Combining global features within a nearest neighbor classifier for content-based retrieval of medical images

Mark O Güld, Christian Thies, Benedikt Fischer, and Thomas M Lehmann
Department of Medical Informatics, RWTH Aachen, Aachen, Germany
mgueld@mi.rwth-aachen.de

Abstract

A combination of several classifiers using global features for the content description of medical images is proposed. Beside two texture features, downscaled representations of the original images are used, which preserve spatial information and utilize distance measures which are robust regarding common variations in radiation dose, translation, and local deformation. No query refinement mechanisms are used. The single classifiers are used within a parallel combination scheme, with the optimization set being used to obtain the best weighing parameters. For the medical automatic annotation task, a categorization rate of 78.6% is obtained, which ranks 12th among 28 submissions. When applied in the medical retrieval task, this combination of classifiers yields a mean average precision (MAP) of 0.0172, which is rank 11 of 11 submitted runs for automatic, visual only systems.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image texture features, Deformation model, Classifier combination

1 Introduction

ImageCLEF 2006 [1] consists of several challenges for content-based retrieval on medical images. [1] A medical automatic annotation task poses a classification problem of mapping 1,000 query images with no additional textual information onto one of 116 pre-defined categories. The mapping is to be learned based on a ground truth of 9,000 categorized reference images. To optimize classifier settings, an additional set of 1,000 categorized images is available.

For the retrieval task, the reference set contains over 50,000 images. Here, 30 queries are given, which encompass 63 images. These tasks reflect the real-life constraints of content-based image retrieval in medical applications, as image corpora are large, heterogeneous and additional

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

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 10,000 reference images grouped into 116 categories and 1,000 images to be automatically categorized. The category definition is based solely on the aspects of

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, four types of features and similarity measures are employed [2].

Castelli et al. propose a combination texture features based on global fractal dimension, coarseness, gray-scale histogram entropy, spatial gray-level statistics and several circular Moran autocorrelation functions [3]. This results in 43 feature values per image. To compare a pair of these feature vectors, Mahalanobis distance with an estimated diagonal covariance matrix Σ is used:

$$d_{\text{Mahalanobis}}(q,r) = (q-r)^T \cdot \Sigma^{-1} \cdot (q-r) \overset{\text{simplified}}{=} \sum_{i=1}^{43} \frac{(q_i - r_i)^2}{\sigma_i^2} \tag{1}$$

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 $H(r)$:

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

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 $h \times h$:

$$s_{\text{CCF}}(q,r) = \max_{|m|,|n| \le d} \left\{ \frac{\sum_{x=1}^{h} \sum_{y=1}^{h} (r(x-m,y-n) - \overline{r}) \cdot (q(x,y) - \overline{q})}{\sqrt{\sum_{x=1}^{h} \sum_{y=1}^{h} (r(x-m,y-n) - \overline{r})^2 \sum_{x=1}^{h} \sum_{y=1}^{h} (q(x,y) - \overline{q})^2}} \right\} \tag{3}$$

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 $d = 4$ is 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]:

$$d_{\mathrm{IDM}}(q,r) = \sum_{x=1}^{X}\sum_{y=1}^{Y} \min_{|x'|,|y'|\leq W_1} \left\{ \sum_{|x''|,|y''|\leq W_2} ||r(x+x'+x'', y+y'+y'') - q(x+x'', y+y'')||_2 \right\} \quad (4)$$

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 are scaled to a fixed height of 32 pixels and the original aspect ratio is preserved.

In all, each image is represented by approximately 1024+1024+384+43 values, or roughly 3 KB memory space, as the scaled representations require one byte per intensity, while the histograms are stored using floating-point numbers.

2.2 Nearest-Neighbor Classifier

To obtain a decision $q \mapsto c \in \{1 \dots C\}$ 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 easy 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 [2]. 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 single classifier distance values $d(q, r_i), i = 1 \dots N$ are first normalized over all references $r_n, n = 1 \dots N$:

$$d'(q,r_i) \quad = \quad \frac{d(q,r_i)}{\sum_{n=1}^{N} d(q,r_n)} \quad (5)$$

For a similarity measure s, $d'(q, r) := 1 - s(q, r)$ is used and the normalization is performed afterwards. The new distance measure based on (normalized) distance measures $d'_m, m = 1 \dots M$

is obtained by weighted summation:

$$d_{\text{combined}}(q,r) = \sum_{m=1}^{M} \lambda_m \cdot d'_m(q,r), \lambda \in [0;1], \sum_{m=1}^{M} \lambda_m = 1 \quad (6)$$

