

A Test Collection for the Evaluation of Content-Based Image Retrieval Algorithms - A User and Task-Based Approach

Marjo Markkula, Marius Tico[#], Bemmu Sepponen, Katja Nirkkonen and Eero Sormunen

University of Tampere, [#]Tampere University of Technology, Finland

{marjo.markkula, eero.sormunen}@uta.fi, tico@cs.tut.fi

Published in *Information Retrieval* 4(3/4), 275-294 (2001)

Abstract: Content-based image retrieval (CBIR) algorithms have been seen as a promising access method for digital photograph collections. Unfortunately, we have very little evidence of the usefulness of these algorithms in real user needs and contexts. In this paper, we introduce a test collection for the evaluation of CBIR algorithms. In the test collection, the performance testing is based on photograph similarity perceived by end-users in the context of realistic illustration tasks and environment. The building process and the characteristics of the resulting test collection are outlined, including a typology of similarity criteria expressed by the subjects judging the similarity of photographs. A small-scale study on the consistency of similarity assessments is presented. A case evaluation of two CBIR algorithms is reported. The results show clear correlation between the subjects' similarity assessments and the functioning of feature parameters of the tested algorithms.

1 Introduction

Content-based image retrieval (CBIR) algorithms have been developed as a promising access method for digital image collections and are a subject of vast research efforts. Content-based image indexing aims to automatic identification and abstraction of the visual content of an image. A common model is that the images in the collection are described with a set of feature vectors. Typical features used are color, shape and texture. The query is made by an example image (e.g. photograph, drawing, sketch). The user may specify which feature parameters are important or default settings are used in matching. The algorithm compares the feature vector of the query image to the feature vectors of collection images. The retrieved images are ranked according to some calculated similarity order by applying best-match techniques. (E.g. del Bimbo 1999, Gong 1998, Gudivada and Raghavan 1997, Gupta and Jain 1997, Picard et al. 1996.) For other approaches in CBIR see e.g. Belongie et al. (1998) and Das et al. (1999).

The usability and usefulness of CBIR algorithms in real image search situations is an unexplored area. One obvious bottleneck of the emerging technologies is that they operate at a very low level of visual abstraction (Eakins 1996, Gupta and Jain 1997). The analysis of user needs in the photograph archives embracing a variety of subject areas (e.g. libraries, museums, photo stock services, mass media) has shown that photographs of named persons and object types are the most common categories of user needs. Further, users define their needs very often using contextual criteria that cannot be derived directly from the photographs but rather from the assigned textual descriptions, e.g. news events. (Armitage and Enser 1997, Enser 1995, Keister 1994, Markkula and Sormunen 1998, 2000.)

CBIR algorithms do not seem very useful for general-purpose photograph databases as a self-contained retrieval method. Rather, they could be exploited as a part of integrated systems, which support both textual and content-based retrieval. The results of our field study on searching behaviors in digital photograph archives suggested that CBIR algorithms could be a potential technology for developing browsing tools for large sets of photographs retrieved by textual queries. Even though users often express their needs at quite high level of abstraction, at the browsing stage they seem to apply lower-level selection criteria. (Markkula and Sormunen 2000.)

From the evaluation viewpoint, content-based image retrieval methods are at 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 (Harman 1993, Tague-Sutcliffe 1992). The performance evaluation of the CBIR systems based on realistic user criteria is nearly an unexplored area in IR. For CBIR algorithms, there are no standard test collections or evaluation frameworks available like TREC (e.g. Voorhees and Harman 1997) in the text retrieval domain (Rasmussen 1997).

One of the major motivations for developing test collections, and conducting laboratory experiments 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 IR 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 (Harman 1993).

A task-oriented approach to create a test collection for the performance evaluation of CBIR algorithms was proposed in Sormunen et al. (1999). The goals and the procedure of building a test collection were illustrated, and the results of a pilot study focusing on the building process were presented. In this paper, we continue by reporting the main results from the implementation project of the test collection.

First, we will summarize the basic ideas of the proposed approach, and describe the building process and the characteristics of the test collection. Next, the test results measuring consistency of similarity assessments made by different persons having different task related information is presented. The main findings of a case evaluation of two CBIR prototypes are also reported. The paper will conclude by discussing the strengths and weaknesses of the proposed task-based evaluation method.

2 The test collection

2.1 Premises

The idea of the test collection and the way of using it for the evaluation of CBIR algorithms is based on three premises (Sormunen et al. 1999):

The function of CBIR algorithms. Our field study on photograph retrieval suggested that CBIR algorithms could play a major role in solving the problem of browsing large query sets in text-based image retrieval systems (Markkula and Sormunen 2000). Thus, the function of CBIR is limited to the problem of ranking large sets of topically related photographs with respect to a user-selected query photograph. If the user identifies one interesting photograph from a large query result set, the CBIR algorithm should help to locate other visually similar photographs from that set.

A realistic retrieval task. Image retrieval tasks defined for the test collection are composed by real users in realistic work contexts. In our case, we focused on photographic needs originating from routine illustration tasks in the newsroom (journalistic work).

Independent, user-defined similarity criteria. The users make content related similarity assessments using criteria perceived appropriate in their work situation. The users are encouraged to express the similarity criteria used so that they are as explicit as possible. The evaluator has to

be aware of the nature of similarity criteria used. The higher the level of abstraction used by real users, the tougher the challenge will be for the CBIR algorithms tested.

2.2 Building the test collection

The test collection was constructed through illustration tasks simulating newsroom practices in illustrating newspaper articles (see Markkula and Sormunen 1998, 2000). Ten newspaper journalists working for the newspaper Aamulehti, the second largest newspaper in Finland, were engaged in the building process. All subjects were accustomed users of the digital photograph archive of Aamulehti. The database used in the building contained a sample of 50 000 photographs from this archive. The retrieval system¹ used was identical to the system at Aamulehti. It is based on standard text retrieval methods with browsing facilities for thumbnail surrogates of photographs (see Markkula and Sormunen 2000). Thus, the subjects were familiar with the environment.

The test collection consists of 45 test sets. Each test set was created through the following steps (Figure 1).

