Analysis of the Effectiveness-Efficiency Dependence for Image Retrieval

Martin Heczko **Daniel Keim** **Roger Weber**

Institute of Computer Science
University of Halle, Germany
{heczko, keim}@informatik.uni-halle.de

Institute of Information Systems
ETH Zurich, Switzerland
weber@inf.ethz.ch

Abstract

Similarity search in image database is commonly implemented as nearest-neighbor search in a feature space of the images. For that purpose, a large number of different features as well as different search algorithms have been proposed in literature. While the efficiency aspect of similarity search has attracted a great interest in the past few years, the effectiveness of the search was often neglected. In this work, however, we argue that these two measures interplay with each other. The longer the feature representation is, the better the quality of the retrieval gets, but the larger the execution costs become. In other words, an improvement in effectiveness leads to a deterioration of performance and vice versa. The aim of this work is to explicitly take both measures into account to optimize the retrieval both form a quality perspective and a performance perspective. To this end, we define a benchmark including a measure for the efficiency and the effectiveness of a feature. Then one can compare different features or feature combinations using simple two-dimensional plots. Based on the quality and performance constraints of a user, the search engine can easily determine the optimal feature or feature combination. Finally, we have applied our benchmark to a large number of different feature types to compare their effectiveness-efficiency relationship.

1 Introduction

Extraction, analysis and preprocessing of feature vectors as well as nearest neighbor search in (probably high-dimensional) feature spaces are the key components in image retrieval systems. Examples of such systems include QBIC [FSA+95], CHARIOT [The00], and MARS [ORC+97]. Feature extraction and nearest neighbor search traditionally belong to different fields: image scientists have designed new effective features without taking retrieval costs into account. On the other hand, database researchers have develop efficient index structure for the nearest-neighbor search problem without considering the effectiveness of the retrieval. ¿From a user's perspective, searching for images in similarity search systems typically involves several steps. In the first few steps, a user refines his or her query with the help of relevance feedback until the query matches the information need sufficiently good (cf. MARS [RHM98], CHARIOT [The00]). In the final step of the search process, the archive is extensively searched for (all) relevant images. Obviously, retrieval effectiveness in the first few steps is not so important as retrieval efficiency. In the final step, on the other hand, result quality plays the key role and a user is ready to tolerate longer response times if more relevant images are retrieved.

In this work, we explicitly investigate the relationship between effectiveness and efficiency of content descriptors in large image databases. As it turns out, their is a strong interaction between these two measures. A large feature vector, for instance, often leads to very good retrieval results. On the other hand, execution costs are known to grow linearly with the length of the feature representation. Given this relationship, we can deploy the effectiveness-efficiency relationship to optimize a query with user constraints on the query processing time and/or on the result quality (cf. CHARIOT [BMW01]). In the following, we present a benchmark to compare image features that are represented by high-dimensional vectors. First, we define the test database and a set of sample queries. Then, we define measures to assess the effectiveness and efficiency of the feature (or feature

combination). With respect to effectiveness, usually one draws precision-recall plots for the features under investigation. In this paper, we propose a more compact effectiveness measure which allows easy summarization and comparison of different features. The efficiency of the retrieval clearly depends on the index structure used to solve the nearest-neighbor search problem. Due to the linear dependency of the execution costs on the number of dimensions and the number of data items, the efficiency of features is simply given by their dimensionality.

Finally, we applied our benchmark for a large number of features coming from two different retrieval systems. The features used in CHARIOT [The00] are based on color moments, texture moments and color histograms. The features of [Hec00] apply the wavelet transformation to represent the color distribution of images. The CHARIOT system further supports the combination of features and the partitioning of images. As such, it offers a large number of feature combinations. Given the benchmark, we have determined for each basic feature type and each feature combination an effectiveness value and an efficiency value. Drawing these values in a two dimensional plot, we are able to easily relate different features and combination of features according to their effectiveness (which is the best combination of features) and their efficiency (which feature allows for a fast retrieval with a relatively good quality).

2 Benchmark

In this section, we describe our benchmark to measure the effectiveness and efficiency of different features and feature combinations. The test database contains about 10,000 color images from miscellaneous areas. To measure the effectiveness of the retrieval, we manually determined all relevant images for 32 sample queries. The restriction to 32 queries is due to the relatively high costs for browsing through the database to identify the relevant images. Some sample queries are depicted in [Hec00].

