Content-Based Retrieval of Medical Images by Combining Global Features

Mark O Güld, Christian Thies, Benedikt Fischer, and Thomas M. Lehmann

Department of Medical Informatics, RWTH Aachen, Aachen, Germany*
{mgueld, cthies, bfischer}@mi.rwth-aachen.de, lehmann@computer.org

Abstract. A combination of several classifiers using global features for the content description of medical images is proposed. Beside well known texture histogram features, downscaled representations of the original images are used, which preserve spatial information and utilize distance measures which are robust with regard to common variations in radiation dose, translation, and local deformation. These features were evaluated for the annotation task and the retrieval task in ImageCLEF 2005 without using additional textual information or query refinement mechanisms. For the annotation task, a categorization rate of 86.7% was obtained, which ranks second among all submissions. When applied in the retrieval task, the image content descriptors yielded a mean average precision (MAP) of 0.0751, which is rank 14 of 28 submitted runs. As the image deformation model is not fit for interactive retrieval tasks, two mechanisms are evaluated with regard to the trade-off between loss of accuracy and speed increase: hierarchical filtering and prototype selection.

1 Introduction

ImageCLEFmed 2005 [1] consists of several challenges for content-based retrieval [2] on medical images. A newly introduced annotation task poses a classification problem of mapping 1,000 query images with no additional textual information into one of 57 pre-defined categories. The mapping is to be learned based on a ground truth of 9,000 categorized reference images. For the retrieval task, the reference set was expanded to over 50,000 images, compared to 8,725 medical images in 2004. These tasks reflect the real-life constraints of content-based image retrieval in medical applications, as image corpora are large, heterogeneous and additional textual information about an image, especially its content, is not always reliable due to improper configuration of the imaging devices, ambiguous naming schemes, and both inter- and intra-observer variability.

2 The Annotation Task

The annotation task consists of 9,000 images grouped into 57 categories and 1,000 images to be automatically categorized. It should be noted that the category definition is based solely on the aspects of

* This work is part of the IRMA project, which is funded by the German Research Foundation, grant Le 1108/4.

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 702–711, 2006.

1. imaging modality, i.e. identification of the imaging device (three different device types)
2. imaging direction, i.e. relative position of the body part towards the imaging device
3. anatomy of the body part examined, and
4. biological system, which encodes certain contrast agents and a coarse description of the diagnostic motivation for the imaging.

Thus, the category definition does not incorporate any diagnosis information, e.g. the detection of pathologies or their quantitative analysis.

2.1 Image Features and Their Comparison

Based on earlier experiments conducted on a similar image set, three types of features and similarity measures were employed [3].

Tamura et al. proposed a set of texture features to capture global texture properties of an image, namely coarseness, contrast, and directionality [4]. This information is stored in a three-dimensional histogram, which is quantized into $M = 6 \times 8 \times 8 = 384$ bins. To capture this texture information at a comparable scale, the extraction is performed on downscaled images of size 256×256, ignoring their aspect ratio. The query image $q(x, y)$ and the reference image $r(x, y)$ are compared by applying Jensen-Shannon divergence [5] to their histograms $H(q)$ and $H(r)$:

$$d_{\mathrm{JSD}}(q, r) = \frac{1}{2} \sum_{m=1}^{M} \left[H_m(q) \log \frac{2H_m(q)}{H_m(q) + H_m(r)} + H_m(r) \log \frac{2H_m(r)}{H_m(q) + H_m(r)} \right] \tag{1}$$

To retain spatial information about the image content, downscaled representations of the original images are used and the accompanying distance measures work directly on intensity values. It is therefore possible to incorporate a priori knowledge into the distance measure by modelling typical variability in the image data, which does not alter the category that the image belongs to. The cross-correlation function (CCF) from signal processing determines the maximum correlation between two 2D image representations, each one of size $h \times h$:

$$s_{\mathrm{CCF}}(q, r) = \max_{|m|,|n| \le d} \left\{ \frac{\sum_{x=1}^{h} \sum_{y=1}^{h} (r(x-m, y-n) - \overline{r}) \cdot (q(x, y) - \overline{q})}{\sqrt{\sum_{x=1}^{h} \sum_{y=1}^{h} (r(x-m, y-n) - \overline{r})^2}} \cdot \frac{1}{\sqrt{\sum_{x=1}^{h} \sum_{y=1}^{h} (q(x, y) - \overline{q})^2}} \right\} \tag{2}$$

Here, $q(x, y)$ and $r(x, y)$ refer to intensity values at a pixel position on the scaled representations of q and r, respectively. Note that s_{CCF} is a similarity measure and the values lie between 0 and 1. CCF includes robustness regarding two very common variabilites among the images: translation, which is explicitly tested

within the search window of size $2d+1$, and radiation dose, which is normalized by subtracting the average intensity values $\overline{q}$ and $\overline{r}$. For the experiments, downscaling to 32×32 pixels and a translation window of size $d = 4$ was used, i.e. translation can vary from -4 to $+4$ pixels in both the x- and the y-direction.

