FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval

Thomas Deselaers, Tobias Weyand, Daniel Keysers,
Wolfgang Macherey, and Hermann Ney

Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, Germany
surname@informatik.rwth-aachen.de

Abstract. In this paper the methods we used in the 2005 ImageCLEF content-based image retrieval evaluation are described. For the medical retrieval task, we combined several low-level image features with textual information retrieval. Combining these two information sources, clear improvements over the use of one of these sources alone are possible.
Additionally we participated in the automatic annotation task, where our content-based image retrieval system, FIRE, was used as well as a second subimage based method for object classification. The results we achieved are very convincing. Our submissions ranked first and the third in the automatic annotation task out of a total of 44 submissions from 12 groups.

1 Introduction

It is known that in content-based image retrieval (CBIR) benchmarking of systems is a major problem. ImageCLEF, as part of the Cross Language Evaluation Forum, is a major step towards creating standard benchmarking tasks and setting up competitions to compare content-based image retrieval systems. One of the main conclusions that can be drawn from the 2004 and 2005 ImageCLEF image retrieval evaluations is that textual information and user feedback, if available, can greatly improve the results. This is especially true if the queries are of semantic nature, as it is intrinsically difficult to solve them using visual information alone.

Particularly in real life applications, for example, in medicine, where textual information is available and pictures alone are not sufficient to describe a medical case, any available information should be used. If, for example, the query image is a microscopy of a bacteria culture, a standard image retrieval system will easily find other pictures of bacteria cultures, but it will hardly be able to distinguish between different kinds of bacteria. With additional textual query information like "Coli bacteria", the query, and thus the result is more precise.

Since we obtained the best score in the category "visual information only, no user interaction" in the 2004 ImageCLEF evaluation, it was an interesting

challenge to extend our FIRE system[1] towards the use of textual information. Other groups had already proposed approaches combining textual information retrieval and content-based image retrieval, e.g. [1–3].

In this paper, we describe the techniques we used for the 2005 ImageCLEF evaluation. In particular, we describe how textual information retrieval and content-based image retrieval were combined.

The 2005 ImageCLEF involved four tasks: automatic annotation, medical image retrieval, bilingual information retrieval, and interactive retrieval. We participated in the automatic annotation task and the medical image retrieval task. Our approach to the medical retrieval task is described in Section 2, the two approaches to the automatic annotation task are described in Section 3.

2 Medical Retrieval Task

For the medical retrieval task in the 2005 ImageCLEF Image Retrieval Evaluation, 25 queries were given. Each query was defined by a short textual query description and one to three example images. One query contained a negative example image, all other example images were positive. A more detailed description of the task and an overview of the results can be found in [4]. In the following we describe the setup of the FIRE-system, for the medical retrieval task.

2.1 Decision Rule

Given a set of positive example images Q^+ and a (possibly empty) set of negative example images Q^- a score $S(Q^+, Q^-, X)$ is calculated for each image X from the database:

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

where $S(q, X) = e^{-D(q,X)}$ is the score of database image X with respect to query q. $D(q, X)$ is a weighted sum of distances calculated according to

$$D(q, X) := \sum_{m=1}^{M} w_m \cdot d_m(q_m, X_m). \tag{2}$$

Here, q_m and X_m are the m^{th} feature of the query image q and the database image 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. Given a query (Q^+, Q^-), the images are ranked according to descending score and the K images X with the highest scores $S(Q^+, Q^-, X)$ are returned by the retriever.

Due to the lack of suitable training data, the weightings w_m were chosen heuristically based on experiences from earlier experiments with other data.

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

2.2 Textual Information Retrieval

To incorporate textual information in FIRE, we decided to use an existing textual information retrieval engine [5]. The text retrieval engine implements a variant of the Smart-2 retrieval metric, which is based on the well-known *term frequency inverse document frequency* (tf-idf) metric. The textual information is preprocessed by removing function words that are considered to be of no importance to the actual retrieval process (so called *stopping*). The stop word list used comprises 319 of the most frequently occurring function words in the English language. After all texts are stopped, the remaining words are reduced to their stems using Porter's stemming algorithm [6]. The stemmed words form the index terms that are used to index the text documents provided in addition to the image data. In our implementation of the Smart-2 retrieval metric we use the following definition of the inverse document frequency:

