Features for Image Retrieval: A Quantitative Comparison

Thomas Deselaers, Daniel Keysers, and Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department,
RWTH Aachen University – D-52056 Aachen, Germany
{deselaers, keysers, ney}@informatik.rwth-aachen.de

Abstract. In this paper, different well-known features for image retrieval are quantitatively compared and their correlation is analyzed. We compare the features for two different image retrieval tasks (color photographs and medical radiographs) and a clear difference in performance is observed, which can be used as a basis for an appropriate choice of features. In the past a systematic analysis of image retrieval systems or features was often difficult because different studies usually used different data sets and no common performance measures were established.

1 Introduction

For content-based image retrieval (CBIR), i.e. searching in image databases based on image content, several image retrieval systems have been developed. One of the first systems was the QBIC system [4]. Other popular research systems are BlobWorld [1], VIPER/GIFT [16], SIMBA [15], and SIMPLIcity [18].

All these systems compare images based on specific features in one way or another and therefore a large variety of features for image retrieval exists. Usually, CBIR systems do not use all known features as this would involve large amounts of data and increase the necessary computing time. Instead, a set of features appropriate to the given task is ususally selected, but it is difficult to judge beforehand which features are appropriate for which tasks. The difficulty to assess the performance of a feature described in a publication is increased further by the fact that often the systems are evaluated on different datasets and few if any quantitative results are reported.

In this work, a short overview of common features used for image retrieval is given and the correlation of different features for different tasks is analyzed. Furthermore, quantitative results for two databases representing different image retrieval tasks are given to compare the performance of the features. To our knowledge no such comparison exists yet, whereas [13] presents a quantitative comparison of different dissimilarity measures.

In the system[1] used, images are represented by features and compared using specific distance measures. These distances are combined in a weighted sum

$$D(Q,X) := \sum_{m=1}^{M} w_m \cdot d_m(Q_m, X_m)$$

[1] http://www-i6.informatik.rwth-aachen.de/˜deselaers/fire.html

where Q is the query image, $X \in \mathcal{B}$ is an image from the database $\mathcal{B}$, Q_m and X_m are the mth features of the images, respectively, d_m is the corresponding distance measure, and w_m is a weighting coefficient. For each d_m, $\sum_{X \in \mathcal{B}} d_m(Q_m, X_m) = 1$ is enforced by normalization. A set $\mathcal{R}(Q)$ of K database images is returned with

$$\mathcal{R}(Q) = \{X \in \mathcal{B} : D(Q, X) \leq D(Q, X') \; \forall X' \in \mathcal{B} \backslash \mathcal{R}(Q)\} \text{ with } |\mathcal{R}(Q)| = K$$

Using only one feature at a time, this architecture allows us to compare the impact of different features on the retrieval results directly. The issue of choosing appropriate weightings of features is addressed in [2, 3]. For the quantification of retrieval results two problems arise:

1. Only very few datasets with hand-labelled relevances are available to compare different retrieval systems and these datasets are not commonly used. A set of 15 queries with manually determined relevant results is presented in [15], and experiments on these data can be used for a first comparison [2]. Nevertheless, due to the small number of images it is difficult to use these data for a thorough analysis. Therefore we use databases containing general images which are partitioned into separate classes of images.
2. No standard performance measure is established in image retrieval. It has been proposed to adopt some of the performance measures used in textual information retrieval for image retrieval [8]. The precision-recall-graph is a common performance measure which can be summarized in one number by the area under the graph. In previous experiments it was observed that the error rate (ER) of the best match is strongly correlated (with a correlation coefficient of -0.93) to this area [3] and therefore we use the ER as retrieval performance measure in the following. This allows us to compare the results to published results of classification experiments on the same data.

2 Features for Image Retrieval

In this section we present different types of features for image retrieval and the method of multidimensional scaling to visualize similarities between different features. We restrict the presentation to a brief overview of each feature and refer to references for further details. In this work, the goal is not to introduce new features but to give quantitative results for a comparison of existing features for image retrieval tasks.

Table 1 gives an overview of the features and comparison measures used.

