

Performance Evaluation of Clustering Algorithms for Scalable Image Retrieval¹

Mohamed Abdel-Mottaleb
Santhana Krishnamachari
Nicholas J. Mankovich
Philips Research
345 Scarborough Road
Briarcliff Manor, NY 10510
{msa, sgk, njm}@philabs.research.philips.com

Abstract

In this paper we present scalable algorithms for image retrieval based on color. Our solution for scalability is to cluster the images in the database into groups of images with similar color content. At search time the query image is first compared with the pre-computed clusters, and only the closest set of clusters is further examined by comparing the query image to the images in that set. This obviates the need to compare the query image with every image in the database, thus making the search scalable to large databases. We have used the hierarchical clustering and the K-means clustering techniques. Performances of these two clustering algorithms are compared when three similarity measures, the histogram intersection measure, the L_1 , and the L_2 measures, are used for image retrieval.

The retrieval accuracy of the clustering algorithms is computed by comparing the results of retrieval with clustering against the results of retrieval without clustering. Our experiments with a database of 2000 color images show that both clustering techniques offer a retrieval accuracy of over 90% with only an average of 300 similarity comparisons (as opposed to 2000 comparisons that are required for retrieval without clustering). Our evaluations show that the hierarchical clustering algorithm outperforms the K-means clustering algorithm for all three similarity measures, although only marginally in some cases.

1: Introduction

Content-based image retrieval has become a prominent research topic in recent years. Research interest in this field has escalated because of the proliferation of video and image data in digital form. The goal in image retrieval is to search through a database to find images that are perceptually similar to a query image. An ideal image retrieval engine is one that can completely comprehend a given image, *i.e.*, to identify the various objects present in the image and their properties. Given the state of the art of research in the image analysis community, such an ideal retrieval system is far from being reality. Moreover retrieval based on human annotation is to no avail, because of the size of the video and image databases and the varying interpretations that different humans can attach to an image. Examples of color content based image retrieval algo-

1. Appeared in IEEE Workshop on Empirical Evaluation of Computer Vision Algorithms, CVPR 1998. Also as a chapter in the book “Empirical Evaluation Techniques in Computer Vision” edited by K. Bowyer and P.J. Phillips

rithms can be found in [1-5, 7, 8].

In a practical scenario, like the Internet, the number of images can be of the order of millions and is ever growing. Even if the time required to compare two images is very short, the cumulative time needed to compare the query image with all the database images is rather long and is probably longer than the time an average user wants to wait. We solve this problem by grouping or clustering the images according to their similarity beforehand, so that at the time of the query, it is not necessary to perform an exhaustive comparison with all the images in the database. The clustering is performed based on visual features extracted automatically from the images.

Performance evaluation has been a challenging issue in the field of content-based retrieval, primarily because of the difficulty associated with calculating quantitative measures to evaluate the quality of retrieval. The *precision* and *recall* measures have been frequently used by many researchers [5] to evaluate the performance of retrieval algorithms. In this paper we introduce a quantitative method to evaluate the retrieval accuracy of clustering algorithms. Our goal is not to subjectively evaluate the quality of retrieval, but to quantitatively compare the performance of retrieval with and without clustering.

In this paper we present clustering techniques for scalable image retrieval from large databases and evaluate the performances of two clustering techniques: the hierarchical and the K-means. Each clustering technique is applied with three different similarity (or distance) measures: the L_1 norm, the L_2 norm, and the histogram intersection measure. The images in the database are clustered into groups based on local color similarity. During retrieval, the query image is initially compared with the representative of each cluster. Then the query image is compared only with the images in the closest clusters and these images are ranked according to their similarity with the query image. Thus it is not necessary to compare the query image with every image in the database. For each clustering technique and similarity measure, the retrieval accuracy is obtained by comparing the results of retrieval with clustering against the results of retrieval without clustering (comparing the query with all the images in the database) using the *same* similarity measure.

