FIRE – Flexible Image Retrieval Engine: ImageCLEF 2004 Evaluation

Thomas Deselaers, Daniel Keysers, and Hermann Ney

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

Abstract. We describe *FIRE*, a content-based image retrieval system, and the methods we used within this system in the ImageCLEF 2004 evaluation. In *FIRE*, various features are available to represent images. The diversity of available features allows the user to adapt the system to the task at hand. A weighted combination of features admits flexible query formulations and helps with processing specific queries. For the ImageCLEF 2004 evaluation, we used the image content alone and obtained the best result in the category "only visual features, fully automatic retrieval" in the medical retrieval task. Additionally, the results compare favorably to other systems, even if they 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. Furthermore, a very important aspect is that images can contain information that no words can convey [1]. Thus, even the most complete annotation is useless if it does not contain the details that might be of importance to the actual users in their context. The only way to solve these problems is to use fully automatic, content-based methods.

In this work we describe *FIRE*, a content-based image retrieval system and the methods we used within this system in the ImageCLEF 2004 evaluation. *FIRE* is easily extensible, offers a wide repertoire of features and distance functions. 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].

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

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 688–698, 2005.

2 Retrieval Techniques

In content-based image retrieval, images are searched by their appearance and not by textual annotations. Thus, the appearance of the images is encoded 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 the features of the query image using an appropriate distance measure d.

Given a query image Q and the goal to find images from the database which are similar to the given query image, we calculate a score $S(Q, X)$ for each image $X \in \mathcal{B}$ from the database $\mathcal{B}$:

$$S(Q, X) = \exp\left(-\gamma \sum_{m=1}^{M} w_m \cdot d_m(Q_m, X_m) \right) . \qquad (1)$$

Here, 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, $\gamma = 1$. 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 highest $S(Q, X)$ are returned. When *Relevance Feedback* [2] is used, that is, a user selects a set of relevant images Q^+ and a set of irrelevant images Q^- to refine a query, we calculate the scores for each of the

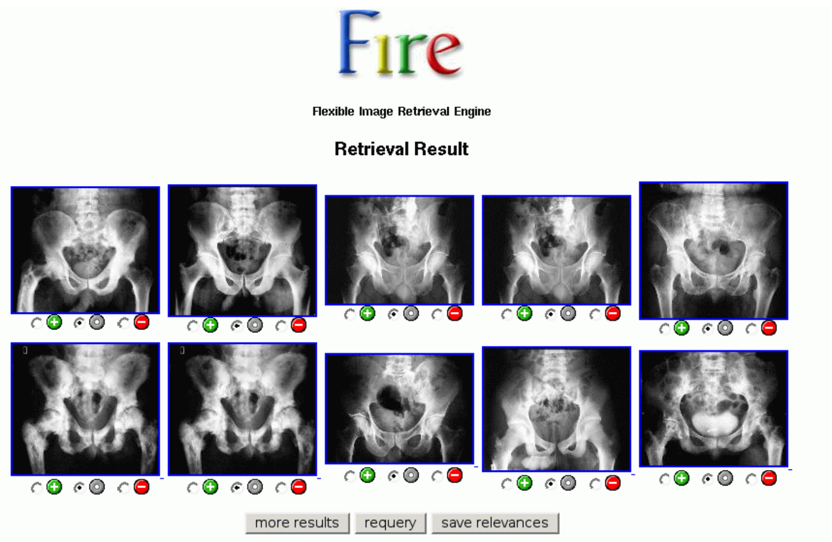

Fig. 1. Interface for relevance feedback. The user is presented the best matches from the database (top left is the query image) and can select for each image whether it is relevant, irrelevant, or neutral

images from the sets of relevant and irrelevant images and combine these into one score

$$S(Q^+, Q^-, X) = \sum_{q \in Q^+} S(q, X) + \sum_{q \in Q^-} (1 - S(q, X)). \tag{2}$$

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

A frequent method for enhancing the query results is *query expansion*. In FIRE, query expansion is implemented as "automatic relevance feedback" [2]. 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 the set of relevant images Q^+ to requery the database and the K best matches of this query 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.

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

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

3.1 Color Histograms

Color histograms are widely used in image retrieval [3, 4, 1]. They are one of the most basic approaches and to show performance improvements, image retrieval systems are often compared to a system using only color histograms. The color space is partitioned and for each bin 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 [5], we use the Jeffrey divergence to compare histograms.

3.2 Appearance-Based Image Features (1,2)

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 [6, 7, 8].

