Using Multiple Image Representations to Improve the Quality of Content-Based Image Retrieval

James C. French   Worthy N. Martin   James V. S. Watson   Jin Xiangyu
Department of Computer Science
University of Virginia
Charlottesville, VA
01-434-982-2213
{french,wnm,jvw3n,xj3a}@cs.virginia.edu

ABSTRACT
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 for the purpose of subsequent retrieval. The present study seeks to capitalize on this work and to extend it by employing content-analysis of multiple representations of an image which we term multiple viewpoints or channels. 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. We show that a simple realization of this strategy can be used to boost the retrieval effectiveness of conventional CBIR.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – image databases; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – search process.

General Terms
Measurement, Performance, Experimentation.

Keywords
Content-based image retrieval, multiple viewpoint systems, multi-channel CBIR, result merging.

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. Good surveys can be found in [1,12,20]. We propose 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. 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 are 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. In earlier work we investigated the application of query diversification in CBIR systems[8].

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 index the corpus with different technologies. Queries are processed in each setting with the results being merged afterwards. The work of Fox and Shaw[6] and Shaw and Fox[14] adopts this strategy. Bartell et al.[2] also look at combing evidence in this framework.

The approach we adopt for extending CBIR systems to combine multiple evidence is analogous to this latter approach. As we describe in the next section, we investigate the use of a diversity of representations to achieve retrieval effectiveness gains over conventional CBIR.

In the remainder of the paper, we describe our approach, our experimental setup, and finally discuss our results.

2. Multiple Viewpoint Systems
A multiple viewpoint system[7] is one which employs more than one organizational approach across a corpus of information. The idea is to provide a user with complementary access strategies under different organizations of the same data and to encapsulate these into a common interface. A conventional library offers the simplest example via subject and author indexes. Each index is a different organizational view with different access properties and both are useful to searchers at different times. In earlier work we have shown the potential of this approach in text IR systems[11]. In this paper we cast CBIR in this framework referring to alternative image representations as *channels*

2.1 Single Channel CBIR
This is the conventional approach to CBIR. We are given a corpus of images. We extract a set of features from each image. The features typically capture color, shape and texture information although spatial and other information might also be used. Image features might be computed globally, or there might be a segmentation and object identification phase before feature extraction. In this case features are typically associated with the objects. After feature extraction the features are generally combined into a feature vector thereby implicitly placing the image (image objects) in a high-dimensional feature space.

In the typical query-by-example approach to retrieval, a query image is presented to the system. The query image is processed in the same way as indexing was done on the stored images. This results in a query vector and subsequent retrieval is done by producing a ranked list of images at increasing distance from the query vector. The distance function is capturing dissimilarity.

Details among individual systems will vary, but the conceptual model is the same: there is a single representation for each image and that representation is consulted when retrieving images. Thus, we have a single channel into the image collection.

2.2 Multi-channel CBIR

Multi-channel CBIR is conceptually a straightforward extension of the single channel case. We create several different representations of the images and consult some or all of them during the retrieval process. What is not so straightforward is deciding how to create these representations. In principle we could extract additional features from each image and create another feature vector for each image, placing the images implicitly in a new vector space. We do not follow this course for three reasons: (1) usually some care has gone into the choice of features used by a particular system, and therefore, any new set of feature choices is in some sense inferior; (2) the newly chosen features could be combined with the old to place the images in a larger single vector space, and thus the "viewpoints" are projections of that larger space; and (3) we are interested in a strategy that considers the CBIR technology as a black box so that we can use the newly defined representations compatibly with several different CBIR technologies.

These considerations led us to a very simple alternative representational scheme. In our approach we transform the images and index the transforms. In the present work we are holding the CBIR system constant although our multi-channel framework does not impose that requirement.[1] For subsequent retrieval we transform the query image to be consistent with each representation and either: (a) present the top *k* results of all channels to the user for inspection or (b) merge the top *k* results from each channel and present the user with *k* or more merged results.