Techniques for fast image retrieval from large databases have been presented in [9-11]. All these techniques require that the similarity (distance) measure used to compare images be a metric, *i.e.*, the similarity measure should satisfy the triangle inequality. However, many similarity measures, like the histogram intersection measure used here, do not satisfy the triangle inequality. The scalable retrieval technique presented here does not require the similarity measure to be a metric and hence is more general.

The rest of the paper is organized as follows. Section 2 carries the details of retrieval based on clustering. Section 3 presents the details of different similarity measures and clustering algorithms that are used. Section 4 presents the experimental set-up and the experimental results. Section 5 presents the conclusions of the performance evaluation and avenues for future work.

2: Image clustering

Searching large databases of images is a challenging task especially for retrieval by content. Most search engines calculate the similarity between the query image and all the images in the database and rank the images by sorting their similarities. One problem with this approach is that it does not scale up for large databases. The retrieval time is the sum of two times: T_{sim} and T_{sort} . T_{sim} is the time to calculate the similarity between the query and every image in the database, and T_{sort} is the time to rank all the images in the database according to their similarity to the query.

$$T_{total} = nT_{sim} + O(n \log n)$$

where n is the number of images in the database, T_{1sim} is the time to calculate the similarity between two images, and $O(n\log n)$ is the time to sort n elements.

When the images in the database are clustered, the retrieval time is the sum of three times, the time to calculate the similarity between the query and the cluster centers, the time to calculate the similarity between the query and the images in the nearest clusters and the time to rank the images. Therefore the total search time is:

$$T_{cluster} = kT_{1sim} + lT_{1sim} + O(l\log l)$$

Here k is the number of clusters, l is the number of images in the clusters nearest to the query. Since $k \ll n$ and $l \ll n$, $T_{cluster} \ll T_{total}$.

3: Image representation, similarity measures and clustering

3.1: Image representation

Several histogram-based approaches have been proposed for image retrieval by color [1, 3, 7]. These approaches are based on calculating a similarity measure between the color histogram of the query image and the images in the database. The difference between these approaches is mainly in their choice of the color space and the similarity measure. Since these approaches use a single image histogram to calculate similarities, the results are expected to reflect only global similarity. For example, if a user submits a query image with a sky at the top and sand at the bottom, the retrieved results would have a mix of blue and beige, but not necessarily with blue at the top and beige at the bottom. This can be achieved only if the image representation reflects the local color information.

In this paper, we use the scheme that we presented in [2], to allow retrieval based on local color features. Images in the database are divided into rectangular regions. Then every image is represented by the set of normalized histograms corresponding to these rectangular regions. It should be noted here that the choice of the rectangular region size is important. In one extreme, the whole image is considered as a single region which reflects the global color information. As the size of the region becomes smaller, the local variations of color information is captured by the histograms. The size of the region should be small enough to emphasize the local color and large enough to offer a statistically valid histogram. In the experiments, images were divided into 16 rectangular regions.

3.2: Similarity measures

The similarity between two images is measured by calculating the similarity between the histograms of the corresponding rectangular regions. Then a single measure of similarity between the two images is calculated by combining the individual similarities. For performance evaluation, we have used three similarity (distance) measures that are frequently used in the literature, namely, the histogram intersection, the L_1 , and the L_2 norms. The L_1 and the L_2 norms are distance measures, whereas the histogram intersection measure is a similarity measure. We converted distance measures to similarity measures by taking the negative values of the distance measures. Given two *normalized* histograms, $P = \{p_1, p_2, \dots, p_m\}$, $Q = \{q_1, q_2, \dots, q_m\}$, the histogram intersection measure is defined as [1]:

$$H(P, Q) = \sum_{i=1}^m \min(p_i, q_i).$$

The L_1 norm is defined as:

$$L_1(P, Q) = \sum_{i=1}^m |p_i - q_i|$$

and the L_2 norm is defined as:

$$L_2(P, Q) = \sqrt{\sum_{i=1}^m |p_i - q_i|^2}$$

3.3: Clustering

We have used two clustering algorithms, the hierarchical and the K-means clustering algorithms to group the images into clusters based on the color content. Both these clustering algorithms have been frequently used in the pattern recognition literature. Brief details on the implementation of these two clustering algorithms are presented below.