While s_{CCF} considers only global displacements, i.e. translations of entire images, and variability in radiation dose, it is suggested to model local deformations of medical images caused by pathologies, implants and normal inter-patient variability. This can be done with an image distortion model (IDM) [6]:

$$d_{\mathrm{IDM}}(q,r) = \sum_{x=1}^{X}\sum_{y=1}^{Y} \min_{|x'|,|y'|\leq W_1} \left\{ \sum_{|x''|,|y''|\leq W_2} || \, r(x+x'+x'', y+y'+y'') - q(x+x'', y+y'')||_2 \right\} \quad (3)$$

Again, $q(x,y)$ and $r(x,y)$ refer to intensity values of the scaled representations. Note that each pixel of q must be mapped on some pixel in r, whereas not all pixels of r need to be the target of a mapping. Two parameters steer d_{IDM}: W_1 defines the size of the neighborhood when searching for a corresponding pixel. To prevent a totally unordered pixel mapping, it is useful to incorporate the local neighborhood as context when evaluating a correspondence hypothesis. The size of the context information is controlled by W_2. For the experiments, $W_1 = 2$, i.e. a 5×5 pixel search window, and $W_2 = 1$, i.e. a 3×3 context patch are used. Also, better results are obtained if the gradient images are used instead of the original images, because the correspondence search will then focus on contrast and be robust to global intensity differences due to radiation dose. It should be noted that this distance measure is computationally expensive as each window size influences the computation time in a quadratic manner. The images were scaled to a fixed height of 32 pixels and the original aspect ratio was preserved.

2.2 Nearest-Neighbor Classifier

To obtain a decision $q \mapsto c \in \{1 \ldots C\}$ for a query image q, a nearest neighbor classifier evaluating k nearest neighbors according to a distance measure is used (k-NN). It simply votes for the category which accumulated the most votes among the k reference images closest to q. This classifier also allows visual feedback in interactive queries.

2.3 Classifier Combination

Prior experiments showed that the performance of the single classifiers can be improved significantly if their single decisions are combined [3]. This is especially true for classifiers which model different aspects of: the image content, such as the global texture properties with no spatial information and the scaled representations, which retain spatial information. The easiest way is a parallel combination scheme, since it can be performed as a post-processing step after the

single classifier stage [7]. Also, no assumptions are required for the application, whereas serial or sieve-like combinations require an explicit construction.

For comparability, the distance values (d_{Tamura}, d_{IDM}) are normalized at first over all distances $d(q, r_i), i = 1 \ldots N$ between sample q and each reference r_i:

$$d'(q, r_i) = \frac{d(q, r_i)}{\sum_{n=1}^{N} d(q, r_n)} \tag{4}$$

Afterwards, a new distance measure can be obtained by a weighted sum of distance measures d_1, d_2.

$$d_{\text{c}}(q, r) = \lambda \cdot d'_1(q, r) + (1 - \lambda) \cdot d'_2(q, r),\ \lambda \in [0; 1] \tag{5}$$

For a similarity measure s, $d(q, r) := 1 - s(q, r)$ is used and the normalization is performed afterwards. Thus, the parallel combination of the three classifiers results in

$$\begin{aligned} d_{\text{combined}}(q, r) = {} & \lambda_{\text{Tamura}} \cdot d'_{\text{Tamura}}(q, r) \\ & + \lambda_{\text{CCF}} \cdot d'_{\text{CCF}}(q, r)) \\ & + \lambda_{\text{IDM}} \cdot d'_{\text{IDM}}(q, r) \end{aligned} \tag{6}$$

with $\lambda_{\text{Tamura}}, \lambda_{\text{CCF}}, \lambda_{\text{IDM}} \geq 0$ and $\lambda_{\text{Tamura}} + \lambda_{\text{CCF}} + \lambda_{\text{IDM}} = 1$.

