphasise that the role of CBIR algorithms as the core of self-contained general purpose photo retrieval systems is not very probable. Rather, we expect them to be exploited as a part of integrated systems which support both textual and content-based retrieval.

Our earlier studies [18-19] indicated that end-users of a digital photo archive

1. had difficulties in formulating focused textual queries
2. often preferred browsing even in very large thumbnail image sets instead of formulating a more refined query
3. browsed thumbnail images to find photos looking different (matching different illustration ideas), and to check if similar photos were available for a photo already found and judged interesting
4. made the final selection by examining enlarged or printed (candidate) photos and their caption texts applying criteria that are difficult to manage by any textual or visual retrieval methods.

We concluded that CBIR algorithms could be a potential technology in developing browsing tools for large sets of thumbnail photos to help in effective collecting of candidate photos.

The above observations help to identify a potential function for CBIR algorithms as a subsystem in an integrated photo retrieval system. The main function is to *retrieve photos similar to the one already found or known*. Before performing the query-by-example, the user does not know if the query photo is the best of the similar photos available in the database. It just has some features that match to the interests of the user. The query photo may have features that the user does not like or features that are unimportant [21]. This leads to three major questions for the evaluation of CBIR algorithms:

1. How effective are the algorithms in similarity matching from the user viewpoint? This gives the general goal for the evaluation, but falls mainly outside the scope of this paper.
2. How can the effectiveness of the CBIR algorithms be measured, i.e. what experimental setting could be relevant for the evaluation? This issue is discussed in Section 2.2.
3. What aspects of similarity are important to users in a realistic search situation? This issue is discussed in Section 2.3 and in the empirical part of this paper.

2.2 The Idea of a Laboratory Test Collection for the Evaluation of CBIR Algorithms

It may sound daring to propose an evaluation framework based on a laboratory test collection. The trend in IR research is towards user-oriented evaluation frameworks emphasising user-system interaction and the essential connection of searching to some background work processes [3, 12]. User-oriented evaluation approaches have also been proposed and experimented with in photo retrieval evaluations (e.g. [14]). We have justified reasons to make a move apart from the general trend.

As pointed out earlier, CBIR algorithms are at an early stage of development. Presently the algorithms are not commonly applied in operational systems, and there is no population of "real" users available for experiments. One possibility is to integrate the CBIR algorithms into an operational text-based photo retrieval system and run user-oriented evaluations by organising a group of potential users to search in simulated task contexts. Because we know so little about the performance characteristics of the CBIR algorithms this approach may turn out to be a time consuming and expensive one. It may also be difficult to make reliable observations about the performance of the CBIR algorithm because of the complex experimental setting. We therefore need an evaluation framework which allows the evaluation of just the CBIR algorithms with respect to human photo assessment - human perceived photo similarity.

The motivation of standard test collections is to provide a common platform for different research groups to conduct performance tests on different algorithms and achieve comparable results. The joint knowledge of CBIR phenomena should cumulate more effectively than in the case of individualistic efforts. One obvious advantage is that new algorithms could be developed and tested more economically and in shorter cycles.

The traditional laboratory framework for evaluation is based on (and has been criticised for) several restrictive assumptions:

1. No interaction takes place between the systems and the user (batch mode queries).
2. Requests and queries are pre-defined and static.
3. Retrieval effectiveness is the primary dimension of performance.
4. Effectiveness measures are based on the relevance of the retrieved documents to requests [11].

These assumptions do not make the laboratory framework worthless, but limit the area where the results are valid. The basic unit observed is a single query and the results can be generalised only within that context. For instance, very little can be inferred about their usefulness to a user or their superiority over other types of retrieval methods (e.g. querying vs. browsing). The results just measure the relative effectiveness of two algorithms having an identical function (e.g. to find similar photos) but applying different matching techniques. Obviously, the relative effectiveness of different algorithms is sensitive to the characteristics of the test

collection, and we need to know them to make justified decisions in constructing integrated photo retrieval systems for the user-oriented experiments.

Our aim is to base the evaluation of CBIR algorithms on the photo similarity judged by the users in the context of real tasks. The idea of solving a realistic illustration task is to guarantee that the user is anchoring similarity assessments to such criteria that are plausible in his/her work. It also means that the sets of similar photos are not randomly selected but rather those that are most appropriate for the users. The illustrators are not looking for average photos but those which capture people's attention. This is a way to build a bridge between the laboratory framework in evaluation and the user needs and work contexts.

In Figure 1 we outline the steps for acquiring photo similarity assessments in creating the test collection.