Simulated illustration task. A journalist was given an illustration task² consisting of a newspaper article and related layout information (the newspaper section, space and location for the photograph on the page). Layout information was included because our field study (Markkula and Sormunen 2000) showed that it has a significant impact on the selection criteria of journalists. The articles had been published in Domestic, Foreign, Economics, Sports, Culture and Current Affairs sections and in the Sunday and Weekend supplements. The subject was asked to read the article and explain what kind of illustration idea(s) she had in mind.

¹ NewsLink by Job Systemintegration AB

² In an *illustration task*, one or more photographs are sought and selected for a particular article. This task is quite open so that different types of photographs can be selected. *Illustration ideas* are specified needs that are potential answers to an illustration task (Markkula and Sormunen 2000).

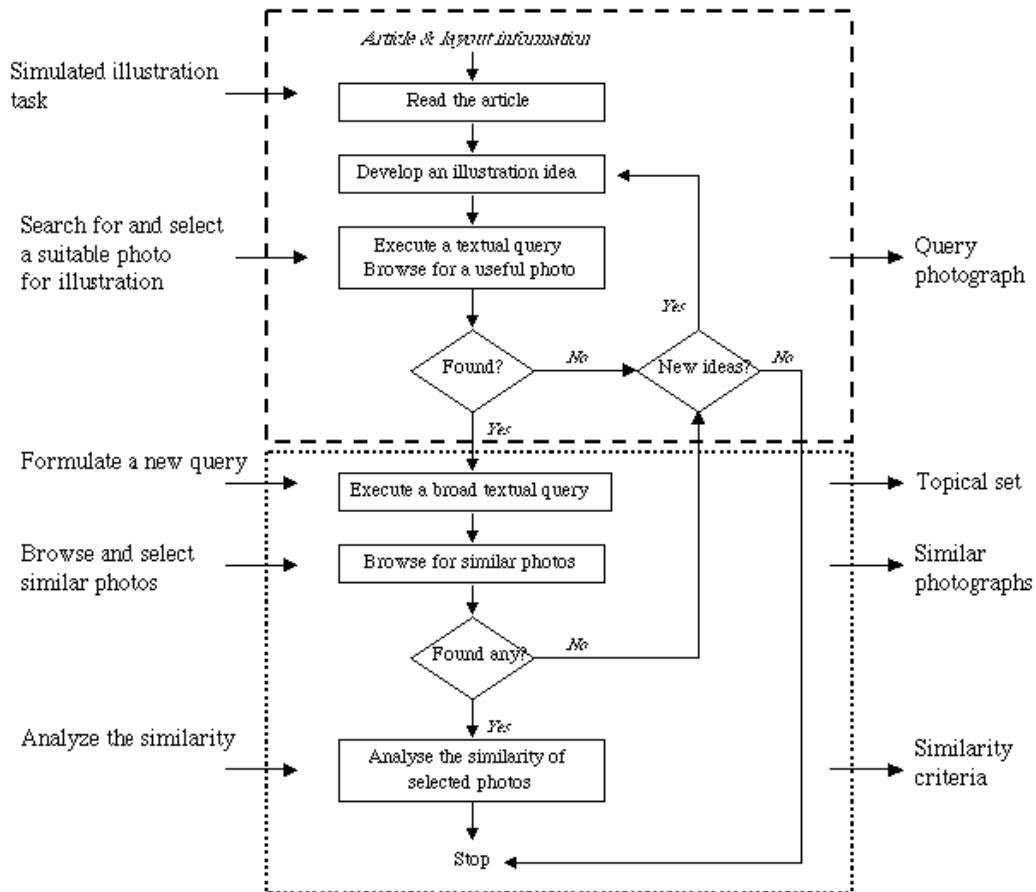


Figure 1. The procedure of building the test collection.

Selecting an illustration photograph. The subject was allowed to search freely the database as well as interactively create new illustration ideas. She made one or more textual queries and browsed thumbnail images to find a photograph to illustrate the article. The search topics of the subjects were various, associated, for example, to named persons (*Princess Diana*), objects (*a mobile phone*) and news themes (*Kosovo refugees*). When the subject found a photograph fulfilling his/her requirements, it was printed on paper.

Retrieving topical sets for similarity assessments. A new textual query was formulated on the search topic of the illustration photograph to obtain a large but not insuperable set of photographs for browsing. The sets were created by simple textual queries typical to end-user practices. In many cases, the original query or union of original queries³ formulated by the

³ The journalists prefer simple queries. Instead of formulating a complex query of synonymous terms or different aspects of the search topic many one-word queries are made. (Markkula and Sormunen 2000).

journalist matched this requirement. If additional query reformulation was required the researcher helped to broaden (by using truncation, adding synonyms) or narrowing (restricting by time) the query. Queries for the above topics were '*Princess Diana*', '*mobile phone# or gsm*' and '*Kosovo and refugees and archiving date=1999-01-01-1999-12-31*'.

Judging the similarity. The subject browsed the topical set of photographs retrieved by the textual query, and selected those photographs she considered as similar to the one selected earlier for illustration. A printed color copy of the illustration photograph was available for reference. Color copies of the photographs judged as similar were printed on paper.

Analyzing the similarity. A reassessment session of similarity in which the subject was asked to justify his/her selections was made with the paper prints. The subject was allowed to eliminate photographs if she wished, and was then asked to group the remaining photographs into similarity subgroups if she found it possible. In the preliminary study, we learned that the reassessment session where photographs are observed side by side is necessary to control the reliability of similarity assessments (Sormunen et al. 1999). The users were encouraged to think aloud during the whole search and selection process and justify the selections and similarity assessments.

The components of the test collection are outlined in Figure 2. Each test set is built up of **a query photograph** (illustration photograph) generated at **step 2**, a **topical set** of photographs retrieved by a textual query at **step 3**, a '**similarity set**', set of photographs assessed as similar to the query photograph (a subset of the topical set) at **step 4**, and respective **similarity criteria** collected at **steps 4 and 5**. Each test set is composed of photographs reflecting some area of user need and retrieved by a textual query. Therefore the photographs in each topical set contain variation in visual attributes. The whole topical sets were judged for similarity. Some of the photographs in the topical set were perceived as similar to the illustration photograph and others fulfill only the textual query criteria.