The hierarchical clustering algorithm [6], is implemented as shown below:

Let n be the number of images in the database, the similarity between all pairs of images is precomputed.

1. The n images in the database are placed in n distinct clusters indexed by $\{C_1, C_2, \dots, C_n\}$.

2. Two distinct unmerged clusters C_k and C_l are picked such that their similarity measure is the largest.

3. These two clusters are merged into a new cluster C_{n+1} . At each step two clusters are merged to form a new cluster. Therefore, the number of clusters is reduced by one.

4. Steps 2 and 3 are repeated until the number of unmerged clusters has reduced to a required number n_c or the largest similarity measure between clusters has dropped to some lower threshold.

The K-means clustering algorithm [6] is implemented as follows:

1. The number of clusters n_c is chosen a priori. The n_c centers are chosen by randomly picking n_c images from the database.

2. For each image in the database, the similarity measure between the image and the clusters centers are computed and the image is assigned to the cluster with which it exhibits the largest similarity measure.

3. New cluster centers are computed as the centroids of the clusters.

4. Steps 2 and 3 are repeated until there is no further change in the cluster centers.

4. Experimental setup

The results presented in this paper are obtained with a database of 2000 images, 200 of which are taken from two collections of COREL Professional Photo CD-ROMs. The rest of the images are obtained from the Department of Water Resources, California. The COREL images are obtained from two CD-ROM collections, the Sampler II - Series 400000 and the Sampler - Series 200000. The list of images that have been used here can be made available to anyone that may be interested in using them by contacting the second author.

The images are of widely varying colors and scene content. The number of clusters n_c is chosen to be 133. The number of clusters, n_c , is chosen such that the average number of images per cluster is 15, i.e., $n_c=2000/15=133$. The number of clusters in both the clustering techniques are the same to ensure fair comparison. The hierarchical and K-means clustering algorithms are applied to the 2000 images in the database using each of the three different similarity measures. For the hierarchical clustering, the number of images in the smallest and the largest cluster is, 2 and 40 with the histogram intersection measure, 2 and 48 with the L_1 measure, and 1 and 63 with the L_2 measure. For the K-means clustering, the corresponding numbers are, 1 and 49 with the histogram intersection measure, 1 and 52 with the L_1 measure, and 1 and 69 with the L_2 measure.

4.1: Retrieval accuracy with clustering

After clustering and selecting the cluster centers, the given query image is first compared with all the cluster centers. The clusters are ranked according to their similarity with the query. Few close clusters from the top of this ranked list are chosen. Then the query image is compared directly with the images in these clusters. Thus the number of comparisons is reduced considerably from comparing the query with all the images in the database. It is shown below that the retrieval accuracy is not compromised in this process. The number of similarity comparisons required depends on the sizes of the clusters and the number of clusters being examined.

We have used a quantitative measure to compare the retrieval results *with* clustering against the retrieval results *without* clustering. A user searching through a large database, is interested in only the top few best matches (say 10 or 20). Hence, if the retrieval with clustering returns the same few best matches as the ones returned by retrieval without clustering, then the retrieval with clustering is very accurate. Assume that the user is interested in only top N best matches and that M is the number of images that are present both in the top N results returned by retrieval with and without clustering. The retrieval accuracy with clustering ψ_i , when the i th image is used as a query is defined as:

$$\psi_i = \frac{M}{N} 100$$

The *average retrieval accuracy* with clustering A_N is obtained by taking the average ψ_i over all the query images.

$$A_N = \frac{1}{n} \sum_{i=1}^n \psi_i$$

4.2: Discussion of results

The experimental results from using the two clustering algorithms with the three similarity measures are presented in Figures 1-7. Each of these plots contain eight points, obtained by examining the top 3, 4, 7, 10, 13, 19, 25, and 31 clusters. In all the figures, the leftmost point corresponds to the result obtained by examining 3 clusters and the rightmost point corresponds to the result obtained by examining 31 clusters. The retrieval accuracy increases as the number of examined clusters is increased.

