

An Empirical Investigation of the Scalability of a Multiple Viewpoint CBIR System

James C. French*, Xiangyu Jin, and W.N. Martin

Department of Computer Science, University of Virginia, Charlottesville, VA, USA,
{french,xj3a,wnm}@cs.virginia.edu,
<http://www.cs.virginia.edu/~cyberia>

Abstract. Our work in content-based image retrieval (CBIR) relies on content-analysis of multiple representations of an image which we term multiple viewpoints or channels. The conceptual idea is to place each image in multiple feature spaces and then perform retrieval by querying each of these spaces and merging the several responses. We have shown that a simple realization of this strategy can be used to boost the retrieval effectiveness of conventional CBIR. In this work we evaluate our framework in a larger, more demanding test environment and find that while absolute retrieval effectiveness is reduced, substantial relative improvement can be consistently attained.

1 Introduction

Content-based image retrieval (CBIR) has been the object of considerable study since the early 90's. Much effort has gone into characterizing the "content" of an image by means of a variety of features for the purpose of indexing and subsequent retrieval. In earlier work [1] we proposed a strategy to capitalize on this work and to extend it by employing content-analysis of multiple representations of an image which we term multiple viewpoints[2]. The idea is to place each image in multiple feature spaces and then effect retrieval by querying each of these spaces and merging the several responses. The impetus for this research comes from work in text IR on combination of evidence strategies that dates back to the early 90's. Two approaches have generally been used. In the first approach a diversity of queries is used to capture an information need more precisely. The several queries can be combined before searching, or issued individually and the results of each query merged afterwards. The work of Belkin et al.[3,4] adopts this approach.

The second strategy is to use a diversity of representations, that is, create several indexes over the same corpus of documents. The typical strategy is to index the corpus with the same technology varying indexing parameters, or to

* This material is based upon work done while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

index the corpus with different technologies. Queries are processed in each setting with the results being merged afterwards. The work of Fox and Shaw[5] adopts this strategy. Bartell et al. [6] also look at combining evidence in this framework. The approach we adopt for extending CBIR systems to combine multiple evidence is analogous to this latter approach.

These ideas are embodied in our *synthetic retrieval model* for CBIR[7] shown schematically in Figure 1. We refer to this as a synthetic retrieval model because we merge the various viewpoints and synthesize a channel for presentation to the user. We increase the number of viewpoints in CBIR systems in three different ways: multiple representations; multiple CBIR systems; and multiple queries. We also employ relevance feedback to further increase retrieval performance. Within this framework we have investigated the use of a diversity of representations that we call *channels* to achieve retrieval effectiveness gains over conventional CBIR[1,8,9,7]. Our approach is exogenous; we treat the CBIR system as a black box. We create additional channels by transforming the images and indexing the transformed images. Our four channels derive from the color positive (C+) and negative (C-) and the black and white positive (B+) and negative images (B-). Note that the C+ channel corresponds to the conventional CBIR system.

In this paper we evaluate our framework in a larger, more demanding test environment. Due to page limitations we refer the reader to [1,8,9,7] for specific details of our framework. In the remainder of this paper we describe the current experimental setup and finally discuss our results.

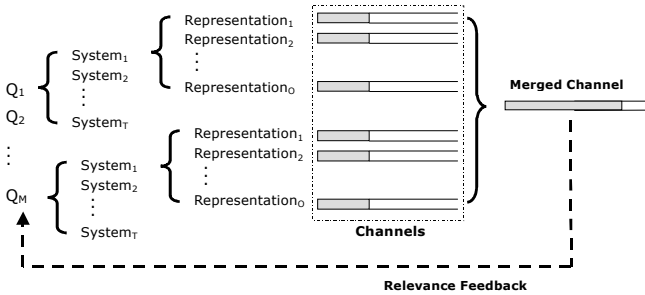


Fig. 1. Synthetic Retrieval Model

2 Experimental Setup

2.1 Basic CBIR Technology

We used a basic CBIR setup similar to that used in the MiAlbum system used in the work of Liu et al. [10]. Our system uses seven image features: three color features and four texture features. For similarity comparisons each feature was compared separately and then combined with equal weight.

2.2 Testbed

Test Data. We used two different image collections in this work.

1. **D34**: this is 3,400 images drawn from 34 categories of the COREL image collection. Each category contains 100 images. The categories were chosen because each of the images has a salient foreground object.
2. **D594**: this is a larger version of the COREL database consisting of 594 image categories each having 100 images each. Thus, **D594** contains 59,400 images. It should be noted that **D34** is a proper subset of **D594**.

