FIRE – Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation

Thomas Deselaers, Daniel Keysers, Hermann Ney
Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University – D-52056 Aachen, Germany
{*deselaers, keysers, ney*}*@cs.rwth-aachen.de*

Abstract

In this paper we present *FIRE*, a content-based image retrieval system and the methods we used in the ImageCLEF 2004 evaluation. In *FIRE*, different features are available to represent images. This diversity of available features allows the user to adapt the system to task specific characteristics. A weighted combination of these features admits very flexible query formulations and helps in processing specific queries. For the ImageCLEF 2004 evaluation, we used content-based methods only and the experimental results compare favorably well with other systems that make use of the textual information in addition to the images.

1 Introduction

Content-based image retrieval is an area of active research in the field of pattern analysis and image processing. The need for content-based techniques becomes obvious when considering the enormous amounts of digital images produced day by day e.g. by digital cameras or digital imaging methods in medicine. The alternative of annotating large amounts of images manually is a very time consuming task. Another very important aspect is that images can contain information that cannot be expressed precisely in textual annotation [17]. Thus, even the most complete annotation is useless if it does not contain the details that might be of importance to the actual users. The only way to solve these problems is to use fully automatic, content-based methods.

Several content-based image retrieval systems have been proposed. One of the first systems was the QBIC system [5]. Other popular research systems are BlobWorld [1], VIPER/GIFT [18], SIMBA [16], and SIMPLIcity [20].

In this work we present *FIRE*, a content-based image retrieval system and the methods we used in the ImageCLEF 2004 evaluation. *FIRE* is easily extendable, offers a wide repertory of available features and distance functions and these varieties allow for assessing the performance of different features for different tasks. *FIRE* is freely available under the terms of the GNU General Public License[1].

2 Retrieval techniques

In content-based image retrieval, images are searched by their appearance and not by textual annotations. Thus, images have to be represented by features and these features are compared to search for images similar to a given query image. In *FIRE*, each image is represented by a set of features. To find images similar to a given query image, the features from the images in the database are compared to those of the query image.

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

2.1 Query by Example

Query by example means that the system is given a query image Q and the goal is to find images from the database which are similar to the given query image. In FIRE, images are represented by features and compared using feature-specific distance measures. These distances are combined in a weighted sum:

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

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 Q and X, 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 re-normalization. The K database images with lowest $D(Q, X)$ are returned.

2.2 Relevance Feedback

Relevance Feedback is a widely used technique [13] that allows for good user interaction and easy query refinements. After a query has been processed, the user is presented a reasonably large set of results. From these, the user can select some images as relevant results Q^+ and some images as irrelevant results Q^- and requery the system with these sets. To process this query, we calculate scores $S(Q, X) = e^{-\gamma D(Q,X)}$ with $\gamma = 1.0$ for the images and combine these into one score

$$S(Q^+, Q^-, X) = \sum_{q \in Q^+} S(q, X) + \sum_{q \in Q^-} (1 - S(q, X)).$$

The set of the K images with the highest scores is returned. The interface used for for relevance feedback is shown in Figure 1.

2.3 Query Expansion

A frequent method for enhancing the query results is *query expansion*. In FIRE, query expansion is implemented as an "automatic relevance feedback" [13]. The user specifies a number of images G that he expects to be relevant after the first query. Then a query is processed in two steps: First the query is evaluated and the first G images are returned. These G images are automatically used as set of relevant images Q^+ to requery the database and the K best matches are returned.

3 Features and Associated Distance Measures

This section gives a short description of each of the features used in the FIRE image retrieval system for the ImageCLEF 2004 evaluation. Table 1 gives an overview of the features and associated distance measures.

3.1 Appearance-based Image Features

The most straight-forward approach is to directly use the pixel values of the images 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 [9, 10, 11].

In this work, we used 32×32 and $32 \times X$(keeping the aspect ration) versions of the images. The 32×32 images are compared using Euclidean distance and the $32 \times X$ images are compared using image distortion model distance (IDM) [9].