Table 1. Used features along with associated distance measures.

Feature X_m	distance measure d_m
image features	Euclidean distance, Image Distortion Model [6]
color histograms	Jeffrey divergence [13]
invariant feature histograms	Jeffrey divergence [13]
Gabor feature histograms	Jeffrey divergence [13]
Tamura texture feature histograms	Jeffrey divergence [13]
local features	direct transfer, LFIDM, Jeffrey divergence [2]
region based features	integrated region matching [18]

Image Features. The most straight forward approach is to directly use the pixel values as features. For example, the images might be scaled to a common size and compared using the Euclidean distance. In optical character recognition and for medical data improved methods based on image features usually obtain excellent results. In this work we use the Euclidean distance and the image distortion model (IDM) [6] to directly compare images.

Color Histograms. Color histograms are widely used in image retrieval, e.g. [4]. It is one of the most basic approaches and to show performance improvement image retrieval systems are often compared to a system using only color histograms. Color histograms give an estimation of the distribution of the colors in the image. The color space is partitioned and for each partition the pixels within its range are counted, resulting in a representation of the relative frequencies of the occurring colors. In accordance with [13], we use the Jeffrey divergence to compare histograms.

Invariant Features. A feature is called invariant with respect to certain transformations, if it does not change when these transformations are applied to the image. The transformations considered here are mainly translation, rotation, and scaling. In this work, invariant feature histograms as presented in [15] are used. These features are based on the idea of constructing features invariant with respect to certain transformations by integration over all considered transformations. The resulting histograms are compared using the Jeffrey divergence [13].

Invariant Fourier Mellin Features. It is well known that the amplitude spectrum of the Fourier transformation is invariant against translation. Using this knowledge and log-polar coordinates it is possible to create a feature invariant with respect to rotation, scaling, and translation [14]. These features are compared using the Euclidean distance.

Gabor Features. In texture analysis Gabor filters are frequently used [5]. In this work we apply the method presented in [10] where the HSV color space is used and hue and saturation are represented as one complex value. From these features we create histograms which are compared using the Jeffrey divergence [13].

Tamura Texture Features. In [17] the authors propose six texture features corresponding to human visual perception: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity*, and *roughness*. From experiments testing the significance of these features with respect to human perception, it was concluded that the first three features are very important. Thus in our experiments we use coarseness, contrast, and directionality to create a histogram describing the texture [2] and compare these histograms using the Jeffrey divergence [13]. In the QBIC system [4] histograms of these features are used as well.

Local Features. Local features are small (square) sub-images extracted from the original image. It is known that local features can yield good results in various classification tasks [11]. Local features have some interesting properties for image recognition, e.g. they are inherently robust against translation. These properties are also interesting for image retrieval. To use local features for image retrieval, three different methods are available [2]:

1. direct transfer: The local features are extracted from each database image and from the query image. Then, the nearest neighbors for each of the local features of the query are searched and the database images containing most of these neighbors are returned. *2. local feature image distortion model (LFIDM)*: The local features from the query image are compared to the local features of each image of the database and the distances between them are summed up. The images with the lowest total distances are returned. *3. histograms of local features*: A reasonably large amount of local features from the database is clustered and then each database image is represented by a histogram of indices of these clusters. These histograms are then compared using the Jeffrey divergence.

Region-based Features. Another approach to representing images is based on the idea to find image regions which roughly correspond to objects in the images. To achieve this objective the image is segmented into regions. The task of segmentation has been thoroughly studied [9] but most of the algorithms are limited to special tasks because image segmentation is closely connected to understanding arbitrary images, a yet unsolved problem. Nevertheless, some image retrieval systems successfully use image segmentation techniques [1, 18]. We use the approach presented in [18] to compare region descriptions of images.

