Baseline Results for the ImageCLEF 2007 Medical Automatic Annotation Task Using Global Image Features

Mark O. Güld and Thomas M. Deserno

Department of Medical Informatics, RWTH Aachen University, Aachen, Germany
mgueld@mi.rwth-aachen.de, deserno@ieee.org

Abstract. This paper provides baseline results for the medical automatic annotation task of CLEF 2007 by applying the image retrieval in medical applications (IRMA)-based algorithms previously used in 2005 and 2006, with identical parameterization. Three classifiers based on global image features are combined within a nearest neighbor (NN) approach: texture histograms and two distance measures, which are applied on down-scaled versions of the original images and model common variabilities in the image data. According to the evaluation scheme introduced in 2007, which uses the hierarchical structure of the coding scheme for the categorization, the baseline classifier yields scores of 51.29 and 52.54 when reporting full codes for 1-NN and 5-NN, respectively. This corresponds to error rates of 20.0% and 18.0% (rank 18 among 68 runs), respectively. Improvements via addressing the code hierarchy were not obtained. However, comparing the baseline results yields that the 2007 task was slightly easier than the previous ones.

1 Introduction

The ImageCLEF medical automatic annotation task (MAAT) 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 MAAT [2] consisted of 10,000 reference images grouped into 116 categories and 1,000 images to be automatically categorized. The categorization is based on a medical code introduced in [3], which has since been refined and extended for new imaging techniques and body regions. In 2007, the hierarchical structure of the code is used to describe the image contents, with the evaluation scheme allowing a finer granularity of the classification accuracy [4]. Sets of 10,000 training images, 1,000 images for parameter optimization, and 1,000 unknown images are used in the experiments, again from 116 categories, i.e. unique codes. As both the task and the participants (i.e. classification algorithms) evolved over the last three years, it is difficult to compare the results. In this paper, we apply one algorithm with a fixed set of parameters to all the tasks, and therefore provide baseline results that allow a rough comparison between any pair of other runs over the last three years.

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 637–640, 2008.

2 Methods

The image retrieval in medical applications (IRMA) framework is used to produce the baseline results [5]. In particular, the image content is represented by global features [6,7], where each image is assigned to one feature vector. Texture properties as proposed by Tamura et al. are extracted for each image (scaled to 256×256 pixels), obtaining a 384-dimensional histogram. The distance between a pair of images is computed via the Jensen-Shannon divergence (JSD) of their respective texture histograms. Down-scaled representations of the images allow to explicitly model frequent, class-invariant variabilites among the images, such as translation, radiation dose, or local deformations. The down-scaled images are compressend to roughly 1KB of size, 32×32 pixels when applying the cross-correlation function (CCF) as a similarity measure, or gradient images of X×32 pixels when applying the image distortion model (IDM), is regarding and acknowledging the original aspect ratio, respectively. The combination of these features is done within a NN scheme: a total distance between a sample image q and a reference image r is obtained by the weighted sum of the normalized distances from the single classifiers. The weighting coefficients where empirically adjusted based on prior (non-CLEF MAAT) experiments.

In 2007, the evaluation is done using the scheme described in [4]. For each image from the test set, an error value $e \in [0..1]$ is obtained, based on the position of classification errors in the hierarchy. By summation over all 1,000 test images, the overall value is obtained. Constantly answering *don't know* yields a value of 500.0, the worst possible value is 1000.0. To address the modified evaluation scheme, the NN decision rule is modified in an additional experiment: From the k neighbors, a *common* code is generated by setting differing parts (and their subparts) to *don't know*, e.g. two neighbors with codes 1121-120-434-700 and 1121-12f-466-700 result in a *common* code of 1121-12∗-4∗∗-700.

3 Results

All results are obtained non-interactively, i.e. without relevance feedback from a human. Tab. 1 contains the baseline error rates. In 2007, the evaluation was not based on the error rate – the table also contains the rank based on the modified evaluation scheme for the corresponding submission of full codes. Runs which were not submitted are displayed marked with asterisks, along with their hypothetical rank. Using the evaluation scheme proposed for 2007, the default weighting for k-NN yields 51.29 and 52.54 for $k = 1$ and $k = 5$, respectively. The *common code* rule yields 80.47 when applied to the 5-NN results.

4 Discussion

The medical automatic annotation task in 2007 is a bit easier than 2006, as the baseline error rate drops from 21.7% to 20.0% and from 22.0% to 18.0% for the 1-NN and the 5-NN, respectively. As the baseline results are reported for

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

the last three years, they allow the a rough comparison between submissions from the years 2005, 2006, and 2007 for algorithms which only participated in one year. Note that the evaluation scheme significantly differs from the error rate: Although the 1-NN yields a better result than the 5-NN, its error rate is actually worse. This depends on the severity of misclassifications, which are captured by the new evaluation scheme.

The *common code* rule, which generates a code fragment that all k nearest neighbors agree on, does not improve the results, but performs significantly worse. In addition, other experiments were performed to either remove neighbors from the NN list by applying a distance threshold, or by modifying the complete decision into *don't know*, again based on a distance threshold. These experiments did not provide any improvement, either. This seems to be consistent with efforts by other groups, which were largely unable to improve their results if they address the code hierachy.

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, 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)
2. 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)
3. 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)
4. 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., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)

5. 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)
6. 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)
7. 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)