IDM is a zero-order image comparison measure that allows local pixel displacements. Local dependencies in the displacement grid are neglected. We consider a deformation grid x_{11}^{IJ}, y_{11}^{IJ} explaining an $I \times J$ image $A = \{a_{ij}\}, i = 1, \ldots, I, j = 1, \ldots, J$ with an $X \times Y$ image $B =$

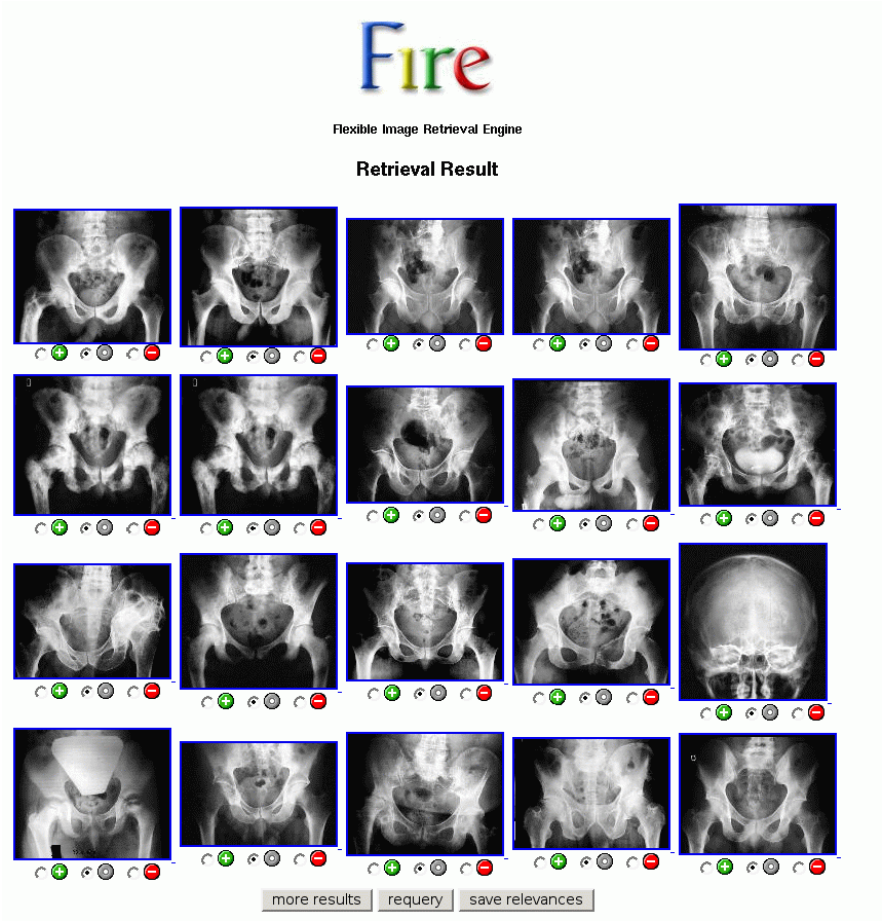

Figure 1: Interface for relevance feedback. The user is presented with the best matches from the database (top left is the query image) and can select for each image whether it is relevant, irrelevant or neutral.

$\{b_{xy}\}, x = 1, \ldots, X, y = 1, \ldots, Y$. Given this deformation grid, the distance between the aligned images is computed as

$$C(A, B, (x_{11}^{IJ}, y_{11}^{IJ})) = \sum_{i,j} ||a_{ij} - b_{x_{ij}, y_{ij}}||^2.$$

and the IDM distance is calculated as

$$D(A, B) = \min_{x_{11}^{IJ}, y_{11}^{IJ}} \left\{ C(A, B, (x_{11}^{IJ}, y_{11}^{IJ})) \right\}$$

taking into account a global warp range limiting the displacement range for the pixels. Instead of using only single pixels, here local context of 3×3 pixels of the horizontal and vertical Sobel derivatives of the images are used [9].

3.2 Color Histograms

Color histograms are widely used in image retrieval [2, 5, 15, 17]. Color histograms are one of the most basic approaches and to show performance improvements, image retrieval systems often are compared to a system using only color histograms. The color space is partitioned and for each partition the pixels with a color within its range are counted, resulting in a representation of the relative frequencies of the occurring colors. In accordance with [15], we use the Jeffrey divergence to compare histograms.