2.1 Correlation of Features

Since we have a large variety of features at our disposal, we may want to select an appropriate set of features for a given image retrieval task. Obviously, there are some correlations between different features. To detect these correlations, we propose to create a distance matrix for a database using all available features. Using a leaving-one-out approach, the distances between all pairs of images from a database are determined for each available feature. For a database of N images with M features this results in an $N' \times M$ distance matrix D obtained from $N' = N \cdot (N-1)/2$ image pairs. From this matrix, the covariances $\Sigma_{mm'}$ and correlations $R_{mm'}$ are determined as

$$\Sigma_{mm'} = \frac{1}{N'} \sum_{n=1}^{N'} D_{nm} D_{nm'} - \Big(\frac{1}{N'} \sum_{n=1}^{N'} D_{nm} \Big) \Big(\frac{1}{N'} \sum_{n=1}^{N'} D_{nm'} \Big), \quad R_{mm'} = \frac{\Sigma_{mm'}}{\sqrt{\Sigma_{mm} \Sigma_{m'm'}}}$$

where D_{nm} and $D_{nm'}$ denote the distances of the nth image comparison using the m-th and m'-th feature, respectively. The entries of the correlation matrix R are interpreted as similarities of different features. A high value $R_{mm'}$ denotes a high similarity in the distances calculated based on the features m and m', respectively. This similarity matrix R is easily converted into a dissimilarity matrix W by setting $W_{mm'} := 1 - |R_{mm'}|$. This dissimilarity matrix W is then visualized using multi-dimensional scaling.

Multi-dimensional scaling seeks a representation of data points in a low dimensional space while preserving the distances between data points as much as possible. Here, the data is presented in a two-dimensional space for visualization. A freely available MatLab library[2] was used for multi-dimensional scaling.

[2] http://www.biol.ttu.edu/Strauss/Matlab/matlab.htm

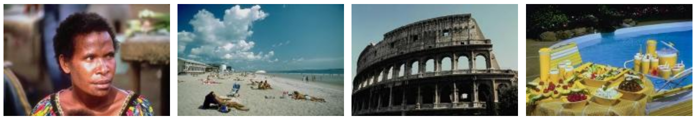

Fig. 1. Example images from the WANG database.

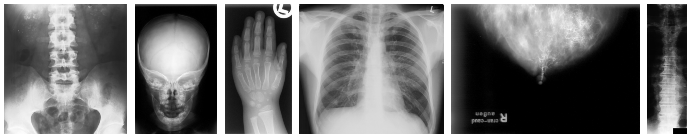

Fig. 2. Examples images from the IRMA database.

3 Databases

Due to the lack of a common database for evaluation in CBIR with known relevances we use two databases where relevances are implicitly given by classifications. These databases are chosen as representatives for two different types of CBIR tasks: The WANG database represents an CBIR task with arbitrary photographs. In contrast, the IRMA database represents a CBIR task in which the images involved depict more clearly defined objects, i.e. the domain is considerably narrower.

WANG. The WANG database is a subset of 1000 images of the Corel database which were selected manually to form 10 classes of 100 images each. The images are subdivided into 10 sufficiently distinct classes (e.g. 'Africa', 'beach', 'monuments', 'food') such that it can be assumed that a user wants to find the other images from the class if the query is from one of these ten classes. This database was created at the Pennsylvania State University and is publicly available[3]. The images are of size 384 × 256 and some examples are depicted in Figure 1.

IRMA. The IRMA database is a database of 1617 medical radiographs collected in a collaboration project of the RWTH Aachen University [7]. The complete data are labelled using a multi-axial code describing several properties of the images. For the experiments presented here, the data were divided into the six classes 'abdomen', 'skull', 'limbs', 'chest', 'breast' and 'spine', describing different body regions. The images are of varying sizes. Some examples are depicted in Figure 2.

4 Results

To obtain systematic results for different features used in CBIR, we first analyze the characteristics of the features using their correlation. Then the performance of the retrieval system is determined for the IRMA and the WANG task. That is, we give leaving-one-out error rates for these two databases. Obviously one can expect to obtain better results using a combination of more than one feature but here we limit the investigation to the impact of single features on the retrieval result. Details about combinations of features are presented in [2, 3].