Even in our simplified approach there is no obvious way to select transforms. As our purpose here is to investigate feasibility of the approach, we make no attempt to argue for optimal transforms. Instead we choose a set of simple transforms with some intuitive justification and proceed from there.

On a final note, we asserted that our approach is analogous to the multiple representation strategy used in text IR. There is a subtle difference. In the text IR approach, the same content (i.e., bit stream) is presented to all the indexing technologies. In our approach, it is the transformed content that is indexed.

2.3 Simple 4-Channel Model

The four representations chosen for our work here are shown in Figure 1. We use the original color image (C+) together with the black and white image (B+) and both the color (C-) and black and white (B-) negatives. So our four channels derive from the color positive and negative and the black and white positive and negative images.

[1] We are investigating multi-channel CBIR systems involving several CBIR technologies, but our work is not far enough along now to report here.

Let *G(I)* denote the gray scale image of image *I* and let *N(I)* denote the negative of image *I*. We can define our channels in terms of transformations more precisely as follows.

C+ = I

C- = *N(I) = N(C+)*

B+ = *G(I) = G(C+)*

B- = *N(G(I)) = G(N(I)) = N(B+) = G(C-)*

The intuition for including black and white channels is to provide channels where shape and texture will not be dominated by color. The multiple channels are intended to be recall enhancing, while the merge operator is precision enhancing.

Note that channel C+ corresponds to conventional single-channel CBIR systems.

We state our expectations for this model as a set of hypotheses to be addressed later in the paper.

Hypothesis 1: The performance of the two color (black and white) channels will be equivalent.

This is based on the fact that the two color (black and white) channels have the same information content.

Hypothesis 2: The color channels will exhibit better performance than the black and white channels.

The black and white channels have less information content.

Hypothesis 3: The responses of the channels can be usefully combined (merged) to synthesize a higher performing channel.

To the extent that the B channels find anything useful that does not overlap the C channels, we hope to be able to merge the outputs to improve overall retrieval effectiveness.

2.4 A Retrieval Example

In this section we discuss a hypothetical retrieval situation to demonstrate the potential utility of multi-channel CBIR systems. We describe two possible uses for the technology. To begin we suppose a scenario in which a user wishes to query an image database for images of roses. The initial query is assumed to be a query-by-example, that is, the user presents an image of a rose. We describe three different scenarios and demonstrate the output using the CBIR system developed for this research.

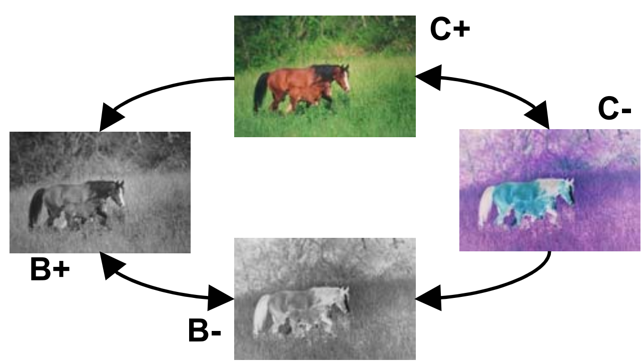


Figure 1. Proposed CBIR Channels

2.4.1 Conventional (Single Channel) CBIR

This is the case depicted in Figure 2(a). The query image is shown in the upper left. The remaining 4 X 10 images are the top 40 responses of the system to the query. The query image is returned as the top ranked image. The 25th ranked response is a rose; the others are not. A user might very well conclude that the database does not contain many relevant images, or possibly that the CBIR technology is not very good.

2.4.2 4-Channel CBIR

Figure 2(b) show the response of our 4-channel CBIR system to the same query. Forty images are also displayed, but this time the user is shown the top 10 images available on each of the four channels. Again, the query image is the top ranked image on each channel. The C+ channel is the same as the first 10 images of the single channel system (Figure 2(a)). Note that neither color channel returns anything useful. However both black-and-white channels return relevant images and, moreover, these are images of a completely different color. There are two useful ways to exploit this new information. Each is described next.