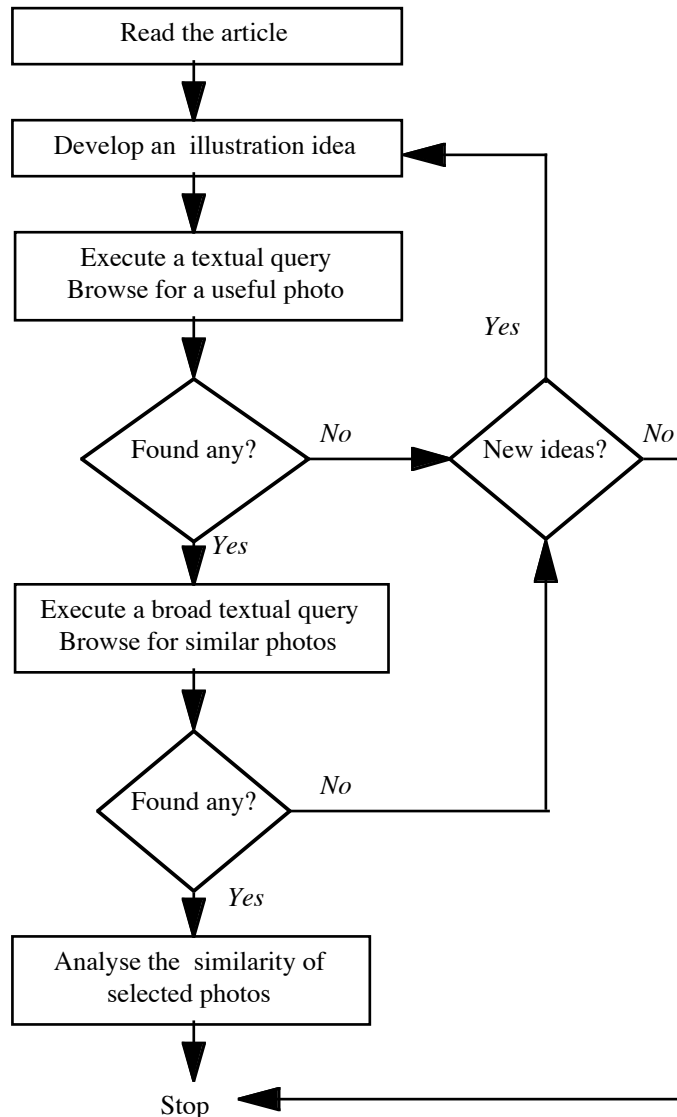


Figure 1: The Steps Involved in Creating the Test Collection Based on the Perceived Similarity of Photos.

We assume a person (here a journalist) receiving an illustration task² (an article) simulating an episode of his/her daily work (see [18-19]). The user develops an illustration idea, makes textual queries and browses thumbnail images (as in real life) to find a photo to illustrate the article. The next step is to identify photos similar to the illustration photo. A new textual query is formulated to retrieve a set of photos matching topically the illustration idea. The user browses the thumbnail images and selects all the photos (s)he considers similar to the illustration photo.

² In an *illustration task*, one or more photos are sought and selected for a particular article. This task is quite open so that quite different types of photos can be selected. *Illustration ideas* are specified photo needs that are potential answers to an illustration task [18].

The users are encouraged to think aloud during the whole search and selection process and justify the selections and similarity assessments. This argumentation is a base for categorising sets of similar photos from different search sessions and users.

The idea of the test collection is outlined in Figure 2. A test collection consists of n test sets each relating to one illustration task. A test set includes a query photo (the illustration photo) and a small database of *topical photos* resulting from the broad textual query. The topical photo set includes a set of s photos perceived similar to the illustration photo and t other photos fulfilling only the textual query criteria. Fig. 2 presents the components of the test set i .

CBIR algorithms are tested by matching the query photo and the topical photo set. Standard performance measures for the ranked output can be used, e.g., precision at the fixed recall levels ($R=0.1-1.0$) or after selected number of top documents (document cut-off values DCV= 5,10,15,20, ...). Also other measures based on the average position of similar photos on the ranked output list can be used. The test collection is very compact and portable because it consists of n test sets, each containing a limited topical photo set.