The advantage of this approach is that the test collection is quite compact. The performance of CBIR algorithms can be tested by matching the query photograph and the topical photograph set. Standard performance measures can be used for the ranked output to measure the algorithms' performance in retrieving the user assessed similar photographs. If appropriate techniques are applied to create the test sets and similarity assessments, there should not arise any major validity or reliability problems in experimental designs. For details, see Sormunen et al. (1999).

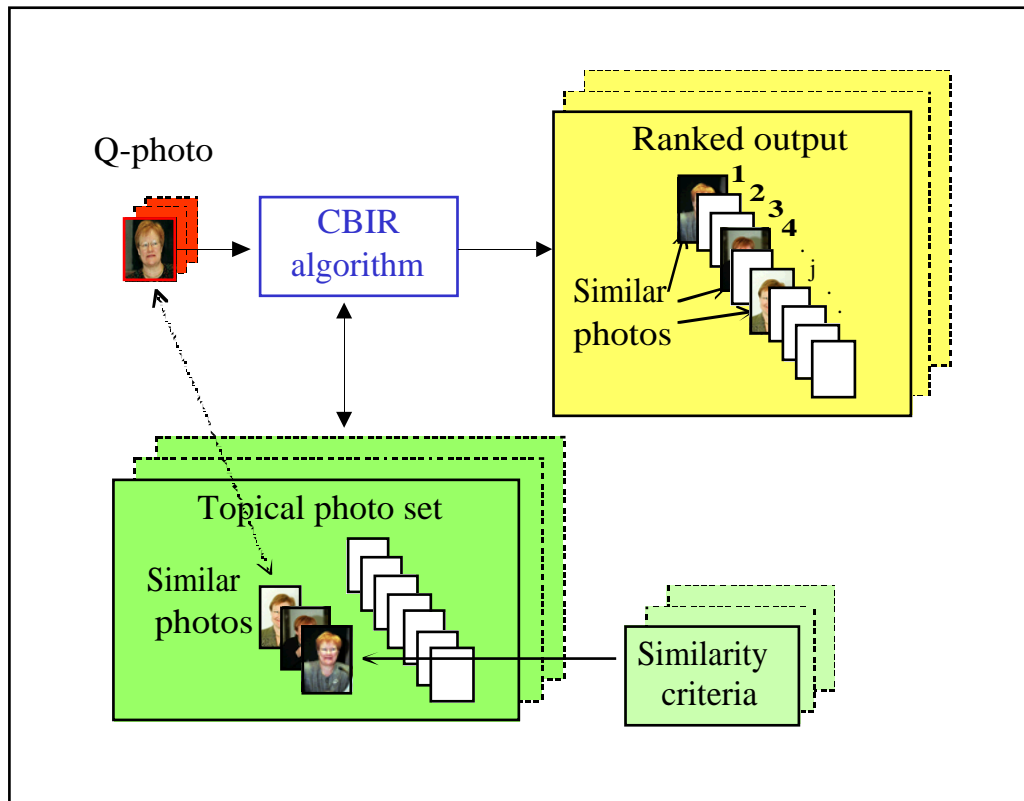


Figure 2. The basic components of the test environment for CBIR algorithms.

The tasks were conducted in two-hour sessions. During a session the journalists completed on average four processes.

In all, 57 illustration tasks were initiated but 12 of them failed to generate a test set. Two tasks did not proceed to the similarity assessment step, because the journalist could not find any acceptable photograph to illustrate the given article. In six cases, the similarity assessment step failed: the journalist did not accept any photographs in the topical set as similar to the illustration photograph. Two tasks were abandoned at step 3, due to the small number of photographs in the database on the topic of the journalist. Two similarity assessment tasks were unsuccessful because the journalist was not able to terminate the illustration process and start the similarity assessment task. The subject continued to select possible photographs relating to different illustration ideas. He realized and explicated this at the analysis session (step 5). Thus, the test collection consists of 45 test sets produced by successfully completed illustration and similarity assessment tasks.

2.3 Characteristics of the test collection

The basic characteristics of the test collection are presented in Appendix A. In all, the test-collection consists of 45 test sets. It is quite compact containing in all 6392 photographs. The task-based approach has some consequences on the test collection:

The sizes of the topical sets vary from 59 to 429 photographs, the average being 142 photographs. The subjects were free to create illustration ideas and the topics they might come up were not restricted in any way. Some topics were well represented in the source collection while on some topics there were only few photographs available. Basically the size of the topical set is not very critical problem for the experimental design. However, sets smaller than 50 photographs were rejected.

Similarity sets are typically small (from 1 to 25 photographs). The subjects were free to select any number of photographs they considered similar enough to pass their criteria. Typically, the journalists applied quite tight criteria and selected only a few photographs (see Table 1). Very small similarity sets may induce uncontrollable deviation in the results on the performance of different feature parameters. The query photograph and one or two other photographs constituting its similarity set may by chance share other visual features than those defined as similar by the user.

The size of the similarity set	Number of test sets
1	4
2	7
3	4
4	10
5	4
6	2
7	3
8	1
9	1
10-	9

Table 1. The number of test sets in different size categories of similarity sets

The criteria applied for similarity sets vary. The subjects mentioned on average 4.7 similarity criteria per similarity set. Usually, various types of criteria were applied for one set. We coded the criteria to 15 different types and organized these according to their level of abstraction (Table 2). The typology is modified from the ideas of Eakins (1996), Panofsky (1970) and Shatford (1986).

	Criterion type	Occurrences (N=210)	Number of test sets (N=45)
Level I	Low-level attributes, e.g. colors, lightning, composition	21	13
	Aspect ratio, e.g. in horizontal size	2	2
	Cyclic time: winter, evening	2	2
	Type of location, e.g. on beach, in ice hockey ring	6	6
	Appearance of objects, e.g., shooting distance & angle	39	22
	Level I criteria in total	70	29
Level II	Type of object, e.g. trains, cars	44	27
	Named object, e.g. members of Spice girls	3	3
	Named location, e.g. in Finland	4	4
	Action, e.g. pointing with finger, fighting	8	8
	Type of event, e.g. rock festivals	3	3
	Level I and II criteria in total	132	39
Level III	Abstract concept, e.g. violence, popularity	26	19
	Emotions of persons, e.g., smiling, looking happy	7	7
	Atmosphere, & feelings, e.g. intensive, sad	22	10
	Symbolism, e.g. winner, icon	20	11
	Reference to actions before the photo was taken	3	3
	Criteria in total	210	45

Table 2. The types of similarity criteria expressed by the subjects distributed to three levels of abstraction, their overall occurrences and the number of test sets in which the criteria type occurred.