[3] http://wang.ist.psu.edu/

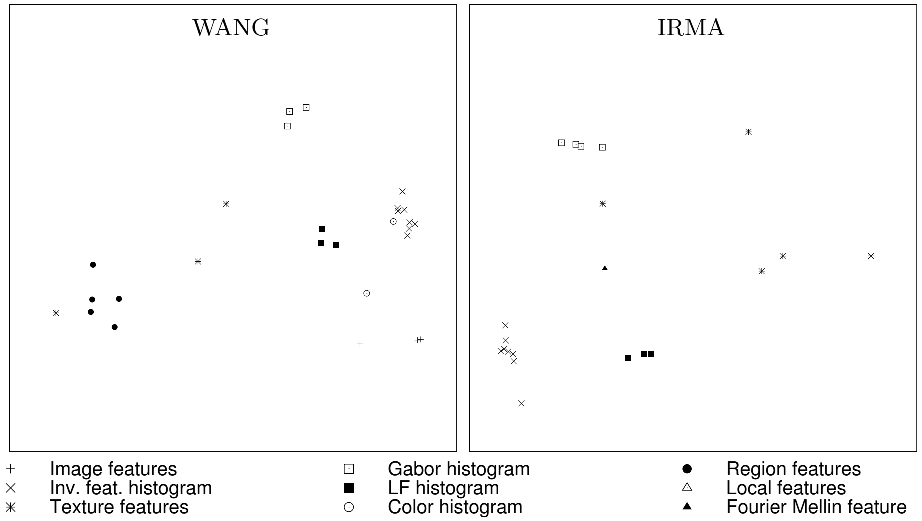

Fig. 3. Two-dimensional representation from multi-dimensional scaling for features on the WANG and IRMA database.

4.1 Correlation of Features

For improved performance, the advantages of different features can be combined. However, it is not clear how to choose the appropriate combination. To analyze which features have similar properties, we perform a correlation analysis as described in Section 2.1 for the WANG and IRMA database in a leaving-one-out manner. The results from multi-dimensional scaling are shown in Figure 3. The points in these graphs denote the different features. Several points of the same type represent different settings for the feature. The distances between the points indicate the correlations of the features. That is, points that are close together stand for features that are highly correlated and points farther away denote features with different characteristics.

The graphs show that there are clear clusters of features. Both graphs have a large cluster of invariant feature histograms with monomial kernel functions. Also, the graphs show clusters of local features, local feature histograms, and Gabor feature histograms. The texture features do not form a cluster. This suggests that they describe different textural properties of the images and that it may be useful to combine them. In contrast, the cluster of invariant features shows that it is not suitable to use different invariant features at the same time. From Figure 3 it can be observed that region features, image features, invariant feature histograms, and Gabor histograms appear to have low correlation for the WANG data and therefore a combination of these features may be useful for photograph-like images. For the radiograph data the interpretation of Figure 3 suggests to use texture features, image features, invariant feature histograms, and Gabor histograms. The combination of these features is addressed in [2, 3].

4.2 Different Features for Different Tasks

As motivated above we use the error rate (ER) to compare the performance of different features. In [3] it has been observed that the commonly used measures

Table 2. Error rates [%] different features for the WANG and IRMA databases.

WANG

Feature	ER [%]
inv. feat. histogram	15.9
color histogram	17.9
pixel values (IDM)	22.3
Tamura histogram	31.0
local feature histogram	32.5
Gabor histogram	48.2
regions	54.3
pixel values (Euclidean)	55.1
local features	62.5

IRMA

Feature	ER [%]
pixel values (IDM)	6.7
local feature histogram	9.3
local features	13.0
pixel values (Euclidean)	17.7
Tamura histogram	19.3
Gabor histogram	24.4
inv. feat. histogram	29.2
Fourier Mellin feature	53.1
extended tangent distance [6]	8.0
pseudo 2D HMM [6]	5.3

precision and recall are strongly correlated to the error rate. Furthermore, the error rate is one number that is easy to interpret and widely used in the context of image classification.