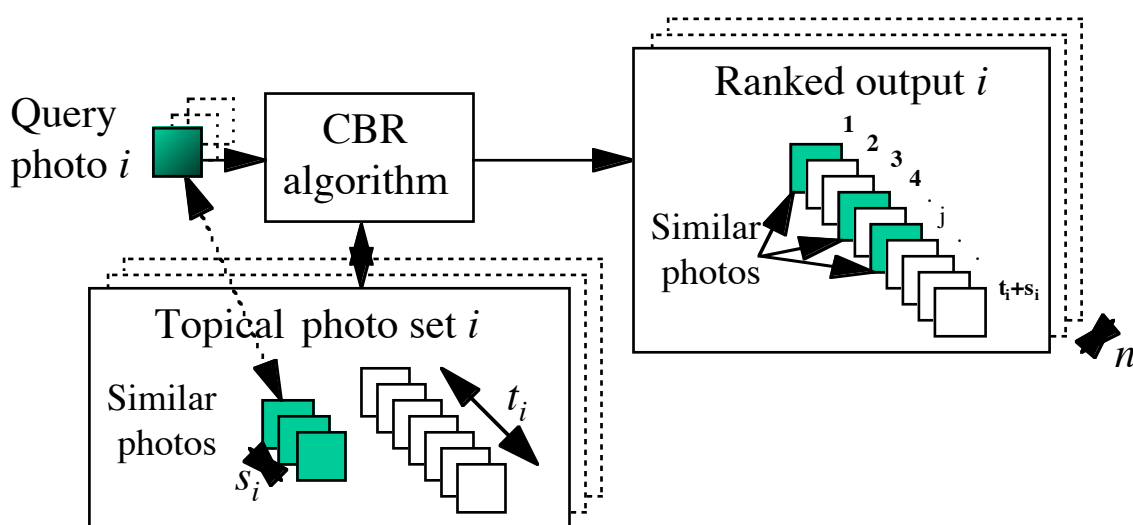


Figure 2: The Test Collection of n Test Sets and Its Use.

2.3 The Concept of Photo Similarity

Photos can be described by various content and context related attributes [15, 24], and thus their similarity is also a multidimensional concept. Two photos may be more or less similar with respect to one or more of these attributes. Similarity may be based both on *visual* and *contextual* attributes.

The visual attributes of similarity are associated to something that can be seen in the photo (the content). They might be quite concrete, such as the existence of a certain object in a photo; shooting distance, e.g. 'face photos' of a person instead of full portraits; the season of the year in the photo etc. The visual attributes might be impressions interpreted by users from visual clues such as a photo with 'a sad atmosphere' or 'action'. The contextual attributes may refer to themes (e.g. 'nuclear power') or news contexts (e.g. 'Bosnia war') with which the photos are associated, or to the attributes of the photo itself, e.g. photo source or cost. In the evaluation of CBIR algorithms the similarity should be based on visual attributes.

We have selected similarity instead of relevance as the primary concept on which to base performance measures. Applying the user-oriented view of relevance (see [20, 23]) to the illustration tasks in the newsroom, a photo can be said to be relevant if it is useful for the user in accomplishing the illustration task. Two relevant photos may be dissimilar since they may originate from different illustration ideas. For instance, a particular news article may be illustrated by a photo of President Clinton giving a statement or by a photo of demonstrating students in Teheran. In our terminology, "similarity" expresses how close the values of defined visual attributes of a photo are to those of another photo, e.g. two photos of a person.

Content-based image retrieval algorithms base measuring the similarity of photos on a low level visual features like colour, shape and texture. People are expected to perceive similarity on a much higher abstraction level [17], but these issues have not been studied empirically. Jørgensen [15] has collected data about the attributes people identify and name when viewing images. Unfortunately, the results do not help us much since the experiment was conducted without linking images to any search task, illustration task or a work context.

Why is it important to study similarity perceptions in a real work context? There are several reasons for doing so. First of all, different people look at or focus on different attributes of a photo [15]. By taking "the man in the street" to assess relevance or similarity, we will probably get more permissive and less consistent assessments than from the expert in a work context. For instance, Spink et al. [26] observed that those persons having a clearer definition of the problem they were to solve, assessed the relevance of documents more precisely (fewer documents were assessed as partially relevant). If we compare a judge who only has a fuzzy idea about the illustration task, and a journalist being an expert at it, the difference could be even more significant.

Although we are studying similarity instead of relevance, the effects of the underlying work processes to the resulting assessments may be equally essential. Observing the whole search process from developing the illustration idea up to the examination of candidate photos helps to collect data about the criteria applied in similarity assessments at different stages of the search process.

3 An Exercise: Building the Test Collection

The goals of the exercise were to

1. test the methodology described for creating a test collection for CBIR, based on user perceived similarity and real illustration tasks
2. find out what criteria journalists apply in assessing the photo similarity.

The exercise was based on simulating typical illustration tasks that journalists perform in the news-room (see [18]). Four journalists illustrating newspaper articles on a daily basis at their work were used as test persons. Each test person had two hours' time to carry out the following tasks for as many articles as (s)he could:

1. *Simulated illustration task.* Illustrate an article. The aim was to produce a photo against which the similarity of other photos would be assessed. We also wanted to find out which photo attributes are exploited in selecting photos for illustration.
2. *Similarity assessment task.* Select photos similar to the illustration photo. The aim was to produce the similarity set and to ascertain how journalists assess photo similarity.
3. *Grouping task.* If possible, group the selected similar photos and describe the difference between groups. The aim was to reveal if there was heterogeneity among the photos selected as similar. Firstly, the criteria applied may have changed during the interactive online selection phase. Secondly, the similarity criteria perceived in the selection phase may relate to different photo attributes, i.e. the photos are selected using varying criteria. The grouping would also reveal the most important criteria applied in each task since all the photos selected similar would supposedly share some criteria while the less important criteria would vary.