At the level I, criteria closest to the attributes commonly exploited by content-based image retrieval algorithms concern the low-level attributes of photographs, i.e. brightness (*dark, bright, strong contrast*), colors (*dark background, red flag*) and composition (*upper halves of the photographs are sky*). Aspect ratio (the ratio between image height and width) is a feature parameter of some CBIR algorithms, e.g. CST-Demo⁴. Seasons (winter) or time of day (evening) as well as types of locations (on beach, in ice hockey ring) are often closely related to low-level features such as colors and brightness of the photograph. Yet another criteria type correlates to the low-level features. These criteria concern objects in the image but does not involve much semantic knowledge. They rather define how the objects are presented in terms of cropping, shooting distance (*close-ups*) and shooting angle (*front*).

Semantic reasoning is required for the level II criteria. These concern types of objects (*trains*), named objects (*members of Spice girls*), actions (*fighting*), event types (*rock festivals*) and named locations (*in Finland*).

The criteria at level III represent the highest level of abstraction. These are abstract ideas (*violence, charity work*), emotions expressed by persons (*happy, smiling*), overall mood or feelings in the photograph (*intensive, uncomfortably crowded, sad*), and symbolic meanings (*winner, irony*). In addition, for three similarity sets, criteria referring to what had happened just before the photo was taken were applied, e.g. "*He (swimmer Jani Sievinen) has just reached the finish*".

The level I criteria correspond closest to the feature parameters of current CBIR algorithms. For level II criteria, the progress of CBIR has been more limited. The recognition of objects in generic contexts is problematic due to the various possible representations of such objects. Conventional object recognition techniques cannot identify general objects, for instance, to classify people and cows to different sets. However, approaches based on rich image descriptions of special object types (e.g. specific material surfaces, horses, naked people) have been successfully applied. (Forsyth et al. 1996, 1997.) The criteria applied by the subjects that define the appearance of objects should facilitate their content-based retrieval. The criteria types at level III might be unattainable by CBIR algorithms since they require a high degree of abstract or subjective reasoning. The subjects mentioned more criteria in the context of small similarity sets (1-2 similar

⁴ by Convera

photographs / 5.3 criteria on average) than for very large sets (ten or more similar photographs / 3.6 criteria). This implies that the small similarity sets were typically selected with tighter criteria than the large ones. Moreover, for small similarity sets, the criteria tend to be from lower-levels of abstraction than for the very large ones. For the small similarity sets, the share of level III criteria was 21% of all the criteria expressed. For similarity sets containing more than ten photographs it was 75%.

For six of the similarity sets, the subjects explicated only level III criteria. The suitability of these test sets for the evaluation of CBIR algorithms should be questioned. However, we have not abandoned any of the test sets at this point. First, we do not know what is the relation between the high-level criteria and the low-level features. Second, the subjects may have not expressed all the criteria they actually applied. Some criteria may have been difficult to put in words, conceptualize or even be aware of.

2.4 The consistency of similarity assessments

The consistency of judges in assessing the relevance of text documents is a thoroughly studied phenomenon. We do not know what level of consistency the similarity assessments of images can reach. Another interesting question is how much the simulated work situation (illustration task) affects the similarity decisions.

Nine of the journalists who previously participated in the building of the test collection were engaged in testing the consistency of similarity assessments. The subjects conducted the similarity assessments between the illustration image (query photograph) and the topical set of 13 test sets taken from the test collection. However, none of the subjects worked with the same illustration tasks and set of photos as in the original building process.

Two kinds of assessments were conducted:

A Similarity assessment with work task information. The subject was given the query photograph, the newspaper article illustrated, the layout information and the topical photograph set. The subject was asked to read the article, browse through the topical set, and select the photographs which (s)he considered as similar to the query photograph.

B Similarity assessment task without work task information. The subject was given the query photograph and the topical photograph set and asked to browse through the topical set and select

those photographs (s)he considered as similar to the query photograph.

Eight subjects conducted six (three of both types) and one subject four (two of both types) similarity assessments. Five of the subjects conducted the assessments in order A, B and four in B, A. Thus, every task was carried out four times, two times with the work task information, two times without it. The tasks in both groups were rotated.

The consistency of photograph similarity selections by subject 1 in relation to subject 2 was calculated using the formula

$$CI_{1,2} = \frac{|I_1 \cap I_2|}{|I_1|} \quad (\text{Saracevic 1984}),$$

where I_1 is the set of photographs selected by subject 1, and I_2 is the set of photographs selected by subject 2. The consistency $CI_{2,1}$ of subject 2 in relation to subject 1 was calculated in a similar way. Pairwise consistency between the two subjects is the average of $CI_{1,2}$ and $CI_{2,1}$. An overall consistency for the whole group was calculated by averaging the pairwise consistencies (Iivonen 1995, Sormunen 2000).

The similarity assessments made with the work task information (type A) appeared to be more consistent (59%) than the assessments of type B without work task information (48%) (Figure 3). In nine test sets, the work task information lead to more consistent assessments. In four test sets, the absence of task information yielded higher consistency. Without work task information, the consistency varied from 0% to 88% between the test sets. The range was slightly smaller, from 16% to 94%, when the work task information was presented. The work task information seems to raise the level of consistency. However, the differences between the two groups were not statistically significant.

The selections made in the consistency test were also compared to the original selections of the similar photographs made in the building of the test collection. The procedure of assessing the similarity was slightly different in the two cases, and one has to be cautious in direct comparisons. However, it is surprising how small the consistency differences were between the original selections and later selections with and without the work task information.