We conducted two sets of experiments. In the first set, we used each of the 2000 images in the database as a query image. For each query, the retrieval accuracy and the number of image similarity comparisons for the eight different cases are calculated. The averages of the retrieval accuracies and the averages number of comparisons are plotted in Figures 1-3. The results are discussed in Section 4.2.1. In the second set of experiment, we used a set of 300 images that are not a part of the database of 2000 images as query images. These 300 images are also obtained from the Department of Water Resources, California. Again, for each query, the retrieval accuracy and the number of image comparisons for the eight different cases are computed. Figures 4-6 carry the results of this experiment. The results of this experiment are discussed in Section 4.2.2. In the case of the K-means clustering, experiments were repeated with different random selections of the initial centers. We found that for different random initializations the variations in the average retrieval accuracy and the number of similarity comparisons were very small.

4.2.1: Results with database images as queries: Figure 1(a) shows the plot of average retrieval accuracy against the number of comparisons for the hierarchical clustering using the three different similarity measures. It is worth mentioning that for each similarity measure, the retrieval results obtained with clustering are compared against the retrieval results obtained without clustering using the same similarity measure. We have not used a common *ground-truth* against which the retrieval results of different similarity measures are compared against. Figure 1(a) shows that for the hierarchical clustering, the histogram intersection measure offers the largest retrieval accuracy for a given number of comparisons, closely followed by the L_1 measure. The retrieval accuracy for the L_2 measure is significantly lower than the rest. Similar results are obtained with the K-means clustering as shown in Figure 1(b).

One inference that can be drawn from Figure 1 is that both clustering algorithms offer a large reduction in the number of comparisons without sacrificing the retrieval accuracy. A retrieval accuracy of over 90% can be obtained for both clustering algorithms with the histogram intersection and the L_1 measure by examining only the top 13 clusters (out of 133 clusters). The average number of comparisons required to examine the top 13 clusters is less than 300, compared to the 2000 comparisons that are required if the retrieval is performed without clustering. We expect that the reduction in the number of comparisons with clustering will be much larger as the size of the database is increased.

Figures 2 shows the retrieval accuracies for the two clustering algorithms with the histogram intersection measures for values of $N=10$ and 20. As expected the average retrieval accuracy for $N=20$ is lower than that for $N=10$, but only a few percentage points. The difference in retrieval accuracies between $N=10$ and $N=20$ reduces as more number of clusters are examined, *i. e.*, the gap between the two plots, in Figure 2, narrows as we move towards the right in the abscissa. The plots are very similar for the other two similarity measures and hence are not presented.

Figure 3(a) shows the comparison between the hierarchical clustering and the K-means clustering with the histogram intersection measure. The hierarchical clustering performs better than the K-means clustering, but only very marginally. Figure 3(b) shows the same comparison for

the L_1 measure and the conclusion here is the same. For the L_2 measure (not shown) the results are similar.

4.2.2: Results with external images as queries: Figures 1-3 showed the results obtained when the query images are taken from the database. A more thorough evaluation should include queries that are not part of the original database. Figures 4-6 show the results obtained from the second set of experiment, where we used a set of 300 images that are not a part of the original database as queries. The performances reported in Figures 4-6 are very similar to the corresponding results in Figures 1-3 showing that there is no degradation in the retrieval performance when the queries are chosen from outside the database.