Effectiveness measure. The typical approach to illustrate the effectiveness of search methods is to draw precision-recall graphs. However, with 32 queries all with different numbers of similar (relevant) images, the precision-recall graph turns out to be of merely little help. To obtain a more compact and comparable effectiveness measure, we followed a different approach: our measure takes the ranking of relevant items in the result list into account and also considers the missed relevant entries. Essentially, the quality measure is defined as the ratio of the sum of ranks of all relevant images (= *SumR*) over the sum of the ranks with an optimal feature (=*SumOptR*), i.e. all relevant images occupy the first ranks of the result list. Clearly, if R is the number of relevant item, *SumOptR* is given by $R \cdot (R+1)/2$. Now consider the result list obtained by an experiment and let *SumR* be the sum of the ranks of all relevant items[1]. The basic effectiveness measure *eff* is then given by $\frac{SumOptR}{SumR}$. Obviously, the better the retrieval, the smaller *SumR* and the greater *eff* becomes. But the range of *eff* depends on the number of returned entries E and the number of relevant entries R and is given by $[\ \frac{R+1}{(2 \cdot E+R+1)}, 1]$. Because R varies for each query, we normalize *eff* and finally obtain the normalized effectiveness measure *EFF* within the range $[0,1]$. Finally, the effectiveness of a feature is given by the average $\overline{EFF}$ of the effectiveness values *EFF* for each query in the test. Using the same database, the same sample queries and the same number E of objects to return, we can compare different features and feature combinations based only on their $\overline{EFF}$-values.

Example 1 *Assume that there are 4 relevant items for a query, i.e.* $R = 4$*. Further, let the number of objects to return be given as* $E = 5$*. The optimal sum would be* $SumOptR = 1+2+3+4 = 10$*. If no relevant item would be returned, then* $SumR_{worst} = 6+7+8+9 = 30$ *and* $eff_{worst} = 10/30 = 0.33$*. Now we compute the effectiveness for a feature A: assume it returns only two relevant objects with ranks 1 and 3. Thus,* $SumR_A = 1+3+(E+1)+(E+2) = 1+3+6+7 = 17$*. Thus, the basic effectiveness of feature A is given as* $eff_A = 10/17 = 0.59$*. After normalization we obtain the actual effectiveness:* $EFF_q = \frac{eff_q - eff_{worst}}{1 - eff_{worst}} = \frac{0.59-0.33}{1-0.33} = 0.38$

[1] If a relevant object is not in the result list, we assign best case ranks to it, i.e. it is assumed that the object would follow right after all entries of the result list.

Efficiency measure. The efficiency is the second important measure of our benchmark. The retrieval costs obviously depend on the index structure with which the benchmark solved the nearest-neighbor search problem. Recent work has shown that search costs in high-dimensional spaces are exponentially dependant on the dimensionality of the features [BBKK97, WSB98, BGRS99]. As such, it becomes obvious that above some dimensionality threshold all data items must be considered to answer the query [WSB98]. Rather surprisingly, this is not only a theoretical phenomenon: in as low as 10-dimensional feature spaces for images, a brute-force sequential scan often performs better than a hierarchical organization of the data set. Newer approaches like the VA-File [WSB98], the IQ-Tree [BBJ+00] or the P-Sphere Tree [GR00] perform better than the sequential scan, but are still linear dependent on the number of dimensions and the number of data items. Consequently, the (total) number of dimensions directly determines the retrieval efficiency of the feature. Absolute response times for the retrieval, however, further depend on the index structure that performed the search and the database size. Our implementation of the benchmark uses the VA-File [WSB98] to search for similar images. A nice property of the VA-File is that it can combine different features on the fly resulting in a still linear dependency on the number of dimensions. Other approaches like Fagin's A0-algorithm [Fag96] suffer from a more than linear dependency.

Effectiveness-efficiency dependence. Finally, we are able to assign an effectiveness and efficiency value to each feature and feature combination. As motivated above, we use $\overline{EFF}$ as the effectiveness measure and the dimensionality as the efficiency measure. To compare the effectiveness-efficiency dependence of different features, we plot the values in a two-dimensional diagram. The dimensions represent the dimensionality of the feature and its effectiveness, respectively.