The results suggest that the availability of the task information increases the consistency of similarity assessments made by experts by focusing the subjects' perception on the features of

the photograph which are relevant to that task situation. The level of consistency was lower than in the relevance assessments of textual documents (e.g. 83% in Kekäläinen 1999). Different subjects seem to select slightly different photograph sets but any of them, together with the query image, reflect professionally perceived photograph needs in a realistic work situation.

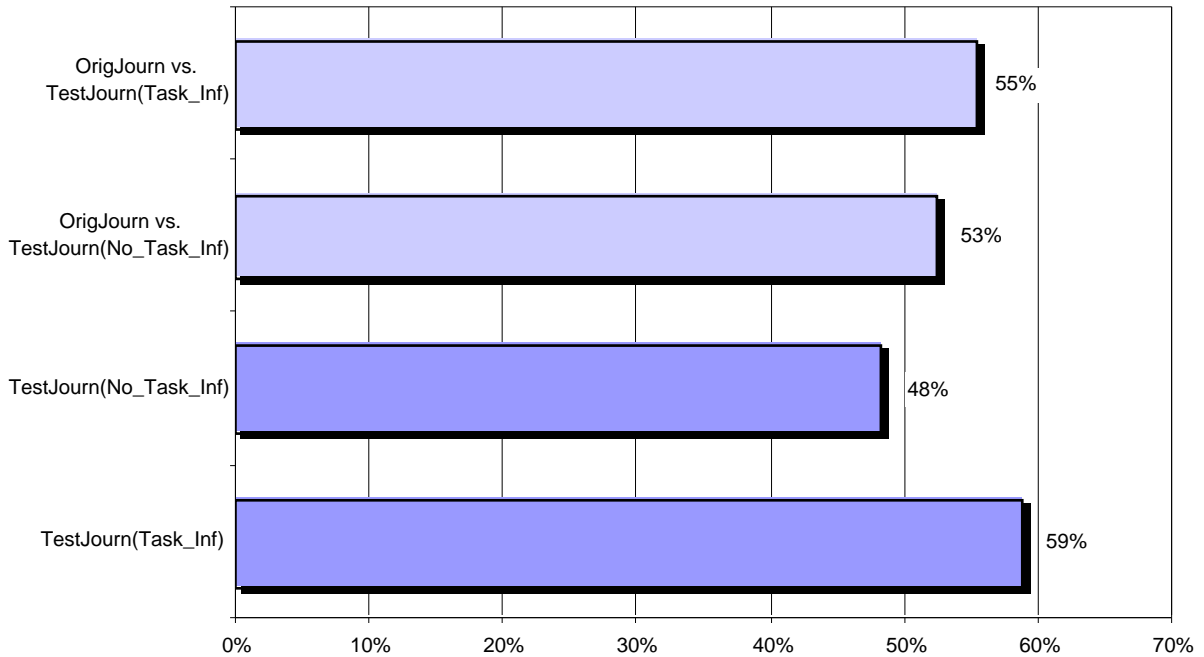


Figure 3. Consistency of similarity assessments between different groups.

3. A case experiment with the test collection

The goal of the case experiment was to demonstrate the use of the test collection and the adopted methodological framework in a typical evaluation situation. The case was expected to reveal some pragmatic strengths and weaknesses of the task-based approach in creating the test collection. Further, we aimed to challenge the CBIR algorithms by the test data that was derived from a realistic use context.

3.1 CBIR algorithms tested

CST Demo is a demonstration version of Excalibur Visual RetrievalWare, a product of Convera. It provides image retrieval based on six feature parameters: Color Content (CC), Shape Content (SC), Texture Content (TC), Color Structure (CS), Brightness Structure (BS) and Aspect Ratio

(AR). The details of the implementation are not publicly available but we regard CST Demo as a typical CBIR search engine of many available (QBIC is another well-known commercial CBIR system).

The other CBIR algorithm was a prototype called *Image retrieval system based on color and orientation* developed in Digital Media Institute (DMI) at Tampere University of Technology. The prototype (from now on DMI Proto) comprises retrieval modules based on color (Tico et al. 1999a, 1999b, 2000), and edge orientation. A short description of the three image features used by DMI Proto is presented in the following.

The Color Histogram (CH) comprises 20 bins, from which 16 are allocated to describe the distribution of color “type” in the image (e.g., red, blue, green, etc.), and the remaining 4 bins are used to describe the distribution of intensity. The type of color can be perceived only in the *chromatic* image region where it is well described by the value of Hue (H) color component. The remaining image pixels, that form the *achromatic* image region, exhibit few intensity levels between black and white, and are characterized by the value of their Intensity (I) component. The standard deviation of the RGB components is used to define the chromatic image region as a fuzzy set, and the contribution of each image pixel to the color histogram is weighted according with the degree of membership to the chromatic image region. The intersection operator, proposed by Swain and Ballard (1991) is used to evaluate the similarity degree between color histograms.

The Color Layout (CL) is meant to mitigate the lack of spatial information in the color histogram. The image is divided in a certain number of rectangular regions (16x16 regions in this implementation of DMI Proto), and the color layout, that may be seen as a low-resolution representation of the digital image, is created by retaining the dominant color inside each image region. The similarity degree between two images, based on their color layouts, is defined as the number of corresponding regions that exhibit the same dominant color in both images.

The Orientation Descriptor (OD) is constructed on the basis of edge orientation, and it is used by the DMI Proto algorithm as an efficient way to capture shape information present in the image. The image is divided in blocks of 8x8 pixels and the dominant edge orientation is estimated inside each block using a Least Squares method. The presence of a dominant edge orientation inside an image block is reflected by a parameter, called *certainty level* that achieves high values in those blocks where a dominant orientation exists. Neglecting the image blocks whose certainty levels are below a certain threshold, the orientation descriptor is created such that to resemble the joint

probability of two orientations lying in neighborhood blocks. In the current implementation of DMI Proto algorithm, the orientation is quantized on 16 levels, and hence the orientation descriptor is represented as a 16x16 co-occurrence matrix. The similarity between orientation descriptors is evaluated based on the intersection operator (Swain and Ballard 1991). It is of importance to mention that the orientation descriptor used in the current implementation of the DMI Proto algorithm, is not rotation invariant. Consequently, it slightly favors the retrieval of images that contain round like shapes (e.g., human faces, wheels, balls, etc.).

