Using Artificial Queries to Evaluate Image Retrieval

Nicholas R. Howe
Cornell University
Department of Computer Science
Ithaca, NY 14853
nihowe@cs.cornell.edu

Abstract

This paper addresses the evaluation and comparison of algorithms for generalized image retrieval. The forms of evaluation currently in vogue are not calibrated with each other and thus do not allow the comparison of results reported by different research groups. We address the problem by proposing a class of tests that are algorithmically defined and relatively independent of the image test set. The proposed tests can be tailored to investigate retrieval performance under specific sets of adverse conditions, allowing additional insight into the strengths and weaknesses of different retrieval mechanisms.

1 Introduction

As researchers continue to develop innovative new approaches to image retrieval, objective comparisons between different techniques become increasingly important. When building retrieval systems, developers will naturally wish to choose the best algorithms for their particular application. Unfortunately, existing means of evaluation and comparison are somewhat *ad hoc*, often relying on the use of a specific image set and subjective determination of the correct response to a set of test queries [2, 3, 4, 5]. As a result, evaluations reported by different researchers are often incomparable, and the relative merits of many proposed retrieval algorithms are unknown.

In fairness, real difficulties hamper the development of consistent evaluation methodologies. A large, standardized, and universally available collection of natural photographic images has not yet developed, although images from the Corel series are often cited [2, 3, 5]. Unfortunately, different researchers use different subsets of the Corel images, and few if any have access to the entire collection. Thus the image set used in evaluation can be expected to vary in both size and content for at least the near future.

Furthermore, even given a standard set of images, there is no agreement on what constitutes a proper set of queries, nor the corresponding correct responses. Researchers in text retrieval have solved this problem by establishing conferences where all work is tested using a common, shared evaluation package [9]. Unfortunately, such a system may be more difficult to set up for image retrieval, where the specification of correct answers actually defines the problem to some extent. For example, image retrieval potentially encompasses similarity based upon (among other things) thematic similarity, the appearance of specific objects or backgrounds, the subtle differences between a series of medical images, and vague notions of general visual similarity. Each of these implies a different answer to the question "What is most similar to image I?" and each has potentially useful applications. To choose a single set of universal queries and correct responses would necessarily favor some approaches and penalize others unfairly.

To address these issues, we propose a set of evaluations using query images that are algorithmically altered in known ways. Using artificially altered data may seem like a step backward for the image retrieval community, but such methods have been successfully used to gain insight in other fields such as machine learning [1]. In our proposal, images from the test set are modified in a specific manner and used as queries. The goal is to retrieve the originals from the test set. Although the specific images used will necessarily affect the result somewhat, relying on artificial queries means that the differences between the query and the target are consistent across test sets. Furthermore, while a specific test may favor one algorithm over another, knowing the nature of the altered query makes the biases transparent. For example, color histograms might be expected to do poorly on queries where the color balance has been altered, but should do well on queries where image elements are moved around.

The next section describes the use of altered-image queries in more detail, introducing three specific types. Section 3 looks at the consistency of these tests, and gives an example of an experiment using them. Finally, the last section concludes with a look at possible further developments and issues in the area of image retrieval evaluation.

2 Altered-Image Queries

Suppose we have a test set $S = \{I_1, I_2, ..., I_n\}$ consisting of n images, and we wish to use these images to compare the performance of several retrieval algorithms. Assume further that the retrieval algorithms all operate in a query-image paradigm, i.e., given a query image Q which may or may not be an element of S, they produce an ordinal ranking on the test images. By convention, lower ranks are better; i.e., $rank(I_j) < rank(I_k)$ implies that I_j is more relevant than I_k. Typically algorithms produce their rankings by sorting on the distance from the query image according to some metric, but other mechanisms are possible.

Historically, algorithms have been evaluated on this task by comparing the ranks they produce to subjectively defined ground truth targets for each query. Human judges designate a target set $R \subset S$ for each query image, using standards that often go unreported. Algorithms are scored based upon the mean rank assigned to the target images, or perhaps the number of target images with rank below a certain threshold. Typically, published results consist of averages over many applications of this basic technique.