Query Sets. We use three different query sets in this work.

1. **Q3400**: Each of the images in **D34** is used as a query. Thus **Q3400** = **D34** and there are 3,400 query images.
2. **Q204**: The 34 categories of **D34** are uniformly sampled and 6 images are included in **Q204** from each category. This is a 6% sample of **D34** with equal representation of each category. Thus, there are 204 query images in **Q204**.
3. **Q3564**: The 594 categories of **D594** are sampled in the same way as **Q204**. Thus, this is a 6% sample of **D594** with 3,564 queries and equal representation of all categories in the sample. Note that **Q204** is not a proper subset of **Q3564** and has no specific relation to it.

Ground Truth. In earlier work we used the “foreground” groundtruth for **D34**[1,8,9], but since we do not have the equivalent for the additional image categories in **D594**, we have used a different but consistent “COREL” groundtruth in the work reported here. This latter groundtruth is defined to mean that all the images in an image category are relevant to all the other images in the category and not relevant to any other images. Thus, any image selected from a test collection to act as a query will have exactly 99 relevant images in the collection.

Our earlier work has shown remarkable consistency between the performance as measured by these two groundtruths and we have never had one contradict the other in an experiment so we believe this choice to be adequate for our purposes.

Indexing the Images. We created four indexes corresponding to each channel in our testbed. The images were transformed into the representation of the channel and then indexed by our CBIR system. Thus, for each testbed we have a single corpus of images over which we have four separate indexes.

Experiment Notation. We denote a particular experiment by **Q/D** where **Q** is the query set and **D** is the testbed data set. For example, **Q204/D594** denotes the 204 queries of the 6% sample of **D34** processed against the 59,400 images in the large data set.

User Model for Relevance Feedback. In an earlier study [7] we observed a significant improvement in retrieval performance when using relevance feedback, that is, providing images identified as relevant in one iteration of the search to the query set in the next iteration. This is consistent with the results of other studies of relevance feedback. Our approach is to issue each feedback query

independently and then merge the results for presentation to the user. This leaves open the issue of how the feedback queries are chosen. We use two strategies:

1. Identify the top k images (Top- k); and
2. Take k images at random from among the relevant images (Random- k).

The former strategy is customarily used in text IR experiments. However, the latter strategy seems more appropriate for CBIR given the relative ease with which a user may judge the relevance of images. We feel that the Random- k user model more accurately reflects user behavior. Earlier work [7] has shown that this strategy will result in higher retrieval performance because it defeats self-similarity in feedback images and therefore achieves a greater visual diversity among the feedback images. Note that there are at most k images chosen by either strategy because in some cases fewer than k images are present in the retrieval result. Further, $k = 8$ in all the experiments reported here.

2.3 Methodology

The query processing is the same in all experiments. One query set is processed against one testbed. Each query is processed separately and the precision¹ at 100 images seen (P100) is calculated for each. The average P100 is calculated over the entire query set and that result is reported. We note that P20 has become a very common metric for reporting results in text-based IR. A typical CBIR UI displays 30-50 thumbnails at a time in response to a query. Because of the ease of evaluating images for relevance relative to text documents, we feel that P100 is a more appropriate performance measure. One hundred images is also the first time we could conceivably achieve recall² of 1.0 for any of the queries in our query sets.

Our merging results in [1,8,9] were produced using the *combSUM*[5,11] approach, that is, we summed the similarity values for images across the channels in which the image was included in the response set. (The conditions set out by Vogt[12] for linearly combining relevance scores apply here: our channels do have reasonable performance and they do not rank relevant documents similarly.) We have also used a rank sum approach, midrank merge³, for merging and found that to perform comparably with *combSUM*. We use that technique here.

3 Results

The four plots show in Figure 2(a-d) each show five experiments, **Q204/D34**, **Q3400/D34**, **Q204/D594**, **Q3400/D594** and **Q3564/D594** respectively. The

¹ Precision is the ratio of the number of relevant images retrieved to the total number of images retrieved.

² Recall is the ratio of the number of relevant images retrieved to the total number of relevant images.

³ Each channel assigns a rank to each image retrieved and not retrieved. The assigned ranks are summed to determine the image's rank in the final result. When a channel does not retrieve an image, it is assigned a rank higher than 100.

plots are paired vertically by user model (Top- k , Random- k) and are paired horizontally by channel configuration (one channel, four channels). The Random- k user model is higher performing than the Top- k model. We have observed this in earlier experiments [7] and attribute it to greater visual diversity in the feedback images. The topmost pair of lines show that the small sampled query set (Q204) has very similar performance to the larger query set (Q3400) in the smaller testbed (D34). The next two lines show that the small sampled query set also has very similar performance to the larger query set in the larger testbed (D594). This is consistent in all four plots.