Here, we use 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) [6].

3.3 Global Texture Descriptor (3)

In [3] a texture feature consisting of several parts is described: *Fractal dimension* measures the roughness or the crinkliness of a surface. Here, the fractal dimension is calculated using the reticular cell counting method [9]. *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 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 [10]. The *Circular Moran autocorrelation function* measures the roughness of the texture [11].

3.4 Tamura Features (4)

In [12] the authors propose six texture features corresponding to human visual perception: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity*, and *roughness*. From experiments that tested 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 [3] and compare these histograms using the Jeffrey divergence [5]. In the QBIC system [4] histograms of these features are used as well.

3.5 Invariant Feature Histograms (5,6)

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 by Siggelkow [13] 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 [5]. Previous experiments have shown that the characteristics of invariant feature histograms and color histograms are very similar but that invariant feature histograms often outperform color histograms [14]. Thus, in this work color histograms are not used.

3.6 Color/Gray Binary Feature (7)

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 image, 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 [15] 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.

We participated in the bilingual ad-hoc task and the medical retrieval task. For the experiments, 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 [16, 14].

4.1 Medical Retrieval Task

In the *Medical Retrieval Task* [15] we submitted results in three different categories: 1. fully automatic visual retrieval, 2. query expansion using visual data only, 3. manual relevance feedback using only visual information. We would like to emphasize that no textual data was used at all during the experiments.

4.2 Fully Automatic Visual Retrieval

Fully Automatic Retrieval means that the system is given the query image and must 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 along with the obtained mean average precision and 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 [17], labelled `i6-020500`.
- Use the features in a combination which was optimized towards the given task. See Section 4.5 on how we optimized the parameters towards this task. This run is labelled with the run-tag `i6-025501`.

Table 2 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. Two example queries are given in Figure 2.

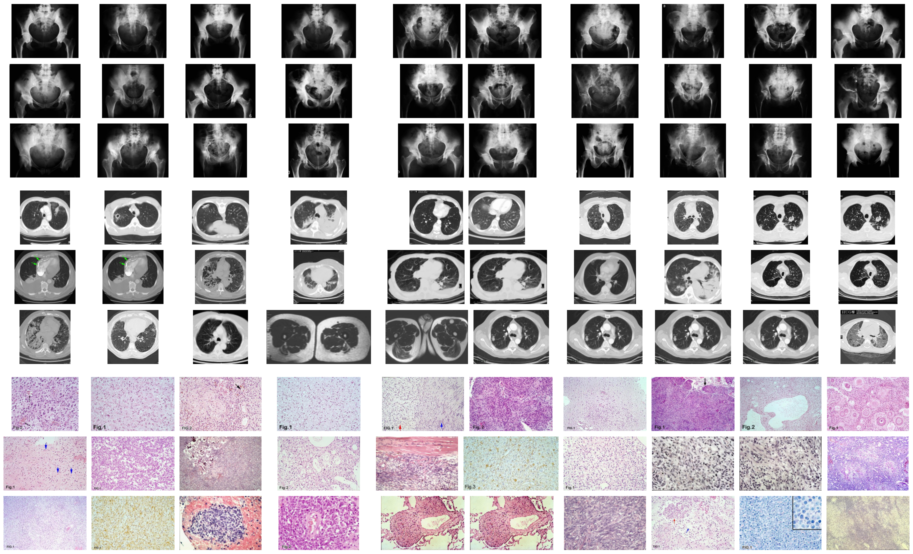

Fig. 2. Two example queries with results from the fully automatic medical retrieval task

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 with and without query expansion (QE) and for the run with relevance feedback (RF)

	weight for feature number							MAP		
run-tag	1	2	3	4	5	6	7	w/o QE	w/ QE	w/ RF
i6-111111	1	1	1	1	1	1	1	0.318	0.278	-
i6-020500	0	5	0	2	0	0	0	0.308	0.354	-
i6-025501	5	5	0	2	1	0	0	**0.386**	0.374	-
i6-rfb1	10	0	0	2	1	0	0	-	-	0.394

4.3 Fully Automatic Queries with Query Expansion

This task is similar to the *fully automatic task*. The system is given the query image only and can perform the query in two steps, but without any user interaction as described in Section 2:

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 was a relevant one. In our opinion, this method slightly enhanced the retrieval result visually, but the results are worse than the single-pass runs in two of three cases in the ImageCLEF 2004 evaluation. In Table 2 the results for these runs are given in comparison to the fully automatic runs without query expansion. For these experiments we used the same three settings as for the fully automatic runs with and without query expansion. The fact that the results deteriorate (against our expectation) can be explained by the missing medical relevance of the first query result. Another reason might be that we looked only at the first 30 results, but for the evaluation the first 1000 results were assessed.