While results from such tests can provide valuable information about retrieval performance, they can implicitly incorporate biases that unfairly favor one algorithm over another. The criteria used to choose the target set R are usually difficult to specify or even express, muddying the interpretation of the results. They are also difficult to duplicate, preventing other research groups from conducting an equivalent test using a different set of images. Finally, they are arbitrary to a certain degree, since there may be contexts in which the images in R are not the most relevant images to the query Q. (As an example, given a picture of a flowerpot on a windowsill, should a system retrieve pictures of windowsills, or flowerpots?) A good test should make the relationship between Q and R transparent, and hence also the context in which the algorithms are compared.

One solution is to use Q and R with a known, quantifiable relationship. We propose to do this by picking an image $I_j \in S$ at random, setting $R = \{I_j\}$ and $Q = f(I_j)$ where f is an easily computable transformation of the original image. To make the test effective, f is chosen to produce some change to the image that leaves a clear relation to the original. By using a variety of choices for f, we can see how different algorithms respond to diverse sorts of image variations, alone or in combination. If the choices are made carefully, the variations will correspond to those that might arise in a real application. The remainder of this section describes three choices of f that probe at different aspects of retrieval performance in action.

2.1 The Crop Test

Image databases often contain multiple views of the same scene or subject taken from different distances. A valid expectation for an image retrieval system would be to retrieve all views of the subject given a single view as the query. The *Crop* test attempts to simulate this task in a controlled manner. Query images are created by trimming a margin off the edges of the target image, resulting in a "close-up" of the center section. Because photographs are often centered on a subject, this will often be a closeup of the subject of the photo, but in some cases it may merely be a detailed view of some part of a scene. In either case it is reasonable to expect the original image to be retrieved as a related image.

The function f_{Crop} takes an additional parameter k, which is the percentage of the original image area to retain. Empirically, we have found $k = 50\%$ to be a good value, resulting in query images that are easily recognizable to humans and moderately challenging for machine algorithms. We crop the image such that the area remaining is centered and has the same aspect ratio as the original. (The area remaining may be slightly more than $k\%$ of the original, because any fractional pixels left after cropping are rounded out to whole ones.) Figure 1(b) shows an image that has been subjected to the *Crop-50* transformation.

2.2 The Jumble Test

In many applications, the identity of objects in a photograph is much more important than their arrangement. For example, pictures taken from different angles in the same room will show the same objects in different locations and orientations. (Although such photos would usually be judged as relevant images, this may not always be the case. For example, we may wish to search for all shots taken from a particular camera angle.) Nevertheless, in many cases we expect retrieval algorithms to ignore the details of object placement. The *Jumble* test simulates this condition, albeit imperfectly, by splitting the image into rectangular sections and exchanging them randomly. Although this procedure leads to artificial boundaries in the image, it can nevertheless prove an effective test. If the sections moved are large enough, recognizable object features will be preserved within them and moved *en masse* to a new location.

The function f_{Jumble} takes two additional parameters representing the number of divisions to make along each axis. Setting both to four (*Jumble-4x4*) divides the image into sixteen rectangular areas, which are shuffled randomly. Human observers can identify the photograph with some effort. Figure 1(c) shows an image that has been subjected to the *Jumble-4x4* transformation.

2.3 The Low-Con and Gain Tests

Photography is subject to different lighting conditions, variations in development and scanning, and other processes that can make two otherwise identical pictures appear different. In almost all cases we wish to ignore differences in lighting, camera gain, contrast, and color balance for purposes of image retrieval. The *Low-Con* and *Gain* tests measure sensitivity to these factors. *Low-Con* decreases the image contrast, while *Gain* simulates a picture taken through a camera with a different gain. In general, human viewers are quite insensitive to these sorts of changes, but many color-based algorithms find them challenging.