3.3 Invariant Feature Histograms

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 translation, rotation, and scaling. In this work, invariant feature histograms as presented in [16] 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 [15]. Previous experiments have shown that the characteristics of invariant feature histograms and color histograms are very similar and that invariant feature histograms often outperform color histograms [4]. Thus, in this work color histograms are not used.

3.4 Tamura Features

In [19] 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 [15]. In the QBIC system [5] histograms of these features are used as well.

3.5 Global Texture Descriptor

In [2] a texture feature consisting of several parts is described: *Fractal dimension* measures the roughness or the crinkliness of a surface. In this work the fractal dimension is calculated using the reticular cell counting method [7]. *Coarseness* characterizes the grain size of an image. Here it is calculated depending on the variance of the image. *Entropy* is used as a measure of unorderedness or information content in an image. The *Spatial gray-level difference statistics* (SGLD) describes the brightness relationship of pixels within neighborhoods. It is also known as co-occurrence matrix analysis [8]. . The *Circular Moran autocorrelation function* measures the roughness of the texture. For the calculation a set of autocorrelation functions is used [6].

3.6 Region-based Features

Another approach to representing images is based on finding image regions which roughly correspond to objects or parts of objects in the images. To this purpose, the image is segmented into regions. The task of segmentation has been thoroughly studied [14], 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, 20]. We use the approach presented in [20] to compare region descriptions of images.

3.7 Color/Gray Binary Feature

Since the databases contain both color images and gray value images, an obvious feature is whether the image is a color or gray valued image. This can be extracted easily by examining a reasonably large amount of pixels in the image. If all of these pixels are gray valued, the image is considered to be a gray valued images, otherwise it is considered to be a color image. This feature can easily be tested for equality.

4 Submissions to the ImageCLEF 2004 Evaluation

The ImageCLEF 2004 evaluation covered 3 tasks: 1. *Bilingual ad-hoc task* using the St. Andrews database of historic photographs, 2. *Medical Retrieval Task* using the Casimage database of medical images, and 3. *Interactive Retrieval task* using the St. Andrews database.

Table 1: Features extracted for the ImageCLEF 2004 evaluation and their associated distance measures.

number	feature	associated distance measure
0	32×32 down scaled version of the image	Euclidean
1	$32 \times X$ down scaled version of the image (keeping aspect ratio)	IDM
2	global texture descriptor	Euclidean
3	Tamura texture histogram	Jeffrey divergence
4	invariant feature histogram with monomial kernel	Jeffrey divergence
5	invariant feature histogram with relational kernel	Jeffrey divergence
6	binary feature: color/gray	equal/not equal

We participated in the bilingual ad-hoc task and the medical retrieval task. First a set of features was extracted from each of the images from both databases and the given query images. Table 1 gives an overview of the features extracted from the databases and the distance measures used to compare these features. The features extracted were chosen based on previous experiments with other databases [3, 4].

4.1 Medical Retrieval Task

The *Medical Retrieval Task* consisted of 26 query images for which similar images had to be retrieved from the Casimage database, a database of 8725 medical images from various medical domains. Along with the images a set of 2078 text documents describing the medical cases is available, which were not used in the system, however. Each of the images belongs to one of the cases, thus several images may belong to one case.

In ImageCLEF 2004 it was possible to submit results to the medical retrieval task under different conditions:

- only visual retrieval
- query expansion textual/visual
- manual feedback from the first 20 results images visual
- manual feedback from the first 20 results images visual/textual

We submitted results to the first three categories using visual information only. We did not make use of the textual data at all.

4.1.1 Fully Automatic Queries / Only visual retrieval

Fully Automatic Query means that the system is given the query image and has to return a list of the most similar images without any further user interaction.

To this task we submitted 3 runs differing in the feature weightings used. The precise feature weightings are given in Table 2. The table clearly shows that the parameters optimized for this task outperformed the other parameters and thus that optimizing the feature weightings in image retrieval for a given task improves the results. The feature weightings were chosen on the following basis:

- Use all available features equally weighted. This run can be seen as a baseline and is labelled with the run-tag `i6-111111`.
- Use the features in the combination that produces the best results on the IRMA database [12], labelled `i6-020500`.
- Use the features in a combination which was optimized towards the given task. See Section 4.1.4 on how we optimized the parameters towards this task. This run is labelled with the run-tag `i6-025501`.

Three example queries are given in Figure 2.

