Baseline results for the ImageCLEF 2006 medical automatic annotation task

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

The ImageCLEF 2006 medical automatic annotation task encompasses 11,000 images from 116 categories, compared to 57 categories for 10,000 images of the similar task in 2005. As a baseline for comparison, a run using the same classifiers with the identical parameterization as in 2005 is submitted. In addition, the parameterization of the classifier was optimized according to the 9,000/1,000 split of the 2006 training data. In particular, texture-based classifiers are parallel combined with classifiers, which use spatial intensity information to model common variabilities among medical images. However, all individual classifiers are based on global features, i.e. one feature vector describes the entire image. The parameterization from 2005 yields an error rate of 21.7%, which ranks 13$^{\text{th}}$ among the 28 submissions. The optimized classifier yields 21.4% error rate (rank 12), which is insignificantly better.

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

Content-based image retrieval, Pattern recognition, Classifier combination

1 Introduction

The ImageCLEF medical automatic annotation task was established in 2005 [1], demanding the classification of 1,000 radiographs into 57 categories based on 9,000 categorized reference images. The ImageCLEF 2006 annotation task [2] consists of 10,000 reference images grouped into 116 categories and 1,000 images to be automatically categorized. This paper aims at providing a baseline for comparison of the experiments in 2005 and 2006 rather than presenting an optimal classifier.

2 Methods

2.1 Image Features

The content of a medical image is represented by texture features proposed by Tamura et al. [3] and Castelli et al. [4], denoted as TTM and CTM, respectively. Down-scaled representations of the original images were computed to 32×32 and $X \times 32$, pixels disregarding and according to the original aspect ratio, respectively. Since these image icons maintain the spatial intensity information, variabilities which are commonly found in a medical imagery are modelled by the distance measure. These include radiation dose, global translation, and local deformation. In particular, the cross-correlation function (CCF) which is based on Shannon and the image distortion model (IDM) suggested by Keysers et al. [5] is used. In particular, the following parameters were set:

- **TTM:** texture histograms, 384 bins, Jensen-Shannon divergence
- **CTM:** texture features, 43 bins, Mahalanobis distance with diagonal covariance matrix Σ.
- **CCF:** 32×32 icon, 9×9 translation window
- **IDM:** $X \times 32$ icon, gradients, 5×5 window, 3×3 context

2.2 Classifiers

The single classifiers are combined within a parallel scheme, which performs a weighting of the normalized distances obtained from the single classifiers C_i, and applies the nearest-neighbor-decision function C to the resulting distances:

$$d_c(q,r) = \sum_i \lambda_i \cdot d_i(q,r) \tag{1}$$

where λ_i, $0 \leq \lambda_i \leq 1$, $\sum_i \lambda_i = 1$ denotes the weight for the normalized distance $d_i(q,r)$ obtained from classifier C_i for a sample q and a reference r.

2.3 Submissions

Based on a combination of classifiers used for the annotation task in 2005 [6], a run using the exact same parameterization is submitted. Additionally, a second run is submitted which uses a parameterization obtained from an exhaustive search (using a step size of 0.05 for λ_i) for the best combination of the single classifiers. For this purpose, the development set of 1,000 images is used and the system was trained on the remaining 9,000 images.

3 Results

All results are obtained non-interactively, i.e. without relevance feedback by a human user. Table 1 shows the error rates in percent obtained for the 1,000 unknown images using single k-nearest neighbor classifiers and their combination for both $k = 1$ and $k = 5$. The error rate of 21.4% ranks 12^{th} among 28 submitted runs for this task. The weights that were optimized on the ImageCLEF 2005 medical automatic annotation task (10,000 images from 57 categories) yield an error rate of 21.7% and rank 13^{th}.

4 Discussion

The comparability of similar experiments such as ImageCLEF 2005 and ImageCLEF 2006 is difficult since several parameters such as the images and the class definitions were changed. In general, the difficulty of a classification problem is proportional to the error rate, increases with

Table 1: Error rates for the medical automatic annotation task.

Classifier	λ_{TTM}	λ_{CTM}	λ_{CCF}	λ_{IDM}	$k = 1$	$k = 5$
TTM	1.00	0	0	0	44.4%	44.9%
CTM	0	1.00	0	0	27.9%	25.7%
CCF	0	0	1.00	0	23.0%	23.4%
IDM	0	0	0	1.00	57.2%	54.9%
ImageCLEF 2005	0.40	0	0.18	0.42	21.7%	22.0%
Exhaustive search	0.25	0.05	0.25	0.45	21.5%	21.4%

the number of classes, and decreases with the number of references per class. Since the number of references varies for the classes, this relation is rather complex than linear. However, linearity can be used as a rough but first approximation.

Since the error rates are not substantially improved by a training on the 2006 data, the classifier that has been optimized for 2005 is also suitable for the ImageCLEF 2006 medical automatic annotation task. The parameterization optimized for ImageCLEF 2005 and 2006 yields $E^{2005} = 13.3\%$ and $E^{2006} = 21.7\%$, respectively, and the quotient E^{2005}/E^{2006} can be used to relate the results from the 2005 and 2006 campaigns. In other words, the best $E = 16.2\%$ from 2006 would approximately have resulted in $E = 10,0\%$, which is better than the best rate obtained in 2005. This is in accordance with the rank of submission, that drops from the 2^{nd} to the 13^{th} place. In conclusion, general improvements for automatic image annotation have been made during the last year.

5 Acknowledgment

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

References

[1] Clough P, Müller H, Deselaers T, Grubinger M, Lehmann TM, Jensen J, Hersh W. The CLEF 2005 cross-language image retrieval track. LNCS 2006; 4022: in press.

[2] Müller H, Deselaers T, Lehmann TM, Clough P, Hersh W. Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks. In: CLEF Working Notes; 2006, Sep; Alicante, Spain; in press.

[3] Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics 1978; 8(6): 460-73.

[4] 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 1998; 42(2): 253-68.

[5] Keysers D, Dahmen J, Ney H, Wein BB, Lehmann TM. A statistical framework for model-based image retrieval in medical applications. Journal of Electronic Imaging 2003; 12(1): 59-68.

[6] Güld MO, Thies C, Fischer B, Lehmann TM. Content-based retrieval of medical images by combining global features. LNCS 2006; 4022: in press.