3 Results

This section investigates the effectiveness and efficiency of the features used by CHARIOT [The00] and by [Hec00]. We are interested in determining the best feature or feature combination and to explicitly relate improved retrieval quality to additional execution costs. First, we consider single features, then we investigate feature combinations. In all experiments, we searched for the first 20 answers, i.e. we set $E = 20$.

3.1 Single features

The following list describes the basic feature types and denotes their effectiveness and efficiency values.

CHARIOT/Color Histograms: We used color histograms in the RGB-space with 64 reference colors [SK97]. The distance measure was a quadratic distance function taking correlation between reference colors explicitly into account.
$\overline{EFF}_{Hist64} = 0.19$ (64 dimensions)

CHARIOT/Color Moments: The color moment feature of Stricker et. al [SO95] first transforms the pixels from the RGB-space to a perceptually uniform space such as the Lab-space. As most of the information is concentrated in the first few moments, they determine only the mean value, the variance and the skewness of the color distribution for each channel.
$\overline{EFF}_{LabCovR1} = 0.26$ (9 dimensions)

CHARIOT/Texture Moments: A widely used representation for texture is based on Gabor filters [MM96]. Essentially, the Gabor filter measure the presence of patterns in various directions and at various scales. Our implementation uses 5 directions at 3 different scales and determines the first two moments.
$\overline{EFF}_{TextureGaborR1} = 0.20$ (30 dimensions)

CHARIOT/Layout of Image: The approach of Stricker et. al [SD97] is to divide the image into several possibly overlapping regions, to determine a feature vector for each region, and to concatenate these vectors

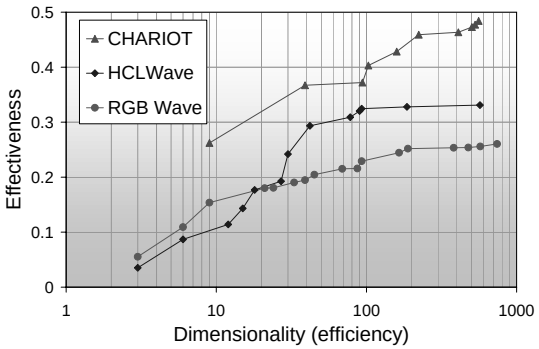


Figure 1: Best effectiveness-efficiency dependence.

to a much longer feature vector for the entire image. We applied this approach to Color Moments (35 hierarchical, overlapping regions) and to Texture Moments (5 fuzzy regions).
$\overline{EFF}_{LabCovR35} = 0.37$ $(9 \times 35 = 315$ dimensions)
$\overline{EFF}_{TextureGaborR5} = 0.32$ $(30 \times 5 = 150$ dimensions)

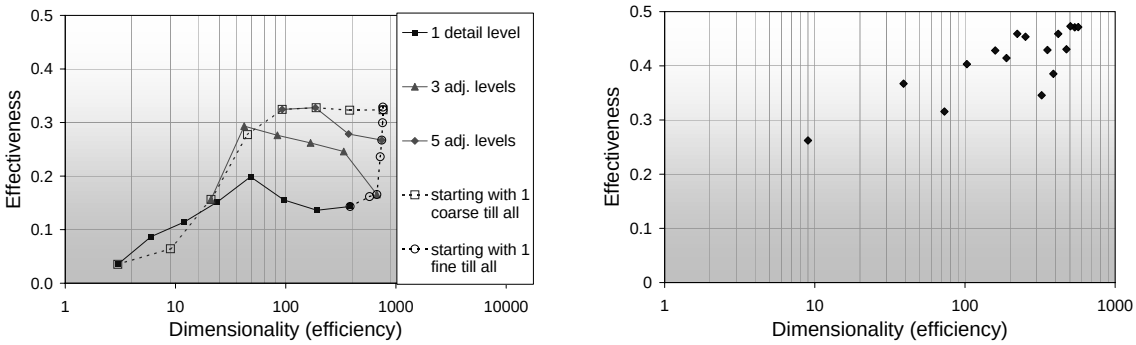
Color Wavelet Coefficients: We processed he pixels of an image either in the *RGB*, *HCL* or the *Lab* color space. For each channel, we determined a density histogram (256 values) and applied the Haar-Wavelet transform to each histogram [HHK00]. The different scales in the wavelet transform describe different aspects of the content. For instance, while the higher levels (more coefficients) describe the finer nuances in color distribution (textural structure) of the image, the lower levels carry coarse information about the color distribution.
$\overline{EFF}_{RGB\_Wave} = 0.26$ $(3 \times 255 = 765$ dimensions)
$\overline{EFF}_{Lab\_Wave} = 0.32$ $(2 \times 255 + 127 = 637$ dimensions)
$\overline{EFF}_{HCL\_Wave} = 0.33$ $(2 \times 255 + 127 = 637$ dimensions)