2.4.2.1 Interactive Retrieval with Feedback

In an interactive retrieval setting we can employ user feedback to help guide the search. To demonstrate the utility of this strategy we have selected the second ranked image off either B channel as a new query to the system. Figure 2(c) shows the system response to this query. It is strikingly different (14 relevant images) from the system response of the conventional system (one relevant image) shown in Figure 2(a) and arguably is of more potential use to the searcher. Again, this is the main tenet of multiple viewpoint systems – the viewpoints provide alternative interpretations of the data that might be profitably employed in search strategies.

2.4.2.2 Retrieval with Merged Result

In this strategy we assume that a new channel is synthesized by merging one or more of the base channels. The interface to such a system might expose the underlying base channels as well or might simply provide the merged result as the system's output. Other configurations are also possible. In Figure 2(d) we have chosen to demonstrate this with two merged channels, M1 and M2, that we have been experimenting with. Since Figure 2(c) shows 14 relevant images, it is possible for a merge operator to have 10 relevant images whereas the two we show only have 8 relevant images in the top 10 ranked merged images. Note that in this example channel C- also had 8 relevant images.

2.4.3 Discussion

The example of Figure 2 clearly demonstrates the potential of multi-channel CBIR systems. The technology could be used in a variety of search interfaces to provide additional functionality to existing systems. While we have hinted at the potential for use in interactive systems, we will not explore that further in this paper. The remainder of the paper is focussed on assessing the potential of merging the separate channels into a coherent single response to user queries. We undertake that assessment via a set of experiments described in the next sections.

3. EXPERIMENTAL SETUP

3.1 Basic CBIR Technology

We used a basic CBIR setup similar to that used in the *MiAlbum* system used in the work of Wenyin et al. [19]. 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. This CBIR system was used in all experiments and generated the output used in the example of Figure 2.

3.2 Testbed

3.2.1 Test Data

Our test data consisted of 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.

3.2.2 Ground Truth

Each of the images in our testbed was labeled as to foreground and background objects. The image labeling is described in [16].

3.2.3 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, we have a single corpus of images over which we have four separate indexes.

3.3 Methodology

Each image in our test data collection was used as a test query in each channel of our multi-channel testbed. Since each image is annotated with labels denoting foreground and background objects, we had de facto relevance assessments. For the results reported here, we declared images to be relevant to a query image if it had *any* foreground[2] label in common with the query image. We used *treceval* to generate the performance results.

Our merging results were produced using the combSUM[6,14] 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[17] for linearly combining relevance scores apply here: our channels will be seen to have reasonable performance and they do not rank relevant documents similarly.) We made no attempt to optimize the merging algorithm although intuition suggests that a weighted sum is almost certainly appropriate. Merging algorithm M1 of Figure 2 is a visual example of this algorithm.

Finally, in our discussion of results, we adopt Sparck Jones' standard of assessing significance [15, page 397]: an effect will be *noticeable* if the performance increase is 5-10%; it will be *material* if greater than 10%.

[2] We also collected data for two other task scenarios: (1) when the target and query images had any background label in common; and (2) when they had either a foreground or background label in common. The results were similar to those reported here and are omitted because of space limitations.

Figure 2. Example of Multi-Channel CBIR

4. RESULTS

To better understand why a multiple channel CBIR system might have potential to improve retrieval effectiveness, we took the result lists for each query image and computed the overlap of the result lists on each pair of channels in the system. We calculated the overlap as a Jaccard coefficient.[3] Table 1 shows the average Jaccard coefficient obtained when the lists were truncated to

Table 1. Channel Overlap at Various List Lengths

k	C+C-	C+B+	C+B-	C-B+	C-B-	B+B-
10	0.34	0.13	0.12	0.11	0.14	0.29
20	0.35	0.11	0.11	0.10	0.13	0.31
50	0.39	0.12	0.12	0.10	0.14	0.35
100	0.44	0.13	0.14	0.12	0.16	0.40

length k, for k = 10, 20, 30 and 100. (Note that the overlap for any pair of channels will be 1.0 when k=3,400, i.e., all the images.)