During the tasks the test persons were encouraged to think aloud and all their utterances were tape recorded. The researcher participated actively during the tasks by asking questions.

3.1 Simulated Illustration Task

As illustration tasks, we used the following six articles published in the *Aamulehti* newspaper, the second largest newspaper in Finland.

Article		Journalist			
	Title	1	2	3	4
1	Clinton's success rose despite the sex scandal	Task 1	Task 4		
2	Internet home page for Yeltsin	Task 2			
3	Ronaldo elected as the best European player of the year		Task 5	Task 7	
4	Ajax plays for record	Task 3			
5	Ice hockey in the American way (increasing violence in ice hockey)		Task 6		Task 9
6	The innocent victims of the 'holy war'			Task 8	Task 10

Table 1: The Illustration Tasks Conducted by the Test Persons

Two test persons carried out the whole procedure for three articles in the two hours time allowed. Article 6 seemed to take more time than the others. The two test persons working with this article managed to conduct the procedure for only two articles. In all, 10 illustration tasks were performed. Articles 1, 3, 5 and 6 were handled twice while Articles 2 and 4 were handled only once.

One article was given to the test person at a time. The test person was asked to read it and then explain the illustration idea (s)he had in mind. In most cases, the test persons gave quite concrete definitions of the photo they wished for the illustration. For example, "*I'm going to look for an informal photo of President Yeltsin. He should be alone and be doing something. I don't want any portraits or protocol photos in which he is giving a speech.*"

Article 6 was an exception. This article did not produce as concrete descriptions of the illustration ideas as others. The illustration of this article was performed more iteratively and ideas were created during browsing the thumbnail images.

The image retrieval system, based on thumbnail images, textual representations of photo attributes and textual queries, and the interface used in the tasks were the same as used in test persons' work-place in *Aamulehti* newspaper. Thus, no instruction sessions on the use were needed. The photo collection consisted of 25 000 low resolution images and was divided into sub-collections by photo source and archiving year.

The test persons were allowed to search freely for the illustration. They formulated queries and browsed thumbnail images until they were satisfied with the result. Often some candidate photos were printed on paper and the final selection was made from these photos. The searching behaviour was similar to that observed in the newsroom (see [18]).

The test persons were asked to explain why the selected photo was suitable for illustrating the article, which photo attributes they considered important in it and why it was considered a 'good photo'.

3.2 Similarity Assessment Task

The test persons were asked to pick up photos similar to the one they selected for illustration. The concept *similar* was explained in the following way: "The photos may be similar to each other in many ways. For example, there may be the same persons in the photos, the photos may visually resemble each other, they may express the same feeling or atmosphere or they may document the same news event. You may yourself define what you consider similar." We emphasised that we were mainly interested in similarity perceived from the photos, not from the associated information given, for example, in captions. However, at this stage the test persons were allowed to select photos on what ever basis they preferred.

The photo sets used in the similarity assessments (i.e. topical sets, see Figure 2) were retrieved from the database individually for each similarity assessment session. We did not want to exhaust the test persons but wished to offer them enough photos. The researcher assisted the journalist in formulating one or more textual queries to retrieve browsable sets of photos on the topic of the illustration idea. Differences in the size of the photo sets were inevitable, since some topics were poorly represented in the collection (see Table 2). While the test persons were working on the task they were asked to explain why they considered the photos they looked at similar or dissimilar. The photos judged similar to the illustration photo were printed on paper.

3.3 Grouping Task

The paper prints of photos judged as similar were put side by side on the table and the test persons were asked to recheck if the photos were similar to the illustration image. Then we asked the test person to say if there were any qualities common to all the photos selected. After that, we offered the opportunity to group the photos. If groups were formed, the test persons were asked to describe the grounds for the groupings made. Last, the test persons were asked to explain which photos were (un)suitable for illustrating the article and why.

4 Results on User-Perceived Similarity

The test persons selected different photos for illustration, even though illustrating the same article. Thus, the photos against which the similarity assessments were made were different for each test person. The test persons created similarity sets of three to nine photos (see Table 2). The similarity criteria expressed by them varied according to the article and illustration idea generated. However, each test person applied mainly the same criteria in both selecting the illustration photo (illustration task) and selecting photos similar to that (similarity assessment task). This implies that the actual work context, illustrating an article, was kept in mind when assessing similarity.