Figure 4 shows the retrieval performance for the three different similarity measures. Similar to Figure 1, the histogram intersection measures performs slightly better than the L_1 measure. The L_2 measure exhibits the lowest retrieval accuracy among the three measures. Figure 5 shows the comparison of retrieval accuracies for values of $N=10$ and 20. The results are similar to Figure 2, with the retrieval accuracy for $N=20$ slightly lower than $N=10$ and the difference reduces as the number of clusters examined is increased. Figure 6 shows the comparison of the hierarchical and the K-means clustering algorithms. Again, the results are similar to Figure 3, with the hierarchical clustering performing marginally better than the K-means clustering.

Figure 7 shows the comparison between the two sets of experiments. The average retrieval accuracies obtained with the 2000 images as queries (*internal queries*) and with the 300 images (external to the database) as queries are plotted in Figure 7. It is interesting to see that when less than seven clusters are examined (the third point in each plot), the retrieval accuracy for external queries is less than that of the internal queries. But, as more clusters are examined, the retrieval accuracy for external queries outperformed the retrieval accuracy for internal queries. The results are similar for all three similarity measures (the L_2 measure plots are not shown).

5: Conclusions and future work

The results obtained with the hierarchical and the K-means clustering algorithms show that both clustering algorithms drastically reduce the number of required similarity comparisons without sacrificing the retrieval accuracy. These clustering techniques can be used for scalable image and video retrieval from large databases. The hierarchical clustering algorithm outperforms the K-means clustering algorithm in all cases, even though only marginally in some cases. Out of the three similarity measures that are investigated, the histogram intersection and the L_1 measures perform very similarly, whereas the L_2 measure is significantly worse. We also found that the histogram intersection and the L_1 measures result in more uniform clusters, whereas the L_2 measure results in many very small and very large clusters.

We chose the histogram intersection, the L_1 , and the L_2 measures, because they are well known measures. Our goal is to show the effectiveness of clustering for scalable retrieval and to compare the results when using different clustering algorithms. We are presently experimenting with more sophisticated similarity measures that have been used in the content-based retrieval community and these results will be presented in the near future.

In our computations of retrieval accuracy for each similarity measure, we have obtained the retrieval accuracy by comparing the results of retrieval with clustering against the results of retrieval without clustering using the same similarity measure. One direction of future work, is to identify a benchmark against which the performances of different similarity measures can be compared against.

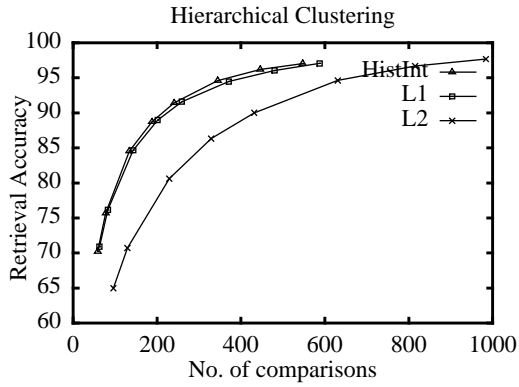
The clustering technique presented here for image retrieval can potentially be used for fast browsing of large image and video databases. From each cluster, one or more representative images can be chosen and thumbnail representations of these images can be used to navigate and browse through the database.

Acknowledgments

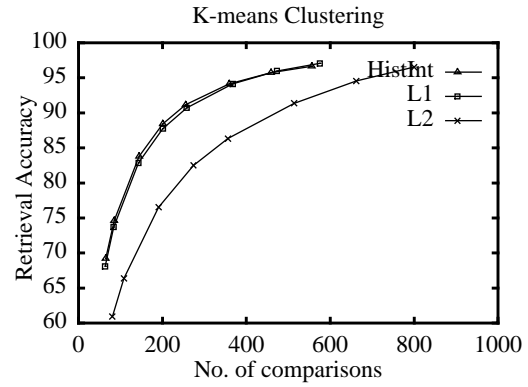
We wish to thank Dave Kearney of the California Department of Water Resources and Ginger Ogle of the University of California, Berkeley, for their assistance in providing us with the Cypress database images.