2.4 Training and Evaluation on the Reference Set

The combined classification process relies on three parameters: λ_{Tamura}, λ_{CCF} and k for the number of nearest neighbors to be evaluated (λ_{IDM} is linearly dependent). To obtain suitable values for the parameters, the reference set of 9,000 images was split at random into a static training set of 8,000 images and a static test set of 1,000 images. The best parameter values found for this configuration are then applied to the 1,000 query images. For practical reasons, the matrices $D_{\text{Tamura}} = (d_{\text{Tamura}}(q_i, r_j))_{ij}$, $S_{\text{CCF}} = (s_{\text{CCF}}(q_i, r_j))_{ij}$, and $D_{\text{IDM}} = (d_{\text{IDM}}(q_i, r_j))_{ij}$ are computed once. Afterwards, all combination experiments can be performed rather quickly by processing the matrices.

2.5 Use of Class Prototypes

Since the distance computations for the scaled representations are rather expensive, there is – in general – great interest for prototype selection which reduces the required computation time, storage space and might even improve the categorization rate by removing possible outliers in the reference set.

Prototype sets can be obtained in various ways [8]. For simplicity, only random prototype selection and KCentres for K=1 and a simplified variation of it were used. Based on the empirically optimized d_{combined}, a set of category prototypes $R_{\text{prototypes}} \subset R = \bigcup_{c=1\ldots C} R_c$, with R_c being the set of all references belonging to class c, is computed by using KCentres:

$$R_{\text{prototypes}} = \bigcup_{c=1\ldots C} \left\{ \arg \min_{r' \in R_c} \left\{ \sum_{r \in R_c} d_{\text{combined}}(r, r') \right\} \right\} \tag{7}$$

These elements $\{r'_c\}, c = 1..C$ yield the smallest sum of distances to all members of their respective category.

The prototypes are used to obtain a dissimilarity-space representation of the reference images and the unknown images [8]:

$$r \mapsto (d(r, r'_1), \ldots, d(r, r'_C))^{tr} \in I\!R^C \tag{8}$$

$$q \mapsto (d(q, r'_1), \ldots, d(q, r'_C))^{tr} \in I\!R^C \tag{9}$$

For the classification, two representations are compared using Euclidian distance.

3 The Retrieval Task

The retrieval task uses 50,024 images for reference and consists of 25 queries, which are given as a combination of text information and query images, with some queries specifying both positive and negative example images. While the image data for the annotation task only contains grayscale images from mostly x-ray modalities (plain radiography, fluoroscopy, and angiography), the image material in this task is much more heterogeneous: It also contains photographs, ultrasonic imaging and even scans of illustrations used for teaching. Note that the retrieval task demands a higher level of image understanding, since several of the 25 queries directly refer to the diagnosis of medical images, which is often based on local image details, e.g. bone fractures or the detection of emphysema in computed tomography (CT) images of the lungs.

3.1 Image Features and Their Comparison

The content representations described in the previous section only use grayscale information, i.e. color images are converted into grayscale by using color weighting recommended by ITU-R:

$$Y = \frac{6969 \cdot R + 23434 \cdot G + 2365 \cdot B}{32768} \tag{10}$$

In general, however, color is the single most important discriminate feature type on stock-house media and the image corpus used for the retrieval task contains many photographs, color scans of teaching material, and microscopic imaging. Therefore, a basic color feature was employed to compute mean, variance and third order moments for each color channel red R, green G, and blue B. This yields a nine-dimensional feature vector. Euclidean distance with equal weights for each color component is used to compute the distance between two vectors.

3.2 Summation Scheme for Queries Consisting of Multiple Images

Some of the queries do not consist of a single example image, but use several images as a query pool Q: positive and negative examples. For such queries, a simple summation scheme is used to obtain an overall distance:

$$d(Q,r) = \sum_{i=1}^{|Q|} w_i \cdot d'(q_i, r), Q = \bigcup_i \{(q_i, w_i)\}, w_i = \begin{cases} 1 : q_i \text{ positive ex.} \\ -1 : q_i \text{ negative ex.} \end{cases} \quad (11)$$

4 Results

All results were obtained non-interactively, i.e. without relevance feedback by a human user, and without using textual information for the retrieval task.

4.1 Annotation Task

Table 1 shows the categorization results obtained for the 1,000 unknown images using single classifiers. As IDM is very expensive (see running times below), a serial combination with a faster, but more inaccurate classifier as a filter was also tested. For this, Euclidian distance on 32×32 scaled representations was used and only the 500 closest references were passed to IDM. This cuts computation time down to 1/18, as the costs for the filtering step are negligible.