4.1 The Similarity Criteria Applied in Online Browsing Sessions

Assessment Stage One: Browsing the Thumbnail Images

The first impression seemed to be crucial since only a small share of photos was examined more closely (i.e. thumbnail images enlarged and captions read). The selection of potential thumbnail images for illustration as well as potential "similar" photos depended solely on the visual attributes (please remember that the set of thumbnail images was already fixed by the attributes expressed in the initial query). The test persons seemed to concentrate on the essential attributes when browsing thumbnail image sets. One test person explained:

"What draws your attention...well, you keep the story and the idea you have in mind. It's quite quick to separate those...for example, there's a single goal-keeper, there's the stand, a single player... if you are looking, as I am, for a mass, for more players, those photos catch your eye."

Some of the attributes required of a suitable or a similar photo, often spontaneously defined before the selection session, were recognised in the small thumbnail images. The *number of persons*, *shooting distance* and *colours* often seemed to be the perceived and exploited photo attributes in the first place. The number of persons in the photo (e.g. "*Ronaldo alone*", "*Bill and Hillary Clinton alone*", "*a mass of ice hockey players*") seemed to be important since this attribute was mentioned frequently both in the context of the illustration task and the similarity assessment task. The shooting distance was the next most often expressed criterion at this stage. This criterion came up often when the test person was asked why some photos were discarded ("*These are not at all similar. They are shot from a much longer distance*"). Photos with few and bright colours were looked for. Often the colours were deemed to be too light or dark. The colours were also exploited in indicating the background (e.g. "*there's green...also a football field in the background*").

The test persons also tended to examine photos which did not possess the attributes desired but which differed from the other photos in the set. These photos were claimed to look so interesting that the test persons wanted to look at them more closely (e.g. "*No, this is not similar. It just seemed so interesting that I wanted to see it*"). These photos were explained to be untypical in regard to their colours or *composition* and thus stood out from the set.

Assessment Stage Two: Examining the Enlarged Photos and Captions

When photos were examined closely, the *gestures and facial expressions* (e.g. "*a triumphant smile*") of persons emerged as important criteria in both the illustration and similarity assessment tasks. In all, these attributes were referred to in nine tasks (out of ten). Not only were the faces desired to be seen clearly but the expressions on these faces and the gestures created the desired *atmosphere and feeling* in the photo. On the contrary, for Article 6, which concerned Islamic countries, the desired atmosphere, a feeling of threat, was generated from the covered faces of Muslims.

The specified *action taking place* (e.g. playing) in the photos seemed to be a criterion of similarity in three tasks. The *background* (e.g. a football field) of persons and *other objects* (e.g. flags, posters) in the photo were said to be important in the context of three tasks.

Some attributes such as the *photo composition* were mentioned more often during the similarity assessment task than during the illustration task. We did not find out exactly what the test persons meant by the 'composition', but at least criteria such as "*the persons are in the middle in the photo*" were put under this term. Other attributes possibly gathered under the label 'composition' were also mentioned frequently: *similar cropping*, *shooting angle* and *photo direction*. These were also the photo attributes mentioned when the test persons described "*a good photo*".

The photos were looked at more closely because they visually fulfilled the requirements or were otherwise attractive. However, the test persons also applied contextual criteria. *Abstract ideas* (e.g., democracy, religion) were referred to as the basis for similarity in tasks 8 and 10, both concerning Article 6.

4.2 Grouping the Photos

In the grouping task the test persons were asked 1) to describe what was similar, i.e. shared by all the photos judged as similar, 2) to further group the photos judged as similar if possible, 3) to describe which attributes made the difference between groups.

In most cases, the similarity of selected photos seemed to be based on two or three photo attributes. These were considered important in the context of the article and the illustration idea. Regarding other attributes, the photos might be dissimilar. Again, the number of people, the gestures and facial expressions, shooting distance, the background and action were mentioned most often as the criteria shared by all the photos judged as similar. In two cases, the criteria shared by all similar photos were abstract concepts or atmosphere interpreted from the photos or from the context information. Both of these tasks related to Article 6. The photos selected as similar in the context of Article 6 did not need to share the same visual attributes even if this was sometimes the case.

In nine tasks, some groupings were made, revealing that there was some heterogeneity in the set of photos considered as similar (Table 2). We would like to point out that for a given article, the photo against which the similarity assessments were made, was different for each test person.

Task	Article	Photos in Topical sets	Photos selected similar	Groups formed
1	1	203	6	2
2	2	185	4	2
3	4	59	3	2
4	1	203	9	3
5	3	61	8	2
6	5	427	6	3
7	3	61	6	3
8	6	323	4	2
9	5	427	3	-
10	6	199	9	3

Table 2: The Number of Photos Retrieved by Broad Textual Queries, Number of Photos Selected Similar and the Number of Groups Created in Each Task

The number of people, the gestures and facial expressions, shooting distance, the background, action, photo direction, shooting angle and the direction of movement were used as the bases of groupings. For example, when the number of persons (“*alone*”), background (“*football field*”) and composition were mentioned as criteria shared by all the photos in the similarity set, action was the basis of the grouping, i.e., “*playing*” in Group 1 vs. “*posing*” in Group 2 (Article 3). The context information was referred to three times as a basis for forming groups of photos unsuitable for illustration. The *actual news context* (e.g. specific event, country) and the *shooting time* determined if the photos were usable or useless in the context of the article.

