Baseline Results for the CLEF 2007 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 paper provides baseline results for the medical automatic annotation task of CLEF 2007. Therefore, the algorithms initially used for the corresponding tasks in 2005 and 2006 are applied, using the same parameterization. Three classifiers based on global image features are used and combined within a nearest neighbor approach. In 2007, a hierarchical code is introduced to describe the image contents, with the evaluation scheme allowing a finer granularity of the classification accuracy. We therefore evaluate some techniques for estimating the confidence in the classifier decision, which stop or alter classifier reports at code levels with uncertain classifier reports.

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

This paper provides baseline results for the medical automatic annotation challenge of CLEF 2007. By using the same methods and the same parameterization, the obtained results allow to roughly estimate the complexity of the annotation tasks from 2005, 2006, and 2007 relative to each other. In 2007, the evaluation scheme addresses the hierarchical structure of the IRMA code [1] by allowing the classifier to decide *don't know* at any level of the code, independently for each of the four axes [2]. To effectively support this scheme, models which estimate the classifier's confidence in its decision are required.

2 Methods

2.1 Features, classifiers, and their combination

The image content is described by global features, i.e. the intensity information from the images is drastically reduced to few (numerical) values [3].

TAMURA ET AL. proposed histograms of the coarseness, directionality, and contrast to capture texture properties [4]. The histograms have $6\times88\times8 = 384$ bins and are compared by using Jensen-Shannon divergence as a distance measure. This approach is denoted as Tamura texture measures (TTM). By using down-scaled representations of the original images, a-priori knowledge about common variabilities can be integrated into the distance measure. Here, the cross correlation-function (CCF) is used to measure similarity between 32×32 representations. It is robust to global translations (by using a 9×9 translation window) and varying radiation dose (by normalizing the intensity values). The image distortion model (IDM) models local deformations by allowing pixel warping within a neighborhood [5]. It uses $X\times32$ representations, a 5×5 search window, 3×3 contexts, gradients (instead of intensities) and a distance threshold.

The three classifiers are combined within a nearest neighbor classifier. The overall distance between a sample q and a reference r is computed by

$$d_c(q,r) = \lambda_{\text{IDM}}\cdot d'_{\text{IDM}}(q,r) + \lambda_{\text{CCF}}\cdot d'_{\text{CCF}}(q,r) + \lambda_{\text{TTM}}\cdot d'_{\text{TTM}}(q,r)$$

where $d'(q,r)$ denotes a normalized distance between q and r. Normalization is done by dividing the individual distance by the sum of distances between the sample q and all references. For CCF, distances are obtained by transforming similarity s to $d = 1 - s$. The nearest neighbor classifier then decides based on the class information of the k best references.

2.2 Code hierarchy and confidence

To address the modified evaluation scheme of the 2007 challenge, the nearest neighbor decision rule is modified. Three options are implemented:

1. 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 1121-12X-4XX-700.

2. For the k neighbors, a threshold t_d is applied to the majority vote decision. If the distance for the best neighbor from the decided class is greater than t_d, the decision is rejected, i.e. the reported code is XXXX-XXX-XXX-XXX.

3. A threshold t_n is applied to the k neighbors. A neighbor is excluded from the decision if its distance is greater than t_n. For the remaining neighbors, majority vote is used to obtain the decision. If all neighbors are excluded, XXXX-XXX-XXX-XXX is reported.

To keep the number of combinations at bay, only combinations 1+2 and 1+3 are evaluated.

3 Results

The classifier weights are the same as 2005 and 2006: $\lambda_{\text{IDM}} = 0.42$, $\lambda_{\text{CCF}} = 0.18$, $\lambda_{\text{TTM}} = 0.40$. To obtain estimates for t_d and t_n, the development set is used: inspecting the classifiction for this set, the results are sorted based on the best distances from neighbors from the decision class for each sample. Based on this sorted list, the thresholds are chosen from the 1^{st}, 5^{th}, 10^{th}, 25^{th}, and

		t_d (index)					
decision	k	-	1	5	10	25	50
majority vote	1	51.29	51.34	52.06	56.32	59.26	71.46
	5	52.54	52.82	56.90	62.94	72.69	101.35
common code	5	80.47	80.77	81.45	84.49	86.59	97.19

Table 1: Results for decision threshold t_d. The results in the left column are obtained without t_d.

50^{th} worst distances encountered. Both thresholds are evaluated with the normal majority vote decision rule first and afterwards with the policy to obtain the *common* code parts.

The evaluation is done using the scheme described in [2]. 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 $1,000.0$.

Results for applications of the decision threshold t_d are summarized in Tab. 1. The neighbor threshold t_n is used in combination with $k \in \{1, 5, 10, 25, 50, 100\}$, because otherwise the number of considered neighbors would be so high that small classes are never reported. Tab. 2 contains the results for the application of t_n.

	t_n (index), majority vote					t_n (index), *common* code				
k	1	5	10	25	50	1	5	10	25	50
1	51.34	52.06	56.32	59.26	71.46	51.34	52.06	56.32	59.26	71.46
5	52.25	53.32	56.10	57.88	70.89	80.51	80.24	83.79	83.06	93.20
10	54.45	55.38	59.21	61.24	72.51	110.65	109.90	111.76	109.03	115.11
25	62.78	62.56	66.91	68.82	79.72	161.10	156.95	154.27	147.81	147.69
50	87.50	82.81	83.84	80.60	86.59	201.71	193.63	186.34	176.36	166.44
100	114.85	104.59	101.79	94.77	94.88	236.91	225.93	213.50	197.61	179.73

Table 2: Results for neighbor threshold t_n. Results for $k = 1$ can be found in Tab. 1 as well.

For comparsion with the the medical automatic annotation task from the previous years, Tab. 3 contains baseline error rates.

			error rate	
year	references	classes	$k = 1$	$k = 5$
2005	9,000	57	13.3%	14.8%
2006	10,000	116	21.7%	22.0%
2007	11,000	116	20.0%	18.0%

Table 3: Error rates for the medical automatic annotation task.

4 Discussion

The proposed mechanisms for the estimation of classifier results and the modification of reported codes do not improve the baseline results of 51.29 for 1-NN and 52.54/rank 18 for 5-NN. This seems to have been observed by the other groups as well [2]. In our case, *common* code performs generally worse than the majority vote decision. The results become drastically worse for bad parameter sets, especially when the number of considered neighbors is too high.

Comparing the baseline error rates to the ones from the previous year, the medical automatic annotation task in 2007 is a bit easier than 2006. This can be taken into account when comparing methods by groups who participated in only one of the past years.

References

[1] Lehmann TM, Schubert H, Keysers D, Kohnen M, Wein BB. The IRMA code for unique classification of medical images. Procs SPIE 2003; 5033: 109-117

[2] Müller H, Deselaers T, Kim E, Kalpathy–Cramer J, Deserno TM. Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks. Working Notes of the 2007 CLEF Workshop (in this book)

[3] Lehmann TM, Güld MO, Thies C, Fischer B, Spitzer K, Keysers D, Ney H, Kohnen M, Schubert H, Wein BB. Content-based image retrieval in medical applications. Methods of Information in Medicine 2004; 43(4): 354-361

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

[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