We can see from the table that the color (black and white) channels have more overlap (29-44%) than any combination of a color channel with a black and white channel (10-16%). This accords with intuition. The two promising features of the data are: (a) the color channels have some overlap with the black and white; and (b) they color channels do not overlap each other by a large amount. This is at least a circumstantial case for the possibility that each channel might contribute usefully to a retrieval result.

Now we consider the hypotheses posed in Section 2.3.

Hypothesis 1: The performance of the two color (black and white) channels will be equivalent.

This seems intuitive given that the two color channels have the same information content, likewise the two black and white channels. There is no reason *a priori* to expect their retrieval effectiveness to be significantly different. However, our experiments did not bear this out. Figure 3 shows the retrieval performance of each channel. The counterintuitive aspect of Figure 3 is that the negative channels seem to outperform the positive channels. While the color channels might not be significantly different, the fact the the C- channel is almost everywhere greater that the C+ channel is unusual. The B- channel is clearly more effective than the B+ channel.

The comparisons of the baseline channels in Table 2 show this in more detail. The improvement in performance of the C- channel over the C+ channel is noticeable (>5%) for interpolated recall greater than .4. However the improvement in average non-interpolated precision is only about 4% for the color channels. The B- channel is clearly materially better than the B+ channel. We have no explanation for this phenomenon at the present time, but the hypothesis is clearly false.

[3] This is simply the list intersection divided by the list union. It is the *symmetric overlap* of Das-Gupta et al.[5].

Hypothesis 2: The color channels will exhibit better performance than the black and white channels.

The color channels clearly have better retrieval performance than the black and white channels and that accords with our intuition since the color channels have more information content. This can be clearly seen in Figure 3 and the data for the baseline channels in Table 2. Thus, the hypothesis is true.

Hypothesis 3: The responses of the channels can be usefully combined (merged) to synthesize a higher performing channel.

We tackled this question in two ways. First we considered performance over the full lists. Since these full result lists amount to permutations of all the images in the collection, we could have used the evaluation approach suggested by Narasimhalu[10] (described also in Santini[13]) for comparing these permutations. However, as will be seen, we found it more convenient to use a standard IR evaluation via recall and precision to more easily compare with merged results involving truncated result lists.

Although we considered five different channel combinations, the combination of all four channels always had the best performance so we confine our comparisons to the 4-channel configuration. First we show the merge results from the point of view of the full list (i.e., all images ranked). This is shown in Figure 4. The 4-channel configuration is marginally more effective than C- but considerably more effective than C+. This can be seen more clearly from the data in Table 2. Recall that C+ is the conventional case, so it may be inferred that 4-channel system is more effective, thus supporting the hypothesis.

The second approach we used to make the assessment is to consider the results of merging the truncated lists. Our approach was as follows. We took the top-ranked k images from each channel and merged them into a single results list. The merged list had from k to $4k$ images in it. (In the experiments we use k =10, 20, 50 and 100. Only k = 100 is reported here.) The first merge algorithm was the so-called perfect merge (i.e., merge by oracle). In this approach, we sorted the merged list by known relevance, that is, we assumed an oracle would place the r relevant images in position 1 through r while placing the non-relevant images in positions $r+1$ and so on. This merge represents the maximum possible performance achievable by any merge algorithm and gives us an operational upper bound on performance. The results of five channel combinations using this perfect merge are shown in Figure 5 superimposed over the full baseline channel performance reproduced from Figure 3. The potential for performance improvement can be seen. Note that the merged results become poorer than the full results because the truncated lists impose limits on possible recall performance.