3.2 Feature Combinations

In the following, we consider combinations of features. With the CHARIOT system, we are able to freely combine features, e.g. color and texture with 5 fuzzy regions. The wavelet based features, can be split into 8 detail levels. Hence, we can combine these levels in arbitrary ways.

With the first experiment, illustrated by Figure 1, we compared the effectiveness-efficiency dependence of: 1) the feature combination based on the wavelet approach (for the color models *HCL* and *RGB*; *Lab* performs comparable to *HCL*), and of 2) the feature combinations in the CHARIOT system. For each combination, we determined its efficiency (i.e. the dimensionality) and its effectiveness. For the figure, we only plotted the results of larger feature combinations, if no smaller feature vector had a better effectiveness. This way, one can see by how much the feature vector must be enlarged to gain a better effectiveness. For instance, the best feature combination of the CHARIOT system with 559 dimensions had an effectiveness of 0.48. The best 9-dimensional feature yielded only an effectiveness of 0.26. On the other hand, to improve the effectiveness of the retrieval by 0.1, we have to use features with roughly 8 times more dimensions. With other words, the improved effectiveness comes at much higher retrieval costs.

In the following, we elaborate on the optimal combination of wavelet levels and feature types in the CHARIOT system. For the wavelet based approach, there is the problem of choosing the right combination of detail levels to search on. We used our benchmark to explore the contribution of the individual detail levels and their combinations to the effectiveness of the search (*HCL* color model). On the other hand, we are also interested in the additional costs involved by the improvement of the search. Figure 2 (a) compares the effectiveness and



(a) Combinations of Wavelet levels

(b) Feature combination with *LabCovR*1

Figure 2: Effectiveness-efficiency dependence of feature combinations.

efficiency of combinations with adjacent levels (1, 3, and 5 levels). The figure further contains the combinations starting with the finest/coarsest levels and continuously adding the next coarser/finer level. As one can see, the coarser levels are significant for similarity. Adding finer levels can increase the effectiveness. The second detail level alone exposed to be better than the first coarse levels together. The first 4 coarsest levels are a little bit less effective than the same combination without the coarsest detail level (second 3 adjacent levels combination). This does not hold for larger combinations (compare 5 adjacent levels).

Figure 2 (b) compares different feature combinations containing the color moment feature *LabCovR*1. For each possible combinations with other feature types (i.e. 16 combinations), we determined the effectiveness and plotted this value together with the dimensionality of the feature combination in the figure. An interesting observation is that not each feature combination leads to a better retrieval compared to using *LabCovR*1 only. Especially, if one combines features of the same type but using a different partitioning of the images (e.g. *LabCovR*1 and *LabCovR*35 with $\overline{EFF}_{LabCovR1,LabCovR35} = 0.34$), the result is often not much better than using only the feature (e.g. *LabCovR*35 with $\overline{EFF}_{LabCovR35} = 0.37$). The best results are obtained, when combining different feature types like, for instance, color moments and texture moments. A further observation not depicted by the figures is that features with a partitioning of the image obtain considerably better retrieval effectiveness values.

4 Conclusions and Outlook

The variety of features to describe images is very large. So far, only little work exists to compare the different feature types in terms of effectiveness. In this paper, we not only considered the effectiveness of different feature types but also took the efficiency of the features into account. A first contribution was the presentation of a benchmark for the evaluation of different features and their combinations. The proposed effectiveness measure simplifies the comparison of different features and avoids the problems incurred by precision-recall plots. Our preliminary investigation has shown that the effectiveness of high-dimensional features can be better than the one of low-dimensional features, but this is not always the case. Recently, it was questioned whether such high-dimensional feature combinations are useful at all [BGRS99]. Our experiments, however, show that very high-dimensional (more than 500 dimensions) features are not harmful as concluded in [BGRS99]. For instance, the best feature combination with 9 dimensions achieved an effectiveness of 0.26. A combination with 559 components obtained an effectiveness of 0.48. In other words, increasing the dimensionality by a factor of 62 lead to a 80% better retrieval. However, the improved effectiveness comes at much higher execution costs: executions costs for a retrieval with the 559-dimensional feature are 62 times higher than the ones for the 9-dimensional feature. The relationship between efficiency and effectiveness can be exploited to optimize query