Table 2 shows error rates for different features for the WANG and IRMA databases. From this table it can be observed that these different tasks require different features. For the WANG database, consisting of very general photographs, invariant feature histograms and color histograms perform very well but for the IRMA database, consisting of images with mainly one clearly defined object per image, these features perform badly. In contrast to this, the pixel values as features perform very well for the IRMA task and badly for the WANG task. Also, the strong correlation of invariant feature histograms with color histograms is visible: for the WANG database the invariant feature histograms yield only small improvement. In both cases, the Tamura histograms obtain good results taken into account that they represent the textures of the images only. It is interesting to observe that for the IRMA database the top four methods are different representations and comparison methods based on the pixel values, i.e. appearance-based representations.

5 Conclusion

In this work, we quantitatively compared different features for CBIR tasks. The results show clearly that the performance of features is task dependent. For databases of arbitrary color photographs features like color histograms and invariant feature histograms are essential to obtain good results. For databases from a narrower domain, i.e. with clearly defined objects as content, the pixel values of the images in combination with a suitable distance measure are most important for good retrieval performance. Furthermore, a method to visualize the correlation between features was introduced, which allows us to choose features of different characteristics for feature combination. In the future, the observations regarding the suitability of features for different tasks have to be experimentally validated on further databases.

References

1. C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *Int. Conf. Visual Information Systems*, pp. 509–516, Amsterdam, The Netherlands, June 1999.
2. T. Deselaers. Features for image retrieval. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, Germany, Dec. 2003.
3. T. Deselaers, D. Keysers, and H. Ney. Classification error rate for quantitative evaluation of content-based image retrieval systems. In *Int. Conf. on Pattern Recognition*, Cambridge, UK, Aug. 2004. In press.
4. C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3(3/4):231–262, July 1994.
5. Q. Iqbal and J. Aggarwal. Using structure in content-based image retrieval. In *Int. Conf. Signal and Image Processing*, pp. 129–133, Nassau, Bahamas, Oct. 1999.
6. D. Keysers, C. Gollan, and H. Ney. Classification of medical images using non-linear distortion models. In *Bildverarbeitung für die Medizin*, pp. 366–370, Berlin, Germany, Mar. 2004.
7. T. Lehmann, M. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. Wein. The IRMA project – A state of the art report on content-based image retrieval in medical applications. In *Korea-Germany Workshop on Advanced Medical Image*, pp. 161–171, Seoul, Korea, Oct. 2003.
8. H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun. Performance Evaluation in Content-based Image Retrieval: Overview and Proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.
9. N.R. Pal and S.K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, Nov. 1993.
10. C. Palm, D. Keysers, T. Lehmann, and K. Spitzer. Gabor filtering of complex hue/saturation images for color texture classification. In *Int. Conf. on Computer Vision*, volume 2, pp. 45–49, Atlantic City, NJ, Feb. 2000.
11. R. Paredes, J. Perez-Cortes, A. Juan, and E. Vidal. Local representations and a direct voting scheme for face recognition. In *Workshop on Pattern Recognition in Information Systems*, pp. 71–79, Setúbal, Portugal, July 2001.
12. M. Park, J.S. Jin, and L.S. Wilson. Fast content-based image retrieval using quasi-Gabor filter and reduction of image feature. In *Southwest Symposium on Image Analysis and Interpretation*, pp. 178–182, Santa Fe, NM, Apr. 2002.
13. J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Int. Conf. on Computer Vision*, volume 2, pp. 1165–1173, Corfu, Greece, Sept. 1999.
14. B.S. Reddy and B. Chatterji. An FFT-based technique for translation, rotation and scale invariant image registration. *IEEE Trans. Image Proc.*, 5(8):1266–1271, Aug 1996
15. S. Siggelkow. *Feature Histograms for Content-Based Image Retrieval.* Ph.D. thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany, 2002.
16. D.M. Squire, W. Müller, H. Müller, and J. Raki. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In *Scandinavian Conference on Image Analysis*, pp. 143–149, Kangerlussuaq, Greenland, June 1999.
17. H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. Systems, Man, and Cybernetics*, 8(6):460–472, June 1978.
18. J.Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947–963, Sept. 2001.