Any combination of the features provided by CST Demo and DMI Proto may be exploited in searching. Both systems are best-match systems and query results are sorted in the calculated order of similarity.

3.2 The testing procedure and measures applied

The 6329 low-resolution photographs and their thumbnail copies of the test collection were imported to both CBIR algorithms. Each photograph in the test collection is named uniquely. The test sets are numbered, and the name of each photograph is coded to start with the test set number. This allows us to restrict the search to one test set at a time although the whole collection is in one database. The median size of the low-resolution photographs used in the indexing process was $267 \times 350 = 93450$ pixels.

For each test set, the query image of the test set was used for the query and all single feature parameters and their possible combinations were applied. The proportional weights of parameters were freely adjustable but we used only binary weights. Either the parameter was "on" or "off" meaning that all the feature parameters applied ("on") had an equal proportional weight in ranking ($= 1/n$, where n is the number of parameters applied). In addition, for each test set a search with all parameters "off" was performed and used as a baseline to which the results of the parameter searches were compared.

For CST Demo, 63 parameter searches were performed for each of the 45 test sets. These included all combinations of the feature parameters available starting by every single parameter "on" in turn, and ending with all parameters "on". In total, 2835 searches were performed in the test program. For DMI Proto, the seven feature parameter combinations and 45 test sets produced in all 315 searches.

Following evaluation measures were used: **Average precision** is the mean of precision values calculated at the positions of the similarity set photographs in the ranked result set. **Average rank** is the average of ranks (positions) of similar photographs in the result set. Precision was also calculated at document cutoff values DCV_5 , and DCV_{20} . Average precision and precision at DCV:s are standard measures used in TREC (see, e.g. Voorhees & Harman 1997). Average rank has seldom been applied in text retrieval experiments but it seems to be quite concrete and illustrative measure for CBIR testing where the sets of similar photographs are quite small. Precision at standard recall levels was also calculated but not reported here since the small number of similar photographs led to heavy interpolation.

Ideally, the baseline would be the original order of the images in the file but in importing the photographs to CST Demo this order subtly changed. The baseline of DMI Proto was slightly better than that of CST Demo. Therefore, we focus on the differences between the measured performance and the baseline to make the figures as comparable as possible.

3.3 Results

An overall summary of results averaging the performance over all feature parameter combinations (63 combinations for CST Demo and 7 combinations for DMI Proto) are presented in Table 3 and 4. Both algorithms tested improved the order of photographs in the ranked output when compared to the baseline. In CST Demo, the position of the user assessed similar photographs raised on average 19 ranks and in DMI Proto 16 ranks. Other measures showed similar improvements. The results imply that there is a correlation between the photograph similarities computed by the algorithms and the similarities assessed by the users.

'Color Content and Shape Content' (CC&SC) was the best performing combination of CST Demo improving the average rank of retrieved images from 73 to 47, i.e. by 28 ranks. In terms of average precision, the combination 'Color Content and Shape Content and Color Structure and Aspect Ratio' (CC&SC&CS&AR) was the best increasing precision from the baseline of 8.9% up to 21%, i.e. precision more than doubled. However, many parameter combinations produced almost as good results. For instance, the average precision for the above-mentioned 'CC&SC' combination was 20.4%. Even the worst single feature parameter 'Aspect Ratio' (AR) could demonstrate positive effects although the average position of photographs raised only by 7 ranks (precision up 1.4 %).

Measure	Baseline	Average performance of parameters		Performance of the best parameter combination		
	Score	Score	Change	Combination	Score	Change
Average rank	72.8	53.6	-19.2	CC&SC	47.1	-25.7
Average Precision	8.9%	18.3%	+9.4%	CC&SC&CS&AR	21.0%	+12.1%
DCV5	5.4%	13.6%	+8.2%	CC&CS&AR	18.1%	+12.7%
DCV20	5.6%	10.4%	+4.8%	TC&CS&AR	12.1%	+6.5%

Table 3. The average performance and the performance of the best feature parameter combination of CST Demo (45 test sets).

Measure	Baseline	Average performance of parameters		Performance of best parameter combination		
	Score	Score	Change	Combination	Score	Change
Average rank	69.1	53.2	-15.9	CH&OD&CL	48.5	-20.6
Average Precision	10.9%	17.2%	+6.3%	CH&OD&CL	19.5%	+8.6%
DCV5	5.5%	13.9%	+8.4%	OD&CL	15.8%	+10.3%
DCV20	5.5%	10.2%	+4.7%	OD&CL	11.4%	+5.9%

Table 4. The average performance and the performance of the best feature parameter combination of DMI Proto (45 test sets).

The Friedman two-way analysis by ranks was based on the precision and rank scores to test the statistical significance of the observed performance differences. In CST Demo, the analysis between 64 rank and precision scores (baseline included) over 45 test sets showed that the baseline scored significantly worse than the parameter searches (average precision $p < 0.005$ and average rank $p < 0.0001$). Significant differences appeared also between the performance of many parameter combinations, which are not specified in this paper. For the DMI Proto, the test showed similar results (average precision $p < 0.05$ and average rank $p < 0.0005$).

We designed an additional test to ascertain that the positive results originated from the parallel interpretations of similarity by the feature parameters and the users, not from some unknown variable. A complete series of test searches was performed for CST Demo. The query image was

kept the same but the similar photographs were replaced with randomly selected photographs. For each test set, we selected as many random photographs as were in the original similarity set. The random photographs were selected next to the original similar photographs in the baseline. Thus, the average of ranks in the baselines of the new random sets and the original similarity sets were about the same.

The results from this test verified that the feature parameters of CST Demo correlated with the users' similarity assessments. The parameter searches on randomly selected sets did not achieve any improvement to the baseline. In fact, they slightly worsened the order of the images from the user point of view (Table 5).

Measure	Baseline	Average performance of parameters	
	Score	Score	Change
Average rank	72.1	72.3	+0.2
Average Precision	9.4%	8.4%	-1.0%
DCV5	6.3%	4.4%	-1.9%
DCV20	6.0%	5.1%	-0.9%

Table 5. The *CST-Demo* results from a test run executed with randomly selected images representing the similarity sets (45 test sets).