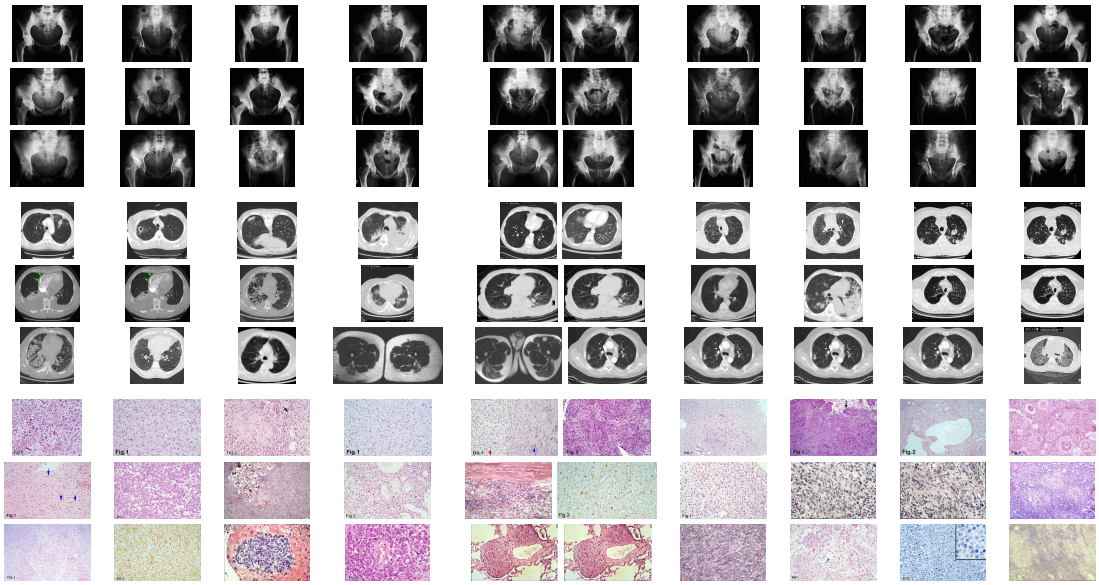

Figure 2: Three example queries with results from the fully automatic medical retrieval task.

Table 2: Different feature weightings and the mean average precision (MAP) from the ImageCLEF 2004 evaluation used for the medical retrieval task for the fully automatic runs.

	feature number							
run-tag	0	1	2	3	4	5	6	MAP
i6-111111	1	1	1	1	1	1	1	0.2857
i6-020500	0	5	0	2	0	0	0	0.2665
i6-025501	5	5	0	2	1	0	0	0.3407

4.1.2 Fully Automatic Queries with Query Expansion

This task was similar to the *fully automatic task*. The system was given the query image only and could perform the query in two steps, but without any user interaction. This method is described in Section 2.3:

1. normal query
2. query expansion, i.e. use the query image and its first nearest neighbor to requery the database.

We decided to use this method after we observed that for most query images the best match is a relevant one. In our oppinion, this method slightly enhanced the retrieval result, but the results are worse than the single-pass runs in two of three cases in the ImageCLEF 2004 evaluation. In Table 3 the results for these runs are given in comparison to the fully automatic runs without query expansion described in Section 4.1.1. For these experiments we used the same three settings for the fully automatic runs with and without query expansion. The fact that the results deteriorate (against our expectation) might be explained by missing medical relevance of the first query result. Another reason might be that we only looked into the first 20 to 30 results, but for the evaluation the first 1000 results were assessed.

4.1.3 Queries with Relevance Feedback

In the runs described in the following, *relevance feedback* was used. The system was queried with the given query image and a user was presented the 20 most similar images from the database. Then the user marked one or more of the images presented (including the query image) as relevant,

Table 3: Results from the ImageCLEF evaluation for the experiments with query expansion in comparison to the fully automatic runs.

run-tag	fully automatic	with query expansion
i6(qe)-020500	0.2665	0.3115
i6(qe)-025501	0.3407	0.3323
i6(qe)-111111	0.2857	0.2495

Table 4: Feature weighting used for the experiments with relevance feedback in the medical retrieval task.

	feature number							
run-tag	0	1	2	3	4	5	6	MAP
`i6-rfb1`	10	0	0	2	1	0	0	0.3437

irrelevant or neutral. The sets of relevant and irrelevant images were then used to requery the system as described in Section 2.2. Although in some scenarios several steps of relevance feedback might be useful, here only one step of query refinement was used.

