

## IMAGE RETRIEVAL EVALUATION

*John R. Smith*

IBM T.J. Watson Research Center  
30 Saw Mill River Road  
Hawthorne, NY 10532  
jrsmith@watson.ibm.com

### ABSTRACT

One of the most significant problems in content-based image retrieval results from the lack of a common test-bed for researchers. Although many published articles report on content-based retrieval results using color photographs, there has been little effort in establishing a benchmark set of images and queries. Doing so would have many benefits in advancing the technology and utility of content-based image retrieval systems. In this paper, we address the growing need for establishing a common content-based image retrieval test-bed.

### 1. INTRODUCTION

There are several current obstacles to advancing image retrieval capabilities that have little to do with technology. Consider the following questions:

1. How do we compare the performance of image retrieval methods?
2. How do we select image features and evaluate automatically generated indices?
3. How do we review papers reporting on new content-based retrieval algorithms?
4. How can we collaborate in advancing image retrieval technologies and system capabilities?

Many of these same questions have emerged in the past decade in fields as diverse as text retrieval, data bases, and computer vision. In some cases, these questions have been addressed by establishing a common evaluation test-bed. For example, by creating a text retrieval test-bed, the Text Retrieval Conference (TREC) has become an important forum for evaluating and advancing the state-of-the-art in text retrieval systems.

In this report, we examine the elements of an image retrieval test-bed. We survey some of the image collections, benchmark queries and evaluation methods used by image retrieval researchers. We make some recommendations for establishing a common image retrieval test-bed that consists of a standard image collection, benchmark queries, relevance assessments, and evaluation methods.

### 2. IMAGE RETRIEVAL EVALUATION

The objective of an image retrieval system is to retrieve images (documents) in rank order, where the rank is deter-

mined from the relevance to the query at hand. The overall retrieval effectiveness of the system can be gauged only if the actual relevances are known. In general, an information retrieval system evaluation test-bed consists of

1. a collection of  $N$  documents,
2. a set of  $M$  benchmark queries,
3. a set of ground-truth relevance scores for the benchmark queries,
4. a set of evaluation metrics.

The standard practise in information retrieval for evaluating retrieval effectiveness is as follows: a benchmark query is issued to the system, the system retrieves the documents in rank order, then, for each cut-off value  $k$ , the following values are computed, where  $V_n \in \{0, 1\}$  is the relevance of the document with rank  $n$ :

- Detections:  $A_k = \sum_{n=0}^{k-1} V_n$ ,
- False alarms:  $B_k = \sum_{n=0}^{k-1} (1 - V_n)$ ,
- Misses:  $C_k = \sum_{n=0}^{N-1} V_n - A_k$ ,
- Correct dismissals:  $D_k = \sum_{n=0}^{N-1} (1 - V_n) - B_k$ .

From these values, a number of standard information retrieval measures are computed, such as

- Recall:  $R_k = \frac{A_k}{A_k + C_k}$ ,
- Precision:  $P_k = \frac{A_k}{A_k + B_k}$ ,
- Fallout:  $F_k = \frac{B_k}{B_k + D_k}$ .

The retrieval systems can then be evaluated and compared to other systems based on recall, precision and fallout. For example, a more effective system shows a higher precision for all values of recall.

#### 2.1. Evaluation methods

Currently, the image retrieval research community is using a wide variety of measures for evaluating image retrieval performance. Some examples include:

1. Retrieval effectiveness:  $P_k$  vs.  $R_k$ ,
2. Receiver operating characteristic:  $A_k$  vs.  $B_k$ ,
3. Relative operating characteristic:  $A_k$  vs.  $F_k$ ,
4. R-value:  $P_k$  at cut-off  $k = \text{int}(\sum_{n=0}^{N-1} V_n)$ ,

5. 3-point average: average  $P_k$  at  $R_k = 0.2, 0.5, 0.8$ ,
6. 11-point average: average  $P_k$  at eleven recall points,
7. AVRR: average relevant rank in top  $k$ ,  $(\frac{1}{k} A_k)$ ,
8. MT: relevant and missing in top  $k$ ,  $(C_k)$ ,
9. "First page score": R-value at  $k = 20$ ,  $(P_{20})$ ,
10. Visual inspection of top  $k$ ,  $(P_k)$ ,
11. "Response ratio":  $\frac{A_k}{B_k}$  vs.  $A_k$ .