In two cases the grouping revealed that the photos selected as similar were actually for two different page layouts. In the first case, the grouping was based on the number of persons in the photo. The test person claimed that one of the photos with two persons (Group 1) could be chosen for the layout where the photo would have a large space on the page. If the photo was used in a small size, a photo of a person alone (Group 2) would be a better choice. The shared attributes were the facial expressions and gestures of the person(s) in the photo (Figure 2).

In the second case, the grouping was based on which direction the photo was “facing”. The meaning of this grouping related again to the layout: the journalist explained that which group to use depended on which side of the article the photo would be: The photo should be facing the article. These cases revealed a problem in our methodology. We had not considered the role of the whole page and its lay out as important as it was.

In addition, some groupings indicated that some photos selected as similar were actually not considered to be very similar. This was due to the data gathering method. To ascertain as many of the criteria applied as possible, the researcher asked why some photos were considered to be dissimilar. The effects of this disturbance can be seen from the following sample in our data: “*No, there's nothing similar...well, they are also walking and the shooting distance is about the same and the angle...*” and soon after “*well, there's a lot of similarity.*”

Some photos selected as similar were judged to be unsuitable for illustrating the article. The explanations were that the photo was connected to some other news event or country than the one the article was dealing with, the photo expressed another abstract idea than desired (e.g., “*army*” instead of “*terrorism*”) or that the photo was not considered “*good enough*”. However, it is difficult to give any numbers on this matter since the test persons sometimes expressed uncertainty. They explained that in the work-place they would consult the writer of the article or someone in the news-room. For example: “*if this ice hockey player is known as a nice guy on the ice and never offensive, then I couldn't use this*”. Illustration is sometimes teamwork.



Figure 2: Photos Selected as Similar in Task 1. Photo selected to illustrate the article in the upper left corner. In the grouping task, the photos were grouped for different page layouts: Group 1: photos 1 and 2, Group 2: photos 3, 4, 5, 6, and 7. Photo 3 formed a subgroup of group 2 based on the shooting distance. Photo 7 was deemed “not good” and thus unsuitable for illustration.

5 Conclusions

In this paper we have introduced an evaluation framework and a test collection for CBIR. We also have reported an exercise testing the proposed methodology in building the test collection, and how journalists perceive the similarity of photos.

The proposed evaluation framework has the following advantages:

1. So far there has been no evaluation framework of CBIR algorithms available.
2. The performance testing of CBIR algorithms is based on photo similarity perceived by typical (expert) users in the context of realistic illustration tasks and environment.
3. The laboratory model and a standard, portable test collection supports efficient prototyping in short development cycles.

In the evaluation framework, the test database from which the CBIR algorithms are supposed to retrieve and rank photos, and the recall base are defined in a unique way, based on concept similarity, not relevance. The test database for each illustration task is built through textual queries and is therefore independent of the algorithms. They may be relatively small (e.g. 100-300 photos), and still the retrieval and ranking capabilities of the algorithms can be measured. If test users are making the broad textual queries in a large operational news photo archive, the resulting test databases are large enough and contain enough variation in colours, shapes and textures. The number of test sets n has, of course, to be high enough to achieve statistically testable results.

The recall base is built on the test user's similarity judgements which form sets of photos judged as similar on explicable criteria. These criteria are based on a real illustration task and the test persons' experience in performing these tasks as a professional work routine. We do not necessarily need to know the similarity criteria applied by the test users in different test queries, but that information is useful. The assessments alone are enough to evaluate average performances, but knowing the similarity criteria helps in making comparisons between differently profiled sets of similar photos.

The recall bases also contain enough variation in visual attributes, which may be secondary to human users, in order to reveal whether the CBIR algorithms too easily estimate as dissimilar such photos which for humans are similar. This may happen, e.g. if portraits of a single person have radically different backgrounds. This algorithm property may also be tested by systematically varying the starting image for the algorithms.

If the CBIR algorithms were to retrieve photos from the whole photo database, they would undoubtedly retrieve visually similar photos not retrieved by the textual queries before. However, at this stage of development, we think that the smaller task proposed in the evaluation framework, based on smaller sets of illustration task dependent photos, is sufficient. We need to learn the capabilities and pitfalls of the CBIR algorithms and to learn to apply them properly. The proposed evaluation framework serves this purpose well.