References

- [1] M. J. Swain and D. H. Ballard, "Color Indexing", *Intl. J. of Computer Vision*, 7(1), pp. 11-32, 1991.
- [2] M. Abdel-Mottaleb, N. Dimitrova, R. Desai, and J. Martino, "CONIVAS: CONtent-based Image and Video Access System", *Proc. of ACM Intl. Multimedia Conference*, Nov. 1996.
- [3] W. Niblack, R. Barber, *et.al.*, "The QBIC Project: Querying Images by Content Using Color, Texture and Shape", In *Storage and Retrieval for Image and Video Databases I*, Vol. 1908, SPIE Proceedings, Feb. 1993.
- [4] J. Smith and S.-F. Chang, "A Fully Automated Content-based Image Query System", *Proc. of ACM Intl. Multimedia Conference*, Nov. 1996.
- [5] H. Zhang, Y. Gong, C. Y. Low and S. W. Smoliar, "Image Retrieval Based on Color Features: An Evaluation Study", *Proc of SPIE*, Vol 2606, pp. 212-220, 1995.
- [6] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.
- [7] M. Stricker and M. Orengo, "Similarity of Color Images", *SPIE Proceedings*, Vol 2420, pp. 381-392, 1995.
- [8] W.Y. Ma and B.S. Manjunath, NeTra: "A Toolbox for Navigating Large Image Databases", pp 568-571, *Proc. of ICIP*, Vol 1, 1997.
- [9] J-Y. Chen, C. A. Bouman, and J. P. Allebach, "Fast Image Database Search using Tree-Structure VQ", pp. 827-830, *Proc. of ICIP*, Vol II, 1997.
- [10] J. Barros, J. French, W. Martin, P. Kelly, and M. Cannon, "Using the triangle inequality to reduce the number of computations required for similarity-based retrieval", *Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases IV*, Vol. 2670, 1996.
- [11] A. Berman and L. Shapiro, "Efficient image retrieval with multiple distance measures", *Proc. of SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases V*, Vol. 3022, 1997.

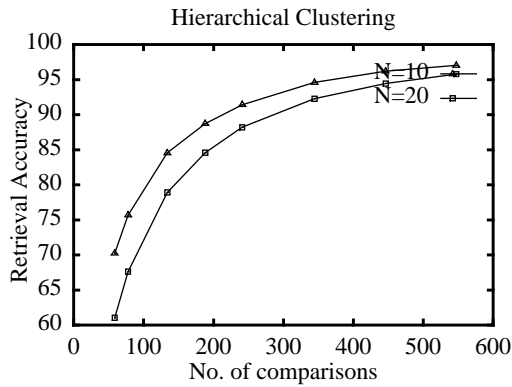


(a)

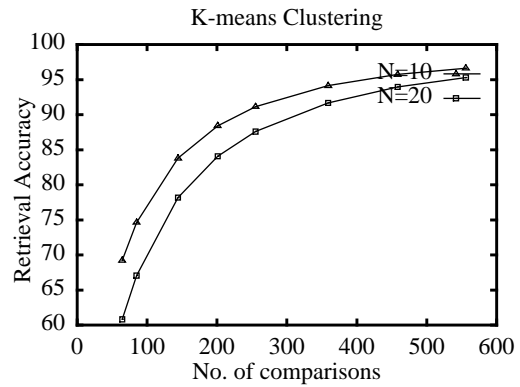


(b)

Figure 1: Average retrieval accuracy of (a) the hierarchical clustering and (b) the K-means clustering with $N=10$. (The eight points in these and the rest of the plots are obtained by examining the top 3, 4, 7, 10, 13, 19, 25, and 31 clusters. The retrieval accuracy and the number of comparisons required increase with the number of clusters examined).

