Baseline Results for the ImageCLEF 2008 Medical Automatic Annotation Task in Comparison over the Years

Mark O. Güld, Petra Welter, and Thomas M. Deserno

Department of Medical Informatics, RWTH Aachen University, Aachen, Germany
mgueld@mi.rwth-aachen.de, pwelter@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. A nearest-neighbor (NN) classifier is used, which uses a weighted combination of three distance and similarity measures operating on global image features: Scaled-down representations of the images are compared using models for 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. 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.

1 Introduction

In 2008, the Medical Automatic Annotation Task (MAAT) [1] is held for the fourth time as part of the annual challenge issued by the Cross-Language Evaluation Forum (CLEF). It demands the non-interactive classification of a set of 1,000 radiographs according to a hierarchical, multi-axial code [2]. For training, a separate set of radiographs is given along with their code, which was defined by expert physicians. Over the four years, the task difficulty changed: the challenge in 2005 used a grouping based on the code hierarchy, 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. It penalizes misclassification in upper (broader) hierarchy levels more than errors on lower, more detailed levels. The participants in the task also varied over the years.

It is therefore desirable to have baseline results for the CLEF MAATs, which allow a rough estimation of the task difficulties. Based on the Image Retrieval in Medical Applications (IRMA) framework [3,4], we provide this baseline computations.

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 752–755, 2009.

2 Methods

The content of each radiograph is represented by TAMURA's texture measures (TTM) proposed in [5] 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 that are commonly found in a medical imagery are modeled 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 [6] 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 NN 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; and
- IDM: $X \times 32$ icon, gradients, 5×5 window, 3×3 context

The weighting coefficients were set empirically during CLEF MAAT 2005:

$$\begin{aligned} \lambda_{\text{IDM}} &= 0.42, \\ \lambda_{\text{CCF}} &= 0.18, \quad \text{and} \\ \lambda_{\text{TTM}} &= 0.40. \end{aligned}$$

3 Results

Tab. 1 lists the baseline results for the four years [7,8,9]. Runs which were not submitted are marked with asterisks, along with their hypothetic rank. In 2007 and 2008, the evaluation was not based on the error rate. Therefore, the table lists the rank based on the modified evaluation scheme on full codes. In average, the $k = 1$ NN classifier is better than $k = 5$. Disregarding the hierarchical information, which was made essential to solve the 2008 task, we obtain an error rate of 51,3%.

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

			$k = 1$		$k = 5$	
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

4 Discussion

The baseline error rates allow a rough estimation of the task difficulty: Comparing 2005 and 2006, the number of classes increased by 103%, while the error rates 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 is therefore considered to be more difficult than the 2007 task. With similar estimation, the 2008 task is also found to be more difficult than the 2005 task, as the number of classes increased by 246%, but the error rate increased by 286% and 257%, respectively.

Acknowledgment

This work is part of the IRMA project, which is funded by the German Research Foundation (DFG), grants Le 1108/4, Le 1108/6, and Le 1108/9.

References

1. Deselaers, T., Deserno, T.M.: Medical Image Annotation in ImageCLEF 2008. In: Peters, C., et al. (eds.) Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum. LNCS, vol. 5706, pp. 523–530. Springer, Heidelberg (2009)
2. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The IRMA code for unique classification of medical images. In: Proceedings SPIE, vol. 5033, pp. 109–117 (2003)
3. 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), 354–361 (2004)
4. Güld, M.O., Thies, C., Fischer, B., Lehmann, T.M.: A generic concept for the implementation of medical image retrieval systems. International Journal of Medical Informatics 76(2-3), 252–259 (2007)
5. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics, B 8(6), 460–473 (1978)

6. Keysers, D., Dahmen, J., Ney, H., Wein, B.B., Lehmann, T.M.: A statistical framework for model-based image retrieval in medical applications. Journal of Electronic Imaging 12(1), 59–68 (2003)
7. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
8. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)
9. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D. (eds.) CLEF 2007. LNCS, vol. 5152, pp. 472–491. Springer, Heidelberg (2008)