$$\mathrm{idf}(t) := \log \left\lfloor \frac{K}{n(t)} \right\rfloor \tag{3}$$

Here, t denotes an index term, and K is the number of text documents. Due to the floor operation in Eq. (3) a term weighting will be zero if it occurs in more than half of the documents. According to [7], each index term t in a document $\mathbf{d}$ is associated with a weighting $g(t, \mathbf{d})$ which depends on the ratio of the logarithm of the term frequency $n(t, \mathbf{d})$ to the logarithm of the average term frequency $\overline{n}(\mathbf{d})$

$$g(t, \mathbf{d}) := \begin{cases} \big[1 + \log n(t, \mathbf{d})\big] \big/ \big[1 + \log \overline{n}(\mathbf{d})\big] & \text{if } t \in \mathbf{d} \\ 0 & \text{if } t \notin \mathbf{d} \end{cases} \tag{4}$$

with $\log 0 := 0$ and

$$\overline{n}(\mathbf{d}) = \frac{\sum_{t \in \mathcal{T}} n(t, \mathbf{d})}{\sum_{t \in \mathcal{T} : n(t, \mathbf{d}) > 0} 1} \tag{5}$$

The logarithms in Eq. (4) prevent documents with high term frequencies from dominating those with low term frequencies. In order to obtain the final term weightings, $g(t, \mathbf{d})$ is divided by a linear combination of a pivot element c and the number of singletons $n_1(\mathbf{d})$ in document $\mathbf{d}$:

$$\omega(t, \mathbf{d}) := \frac{g(t, \mathbf{d})}{(1 - \lambda) \cdot c + \lambda \cdot n_1(\mathbf{d})} \tag{6}$$

with $\lambda = 0.2$ and

$$c = \frac{1}{K} \sum_{k=1}^{K} n_1(\mathbf{d}_k) \quad \text{and} \quad n_1(\mathbf{d}) := \sum_{t \in \mathcal{T} : n(t, \mathbf{d}) = 1} 1 \tag{7}$$

Unlike tf-idf, only query terms are weighted with the inverse document frequency $\mathrm{idf}(t)$:

$$\omega(t, \mathbf{q}) = \big[1 + \log n(t, \mathbf{q})\big] \cdot \mathrm{idf}(t) \tag{8}$$

The SMART-2 retrieval function is then defined as the product over the document and query specific index term weightings:

$$f(\mathbf{q}, \mathbf{d}) = \sum_{t \in \mathcal{T}} \omega(t, \mathbf{q}) \cdot \omega(t, \mathbf{d}) \tag{9}$$

To use the textual information for image retrieval, each image has to be attached to at least one (possibly empty) text document. These text documents are used in the image retrieval process described above. To determine the distance $d_{\text{text}}(q_m, X_m)$ between a query image q with query text q_m and a database image X with attached text X_m, first the textual information retriever is queried using the query text. Then, the textual information retriever returns the list of all documents from the database that it considers relevant. These documents are ranked by the retrieval status value (RSV) R which is high for documents similar to the query and low for dissimilar documents. The distance $d(q_m, X_m)$ is then calculated as

$$d_{\text{text}}(q_m, X_m) = \begin{cases} R_{\max} - R_X & \text{if } X \in \text{list of relevant documents} \\ \rho & \text{otherwise} \end{cases} \tag{10}$$

where $R_{\max}$ is the maximum of all returned RSVs, R_X is the RSV for image X, q_m and X_m are the query text and the text attached to image X, respectively, and ρ is a sufficiently large constant, chosen so as to make sure that images whose texts do not appear in the list of relevant objects have high distances. Note that the case where $\rho = R_{\max}$ corresponds to assigning an RSV of 0 to all non-relevant texts. The resulting distances $d_{\text{text}}(q_m, X_m)$ are used in the retrieval process described in the previous section.

2.3 Image Features

In the following we describe the visual features we used in the evaluation. These features are extracted offline from all database images.

Appearance-based Image Features. The most straightforward 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 [8–10].

In this work, we have used 32×32 versions of the images. These have been compared using Euclidean distance. It has been observed that for classification and retrieval of medical radiographs, this method serves as a reasonable baseline.