As user interaction was involved here, a fast system was desirable. To allow for faster retrieval, the image distortion model was not used for the comparison of images. The feature weighting used is given in Table 4.

The mean average precision of 0.3437 reached here is slightly better than in the best of the fully automatic runs (0.3407).

4.1.4 Manual selection

To find a good set of parameters for this task, we performed some manual experiments. To be able to compare different parameter sets, we manually created relevance estimates for some of the images. These relevance estimates were submitted as "human visual system (of a computer scientist)". These experiments were carried out as follows:

1. Start with an initial feature weighting.
2. Query the database with all query images using this weighting.
3. Present the first 30 results for each query image to the user.
4. The user marks **all** images as either relevant or irrelevant. The system calculates the number of relevant results in total.
5. Slightly change the weighting and go back to 2.

We performed experiments to assess the quality of particular features, i.e. we used only one feature at a time (cf. Table 5). With this information in mind we started to combine different features. First we tried to use all features with identical weight at the same time and the setting which proved best on the IRMA task. Then we modified these settings to improve the results. In this way we could approximately assess the quality of the results for different settings. We tried 11 different settings in total. The complete results from these experiments are given in Table 6.

The mean average precision for this run is very low (0.279) because not enough images were viewed and assessed, and thus the number of returned images was far too small. In average only 53 results were returned with a minimum number of 6 images and a maximum of 142 images for the 26 queries.

Table 5: The subjective performance of particular features on the medical retrieval task.

feature number	0	1	2	3	4	5	6
precision of the first 30 images	0.55	0.44	0.31	0.54	0.40	0.36	0.03

Table 6: Effect of various feature combination on precision.

feature number							precision of the
0	1	2	3	4	5	6	first 30 results
1	1	1	1	1	1	1	0.60
0	5	0	2	0	0	0	0.65
0	5	0	2	2	0	0	0.61
0	10	0	2	2	0	0	0.63
0	5	0	2	0	2	0	0.59
10	0	0	2	2	0	0	0.65
0	10	0	2	0.5	0	0	0.63
5	0	0	2	0	0	0	0.65
0	10	0	2	1	0	0	0.65
5	5	0	2	1	0	0	0.67
10	0	0	2	0.5	0	0	0.65

Table 7: Different feature weightings used for the bilingual retrieval task for the fully automatic runs and the run with relevance feedback.

	feature number							
run-tag	0	1	2	3	4	5	6	MAP
`i6-111111`	1	1	1	1	1	1	0	0.0859
`i6-010012`	0	0	0	1	2	1	0	0.0773
`i6-010101`	0	1	0	1	1	0	0	0.0859
`i6-rfb1`	0	0	0	1	1	0	0	0.0839

4.2 Bilingual Retrieval Task

The *Bilingual Retrieval Task* consisted of 25 queries given as a short textual description in several languages, a slightly longer textual description in English, and an example image fitting the query. In our system we only used the 25 example images to query the database. The database is the *St. Andrews Image Collection* consisting of approximately 30 000 images.

In the experiments described in the following, the provided example images were used to query the database. Unfortunately, no other group participated in this track.

4.2.1 Fully Automatic Queries

Here, the example images given were used to query the database. Different feature weightings were used:

1. equal weight for each feature (run-tag `i6-111111`)
2. two weightings which have proven to work well for general purpose photographs [2] (run-tags `i6-010012` and `i6-010101`).

The exact weightings are given in Table 7 together with the results from the ImageCLEF 2004 evaluation.

A look at the query topics clearly showed that pure content-based image retrieval would not be able to deliver satisfactory results as queries like "Portrait pictures of church ministers by Thomas Rodger" are not processible by image content only (church ministers do not differ in their appearance from any other person, and it is usually not possible to see from an image who made it). The mean average precision values clearly show that visual information alone is not sufficient to obtain good results, although the results from queries are visually quite promising as shown in Figure 3. Due to the fact that this task was quite futile we did not focus on this task.

Figure 3: Query results for the bilingual retrieval task for three different queries using only visual information.