In the analysis session (see Figure 1) of the similarity selections, the subjects were asked to group the selected photographs if they found it possible. More than one group was created for 29 of the similarity sets. The group into which the query image was placed was considered to consist of the most similar photographs to the query image, and called the *Core group*.

A test was designed to test whether the feature parameters of CST Demo rank the Core group photographs (perceived as most similar by the journalists) higher than other photographs in the similarity sets. Since the number of photographs in the similarity set affects the measured performance, two sets (one of Core group photographs and the other of other photographs in the similarity set) of equal sizes and about the same average ranks were created for each of the 29 test sets. In seven test sets, the number of Core group photographs was reduced, in 19 test sets the number of other similar photographs was reduced.

The results showed that the feature parameters of CST Demo managed slightly better with the Core group photographs than with the other photographs in the similarity set (Table 6). Average rank was 6.6 ranks and average precision 0.8% better for the Core group photographs than for the other similar photographs. This indicates that CST Demo construed the change in the degree of similarity to a parallel direction with the journalists. However, both groups of similar photographs worked in the same direction. This means that whole similarity sets may be used in testing. This is important if we wish to keep the sizes of similarity sets as large as possible.

Measure	Core group			Other similar photographs			Change ¹ - Change ²
	Baseline	Parameters/ average	Change ¹	Baseline	Parameters/ average	Change ²	
Ave rank	65.8	55.8	-10.0	67.1	63.7	-3.4	-6.6
Ave Precision	8.9%	11.9%	+3.0%	7.7%	9.9%	+2.2%	0.8%
DCV5	5.2%	9.3%	+4.1%	3.9%	6.9%	+3.0%	1.1%
DCV20	4.8%	7.7%	+2.9%	4.2%	6.2%	+2.0%	0.9%

Table 6. The performance of *CST Demo* on the Core group photographs assessed as most similar by the subjects compared to the performance on other photographs in the similarity sets.

4 Conclusions

The test-collection introduced in this paper is an effort to bridge the gap between the development of CBIR algorithms and user demands. So far, the research comparing human and algorithm judgements of image similarity has been few. However, this kind of research is essential for the development of CBIR algorithms to benefit the users of digital image collections.

In the present test collection, the performance of CBIR algorithms is tested against the similarity assessments of photographs made by professional end-users in the context of simulated but realistic illustration tasks and operational environment. Therefore, the performance of CBIR algorithms can be evaluated on realistic criteria. This gives a solid basis to judge whether they can be successfully integrated into systems, which support real users in their tasks.

As far as the authors know, the task-based approach in building test collections has not been applied before. This study showed that this approach can be successfully used but has some

weaknesses when compared to traditional test collections for text retrieval. The building process, though not very laborious, is vulnerable. In our study, about one fifth of all the tasks initiated failed to construct a valid test set.

The journalists were allowed to work freely when creating illustration ideas, searching the database and selecting the illustration/query photograph and judging the similarity. Therefore, the size of the resulting test sets, the number of photographs judged as similar in each test set as well as the similarity criteria applied vary greatly. In the task-based approach, it is difficult to guarantee that the collection is balanced in respect of variables mentioned.

The subjects applied similarity criteria ranging from low-level features such as colors to abstract ideas. CBIR algorithms perform on low-level features, and some of the test sets might be too challenging for the algorithms at their current stage of development. However, all the test sets reflect realistic, professionally perceived photograph needs in newspaper illustration domain giving an appropriate goal for the development of algorithms.

The case evaluation performed showed that the test-collection can be successfully used to test the performance of CBIR algorithms. In this paper, we have presented mainly the results based on the overall performance of the feature parameters over all 45 test sets. The work will be extended to the performance analysis of different feature parameters on individual test sets.

An interesting question for our future work is how the similarity criteria of users and the feature parameters of CBIR algorithms correlate. For users this information is essential to benefit of the CBIR algorithms. The averaged results showed clear correlation between the human similarity judgements and the photograph similarity computed by the algorithms. These results are encouraging news for the developers of CBIR methods. On the other hand, similarity criteria and their level of abstraction varied from one search to another. It is impossible to identify one parameter combination, which works best, or have a positive effect on the ranks of retrieved photographs in all searches. A challenging goal would be to develop an image retrieval system, which allows the user to select which kind of similarity (s)he is emphasizing in a particular search situation.

Acknowledgements

This work was funded by the Academy of Finland. We also thank the newspaper Aamulehti and photo agency Lehtikuva for their co-operation. Thanks are also due to the journalists participating in the study.

References

- Armitage L and Enser P (1997) Analysis of user needs in image archives. *Journal of Information Science*, 23:287-299.
- Belongie S, Carson C, Greenspan H and Malik J (1998) "Color- and texture-based image segmentation using the expectation-maximization algorithm and its application to content-based image retrieval. Sixth International Conference on Computer Vision, pp. 675-682. <http://www.cs.berkeley.edu/projects/vision/publications.html> (visited June 18th, 2001).
- Del Bimbo A (1999) *Visual information retrieval*. Morgan Kaufmann, San Francisco.
- Das M, Manmatha R, Greenspan H and Malik J (1999) Indexing flowers by color names using domain knowledge-driven segmentation. *IEEE Intelligent Systems*, 14(5):24-343.
- Eakins J (1996) 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, pp. 123-135.
- Enser P (1995) Pictorial information retrieval. *Journal of Documentation*, 51:126-170.
- Forsyth DA, Malik J, Fleck MM, Greenspan H, Leung T, Belongie S, Carson C, Bregler C (1996) Finding pictures of objects in large collections of images. In: *Proceedings of the International Workshop on Object Recognition*, Cambridge. <http://www.cs.berkeley.edu/~daf/> (visited June 4th 2001).
- Forsyth DA, Fleck MM (1997) Body plans. In: *Proceeding of IEEE Conference on Computer Vision and Pattern recognition*, Puerto Rico.
- Gong Y (1998) *Intelligent image databases : towards advanced image retrieval*. Kluwer Academic Publishers, Boston.
- Gudivada V and Raghavan V (1997) Modeling and retrieving images by content. *Information processing & Management*, 33:427-452.
- Gupta A and Jain R (1997) *Visual information retrieval*. Communications of the ACM, 40:71-79.
- Harman D (1993) *The First Text Retrieval Conference (TREC-1)*. Gaithersburg: National Institute of Standards and Technology. (NIST Special Publication 500-207).