2.4 Training and Evaluation on the Reference Set

The optimization (or development) set of 1,000 images is used to estimate the weights $\lambda_{\text{Castelli}}$, λ_{Tamura}, λ_{CCF}, and λ_{IDM}. The corresponding matrices $D_{\text{Castelli}} = (d_{\text{Mahalanobis}}(q_i, r_j))_{ij}$, $D_{\text{Tamura}} = (d_{\text{JSD}}(q_i, r_j))_{ij}$, $S_{\text{CCF}} = (s_{\text{CCF}}(q_i, r_j))_{ij}$, and $D_{\text{IDM}} = (d_{\text{IDM}}(q_i, r_j))_{ij}$ are only computed once for the single classifiers. Afterwards, all combination experiments can be performed efficiently by processing the matrices. The stepsize during the search for the best weight combination is 0.05. For comparison with the experiments from the ImageCLEF 2005 annotation task, a run with the weights used in [8] is also submitted.

3 The Retrieval Task

The retrieval task uses 50,024 images for reference and consists of 30 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. The retrieval task demands a higher level of image understanding, since several of the 30 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:

$$Y = \frac{6969 \cdot R + 23434 \cdot G + 2365 \cdot B}{32768} \quad (7)$$

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.

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:

$$d(Q,r) = \sum_{i=1}^{|Q|} w_i \cdot d'(q_i,r), Q = \bigcup_i \{(q_i, w_i)\}, w_i = \begin{cases} 1 : q_i \text{ positive example} \\ -1 : q_i \text{ negative example} \end{cases} \quad (8)$$

4 Results

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

Table 1: Categorization rates (in percent) for the medical automatic annotation task.

Content representation	k=1	k=5
CASTELLI texture features, Mahalanobis distance, diagonal Σ	42.8	45.1
TAMURA texture histogram, Jensen-Shannon divergence	55.6	55.1
32×32, CCF (9×9 translation window)	72.1	74.3
X×32, IDM (gradients, 5×5 window, 3×3 context)	77.0	76.6
ImageCLEF2005: w_{Tam}=0.4, w_{CCF}=0.12, w_{IDM}=0.48	78.3	78.0
Exhaustive search: $w_{Castelli}$=0.05, w_{Tamura}=0.25, w_{CCF}=0.25, w_{IDM}=0.45	78.5	78.6

4.1 Annotation Task

Table 1 shows the categorization rates obtained for the 1,000 unknown images using single classifiers and their combination, both for 1-NN and a 5-NN. The categorization rate of 78.6% ranks 12th among 28 submitted runs for this task. The weights used in the ImageCLEF 2005 medical automatic annotation task (10,000 images from 57 categories) yield a categorization rate of 78.3%.

4.2 Retrieval Task

Since no ground truth for the automatic optimization of the parameters is available, a run using the optimized weighing parameters from the annotation task is submitted. The run yields a mean average precision (MAP) of 0.0172 and is ranked 11th among 11 submitted runs in the "visual only, automatic" category of this task. For comparison, the best run submitted for this category yields 0.0753 MAP.

5 Discussion

The weighing coefficients used for the medical automatic annotation task of ImageCLEF 2005 are also suitable for the 2006 task. The exhaustive search for the optimal weighing coefficients does not yield a combination which provides significantly better results.

While results for the retrieval task are satisfactory in queries based on grayscale radiographs, other queries, especially from photograpy imaging, have rather poor results, partly due to the lack of color features employed. Furthermore, a detailed visual evaluation might result 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. Several queries from the retrieval task 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. 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. For a better query completion, subsequent image abstraction steps are required.

References

[1] Müller H, Deselaers T, Lehmann TM, Clough P, Hersh W: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. CLEF working notes, Alicante, Spain, September 2006.

[2] Güld MO, Keysers D, Deselaers T, Leisten M, Schubert H, Ney H, Lehmann TM: Comparison of global features for categorization of medical images. Proceedings SPIE **5371** (2004) 211–222

[3] Castelli V, Bergman LD, Kontoyiannis I, Li CS, Robinson JT, Turek JJ: Progressive Search and Retrieval in Large Image Archives. IBM Journal of Research and Development 42(2): 253-268, 1998.

[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 AK, Duin RPW, Mao J: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1) (2000) 4–36

[8] Güld MO, Thies C, Fischer B, Lehmann TM: Content-based Retrieval of Medical Images by Combining Global Features. LNCS 2006; 4022, in press.