4.4 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, irrelevant or neutral. The sets of relevant and irrelevant images were then used to requery the system as described in Section 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, 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 2.

The mean average precision of 0.394 reached here is slightly better than in the best of the fully automatic runs (0.386).

4.5 Manual Selection

To find a good set of parameters for this task, we manually compared some parameter combinations. Therefore, we manually created relevance estimates for some of the images. 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. The user marks **all** images as either relevant or irrelevant. The system calculates the number of relevant images in total.
4. Slightly change the weighting and go back to 2.

As starting point, we performed experiments to assess the quality of particular features, i.e. we used only one feature at a time (cf. Table 3(a)). With this information in mind we combined 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 and manually chose the best one for submission. The complete results for these experiments are given in Table 3(b).

Table 3. a) The subjective performance of particular features on the medical retrieval task measured as precision of the first 30 results, b) Effect of various feature combinations on the precision for the medical retrieval task

a)

	precision of the
feature no	first 30 results
1	0.55
2	0.44
3	0.31
4	0.54
5	0.40
6	0.36
7	0.03

b)

weight for feature no							precision of the
1	2	3	4	5	6	7	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 4. Different feature weightings used for the bilingual retrieval task for the fully automatic runs and the run with relevance feedback

	weight for feature number							
run-tag	1	2	3	4	5	6	7	MAP
`i6-111111`	1	1	1	1	1	1	0	0.086
`i6-010012`	0	0	0	1	2	1	0	0.077
`i6-010101`	0	1	0	1	1	0	0	0.086
`i6-rfb1`	0	0	0	1	1	0	0	0.084

4.6 Bilingual Retrieval Task

For the *Bilingual Retrieval Task* [15] we used only the 25 example images to query the database. That is, we used only the visual information provided and not the textual information at hand.

4.7 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 had been proven to work well for general purpose photographs [3] (run-tags `i6-010012` and `i6-010101`).

The exact weightings are given in Table 4 along with the results from the ImageCLEF 2004 evaluation.

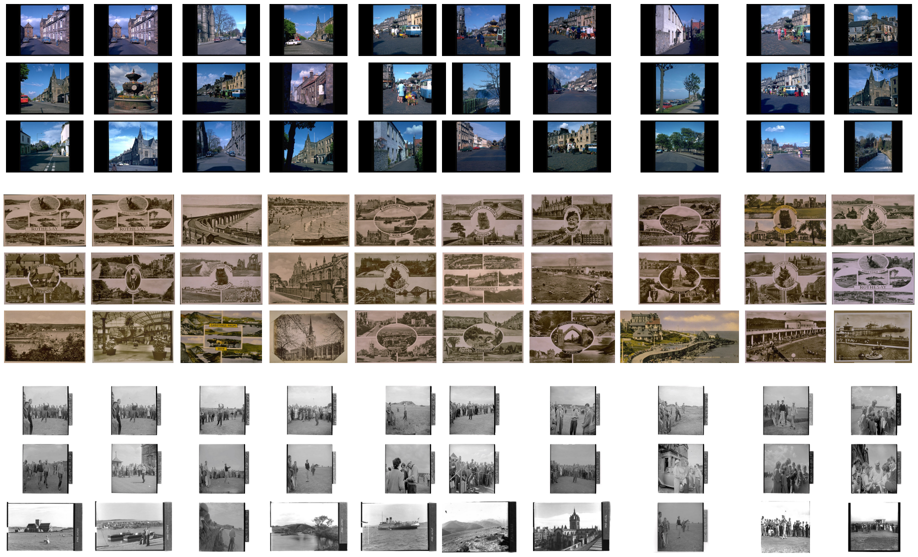

Fig. 3. Query results for the bilingual retrieval task for two different queries using only visual information

A look at the query topics clearly showed that pure content-based image retrieval would not be able to deliver satisfactory results because queries such as "Portrait pictures of church ministers by Thomas Rodger" are not processible by image content only (church ministers do not differ significantly 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. As this task was quite futile we did not focus on this task.

4.8 Queries with Relevance Feedback

Using the feature weighting given in Table 4, i.e. column `i6-rfb`, we submitted one run using relevance feedback for this task. No improvement can be observed: A mean average precision of 0.084 was measured. This is even worse than the best of the fully automatic runs.