Color Histograms. Color histograms are widely used in image retrieval [11–13], and constitute one of the most basic approaches. To demonstrate performance improvements, image retrieval systems are often 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 [12], we use the Jeffrey divergence to compare histograms.

Tamura Features. Tamura et al. propose six texture features corresponding to human visual perception: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity*, and *roughness* [14]. From experiments testing the significance of these features with respect to human perception, it has been concluded that the first three features are the most important. Thus in our experiments we use coarseness, contrast, and directionality to create a histogram describing the texture [11] and compare these histograms using the Jeffrey divergence [12].

Global Texture Descriptor. In [11] a texture feature consisting of several parts is described: *Fractal dimension* measures the roughness or the crinkliness of a surface [15]. *Coarseness* characterizes the grain size of an image. *Entropy* is used as a measure of disorderedness 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 [16]. The *Circular Moran autocorrelation function* measures the roughness of the texture. For the calculation a set of autocorrelation functions is used [17].

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 [18] 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 [12]. 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 [19].

3 Automatic Annotation Task

In the automatic annotation task, the objective was to classify 1,000 images into one of 57 classes using 9,000 training images. We participated in the automatic annotation task using two different methods. Method A is identical to the approach we have chosen for the medical retrieval task, except that here no textual information was available, and that we used appearance-based image features and Tamura Texture Features only, as we know from earlier experiments that these features perform well on medical radiographs [20].

Method B is a general object recognition method using histograms of image patches and discriminative training of log-linear models [21, 22].

Table 1. Error rates [%] using different features on the IRMA 10,000 validation data.

feature	distance	dev corpus	test corpus
32×32 thumbnails	Euclidean	25.3	36.8
X×32 thumbnails	IDM	13.0	12.6
Tamura texture histogram	JSD	33.1	46.0

The parameters of method A were optimized using 1,000 images from the 9,000 training images as a development set and the remaining 8,000 images for training. The parameters of method B were chosen as they work best on the Caltech database [23, 24, 22].

A more detailed description of the task and a detailed analysis of the results can be found in [4].

3.1 Method A: Image Distortion Model

Method A uses our CBIR system FIRE and a subset of the above described features consisting of thumbnails of the images of the sizes 32×32 and $X \times 32$ and Tamura texture histograms. Error rates using these features alone are given in Table 1.

Some experiments with different weightings of Tamura features and thumbnails on our development corpus have shown that using the image distortion model alone outperforms the combinations. In particular, the combination of image distortion model (weighted 5) and Tamura texture features (weighted 2) is interesting, as this performed best in previous experiments on smaller versions of the IRMA database [20]. In our experiments, this combination yielded an error rate of 13.5% on the development corpus. Using the image distortion model alone yielded an error rate of 13.0% for the development data. Thus, we decided to use the image distortion model for our submission.

3.2 Method B: Object Recognition with Subimages and Discriminative Training

For method B an object recognition and classification approach using histograms of image patches and maximum entropy training to classify the 1,000 test images was used [21, 22].

To reduce the time and memory requirements for the clustering process, we used only 4,000 images for estimating the Gaussian mixture model. Nonetheless, we created the histograms for all training images and we used all histograms for the discriminative training of the log-linear model.

The model submitted used multi-scale features where the first PCA component was discarded to account for brightness changes and 4096-dimensional histograms. This combination was reported to work best on the Caltech database [23] and in the PASCAL Visual Object Classes Challenge [25]. The model achieved an error rate of 13.9% and thus is slightly better than the model by Raphaël Marée who follows a similar approach [26].

4 Experiments and Results

In the following the exact setups of the submitted runs for the automatic annotation task and the medical retrieval task are described and the results are discussed. Furthermore, we discuss our methods, point to errors we made, and present results of experiments that were conducted after the evaluation taking into account the lessons learned.

4.1 Automatic Annotation Task

Our submission using model A ranked first in the automatic annotation task. The submission following the object recognition approach ranked third. In total, 44 runs were submitted by 12 groups. The second rank was obtained by the IRMA group[2] using an approach similar to our model A and the fourth rank was obtained by the University of Liège, Belgium using an approach with image patches and boosted decision trees. A clear improvement over the baseline result of 36.8% error rate can be observed. This baseline result is obtained by a nearest neighbor classifier using 32x32 thumbnails of the images and Euclidean distance.