Conclusion 1: The smaller sampled query set, Q204 is representative of the larger query set, Q3400, as regards performance evaluation.

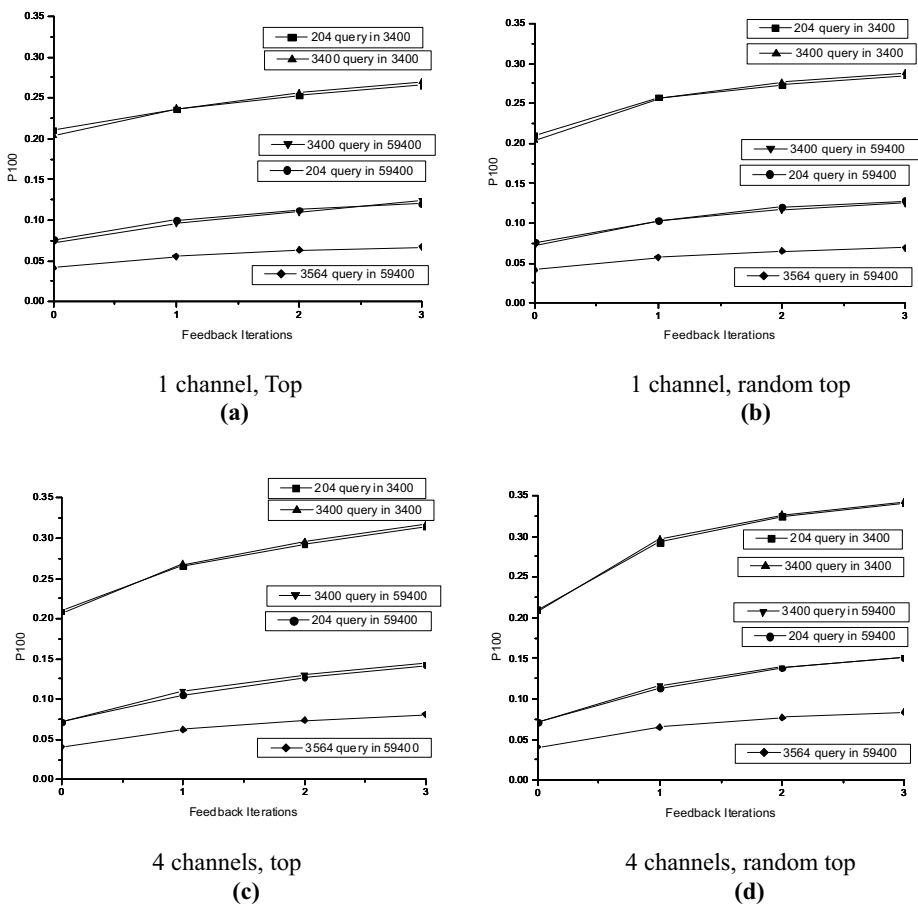


Fig. 2. Retrieval performance as measured by sample vs. all queries in small and large testbeds.

Table 1. Retrieval precision and performance increase after each feedback iteration (Top- k user model). Avg. precision small (large) testbed is 29.3% (61.8%).

	Q204		Q3400		Q3564	
Iteration	D34	D594	D34	D594	D594	
0	.2109	.0759	.2032	.0730	.0413	
1	.2362 12.0%	.1003 32.1%	.2372 16.7%	.0966 32.3%	.0554	34.1%
2	.2528 7.0%	.1126 12.3%	.2563 8.1%	.1102 14.1%	.0632	14.1%
3	.2652 4.9%	.1209 7.4%	.2698 5.3%	.1190 8.0%	.0673	6.5%
Total	25.7%	59.3%	32.8%	63.0%	63.0%	

The 4-channel configurations (Figure 2c,d) are equivalent to the single channel configurations (Figure 2a,b) initially but outperform them considerably in all feedback iterations.

Conclusion 2: Relevance feedback in the multichannel configuration is more effective than in the single channel configuration.

The four plots of Figure 2 clearly show that absolute retrieval effectiveness (as measured by P100) is lower in the larger database (**D594**) as compared with the effectiveness observed in the smaller (**D34**). This occurs in both single and multichannel configurations.

Conclusion 3: CBIR retrieval precision is substantially reduced when the size of the database is increased.