However, without a common measure, there is no way to compare the effectiveness of systems. There would be many benefits in adopting precision *vs.* recall as the basis for image retrieval evaluation since the measures are standard in information retrieval. In addition, many of the measures listed above can be computed from recall, precision and fall-out. In particular, R-value, 3-point average, and 11-point average each quantify performance with a single score. Beyond establishing the evaluation criteria, a test-bed image collection and set of benchmark queries are needed.

## 2.2. Image queries

Currently, researchers have explored a number of content-based image queries, such as for images of sunsets (color), airplanes (shape, color), fish (color, texture, shape), sports teams, faces, indoor/outdoor scenes, buildings, horses, skiing, art paintings, 3-D objects (silhouettes), trademarks, and vases. From our experience with the WebSEEK image search engine, we have found that, generally, users want to search for images at a higher semantic level [SC97]. For example, based on a set of 2,077,818 image queries, the following query topics are some of the most frequent:

Topic	Count
humor	29,709
art/paintings	22,595
animals/dogs	18,304
transportation/ships/titanic	4,853
sports/basketball	3,309

Clearly, there is a significant gap between content-based image retrieval system capabilities and the desired level of the user queries. One strategy for dealing with this gap in the test-bed is to stage the multiple levels of querying. The test-bed could be geared first towards the querying of low-level features such as color, texture, shape and spatial information. As image analysis, indexing and classification technologies advance, the benchmark queries could involve more rich descriptions of the images relating to scene properties, semantics and object recognition. The advantage of this approach is that a single image collection becomes familiar for researchers and serves for querying at all levels.

## 2.3. Image collections

Currently, a great number of image collections are being used for image retrieval research. The collections typically consist of images of animals, plants, architecture, art, people, horses, eagles, flowers, airplanes, faces, portraits, b/w photographs, outdoor images, textures, colored textures,

satellite images, and so forth. Two image collections are being used by multiple research groups: 560,000 images from the State of California Dept. of Water Resources [OS95] and 22,000 professional stock photographs from Corel. In general, the test-bed image collection needs to be large, diverse, free from usage restrictions, and representative of those in practical image retrieval applications.

## 2.4. Image relevance

A major difficulty in creating the test-bed is determining the ground-truth relevances. Given a test-bed of  $N$  documents and  $M$  benchmark queries, assessing the ground-truth relevances requires  $MN$  observations. A large collection ( $N =$  millions) makes exhaustive assessment infeasible.

Faced with a similar predicament, TREC formulated a pooling process that allowed document relevances to be determined from the best results from participating text retrieval systems [Har92]. The pools were sorted and a sub-set of the documents were selected for manual inspection. In the case of the 742,358 TREC documents and 150 benchmark queries, the pooling method obviated the exhaustive assessment of 111 million document-query pairs. One resulting problem was that since by default all non-judged documents were deemed irrelevant, the precision *vs.* recall plots became inaccurate after about 40% recall.

A similar method could be adopted for establishing image retrieval ground-truth relevances. In the case of querying by low-level features such as color and texture, the best results from participating content-based image retrieval systems could be combined, and the relevance of only this smaller sub-set could be assessed manually. Similarly, as image retrieval technologies advance, the relevance to queries involving scene properties and higher-level semantic information could be determined collectively.

## 3. FUTURE DIRECTIONS

We recommend that the content-based retrieval research community establish a standard test-bed for evaluating image retrieval effectiveness. This would entail establishing a large collection of images and benchmark queries, assessing relevances to the benchmark queries, and adopting a set of evaluation measures. By encouraging participation from community at large, and by establishing a friendly evaluation forum, or show-case for state-of-the-art in image retrieval, researchers could better work together to improve image retrieval technologies.

## 4. REFERENCES

- [Har92] D. Harman. Overview of the first Text REtrieval Conference (TREC-1). In D.K. Harman, editor, *Proc. Text Retrieval Conference (TREC)*, pages 1-20, Washington, 1992. NIST.
- [OS95] V. E. Ogle and M. Stonebraker. Chabot: Retrieval from a relational database of images. *IEEE Computer*, 28(9):40 - 48, September 1995.
- [SC97] J. R. Smith and S.-F. Chang. Visually searching the Web for content. *IEEE Multimedia Mag.*, 4(3):12 - 20, July-September 1997.