4.2 Medical Retrieval Task

For the medical retrieval task, we used the features described in Section 2.3 with different weightings in combination with text features. In total, we submitted 10 runs which are briefly described here.

Runs using textual information only: We submitted two fully automatic runs, where only textual information was used. These runs were labelled `En` and `EnDeFr`. In `En` only the English texts were used, for `EnDeFr` the English, the German, and the French texts were used and combined with equal weighting.

Runs using visual information only: We submitted three fully automatic runs, where only visual information was used. The runs `5000215`, `0010003`, and `1010111` only differ in the weighting of the image features. The exact weightings can be seen in Table 2. The run labelled `5000215` uses exactly the same setting as our submission to the 2004 ImageCLEF evaluation which had the best score from all 23 submissions in the category "visual features only, no user interaction". From the bad score of 0.06, it can be seen that this year's tasks differ significantly from the task of the previous year.

Runs using visual and textual information: We submitted three fully automatic runs and two runs with relevance feedback where textual and visual information was used. For the run `i6-3010210111`, the features were combined

[2] `http://www.irma-project.org`

Table 2. Overview of the submitted runs for the medical retrieval task and their setup. For each run, the feature weightings and the achieved MAP with badly chosen ρ and with properly chosen ρ is given. (* as feature weight means that for all features marked with * the distance was calculated and the minimum among those was chosen, - means not used, + means that relevance feedback was used).

	textual information only		visual information only			visual and textual information				+relevance feedback	
run	`En`	`EnDeFr-min`	`1010111`	`5000215`	`0010003`	`3010210111`	`3(3030333)-min(111)`	`3(1010111)-min(111)`	-	`vistex-rfb1`	`vistex-rfb2`
X×32 image features	-	-	1	5	3	3	9	3	1	1	1
32×32 image features	-	-	1	0	0	0	0	3	1	1	1
color histograms	-	-	1	0	1	1	9	3	1	1	1
tamura features	-	-	1	2	0	2	9	3	1	1	1
invariant feat. histo.	-	-	1	1	0	1	9	3	1	1	1
English text	1	*	-	-	-	1	*	*	2	*	*
German text	0	*	-	-	-	1	*	*	0	*	*
French text	0	*	-	-	-	1	*	*	0	*	*
relevance feedback	-	-	-	-	-	-	-	-	-	+	+
score w/ wrong ρ	0.21	0.05	0.07	0.06	0.05	0.07	0.07	0.06	-	0.09	0.08
score w/ properly chosen ρ	0.21	0.15				0.22		0.20	0.25		

in exactly the way described above. For the runs `i6-3(1010111-min(111))` and `i6-3(3030333)-min(111)` before combining text- and visual features, the minimum distance of all three text distances was first taken for each image. This was done to better account for images that have texts in one language only.

The runs `i6-vistex-rfb1` and `i6-vistex-rfb2` used relevance feedback from the first 20 results of the automatic run `i6-3(1010111-min(111))` and differ only in the user feedback. In both cases the feedback was given by a computer scientist familiar with the FIRE system but with little background in medicine. Furthermore, the textual information was not available for the user feedback. Thus, the feedback is based on visual information only.

Table 2 shows an overview of all runs we submitted for the medical retrieval task. Unfortunately, we were unable to test our combination of textual- and visual information retrieval in advance of the competition , which led to a very unlucky choice of ρ in Eq. (10). As a result, any combination with textual information retrieval was adversely affected. The results obtained after the evaluation, where ρ was chosen properly, are significantly improved (Table 2). In particular, using English textual information retrieval only, we could reach a MAP of 0.25 which would have achieved third ranking in the 2005 ImageCLEF evaluation in the category "textual and visual information, no relevance feedback".

5 Conclusion and Outlook

We presented the methods we used in the 2005 ImageCLEF CBIR evaluation. Participating in the automatic annotation task, we obtained the first and third rank. In the medical image retrieval task our results were not satisfying due to improper parameterization. Results with correct settings are presented in this work and results are significantly improved. In particular, the result obtained would have been ranked 3rd in the medical retrieval task in the category "fully automatic runs using textual and visual information".