evaluation with respect to user constraints on the quality and the response times. Furthermore, we can use the benchmark to select the best features and to fine tune parameters of similarity search methods. For instance, it is not always obvious which distance measure implements the best notion for dissimilarity.

As future work, we want to broaden our benchmark in the following ways: 1) enlarging the database as well as the number of sample queries; 2) including more feature types, feature combinations and partitioning schemes; 3) investigating the influence of dimensionality reduction and approximate search on the retrieval effectiveness and efficiency; and 4) taking relevance feedback into account.

References

[BBJ+00] S. Berchtold, C. Böhm, H. V. Jagadish, H.-P. Kriegel, and J. Sander. Independent Quantization: An Index Compression Technique for High-Dimensional Data Spaces. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 577–588, San Diego, CA, Feb./Mar. 2000.

[BBKK97] S. Berchtold, C. Böhm, D. A. Keim, and H.-P. Kriegel. A Cost Model For Nearest Neighbour Search. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 78–86, Tucson, Arizona, USA, May 1997. ACM Press.

[BGRS99] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "Nearest Neighbour" Meaningful? In *Proceedings of the International Conference on Database Theory (ICDT)*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235, Jerusalem, Israel, Jan. 1999. Springer.

[BMW01] K. Böhm, M. Mlivoncic, and R. Weber. Quality-Aware Load-Sensitive Planninf of Image Similarity Queries. In *Proceedings of the International Conference on Data Engineering (ICDE)*, Heidelberg, Germany, Apr. 2001. IEEE Computer Society.

[Fag96] R. Fagin. Combining Fuzzy Information from Multiple Systems. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 216–226, Montreal, Canada, June 1996. ACM Press.

[FSA+95] M. Flickner, H. S. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, Sept. 1995.

[GR00] J. Goldstein and R. Ramakrishnan. Contrast Plots and P-Sphere Trees: Space vs. Time in NN Searches. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 429–440, Cairo, Egypt, Sept. 2000. Morgan Kaufmann.

[Hec00] M. Heczko. Effectiveness test images. http://malta.informatik.uni-halle.de/∼heczko/SimSearch/, 2000.

[HHK00] M. A. Heczko, A. Hinneburg, and D. A. Keim. Multiresolution Similarity Search in Image Databases. Technical report, Department of Computer Science, Martin-Luther-University Halle-Wittenberg, 2000. Submitted for publication.

[MM96] B. Manjunath and W. Ma. Texture Features for Browsing and Retrieval of Image Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8), Aug. 1996.

[ORC+97] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T. S. Huang. Supporting Similarity Queries in MARS. ACM Multimedia. In *Proceedings of the fifth ACM international conference on Multimedia*, pages 403–413, Seattle, WA, USA, Nov. 1997. ACM Press.

[RHM98] Y. Rui, T. Huang, and S. Mehrotra. Relevance Feedback Techniques in Interactive Content-Based Image Retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, San Jose, California, USA, Jan. 1998.

[SD97] M. Stricker and A. Dimai. Spectral Covariance and Fuzzy Regions for Image Indexing. *Machine Vision and Applications*, 10:66–73, 1997.

[SK97] T. Seidl and H.-P. Kriegel. Efficient User-Adaptable Similarity Search in Large Multimedia Databases. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 506–515, Athens, Greece, Aug. 1997. Morgan Kaufmann.

[SO95] M. A. Stricker and M. Orengo. Similarity of Color Images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, volume 2420 of *SPIE Proceedings*, pages 381–392, San Diego/La Jolla, CA, USA, Feb. 1995.

[The00] The Database Group of ETH Zurich. The CHARIOT Project. http://simulant.ethz.ch/Chariot/, 2000.

[WSB98] R. Weber, H.-J. Schek, and S. Blott. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 194–205, New York City, New York, USA, Aug. 1998. Morgan Kaufmann.