Finally, we consider the retrieval effectiveness of a practical merge strategy, combSUM, summing the image similarities. Again, we used the top 100 ranked images from each channel. The results are shown as the dashed line of Figure 6. The other two lines are the conventional system (C+) and the best perfect merge. We see clearly that the merge algorithm has better retrieval effectiveness than the conventional approach. The average non-interpolated precision of the merge (0.1276) is 22% greater than that of the conventional system (0.1049) and, hence, is materially better. In fact, the merge is 15% better in average non-interpolated precision than the C- channel (0.1113) which is noticeably better (6%) than the C+ channel.

Figure 3. Baseline Channel Retrieval Effectiveness

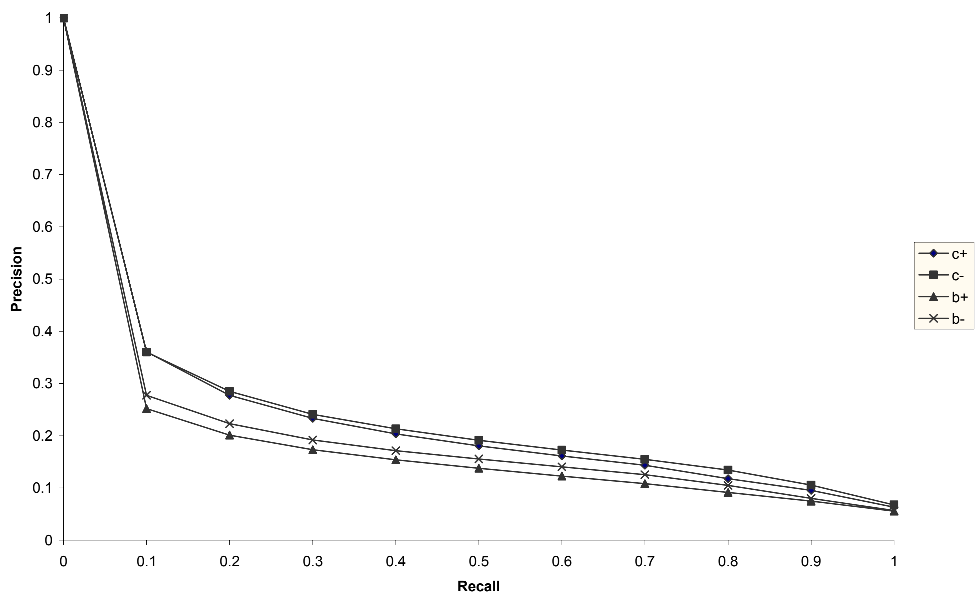


Figure 4. Best Full Merge vs. Best Baselines

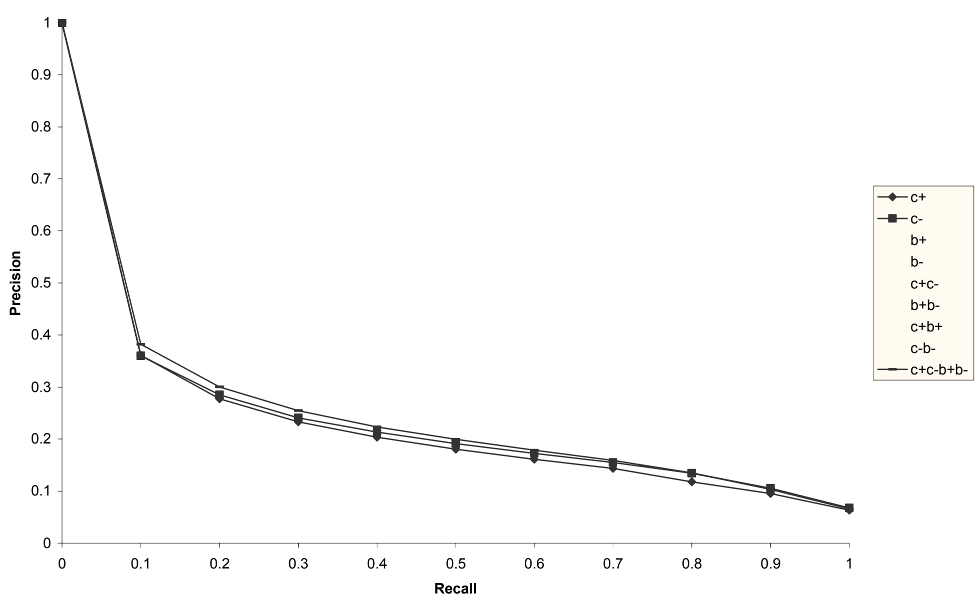


Table 2. Comparison of Best Merged Channel to Best Baseline Channels