References

1. Müller, H., Geissbühler, A.: How to Visually Retrieve Images From the St. Andrews Collection Using GIFT. In: Multilingual Information Access for Text, Speech and Images. Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum. CLEF 2004. Volume 3491 of LNCS., Bath, UK, Springer (2004) 633–642
2. Lin, W.C., Chang, Y.C., Chen, H.H.: From Text to Image: Generating Visual Query for Image Retrieval. In: Multilingual Information Access for Text, Speech and Images. Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum. CLEF 2004. Volume 3491 of LNCS., Bath, UK, Springer (2004) 664–675
3. Alvarez, C., Oumohmed, A.I., Mignotte, M., Nie, J.Y.: Toward Cross-Language and Cross-Media Image Retrieval. In: Multilingual Information Access for Text, Speech and Images. Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum. CLEF 2004. Volume 3491 of LNCS., Bath, UK, Springer (2004) 676–687
4. Clough, P., Mueller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., Hersh, W.: The CLEF 2005 Cross-Language Image Retrieval Track. In Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 - to appear.
5. Macherey, W., Viechtbauer, H.J., Ney, H.: Probabilistic Aspects in Spoken Document Retrieval. EURASIP Journal on Applied Signal Processing Special Issue on "Unstructured Information Management from Multimedia Data Sources"(2) (2003) 1–12
6. Porter, M.F.: An Algorithm for Suffix Stripping, Morgan Kaufmann, 1980, San Francisco, CA.
7. Choi, J., Hindle, D., Hirschberg, J., Magrin-Changnolleau, I., Nakatani, C., Pereira, F., Singhal, A., Whittaker, S.: An Overview of the At&T Spoken Document Retrieval. In: Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Va, USA (1998) 182–188
8. 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
9. 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
10. Keysers, D., Macherey, W., Ney, H., Dahmen, J.: Adaptation in Statistical Pattern Recognition Using Tangent Vectors. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(2) (2004) 269–274

11. Deselaers, T.: Features for Image Retrieval. Diploma thesis, Lehrstuhl für Informatik VI, RWTH Aachen University, Aachen, Germany (2003)
12. 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
13. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval: The End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
14. Tamura, H., Mori, S., Yamawaki, T.: Textural Features Corresponding to Visual Perception. IEEE Transaction on Systems, Man, and Cybernetics **8**(6) (1978) 460–472
15. Haberäcker, P.: Praxis der Digitalen Bildverarbeitung und Mustererkennung. Carl Hanser Verlag, München, Wien (1995)
16. Haralick, R.M., Shanmugam, B., Dinstein, I.: Texture Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics **3**(6) (1973) 610–621
17. 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 Processing **136**(5) (1989) 236–248
18. Siggelkow, S.: Feature Histograms for Content-Based Image Retrieval. PhD thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany (2002)
19. 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
20. Lehmann, T.M., Güld, M.O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., Wein, B.: Automatic Categorization of Medical Images for Content-Based Retrieval and Data Mining. Computerized Medical Imaging and Graphics **29** (2005) in press
21. Deselaers, T., Keysers, D., Ney, H.: Discriminative Training for Object Recognition Using Image Patches. In: CVPR 05. Volume 2., San Diego, CA (2005) 157–162
22. Deselaers, T., Keysers, D., Ney, H.: Improving a Discriminative Approach to Object Recognition Using Image Patches. In: DAGM 2005. LNCS, Vienna, Austria (2005) 326–333
23. Fergus, R., Perona, P., Zissermann, A.: Object Class Recognition by Unsupervised Scale-Invariant Learning. In: Conference on Computer Vision and Pattern Recognition, Blacksburg, VG (2003) 264–271
24. Dreuw, P., Keysers, D., Deselaers, T., Ney, H.: Gesture Recognition Using Image Comparison Methods. In: GW 2005, 6th Int. Workshop on Gesture in Human-Computer Interaction and Simulation, Vannes, France (2005)
25. Everingham, M., Gool, L.V., Williams, C., Zisserman, A.: Pascal Visual Object Classes Challenge Results. Technical report, University of Oxford, Oxford, UK (2005)
26. Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Random Subwindows for Robust Image Classification. In Schmid, C., Soatto, S., Tomasi, C., eds.: IEEE Conference on Computer Vision and Pattern Recognition. Volume 1., San Diego, CA, USA, IEEE (2005) 34–40