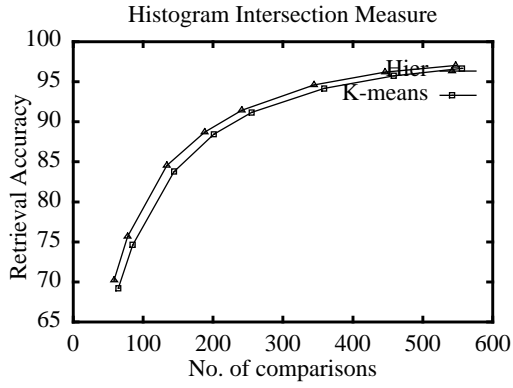


(a)

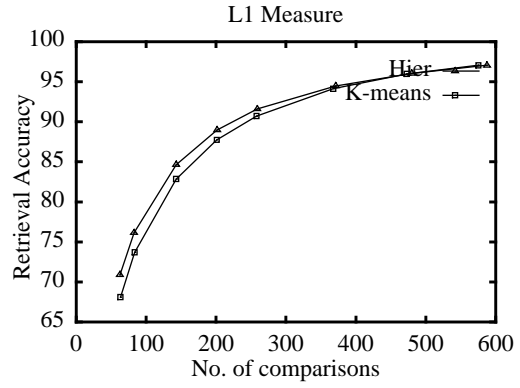


(b)

Figure 2: Average retrieval accuracy of (a) the hierarchical clustering and (b) the K-means clustering with the histogram intersection measure for different values of N .

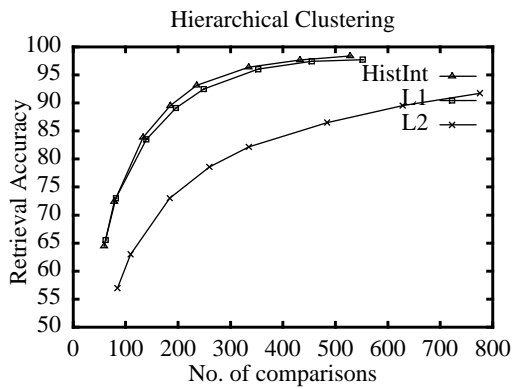


(a)

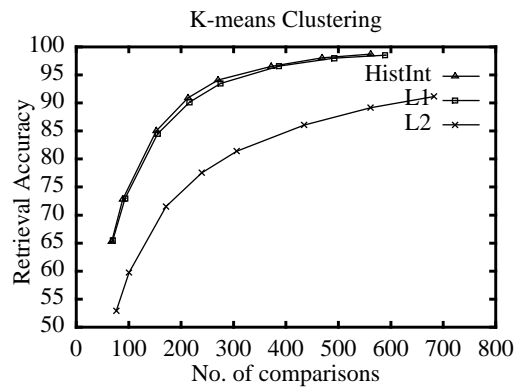


(b)

Figure 3: Comparison of average retrieval accuracies of the hierarchical and the K-means clustering with (a) the histogram intersection measure and (b) the L_1 measure, with $N=10$

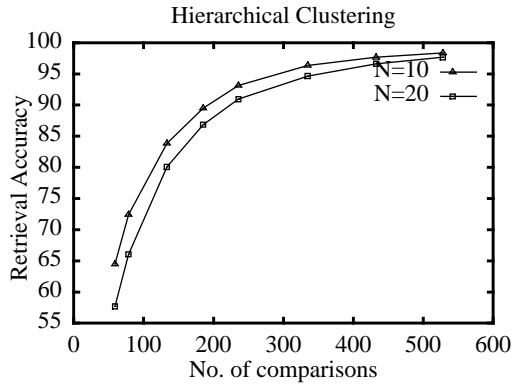


(a)