Interpolated Recall - Precision

Recall	**C+**	**C-**	**%Improve over C+**	**B+**	**B-**	**%Improve over B+**	**C+C-B+B-**	**%Improve over C+**	**%Improve over C-**
0	1	1	0.0	1	1	0.0	1	0.0	0.0
0.1	0.3607	0.3603	-0.1	0.2519	0.2772	10.0	0.3822	6.0	6.1
0.2	0.2774	0.2852	2.8	0.2011	0.2231	10.9	0.3001	8.2	5.2
0.3	0.2333	0.2409	3.3	0.173	0.1916	10.8	0.2546	9.1	5.7
0.4	0.2036	0.2133	4.8	0.154	0.1716	11.4	0.2233	9.7	4.7
0.5	0.1807	0.1914	5.9	0.1378	0.1554	12.8	0.1994	10.3	4.2
0.6	0.1612	0.1726	7.1	0.1226	0.1402	14.4	0.1785	10.7	3.4
0.7	0.1434	0.1549	8.0	0.1082	0.1254	15.9	0.1591	10.9	2.7
0.8	0.1178	0.1344	14.1	0.0913	0.1049	14.9	0.1351	14.7	0.5
0.9	0.0954	0.1055	10.6	0.0747	0.0803	7.5	0.1036	8.6	-1.8
1	0.0634	0.0676	6.6	0.0554	0.0569	2.7	0.0658	3.8	-2.7

Average Precision (Non-interpolated)

	0.2078	0.2160	3.9	0.1564	0.1722	10.1	0.2247	8.1	4.0

Figure 5. Truncated Perfect Merges vs. Baselines

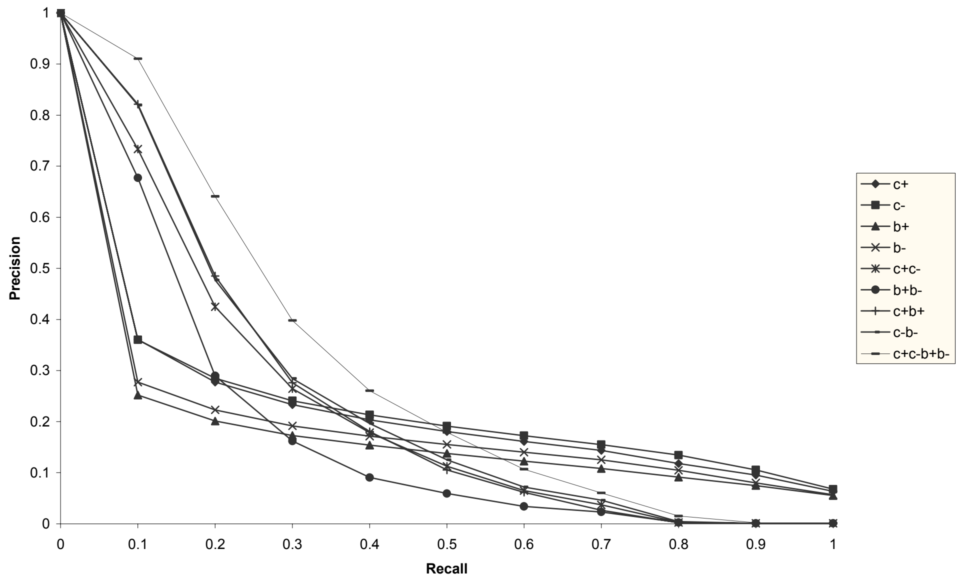
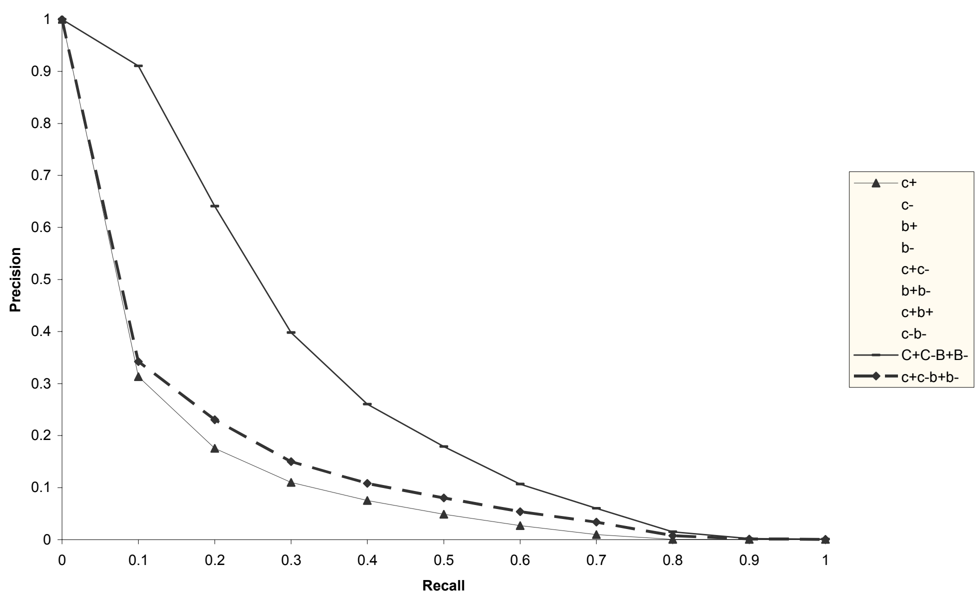