Table 1. Results for single classifiers

	Categorization rate in %	
Content representation	k=1	k=5
TAMURA texture histogram, Jensen-Shannon divergence	69.3	70.3
32×32, CCF (9×9 translation window)	81.6	82.6
X×32, IDM (gradients, 5×5 window, 3×3 context)	85.0	83.8
X×32, IDM (as above, 32x32 Euclid 500-NN as filter)	84.1	83.5

To obtain the optimal empirical weighting coefficients λ for the parallel combination, an exhaustive search would have been necessary. Instead, a more time-efficient two-step process was employed: First, a combination of the two spatial representations was considered. Afterwards, a combination of the combined spatial representations with the Tamura texture feature was investigated. Both runs tested values for λ increasing at a stepsize of 0.1. The results for these two steps on the *testing set*, i.e. 1,000 images of the 9,000 original reference images, are shown in Tab. 2. The resulting empirical parameter configuration is shown in boldface. When using this parameter set for the classification of the 1,000 images to be categorized, a categorization rate of 86.7% for the 1-NN is obtained.

Using the 57 prototypes obtained via (7) as a representation set, a dissimilarity-space representation for the reference set and the unknown set was computed. The dissimilarity representations were then compared using Euclidian distance. In addition, not only the elements with minimum sum of distances were used, but also the ones with the best $n, n = 2 \ldots 5$ elements per category. This yields 114, 171, 228, and 285 components for the representation vectors. For comparison, experiments were also done for a random pick of $1 \ldots 5$ elements per category,

Table 2. Results on the testing subset, combination of IDM, CCF (left), combination of IDM, CCF, and TAMURA texture (right)

Weights		Categorization rate in %		Weights			Categorization rate in %	
λ_{IDM}	λ_{CCF}	k=1	k=5	λ_{IDM}	λ_{CCF}	λ_{Tamura}	k=1	k=5
0.0	1.0	82.5	80.9	0.00	0.00	1.0	70.8	69.0
0.1	0.9	84.0	82.0	0.07	0.03	0.9	78.1	76.9
0.2	0.8	84.8	83.5	0.14	0.06	0.8	81.4	80.7
0.3	0.7	85.6	84.1	0.21	0.09	0.7	83.5	83.1
0.4	0.6	85.6	84.3	0.28	0.12	0.6	84.6	84.0
0.5	0.5	85.5	84.5	0.35	0.15	0.5	85.5	84.6
0.6	0.4	86.2	84.2	**0.42**	**0.18**	**0.4**	**86.7**	**85.2**
0.7	0.3	**86.6**	**84.5**	0.49	0.21	0.3	86.7	85.1
0.8	0.2	85.9	84.0	0.56	0.24	0.2	86.6	85.1
0.9	0.1	85.5	82.8	0.63	0.27	0.2	86.6	84.9
1.0	0.0	84.7	82.6	0.70	0.30	0.1	86.6	84.5

resulting in representation vectors of the same size. The results are shown in Fig. 1, yielding a best categorization rate of 75.4% when using 2 center prototypes per category from the IDM-based 5-NN as the representation set.

4.2 Retrieval Task

Since no ground truth for the automatic optimization of the parameters is available, only a short visual inspection was done and two runs were submitted. The results are listed in Tab. 3. The result quality is measured by mean average precision (MAP).

Table 3. Results for the retrieval task

λ_{IDM}	λ_{CCF}	λ_{Tamura}	λ_{RGB}	MAP
0.4	0.4	0.2	0.0	0.0659
0.36	0.36	0.18	0.1	0.0751

These results are ranked 19th and 14th among 28 submitted runs in the "visual only, automatic" category of this task, reaching half the MAP of the leading competitor in this category.

4.3 Running Times

Table 4 lists the computation times of the algorithms for the annotation task on a standard Pentium 4 PC running at 2.6 GHz. For the retrieval task, extraction times per image are identical and query time is 5.5 times greater as there are 50,000 references compared to 9,000.

