Baseline Results for the CLEF 2008 Medical Automatic Annotation Task

Mark O. Güld, Thomas M. Deserno
Department of Medical Informatics, RWTH Aachen University, Aachen, Germany
mgueld@mi.rwth-aachen.de, deserno@ieee.org

Abstract

This work reports baseline results for the CLEF 2008 Medical Automatic Annotation Task (MAAT) by applying a classifier with a fixed parameter set to all tasks 2005 – 2008. The classifier performs a weighted combination of three distance and similarity measures operating on global image features: Scaled-down representations of the images are compared via metrics that model the typical variability in the image data, mainly translation, local deformation, and radiation dose. In addition, a distance measure based on texture features is used. For classification, a k nearest neighbor classifier is used. In 2008, the baseline classifier yields error scores of 170.34 and 182.77 for k=1 and k=5 when the full code is reported, which corresponds to error rates of 51.3% and 52.8% for 1-NN and 5-NN, respectively. Judging the relative increases of the number of classes and the error rates over the years, MAAT 2008 is estimated to be the most difficult in the four years.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval;

General Terms

Measurement, Experimentation

Keywords

Content-based image retrieval, classifier combination

1 Introduction

In 2008, the Medical Automatic Annotation Task (MAAT) is held for the fourth time as part of the annual challenge issued by the Cross-Language Evaluation Forum (CLEF). This task demands the non-interactive classification of a set of 1,000 radiographs according to a hierarchical, multi-axial code. For training, a set of radiographs is given along with their code, which was set manually by expert physicians. In these four years, the task difficulty changed: the challenge in 2005 used a grouping based on the code, whereas the later challenges use the full code. In addition, a modified error counting scheme is employed in 2007 and 2008 in order to address the severity of classification errors, e.g., whether a misclassification happens in upper (broader) or lower (more detailed) hierarchy levels. The number of participants in the task also varied over the years. It is therefore disirable to have baseline results for the CLEF MAATs, which allow a rough estimation of the task difficulties.

2 Methods

The content of one radiograph is represented by TAMURA's texture measures (TTM) proposed in [1] and down-scaled representations of the original images, 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) that is based on SHANNON, and the image distortion model (IDM) from [2] are used.

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:

$$d_{\text{combined}}(q,r) = \sum_i \lambda_i \cdot d'_i(q,r), \tag{1}$$

$$d'_i(q,r) = \frac{d_i(q,r)}{\sum_{r' \in R} d_i(q,r')} \tag{2}$$

where $0 \leq \lambda_i \leq 1$, $\sum_i \lambda_i = 1$ denotes the weight for the normalized distance $d_i(q,r)$ obtained from classifier C_i for a sample q and a reference r from the set of reference images, R. Values $0 \leq s_i(q,r) \leq 1$ obtained from similarity measures are transformed via $d_i(q,r) = 1 - s_i(q,r)$.

The three content descriptors and their distance measures use the following parameters:

- **TTM:** texture histograms from down-scaled image (256×256), 384 bins, Jensen-Shannon divergence as a distance measure,
- **CCF:** 32×32 icon, 9×9 translation window
- **IDM:** $X \times 32$ icon, gradients, 5×5 window, 3×3 context

The weighting coefficients were set empirically during CLEF MAAT 2005: $\lambda_{\text{IDM}} = 0.42$, $\lambda_{\text{CCF}} = 0.18$, and $\lambda_{\text{TTM}} = 0.4$.

3 Results

Tab. 1 lists the baseline results for the four years [3, 4, 5]. Runs which were not submitted are displayed marked with asterisks, along with their hypothetic rank. In 2007 and 2008, the evaluation was not based on the error rate – the table contains the rank based on the modified evaluation scheme for the corresponding submission of full codes.

			$k=1$		$k=5$	
Year	References	Classes	ER	Rank	ER	Rank
2005	9,000	57	13.3%	2/42	14.8%	*7/42
2006	10,000	116	21.7%	13/28	22.0%	*13/28
2007	11,000	116	20.0%	*17/68	18.0%	18/68
2008	12,089	197	51.3%	*12/24	52.8%	12/24

Table 1: Baseline error rates (ER) and ranks among submissions.

4 Discussion

A rough estimation of the task difficulty can be derived from the baseline error rates: Comparing 2005 and 2006, the number of classes increased by 103%, while the error rate only increased by

63% and 48% for 1-NN and 5-NN, respectively. This suggests that the task in 2006 was easier than in 2005. Since the challenges in 2006 and 2007 use the same class definitions, the obtained error rates are directly comparable and show a slightly reduced task difficulty in 2007. In 2008, the number of classes increased by 70% compared to 2007, while the error rate increased by 157% and 193%, respectively. The 2008 task can therefore be considered to be more difficult than the 2007 task. Applying the same estimation, the 2008 task is also found to be more difficult than the 2005 task, as the number of classes increased by 246% and the error rate increased by 286% and 257%, respectively.

References

[1] Tamura H, Mori S, Yamawaki T. Textural features corresponding to visual perception. IEEE Trans Syst Man Cybern B Cybern 1978; 8(6): 460–73

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

[3] Güld MO, Thies C, Fischer B, Lehmann TM. Content-based retrieval of medical images by combining global features. Lect Notes Comput Sci 2006; 4022:702–11

[4] Güld MO, Thies C, Fischer B, Deserno TM. Baseline results for the ImageCLEF 2006 medical automatic annotation task Lect Notes Comput Sci 2007; 4730: 686–9

[5] Güld MO, Deserno TM. Baseline results for the CLEF 2007 medical automatic annotation task using global image features. Lect Notes Comput Sci 2008, in press