Both $f_{Low-Con}$ and f_{Gain} take an additional parameter. For *Low-Con*, this is the percentage of the original color range that is used in the transformed image. For example, if the original RGB color values are scaled between zero and one, the *Low-Con-80* test will rescale them to run from 0.1 to 0.9. Similarly, *Gain-k* takes RGB values scaled between zero and one, and raises them to the kth power, for typical k values ranging from 0.8 to 1.2. We prefer the *Low-Con* test due to its simplicity, but for algorithms that perform a crude color renormalization before retrieval, the *Gain* test may be more appropriate. Figure 1(d) shows an image that has been subjected to the *Low-Con-80* transformation.

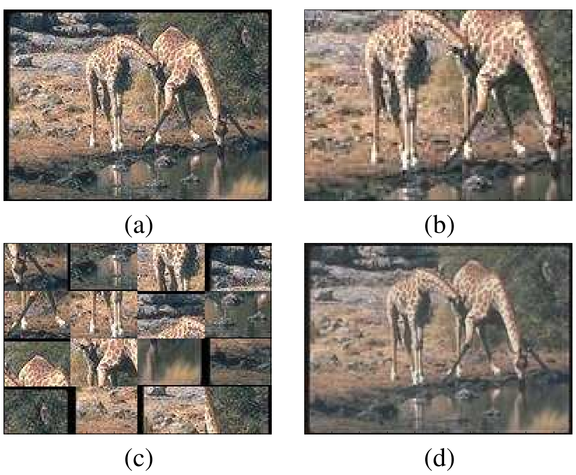

Figure 1. Examples of each type of artificial query. (a) Original. (b) Crop-50. (c) Jumble-4x4. (d) Low-Con-80.

2.4 Other Tests

The remainder of this paper will focus on the types of altered-image queries described above. However, other sorts of artificial queries may be devised to test specific aspects of retrieval performance. For example, a *Blur* test that smooths the original image would test retrieval from a low-resolution or out-of-focus original. A *Gray* test could test whether color images can be retrieved from grayscale queries. The *Low-Con* and *Gain* tests may be applied to individual color channels separately. One can imagine system designers using a large library of altered-image tests to select the algorithm that is best for a particular specialized retrieval task.

3 Evaluation and Example

If altered-image queries are to serve as a yardstick for comparing retrieval algorithms, then it is important to understand their behavior under different conditions. This section explores how results of altered-image query tests change as experimental parameters are varied. We begin with a look at a single retrieval algorithm, and proceed to a look at comparisons between three different ones.

3.1 Stability of Altered-Image Query Tests

Figure 2 shows a set of typical results for an altered-image task. The results were generated for a simple implementation of color histograms [8] on the *Crop* test, at varying parameter settings. The plot shows the average rank of the target image, sorted with the least successful queries on the right. The curves are skewed, with a large region of slowly degrading performance terminating in a sharp tail. In other words, most target images are retrieved at a relatively low rank (which is good), but on a small number of queries the algorithm does much worse. The curves can be concisely summarized by two numbers. The median rank indicates the level of the majority of queries, and thus the performance on a typical case. The mean rank indicates the size of the tail, and thus the performance on the most difficult queries. As the difficulty of the task increases, both numbers generally rise.

One concern when comparing results reported in different places is how much difference the image test set makes. Clearly the test set has some effect: if for example, it contains many shots of the same subject from different distances, then the *Crop* test will be more difficult because many distractors will compete with the target. If on the other hand it contains many dissimilar images, the test will be easier. Most image sets fall short of these two extremes, and the average variation in score is small even when using different test sets.

How much variability should be expected from performing the same test on different image sets? To answer this question, we formed three entirely disjoint test sets of 6000 images apiece and ran the *Crop-50* test. The results, shown in Table 1, display a marked similarity given that the three sets have no images in common. The standard deviation of

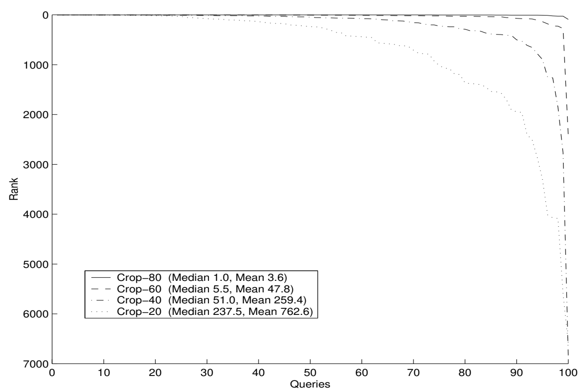

Figure 2. Results of histogram method on Crop task at varying difficulty levels.