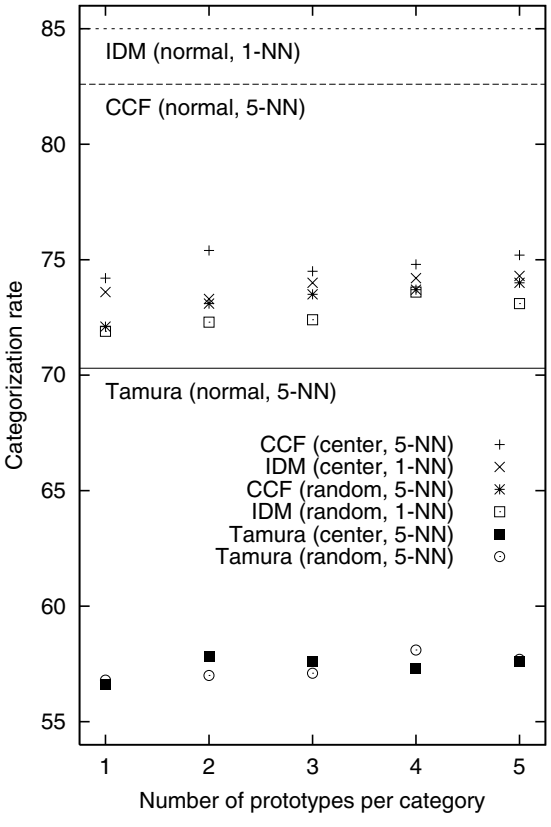

Fig. 1. Results for single classifiers using dissimilarity representation

5 Discussion

The results obtained for the annotation task verify the results obtained on a smaller corpus using leaving-one-out [3]. Note that the rather high weight λ_{Tamura} overemphasizes the role of the texture features in the experiments, as the actual improvement of the categorization rate is statistically insignificant

Table 4. Running times for the annotation task

Content Representation	Extraction [s] (per reference)	Query [s] (per sample)
TAMURA texture histogram, Jensen-Shannon divergence	5	$<<1$
32×32, CCF (9×9 translation window)	3	6
X×32, IDM (gradients, 5×5 window, 3×3 context)	3	190
X×32, IDM (as above, 32x32 Euclid 500-NN as filter)	6	9

for the 1-NN. It marginally improves the quality of the next nearest neighbors as seen in the results for the 5-NN, which produces slightly better results for interactive queries which list a set of nearest neighbors.

While results for the retrieval task were satisfactory in queries based on grayscale radiographs, other queries, especially from photograpy imaging, had rather poor results, partly due to very basic color feature that was employed. Furthermore, a detailed visual evaluation might have resulted in better tuning of the weighing parameters. This was dropped due to time constraints and it is also unrealistic for real-life applications. Therefore, the results can be considered as a baseline for fully automated retrieval algorithms without feedback mechanisms for parameter tuning. It should also be noted that several queries demand a high level of image content understanding, as they are aimed at diagnosis-related information, which is often derived from local details in the image (Tab. 5).

Table 5. Queries in the retrieval task which directly refer to diagnoses

Query	Semantic constraint
2	fracture of the femur
10	emphysema in lung CT
12	enlarged heart in PA chest radiograph
15	gross pathologies of myocardial infarction
16	osteoarthritis in hand
17	micro nodules in lung CT
18	tuberculosis in chest radiograph
19	Alzheimer's desease in microscopic pathologies
20	chronic myelogenous leukemia in microscopic pathologies
21	bone fracture(s) in radiograph
23	differentiate between malignant and benign melanoma
24	right middle lobe pneumonia

Concerning running times, texture features by TAMURA and CCF are fit for interactive use. By pre-filtering with a computationally inexpensive distance measure, computation time can be severly reduced without sacrificing too much accuracy. In the experiments, pre-filtering clearly outperformed dissimilarity space approaches for both random prototype selection and 1Centres. However, further evaluation of algorithms for prototype selection is necessary. The parallel combination of single classifiers proved very useful, as it improves the categorization results considerably and can also be performed as an easy post-processing step on the distance matrices.

The methods used in this work to describe the image content either preserve no spatial information at all (texture features by TAMURA) or capture it at very large scale, omitting local details important for diagnosis-relevant questions. Using only the image information, such queries cannot be processed with satisfactory quality of the results with a one-level approach. Refering to the concepts

described in [9], the methods employed in this paper work on the categorization layer of the content abstraction chain. For a better query completion, subsequent image abstraction steps are required.

References

1. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 Cross-Language Image Retrieval Track. Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer Science (to appear).
2. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380
3. Güld, M.O., Keysers, D., Deselaers, T., Leisten, M., Schubert, H., Ney, H., Lehmann, T.M.: Comparison of global features for categorization of medical images. Proceedings SPIE **5371** (2004) 211–222
4. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics **8**(6) (1978) 460–472
5. Puzicha, J., Rubner, Y., Tomasi, C., Buhmann, J.: Empirical evaluation of dissimilarity measures for color and texture. Proceedings International Conference on Computer Vision, 2 (1999) 1165–1173
6. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non-linear distortion models. Bildverarbeitung für die Medizin 2004, Springer-Verlag, Berlin (2004) 366–370
7. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1) (2000) 4–36
8. Pekalska, E., Duin, R.P.W., Paclik, P.: Prototype selection for dissimilarity-based classification. Pattern Recognition (to appear)
9. Lehmann, T.M., Güld, M.O., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.B.: Content-based image retrieval in medical applications. Methods of Information in Medicine **43**(4) (2004) 354–361