Iivonen M (1995) Consistency in the selection of search concepts and search terms. *Information Processing & Management* 31:173-190.

Keister L (1994) User types and queries: impact on image access systems. In: Fidel R, Hahn T, Rasmussen E, Smith P, eds. *Challenges in indexing electronic text and images*. Learned Information, Medford, New Jersey. pp. 7-22.

Kekäläinen J (1999) The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval. Doctoral thesis. Tampere: University of Tampere. *Acta Universitatis Tamperensis*, ISBN 951-44-4596-1.

Markkula M and Sormunen E (1998) Searching for photos - Journalists' practices in pictorial IR. In: Eakins J, Harper D, Jose J, eds. *The Challenge of Image Retrieval*. *Electronic Workshops in Computing (eWIC)*, 1998. URL: <http://www.ewic.org.uk/ewic/workshop/view.cfm/CIR-98>.

Markkula M and Sormunen E (2000) End-user searching challenges indexing practices in the digital photograph archive. *Information Retrieval* 1: 259-285.

Panofsky E (1970) *Meaning in the visual arts*. Penguin, London.

Picard R, Minka T and Szummer M (1996) 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). http://picard.www.media.mit.edu/cgi-bin/tr_pagemaker (visited June 18th 2001)

Rasmussen E (1997) Indexing images. In: Williams M, ed. *Annual Review of Information Science and Technology* 32. *Information Today*, Medford, New Jersey, 1997. pp. 169-196.

Saracevic T (1984) Measuring the degree of agreement between searchers. In: Flood B, Witiak J, Hogan H, eds. *ASIS '84: proceedings of the American Society for Information Science 47th annual meeting*, vol. 28. White Plains, NY: Knowledge Industry Publications, pp. 227-230.

Shatford S (1986) Analyzing the subject of a picture: a theoretical approach. *Cataloguing and Classification Quarterly*, 6:39-62.

Sormunen E (2000), A Method for measuring wide range performance of boolean queries in full-text databases. Doctoral Thesis. Tampere: University of Tampere. *Acta Electronica Universitatis Tamperensis*, ISBN: 951-44-4732-8. <http://acta.uta.fi/pdf/951-44-4732-8.pdf> (visited June 18th 2001).

Sormunen E, Markkula M and Järvelin K (1999) The Perceived Similarity of Photos - Seeking a Solid Basis for the Evaluation of Content-based Retrieval Algorithms. In: Draper S. et al, eds. *Mira 99: Evaluating interactive information retrieval*. Glasgow, UK, 4-16 April, 1999. *Electronic Workshops in Computing*. URL: <http://www.ewic.org.uk/ewic/workshop/fetch.cfm/MIRA-99/>.

Swain MJ and Ballard H (1991) Color Indexing. *International Journal of Computer Vision*, 7(1):11-32.

Tague-Sutcliffe J (1992) The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management* 28:467-490.

Tico M, Haverinen T and Kuosmanen P (1999a) An unsupervised method of rough color image segmentation. In: *Proceedings of the 33th Asilomar Conference on Signals, Systems and Computers*, vol. 1. Pacific Grove, California, 1999. pp. 58-62.

Tico M and Kuosmanen P (1999b) An efficient sparse data filtering method for image histogram comparison. In: *Proceedings of the 11th Scandinavian Conference on Image Analysis (SCIA'99)*, Kangerlussuaq, Greenland, 1999. pp. 715-722.

Tico M, Haverinen T, and Kuosmanen P (2000) A method of color histogram creation for image retrieval. In: *Proceedings of the Nordic Signal Processing Symposium (NORSIG'2000)*, Kolmarden, Sweden, 2000. pp. 157-160.

Voorhees E and Harman D (1997) *The Fifth Text REtrieval Conference (TREC-5)*. Gaithersburg: National Institute of Standards and Technology. (NIST Special Publication 500-238)

APPENDIX A. The basic characteristics of the test collection

Test set #	Topic name	Photos in topical set	Photos in similarity set	Similarity subgroups
1	Ronaldo	176	6	3
2	Islamic women	61	9	4
3	Car traffic	134	2	1
4	Rock festivals	107	4	2
5	Elections	147	2	1
6	Öcalan	91	3	1
7	Paavo Väyrynen	67	13	4
8	Wembley stadium	77	1	1
9	Bosnian refugees	182	7	2
10	Sauli Niinistö	149	25	4
11	Spice girls	83	4	2
12	Princess Victoria	59	4	3
13	Martti Ahtisaari	200	4	2
14	Tarja Halonen	124	6	2
15	Monica Lewinsky	88	3	1
16	Queen Elisabeth	208	11	1
17	Ice hockey	371	4	2
18	Nuclear power	218	2	1
19	Jani Sievinen	134	2	1
20	Formula1	155	2	2
21	Iraq	110	5	1
22	Mika Häkkinen	148	2	1
23	Algeria	108	4	2
24	Mika Halvari	98	5	2
25	Forrest	60	7	3
26	Railway travelling	188	4	2

27	Mobile phone	107	15	3
28	Bill and Hillary Clinton	79	11	3
29	Speeding	105	2	2
30	Yasser Arafat	141	8	2
31	Mika Myllylä	93	3	2
32	Beach tourism	156	5	2
33	Sauli Niinistö (2)	429	1	1
34	VR (Finnish railways)	63	1	1
35	Princess Diana	125	4	1
36	Finnair	70	1	1
37	Ski jumping	304	4	1
38	Chernomyrdin	132	3	1
39	Yeltsin	146	7	3
40	Arafat (2)	141	10	4
41	Islam	271	5	3
42	Bill Clinton	131	11	3
43	Formula1 (2)	148	12	4
44	Kurds	140	12	4
45	Teemu Selänne	68	4	4
Average		142	5,8	
Median		132	4	2
Sum		6392	260	