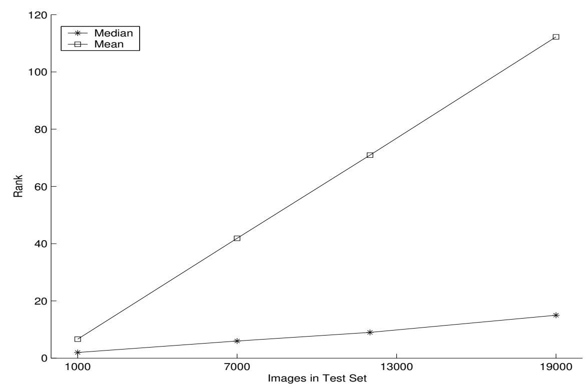

Figure 3. Linear dependence of mean and median ranks on test set size (Crop-50 test).

both the mean and median are around 20% of their values. A comparison with the results presented in Table 2 below shows that this variation is much smaller than the differences that can appear between different algorithms. This key observation suggests that results reported on different image sets may be compared with some hope of drawing accurate conclusions. Naturally it is best if the same images are used; failing that it probably helps if all of the images are drawn from a single source. (The results reported here all use images from the Corel collection.)

Table 1. Results for three disjoint sets sets on the Crop-50 task. Compare with the variation visible in Table 2.

	Set 1	Set 2	Set 3	Mean	Dev.
Median Rank	5	5	7	5.7	1.2
Mean Rank	29.6	33.9	45.5	36.3	8.2

The size of the test set may also make a difference. Figure 3 plots the mean and median ranks for the histogram algorithm as extra images are added to the test set, while keeping the query set constant. The numbers rise linearly with the test set size, suggesting that when comparing different results one should normalize by the total number of images.

The choice of the query images may also make a difference, if the test is not repeated for every image in the test set. With a large set of images, it may be unnecessary to test all the images since the dominant trends become apparent after only a small fraction of the total has been tested. Figure 4 shows the variation in the results as increasing numbers of query images are tested. The numbers show some variability at small numbers of queries, but get steadier as results from more queries are averaged together. On the whole, while increasing the number of queries will give a more accurate result, a surprisingly accurate picture arises after testing fewer than 10% of the total number of images.

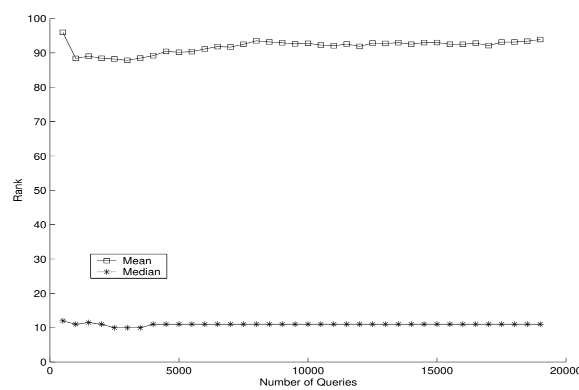

Figure 4. Change in score with size of query set for Crop-50 task (19,000 images total).

3.2 A Comparative Test

As an example of the use of altered-image queries, we present a comparison between three image-retrieval algorithms: color histograms [8], color correlograms [7], and the Stairs algorithm [6]. The first is included as a baseline; we wish to determine the relative strengths and weaknesses of the other two. We evaluate each using as queries a randomly selected 1000-image subset of our collection of 19,000 images. Each algorithm was tested on the *Crop-50*, *Jumble-4x4*, and *Low-Con-80* tests. (The same query set was used for each algorithm on any given test.)