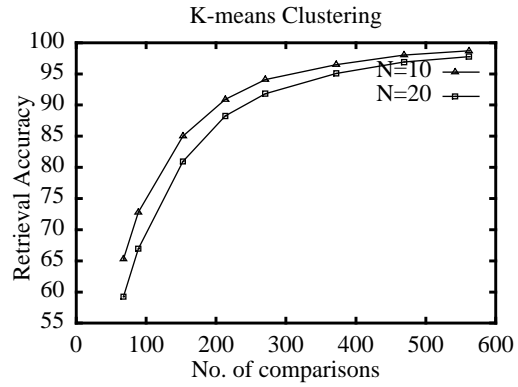


(b)

Figure 4: Average retrieval accuracy of (a) the hierarchical clustering and (b) the K-means clustering with $N=10$ (queries from outside the database).

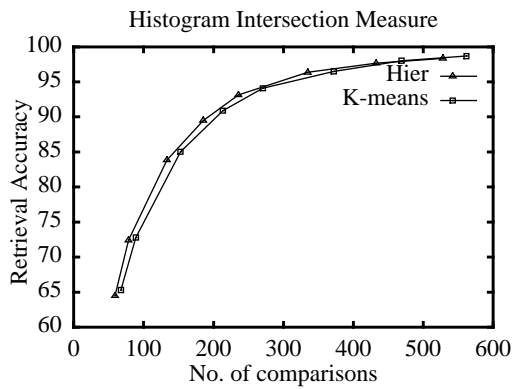


(a)

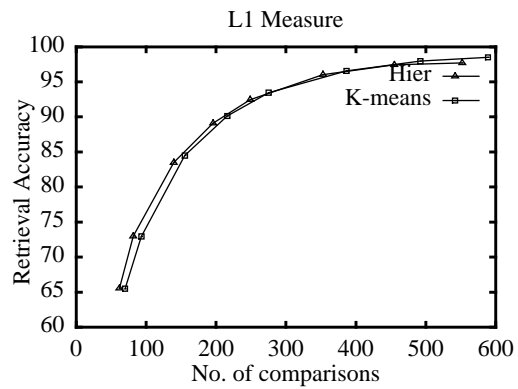


(b)

Figure 5: Average retrieval accuracy of (a) the hierarchical clustering and (b) the K-means clustering with the histogram intersection measure for different values of N (queries from outside the database).

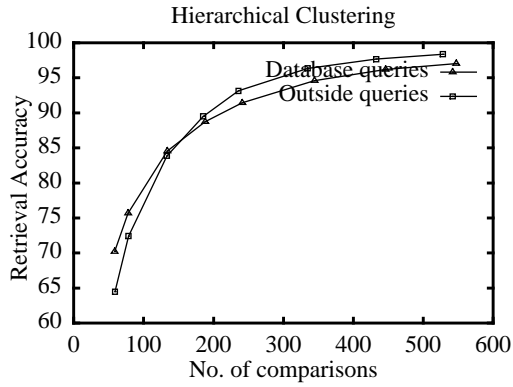


(a)

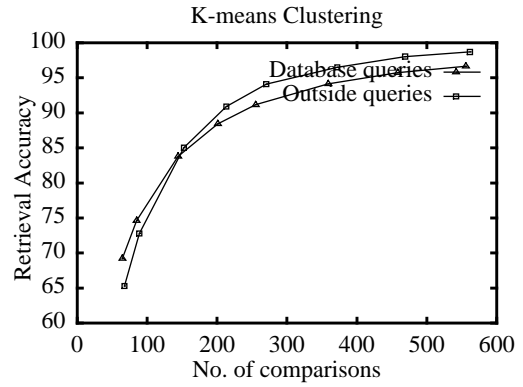


(b)

Figure 6: Comparison of average retrieval accuracies of the hierarchical and K-means clustering with (a) the histogram intersection measure and (b) the L_1 measure, with N=10 (queries from outside the database).



(a)



(b)

Figure 7: Comparison of average retrieval accuracies when the queries come from inside and outside the database for (a) the hierarchical clustering and (b) the K-means clustering with the histogram intersection measure, with $N=10$.