Figure 6. Best Merged Truncated List (k=100) vs. Truncated Conventional System



Although because of space limitations we can only report a small portion of our results here, we have run extensive additional experiments considering different list lengths, different ground truth, different channel combinations and so on. All the evidence suggests that Hypothesis 3 is true. We can, in fact, combine the channels even naively to realize retrieval effectiveness gains over the conventional single-channel CBIR approach. It is even the case that the C- channel is noticeably better than the conventional (C+) channel when truncated lists are used.

5. 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. Our approach offers a simple, cost-effective strategy for boosting the performance of CBIR systems.

We have hinted at the possibilities for using multi-channel CBIR together with relevance feedback (see Figure 2) to enhance retrieval effectiveness.

We have demonstrated that the practical merge performance of an admittedly unoptimized algorithm has materially better retrieval performance (22%) than the conventional approach. Moreover, the performance of the C- channel alone shows noticeable improvement (6%) over the conventional single-channel system.

Although we demonstrated our approach with a 4-channel system, there is some initial evidence that we can get very good performance from three channels (C+C-B-) and from two (C-B-). The latter configuration is easily incorporated into conventional systems. The evaluation of alternative configurations is part of our ongoing investigation.

Our basic CBIR system is typical of conventional approaches. We note that it does not include advanced techniques such as the integrated region matching (IRM) of SIMPLIcity[18] or the keyblocks employed by Zhu et al.[21]. Our future work will consider different CBIR technologies as the black box in the system. In addition we are looking at merging the results of different CBIR technologies over the same and different channels.

6. ACKNOWLEDGMENTS

The authors wish to thank Prof. Zhengxing Sun (Najing University) and Prof. Wenyin Liu (City University of Hong Kong) for the use of their ground truth in this work. Prof. Sun and Prof. Liu were partly supported by the National Natural Science Foundation of China (69903006).

7. REFERENCES

[1] Aslandogan, Y. and Yu, C. Techniques and Systems for Image and Video Retrieval, *IEEE Transactions on Knowledge and Data Engineering*, **11**(1) 1999, 56-63.