The results are summarized in Table 2, which gives the median and mean ranks for each test. Analysis of the results provides more insight into individual strengths and weaknesses than a traditional test. We see that histograms do the worst on every test except *Jumble*, where they by definition have a perfect score. They perform worst on *Low-Con*, because that test directly affects the distribution of color in the histogram. (We do not use any sort of color scaling scheme, but the *Gain* test would reveal a similar weakness in the histogram algorithm, even with scaling.) Thus the results spell out the strengths and weaknesses of color histograms.

Color correlograms also depend on the distribution of color, though to a lesser extent than histograms, and thus show the lowest score in the *Low-Con* test. On the other two tasks, they do quite well. Stairs, by contrast, does well on *Low-Con* but not as well on *Jumble*. This is because it has mechanisms that handle small changes in color, but relies by default on an explicit representation of the location of features in the image. On the other hand, Stairs allows the user to tune parameters relating to how much color, texture, and spatial features are valued. The bottom row of Table 2 indicates that these parameters can be adjusted to make Stairs successful in any of the environments tested.

Table 2. Target rank results in artificial-query tests for three image retrieval algorithms.

		Crop	*Jumble*	*Low-Con*
Histograms	median	18	1	86.5
	mean	126.6	1	350.3
Correlograms	median	1	1	5
	mean	12.4	2.0	83.6
Stairs Default	median	1	26	1
	mean	38.9	205.2	18.2
Stairs Tuned	median	1	1	1
	mean	17.0	1.2	22.6

Introspection might perhaps have led after a while to the insights into each algorithm described above. On the other hand, it would have been difficult to set up a traditional query-target test using only natural images that would have demonstrated the algorithms' respective strengths and weaknesses so clearly. This illustrates the value of using a diverse repertoire of evaluative tools.

4 Conclusion

This paper does not attempt to prescribe comprehensive tests for the evaluation of newly proposed retrieval algorithms or the comparison of existing ones. Rather, it proposes a new class of tests in the hope that they will be adopted and incorporated into the standard practice of the field. Research in image retrieval can only benefit from new evaluative tools, and perhaps their availability will spur new developments. In time, if extensive comparative testing becomes the norm, the field will be ripe for a conference along the lines of TREC [9]. However, a proposal of this sort is beyond the scope of this paper.

Altered-image queries should not completely replace traditional testing methods for image retrieval, but they deserve a place alongside the traditional techniques. The different forms of evaluation serve complementary purposes. Traditional approaches can be seen as field testing under a simulation of real conditions. By contrast, altered-image query testing serves as a diagnostic tool for comparing different retrieval algorithms under more controlled conditions, and for identifying areas where a particular algorithm may be underachieving. Hopefully, the conscientious use of both forms of testing will lead over time to a greater understanding of the issues involved, and ultimately to improved algorithms. The development of more flexible and consistent evaluation tools can only advance that goal.

References

[1] D. W. Aha. A framework for instance-based learning algorithms: Mathematical, empirical, and psychological evaluations. Technical Report 90-42, University of California, Irvine, 1990.

[2] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Recognition of images in large databases using a learning framework. Technical Report 97-939, UC Berkeley, 1997.

[3] J. S. De Bonet and P. Viola. Structure driven image database retrieval. *Advances in Neural Information Processing*, 10, 1997.

[4] U. Gargi and R. Kasturi. Image database querying using a multi-scale localized color representation. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1999.

[5] N. Howe. Percentile blobs for image similarity. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 78–83, Santa Barbara, CA, June 1998. IEEE Computer Society.

[6] N. R. Howe and D. P. Huttenlocher. Integrating color, texture, and geometry for image retrieval. Technical report, Cornell University, 2000. Submitted to CVPR 2000.

[7] J. Huang, S. K. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1997.

[8] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[9] E. M. Voorhees and D. K. Harman, editors. *The Eighth Text REtrieval Conference (TREC-8)*. Department of Commerce, National Institute of Standards and Technology, 1999.