5 Conclusion

In this section, the results are analyzed and compared to the results of other groups in the ImageCLEF 2004 evaluation.

Table 5 shows for each of the tasks the MAP of our best run compared to the best run in this task and to the average MAP in this task and to the best result for

Table 5. Comparison of our results to the results of other groups [15]

	best result from this work			MAP		
Task	MAP	rank	# participants	best result for this task	average	best result using text
Med: Auto (visual only)	0.386	1	23	0.386	0.273	0.390
Med: RF (visual only)	0.394	2	6	0.430	0.336	0.476
AdHoc: Visual	0.086	2	5	0.092	0.081	0.587
AdHoc: Vis,RF	0.084	1	1	0.084	0.084	0.587

the database used. It can clearly be seen that our system compares favorably well with the other systems, e.g. in the task of fully automatic retrieval using visual information only, we obtain the best result and this result is only slightly less precise than the best fully automatic result where textual information was used. The addition of manual feedback did not improve our results further in contrast to the results of other groups. It can clearly be seen that suitable selection and weighting of the features used improves the results strongly. The optimization here is not critical as only a few settings were compared. Comparing the results using only visual information to those using text and user feedback it can be seen that slight improvements are possible in the medical retrieval task and that textual information is indispensable for the ad-hoc retrieval task.

For the future, several things will be improved in the FIRE system. On the one hand, it can be seen that textual information strongly improves the results for some tasks. Thus we are planning to integrate a textual information retrieval component in our content-based image retrieval system. On the other hand, even using visual information only, the results can be strongly improved by using relevance feedback. As our results are only slightly improved using relevance feedback, we are planning to improve the relevance feedback techniques in our system.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval: The end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
2. Müller, H., Müller, W., Marchand-Maillet, S., Squire, D.M.: Strategies for positive and negative relevance feedback in image retrieval. In: International Conference on Pattern Recognition. Volume 1 of Computer Vision and Image Analysis, Barcelona, Spain (2000) 1043–1046
3. Deselaers, T.: Features for image retrieval. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, Germany (2003)
4. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. Journal of Intelligent Information Systems **3** (1994) 231–262

5. Puzicha, J., Rubner, Y., Tomasi, C., Buhmann, J.: Empirical evaluation of dissimilarity measures for color and texture. In: International Conference on Computer Vision. Volume 2, Corfu, Greece (1999) 1165–1173
6. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non-linear distortion models. In: Bildverarbeitung für die Medizin, Berlin, Germany (2004) 366–370
7. Keysers, D., Gollan, C., Ney, H.: Local context in non-linear deformation models for handwritten character recognition. In: International Conference on Pattern Recognition. Volume 4, Cambridge, UK (2004) 511–514
8. Keysers, D., Macherey, W., Ney, H., Dahmen, J.: Adaptation in statistical pattern recognition using tangent vectors. IEEE Trans. on Pattern Analysis and Machine Intelligence **26** (2004) 269–274
9. Haberäcker, P.: Praxis der Digitalen Bildverarbeitung und Mustererkennung. Carl Hanser Verlag, München, Wien (1995)
10. Haralick, R.M., Shanmugam, B., Dinstein, I.: Texture features for image classification. IEEE Trans. on Systems, Man, and Cybernetics **3** (1973) 610–621
11. Gu, Z.Q., Duncan, C.N., Renshaw, E., Mugglestone, M.A., Cowan, C.F.N., Grant, P.M.: Comparison of techniques for measuring cloud texture in remotely sensed satellite meteorological image data. Radar and Signal Proc. **136** (1989) 236–248
12. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. on Systems, Man, and Cybernetics **8** (1978) 460–472
13. Siggelkow, S.: Feature Histograms for Content-Based Image Retrieval. PhD thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany (2002)
14. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval – a quantitative comparison. In: DAGM 2004, Pattern Recognition, 26th DAGM Symposium. Number 3175 in LNCS, Tübingen, Germany (2004) 228–236
15. Clough, P., Müller, H., Sanderson, M.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004). LNCS (2005) in press
16. Deselaers, T., Keysers, D., Ney, H.: Classification error rate for quantitative evaluation of content-based image retrieval systems. In: International Conference on Pattern Recognition. Volume 2, Cambridge, UK (2004) 505–508
17. Lehmann, T., Güld, M., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.: 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 (2003) 161–171