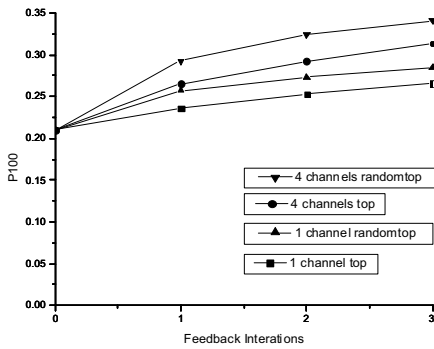
However, even though the absolute effectiveness is reduced in the larger testbed, the rate of improvement with each feedback iteration is roughly constant. In addition the overall improvement in each configuration was also very stable, averaging 62%. Table 1 shows the actual values. This is perhaps the most important feature of the multichannel approach.

*Conclusion 4: The multiple viewpoint techniques demonstrated in the smaller testbed (**D34**) are also effective in the larger testbed (**D594**).*

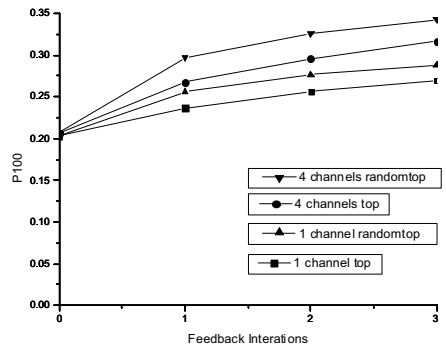
Finally, the **Q3564/D594** experiment has substantially lower performance than the **Q3400/D594**. Recall that all the queries in **Q3400** come from 34 of the 594 categories in **D594** whereas there are 6 queries from each of the categories of **D594** in **Q3564**. We hypothesize that **Q3400** is therefore an “easier” query set than **Q3564**.

Conclusion 5: Q3400 and Q3564 do not have similar retrieval performance in D594.

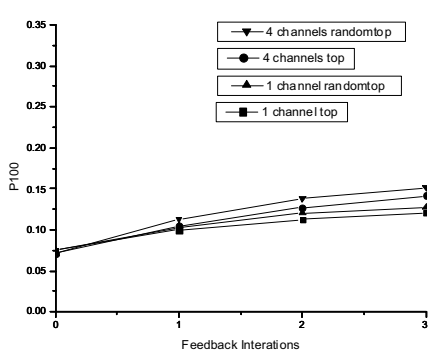
The four plots in Figure 3 are grouped vertically by query set with the smaller query set (**Q204**) on the left and the larger (**Q3400**) on the right. They are grouped horizontally by testbed size with the smaller testbed (**D34**) topmost and the larger (**D594**) on the bottom. In each case four lines are shown corresponding to the two user models (Top- k and Random- k) and the two channel configurations (one, four). Again, the data support *Conclusion 1*. We are also led to the following conclusions.



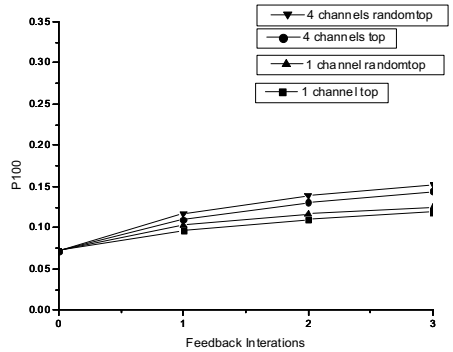
query 204 in 3400
(a)



query 3400 in 3400
(b)



query 204 in 59400
(c)



query 3400 in 59400
(d)

Fig. 3. Performance of single channel vs. multichannel CBIR for two user models.

Conclusion 6: The Random-k user model consistently achieves the highest precision.

This is good news especially since we consider the Random-k user model to more accurately reflect user feedback selection in real CBIR environments.

Conclusion 7: The multichannel configuration is superior to the conventional single channel approach achieving greater precision with each feedback iteration.

Especially noteworthy is the fact that this improvement is attainable in the larger image testbed, 59% (63%) for the smaller (larger) query set. We also observed this in the more difficult large test environment (**Q3564/D594**) discussed earlier where the query set is a 6% sample of the 59,400 images queried (see Table 1). The improvement (63%) was exactly the same as that achieved with the **Q3400** query set (see Figure 3c,d).

The results clearly indicate that even in the more difficult testbed, we can, in fact, combine the channels, even naively, to realize retrieval effectiveness gains over the conventional single-channel CBIR approach.

4 Conclusions

We have described a simple approach for improving the retrieval effectiveness of conventional CBIR systems. Our approach treats the CBIR technology as a black box which can be used to provide different channels of retrieval results for subsequent merging or for use in interactive retrieval interfaces. The channels are implemented as additional indexes over simple image transforms. This offers a simple, cost-effective strategy for boosting the performance of CBIR systems.