In the exercise, the four journalists generated ten sets of 2-10 photos judged to be similar in the context of their illustration task. The results showed that the selection of potential photos while browsing the thumbnail images was based on attributes perceivable at a glance. Attributes like the number of persons, shooting distance and colours as well as composition, cropping and the background were mentioned. Such attributes have a role in the first sight assessment of photos. Thumbnail image browsing is the primary area for applying the CBIR algorithms. At this stage, the criteria applied in similarity assessments are clearly more concrete and tangible than at the stage of examining enlarged photos. This is good news for the developers of CBIR algorithms. However, the level of abstraction was in many cases well above the level of elementary features managed by the present CBIR algorithms.

When examining the enlarged photos, the facial expressions and gestures of persons, the action taking place and the feeling or atmosphere in the photo were often referred to. There was some heterogeneity among photos judged as similar and grouping was often found possible by the test persons. All the photos judged as similar seemed to share some important criteria while variance occurred in the values of other attributes.

The journalists taking part in the exercise were experts in their field and were used to select and judge photos quickly. They did not express any major difficulties in conducting the task.

The exercise gave us some important lessons for future work:

- ⊙ Some visually similar photos may be neglected on the basis of contextual criteria if that is available (i.e. the photo related to a specific event other than that discussed in the article and was thus deemed unsuitable for illustrating the article) or just missed. The effect of contextual criteria in the similarity judgements should be removed (e.g., by preventing the scanning of captions while browsing thumbnail images).
- ⊙ In this study we intervened in the similarity assessments by asking why some photos were found dissimilar. This intervention seemed to sometimes cause judging as similar photos which were not in the first place considered similar. In creating the test collection the test persons should be left free to select photos without any disturbance.
- ⊙ The test persons had difficulties in withdrawing from the illustration task. They continued to select photos for new illustration ideas or photos reminiscent of some of the candidate photos selected earlier. The test persons indicated clearly that these were selected for reasons other than similarity. To exclude these from the set of similar photos, the concluding examination of the selected photos as a complete set (grouping task) is a necessary control step.
- ⊙ The page lay-out seems to be of major significance. The test persons should be given an approximate description of the page lay-out, at least the size of the photo on the page and on which side of the article it would be placed.
- ⊙ There was some heterogeneity among the photos judged as similar. The test persons applied varying criteria in assessing the similarity. This means that the selection of the query photo when testing the CBIR algorithms need further consideration. In our view, there are two possibilities: 1) The query photo should be the illustration photo against which the similarity assessments were originally made. We cannot assume that the CBIR algorithm performs well if any other photo in the similarity set is used as the query photo. 2) The similarity set used in the tests is the group in which the user placed the illustration photo. Now any photo in the similarity set could be used as a query photo.
- ⊙ The source database must be large enough (e.g. similar to the operational ones, 100 000 - 300 000 photos) to guarantee a credible querying environment for different illustration ideas generated by the test users. It is highly probable that the users change their behaviour in small databases where the output overload is not a real problem.

Vakkari and Järvelin [29] have analysed the literature on information seeking and retrieval and suggested a holistic framework for developing task-embedded IR systems. The categories that should be taken into account in systems development include: tasks, actors (users), (organisational) environments, information needs, sources of information, seeking/searching strategies, retrieval techniques, success (including relevance and effort), as well as time. Each of these is multidimensional and may be classified in several ways. Vakkari and Järvelin

propose that systematic relationships between the categories should be studied, i.e., which kinds of actors benefit from which kinds of strategies in what kinds of tasks. They criticise the laboratory model of IR for being limited to an artificial task and for not being responsive toward requirements of tasks, actors, environments, etc.

In the present proposal for an evaluation framework of CBIR algorithms tasks, actors, and environments are taken into account by carefully developing the recall bases with respect to realistic illustration tasks, professional illustrators and operational environment. Therefore the performance of CBIR algorithms can be evaluated on realistic criteria. This helps to judge whether CBIR algorithms can be successfully integrated into systems which support real users in their real tasks.

It is clear that the proposed evaluation framework cannot reasonably produce valid findings concerning the performance of CBIR algorithms in other domains, different from the newspaper article illustration domain. Other illustration domains need to be studied and corresponding test collections designed. Working on a broad spectrum of domains guarantees a solid base for evaluation.