4.2.2 Queries with Relevance Feedback

Using the feature weighting given in Table 7, i.e. column `i6-rfb`, we submitted one run using relevance feedback for this task. No improvement can be seen: A mean average precision of 0.0839 was measured. This is even worse than the best of the fully automatic runs.

5 Summary of the Evaluation

In this section, our results are compared to the results of other groups in the ImageCLEF 2004 evaluation. For the medical retrieval task, the results of the evaluation (cf. Table 8) show that the methods presented here compare favorably well with the other systems. There are three better systems, however the differences are very small and it is not yet clear to which extend the other systems used the textual information in addition to the images. For the bilingual retrieval task, the comparison with the other systems seems to show that the textual information is very important. Note that all results presented in this paper were obtained using visual information only. Furthermore, the results in the medical retrieval task show that suitable selection and weighting of the features used improves the results strongly. The optimization here is not to be seen as "training on the testing data" as only a few different settings were compared.

For the future it is planned to extend the FIRE system to be able to use textual features in the retrieval process.

Table 8: Exemplary results (mean average precision, MAP) from the ImageCLEF 2004 evaluation for the fully automatic runs in the medical retrieval task.

run-tag	UBMedImTxt01	kids_run2	ic_cl04_base	i6-025501	i6-qe0255010	GE_4g_4d_vis	...	mi_combine1	...	enidlrun
MAP	0.35	0.35	0.35	0.34	0.33	0.32	...	0.27	...	0.18

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 *International Conference on Visual Information Systems*, Springer Verlag, Amsterdam, The Netherlands, pages 509–516, June 1999.

[2] T. Deselaers. Features for Image Retrieval. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, Germany, December 2003.

[3] T. Deselaers, D. Keysers, and H. Ney. Classification Error Rate for Quantitative Evaluation of Content-based Image Retrieval Systems. In *International Conference on Pattern Recognition*, Cambridge, UK, August 2004. In press.

[4] T. Deselaers, D. Keysers, and H. Ney. Features for Image Retrieval – A Quantitative Comparison. In *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, LNCS, Tübingen, Germany, September 2004. In press.

[5] 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.

[6] Z. Q. Gu, C. N. Duncan, E. Renshaw, M. A. Mugglestone, C. F. N. Cowan, and P. M. Grant. Comparison of Techniques for Measuring Cloud Texture in Remotely Sensed Satellite Meteorological Image Data. *Radar and Signal Processing*, 136(5):236–248, October 1989.

[7] P. Haberäcker. *Praxis der Digitalen Bildverarbeitung und Mustererkennung*. Carl Hanser Verlag, München, Wien, 1995.

[8] R. M. Haralick, B. Shanmugam, and I. Dinstein. Texture Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, November 1973.

[9] D. Keysers, C. Gollan, and H. Ney. Classification of Medical Images using Non-linear Distortion Models. In *Bildverarbeitung für die Medizin*, Berlin, Germany, pages 366–370, March 2004.

[10] D. Keysers, C. Gollan, and H. Ney. Local Context in Non-linear Deformation Models for Handwritten Character Recognition. In *International Conference on Pattern Recognition*, Cambridge, UK, August 2004. In press.

[11] D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in Statistical Pattern Recognition using Tangent Vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):269–274, February 2004.

[12] 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 Joint Workshop on Advanced Medical Image Processing*, Seoul, Korea, pages 161–171, October 2003.

[13] H. Müller, W. Müller, S. Marchand-Maillet, and D. McG Squire Strategies for positive and negative Relevance Feedback in Image Retrieval In *International Conference on Pattern Recognition*, Barcelona, Spain, 1:1043–1046, September 2000.

[14] N. R. Pal and S. K. Pal. A Review on Image Segmentation Techniques. *Pattern Recognition*, 26(9):1277–1294, November 1993.

[15] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. In *International Conference on Computer Vision*, volume 2, Corfu, Greece, pages 1165–1173, September 1999.

[16] S. Siggelkow. *Feature Histograms for Content-Based Image Retrieval*. PhD thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany, 2002.

[17] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval: The End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[18] 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*, Kangerlussuaq, Greenland, pages 143–149, June 1999.

[19] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transaction on Systems, Man, and Cybernetcs*, 8(6):460–472, June 1978.

[20] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, September 2001.