In [1] we showed that multichannel retrieval could increase CBIR retrieval effectiveness. We demonstrated an 8% increase in non-interpolated average precision with a 4-channel configuration of our CBIR system over the baseline system when ranking all the images of our test database. The average non-interpolated precision increases by 22% in the 4-channel system when we consider result lists of the top 100 images.

In [8] we looked at the potential for performance improvement when two CBIR systems were used to supply the viewpoints for constructing the synthetic channel. Again, the combination of multiple channels (this time from different CBIR systems) resulted in increased retrieval effectiveness. Moreover the combination of the two techniques, multiple systems and multiple representations, were complimentary and resulted in an even greater performance boost.

In [7] we looked at multiple queries as a means of achieving greater retrieval performance by providing more exemplars of the user's information need. We introduced the concept of visual diversity and examined the role of multiple representations and multiple CBIR systems in achieving visual diversity to improve retrieval with multiple queries. We also examined several strategies for accommodating relevance feedback in our synthetic framework. A new feedback evaluation strategy was also proposed and shown to be more effective because it increases the visual diversity of the feedback images.

In this paper we extended our work to validate our approach in a larger, more demanding retrieval environment. This kind of study is hampered somewhat by the lack of suitable testbeds with associated groundtruth. Our work here is extended to a testbed of 59,400 images from our earlier testbed of 3,400 images.

This study confirmed our earlier finding that the multiple viewpoint techniques, singly and in combination, improve retrieval effectiveness of CBIR systems even in more demanding retrieval environments. We found that although the absolute precision was reduced, the rate of improvement held up well in feedback iterations and the overall relative improvement after three feedback iterations was approximately 62%.

Another finding is that we can get extremely accurate performance evaluation with smaller query samples. This will enable us to conduct empirical studies more efficiently with confidence in the results.

Note that the techniques proposed here do not increase the user work in relevance feedback. The user is only concerned with the synthetic channel presented after each feedback cycle. The system transparently feeds the selected images back to all the underlying channels and merges the several results back into a synthetic channel for the user.

The synthetic retrieval framework also makes parallelization a simple matter. Retrieval on each channel is independent of all others. Thus, each channel can be assigned to a separate processor for query processing followed by a merging stage. This strategy can provide high retrieval efficiency in addition to improved retrieval effectiveness.

References

1. French, J.C., Watson, J.V.S., Jin, X., Martin, W.N.: Using multiple image representations to improve the quality of content-based image retrieval. Technical Report CS-2003-10, Dept. of Computer Science, Univ. of Virginia (2003)
2. French, J.C., Chapin, A.C., Martin, W.N.: Multiple viewpoints: A strategy for searching multimedia content. In: Workshop on Multimedia Content in Digital Libraries. (2003)
3. Belkin, N., Cool, C., Croft, W., Callan, J.: The effect of multiple query representations on information retrieval system performance. In: Proc. of ACM SIGIR'93. (1993) 339–346
4. Belkin, N., Kantor, P., Cool, C., Quatrain, R.: Combining evidence for information retrieval. In: Proc. of TREC-2. (1994) 35–44
5. Fox, E., Shaw, J.: Combination of multiple searches. In: Proc. of TREC-2. (1994) 243–252
6. Bartell, B., Cottrell, G., Belew, R.: Automatic combination of multiple ranked retrieval systems. In: Proc. of ACM SIGIR'94. (1994) 173–181
7. Jin, X., French, J.C.: Improving image retrieval effectiveness via multiple queries. In: First ACM Inter. Workshop on Multimedia Databases. (2003) 86–93
8. French, J.C., Watson, J.V.S., Jin, X., Martin, W.N.: Integrating multiple multi-channel cbir systems. In: Proc. Inter. Workshop on Multimedia Information Systems (MIS 2003). (2003) 85–95
9. French, J.C., Watson, J.V.S., Jin, X., Martin, W.N.: An exogenous approach for adding multiple image representations to content-based image retrieval systems. In: Proc. Seventh Inter. Symp. on Signal Processing and its Applications. (2003)
10. Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B.: Semi-automatic image annotation. In: Proc. of Human-Computer Interaction-Interact. (2001) 326–333
11. Shaw, J., Fox, E.: Combination of multiple searches. In: Proc. of TREC-3. (1995) 105–108
12. Vogt, C.: When does it make sense to linearly combine relevance scores. In: Proc. of ACM SIGIR'97. (1997)