6 References

1. Armitage LH, Enser PG. Analysis of user need in image archives. *Journal of Information Science* 1997; 23:287-299.
2. Batley S. Visual information retrieval: browsing strategies in pictorial databases. In: *Online information 88. Proceedings of the 12th International Online Information Meeting, Vol 1*. Learned Information Ltd, Oxford, 1988. pp 373-381.
3. Borlund P, Ingwersen P. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation* 1997; 53(3):225-250.
4. Eakins JP. Automatic image content retrieval - are we getting anywhere? In: *Proceedings of the Third International Conference on Electronic Library and Visual Information Research, De Montfort University, Milton Keynes, May 1996*. pp 123-135.
5. Enser PG. Pictorial information retrieval. *Journal of Documentation* 1995; 51:126-170.
6. Forsyth DA, Malik J, Fleck MM, Greenspan H, Leung T, Belongie S, Carson C, Bregler C. Finding Pictures of Objects in Large Collections of Images. In: *Proceedings of the International Workshop on Object Recognition, Cambridge, April 1996*. Available URL: <http://www.cs.berkeley.edu/~daf/>.
7. Gong Y. *Intelligent image databases : towards advanced image retrieval*. Boston, Kluwer Academic Publishers, 1998.
8. Gudivada VN, Raghavan VV. Modeling and retrieving images by content. *Information processing & Management* 1997; 33(4):427-452.
9. Gupta A, Jain R. Visual information retrieval. *Communications of the ACM*, 1997, 40:71-79.
10. Harman D. *The First Text Retrieval Conference (TREC-1)*. Gaithersburg: National Institute of Standards and Technology. (NIST Special Publication 500-207), 1993.
11. Harter P, Hert C. Evaluation of Information Retrieval Systems. Approaches, Issues, and Methods. In: Williams ME (ed) *Annual Review of Information Science and Technology* 32. Information Today, Medford, New Jersey, 1997. pp 3-62.
12. Hersh W, Pentecost J, Hickam D. A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science* 1996; 47(1):50-56.
13. Holt B, Hartwick L. "Quick, Who Painted Fish?". Searching a Picture Database with the QBIC Project at UC Davis. *Information Services & Use* 1994; 14(2):79-90.
14. Jose JM, Furner J, Harper DJ. Spatial querying for image retrieval: a user-oriented evaluation. In: Croft W B et al. (eds) *SIGIR '98, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Melbourne, Australia, 1998, pp 232-240.
15. Jørgensen C. Attributes of images in describing tasks. *Information Processing & Management* 1998; 34(2/3):161-174.

16. Keister LH. User Types and queries: impact on image access systems. In: Fidel R, Hahn TB, Rasmussen EM, Smith PJ (eds) *Challenges in indexing electronic text and images*. Learned Information, Medford, New Jersey, 1994. pp. 7-22.
17. Lai TS, Tait J. General Photographic Image Retrieval Simulating Human Visual Perception. In: *Multimedia Indexing and Retrieval*. A workshop held in conjunction with the ACM SIGIR Conference, Melbourne, Australia, 28 Aug 1998, pp 17 - 28.
18. Markkula M, Sormunen E. Searching for photos - Journalists' practices in pictorial IR. In: Eakins JP, Harper DJ, Jose JM (eds) *The Challenge of Image Retrieval*. Electronic Workshops in Computing (eWIC), 1998. URL: <http://www.ewic.org.uk/ewic/workshop/view.cfm/CIR-98>.
19. Markkula M, Sormunen E. End-User Searching Challenges Indexing Practices in the Digital Photo Archive. Accepted for publication in *Information retrieval Journal*.
20. Mizarro S. Relevance: The Whole History. *Journal of the American Society for Information Science* 1997; 48(9): 810-832.
21. Picard RW, Minka TP, Szummer M. Modeling user subjectivity in image libraries. M.I.T. Media Laboratory, Perceptual Computing Section Technical Report No. 382, 1996. (also IEEE Int. Conf. On Image Proc., Lausanne, Sept. 1996). Available at URL: http://picard.www.media.mit.edu/cgi-bin/tr_pagemaker
22. Rasmussen EM. Indexing images. In: Williams ME (ed) *Annual Review of Information Science and Technology* 32. Information Today, Medford, New Jersey, 1997. pp 169-196.
23. Schamber L, Eisenberg MB, Nilan MS. A re-examination of relevance: towards a dynamic, situational definition. *Information Processing and Management* 1990; 26:755-776.
24. Shatford S. Analyzing the subject of a picture: a theoretical approach. *Cataloguing and Classification Quarterly* 1986; 6:39-62.
25. Sparck-Jones K. *Information retrieval experiment*. London, Butterworths, 1981.
26. Spink A, Greisdorf H, Bateman J. From highly relevant to not relevant. Examining different regions of relevance. *Information Processing & Management* 1998; 34(5):599-622.
27. Svenonius E. Access to Nonbook Materials: The Limits of Subject Indexing for Visual and Aural Languages. *Journal of the American Society for Information Science* 1994, 45(8):600-606.
28. Tague-Sutcliffe J. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management* 1992; 28(4), 467-490.
29. Vakkari P, Järvelin K. *Tasks, Information Seeking and Retrieval: Elements towards Task-Embedded Information Retrieval*. Tampere, Finland: University of Tampere, Department of Information Studies, manuscript, 25 p.