[2] Bartell, B. Cottrell, G., and Belew, R. Automatic Combination of Multiple Ranked Retrieval Systems in *Proceedings of ACM SIGIR* (Dublin, Ireland, July 1994), 173-181.

[3] Belkin, N., Cool, C., Croft, W., and Callan, J. The Effect of Multiple Query Representations on Information Retrieval System Performance in *Proceedings of ACM SIGIR* (Pittsburgh, PA, 1993), 339-346.

[4] Belkin, N., Kantor, P., Cool, C. and Quatrain, R. Combining Evidence for Information Retrieval in *Proceedings of TREC-2* (Gaithersburg MD, March 1994), 35-44.

[5] Das-Gupta, P. and Katzer, J. A Study of the Overlap Among Document Representations in *Proceedings of ACM SIGIR* (Bethesda, MD, June 1983), 106-114.

[6] Fox, E. and Shaw J. Combination of Multiple Searches in Proceedings of TREC-2 (Gaithersburg MD, March 1994), 243-252.

[7] French, J. C., Chapin, A. C. and Martin, W. N. Multiple Viewpoints as an Approach to Digital Library Interfaces, *Workshop on Document Search Interface Design and Intelligent Access in Large-scale Collections*, Portland, OR, July, 2002.

[8] French, J. C., Martin, W.N and Watson, J. V. S. A Qualitative Examination of Content-Based Image Retrieval Behavior using Systematically Modified Test Images, *45th IEEE International Midwest Symposium on Circuits and Systems*, Tulsa, OK, August 4-7, 2002.

[9] Gravano, L. and Garcia-Molina, H. Merging Ranks from Heterogeneous Internet Sources in *Proceedings of Very Large Data Bases* (Athens, Greece, Aug. 1997), 196-205.

[10] Narasimhalu, A. D., Kankanhalli, M. S. and Wu, J. Benchmarking Multimedia Databases, *Multimedia Tools and Applications*, **4**(3), 1997, 333-355.

[11] Powell, A. L. and French, J. C. Using Multiple Views of a Document Collection in Information Exploration, *CHI'98 Workshop on Innovation and Evaluation in Information Exploration Interfaces*, Los Angeles, CA, April 19, 1998.

[12] Rui, Y., Huang, T. and Chang, S.-F. Image Retrieval: Past, Present, and Future in *Proceedings of International Symposium on Multimedia Information Processing* (Taiwan, Dec. 1997).

[13] Santini, S., Exploratory Image Databases: Content-Based Retrieval, Academic Press, 2001.

[14] Shaw, J. and Fox, E. Combination of Multiple Searches in *Proceedings of TREC-3* (April 1995), 105-108.

[15] Sparck Jones, K. Automatic Indexing in *Journal of Documentation*, **3**(4), 1974, 393-432.

[16] Sun J., Sun, Z., Zhou, R., Wang, H. A Semantic-based Image Retrieval System: VisEngine, *IEEE 1st International Conference on Machine Learning and Cybernetics*, Beijing, China, November, 2002, 349–353.

[17] Vogt, C. When Does It Make Sense to Linearly Combine Relevance Scores? Poster in *Proceedings of ACM SIGIR* (Philadelphia, PA, 1997).

[18] Wang, J. and Du, Yy. Scalable Integrated Region-Based Image Retrieval Using IRM and Statistical Clustering in *Proceedings of ACM and IEEE Joint Conference on Digital Libraries* (Roanoke, VA, June 2001), 268-277.

[19] Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., and Field, B. Semi-Automatic Image Annotation in *Proceedings of Human-Computer Interaction-Interact* (2001), 326-333.

[20] Yoshitaka, A. and Ichikawa, T. A Survey on Content-Based Retrieval for Multimedia Databases, *IEEE Transactions on Knowledge and Data Engineering*, **11**(1), 81-93.

[21] Zhu, L., Rao, A., and Zhang, A. Advanced Feature Extraction for Keyblock-Based Image Retrieval, *Information Systems*, **27**(8), 2002, 537-557.