MIRACLE at ImageCLEFannot 2008: Classification of Image Features for Medical Image Annotation

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3]
José Carlos González-Cristóbal[1,3], José Miguel Goñi-Menoyo[1]
[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid.
[3] DAEDALUS - Data, Decisions and Language, S.A.
slana@diatel.upm.es, jvillena@it.uc3m.es
josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es

Abstract

This paper describes the participation of MIRACLE research consortium at the ImageCLEF Medical Image Annotation task of ImageCLEF 2008. A lot of effort was invested this year to develop our own image analysis system, based on MATLAB, to be used in our experiments. This system extracts a variety of global and local features including histogram, image statistics, Gabor features, fractal dimension, DCT and DWT coefficients, Tamura features and coocurrency matrix statistics. Then a k-Nearest Neighbour algorithm analyzes the extracted image feature vectors to determine the IRMA code associated to a given image. The focus of our experiments is mainly to test and evaluate this system in-depth and to make a comparison among diverse configuration parameters such as number of images for the relevance feedback to use in the classification module.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital libraries.

Keywords

Information Retrieval, medical image, image annotation, classification, IRMA code, axis, learning algorithms, nearest-neighbour, machine learning, ImageCLEF Medical Automatic Image Annotation task, CLEF, 2008.

1. Introduction

MIRACLE team is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a small/medium size enterprise (SME) founded in 1998 as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in many different tracks and tasks, including the main bilingual, monolingual and cross lingual tasks as well as in ImageCLEF, Question Answering, WebCLEF, GeoCLEF and VideoCLEF (VID2RSS) tracks.

This paper describes our participation in the Medical Image Annotation task of ImageCLEF 2008 [6]. Briefly, the objective of this task is to provide the IRMA (Image Retrieval in Medical Applications) code [4] for each image of a given set of 1,000 previously unseen medical (radiological) images covering different medical pathologies. Over 12,000 classified training images were provided this year to be used in any way to train a classifier. This task uses no textual information, but only image-content information.

While in previous participations [5] [7] we approached this task as a machine learning problem, regardless of the domain, as our areas of expertise did not include image analysis research [2] [3], a lot of effort was invested last year to develop our own image analysis system, based on MATLAB, to be used in our experiments. Thus, now the main purpose of our experiments is to test and evaluate this system in-depth and make a comparison among diverse configuration parameters such as number of images for the relevance feedback to use in the classification module.

2. Description of Experiments

The architecture of our system is composed of different functional blocks:

- **Feature extraction module**: in charge of the calculation and extraction of a variety of features of each image, both the training set used to build the classifier and the test set that is to be actually classified. This module has been entirely developed using MATLAB and extracts vectors with a total of 3,741 features for each image.

  Images are converted to gray-scale, rescaled to 256x256 pixels and the following features are extracted:

  - **Global features**: gray histogram (128 levels of gray), image statistics (mean, median, variance, maximum singular value, skewness and kurtosis), Gabor features (4 scales, 6 filter orientations), fractal dimension, Discrete Cosine Transform (DCT) coefficients, Discrete Wavelet Transform (DWT) coefficients, Tamura features (coarseness, contrast, directionality), and coocurrency matrix statistics (energy, entropy, contrast, homogeneity, correlation)
  - **Local features**: images are cut up into 64x64 pixel blocks and then the previous features are extracted for each block.

- **Classifier**: determines the IRMA code associated to a given image, from its feature vector and the feature matrix of the training set. The classifier is internally composed of two blocks: an initial module in charge of selecting those images in the training set whose vectors are at a distance lower than a given threshold from the vector associated to the image to classify, and then a second module that actually generates the IRMA code, depending on the codes and similarity of nearby images.

Finally, we submitted four runs to be evaluated, described in Table 1. For all of them, the returned IRMA code is generated from the combination of the first N images in the training set that are most similar to the image to classify. The combination consists of a simple "addition" of strings characters in which, if both characters are different, the result is the wildcard "*" representing the ambiguity (or "hesitation" to choose). This algorithm actually could be considered as a variation of the classical k-Nearest Neighbour algorithm [8] with a specific definition of the generating the output class.

Additionally, two runs use relevance feedback (RF) with the first two images in the training set that are at a lowest distance. Vectors of those images are added and averaged to build a new vector that is used for querying the system again.

Table 1. Experiment set

Run Identifier	Description
MIRACLE-2I-0F	Merge code of 2 first results
MIRACLE-3I-0F	Merge code of 3 first results
MIRACLE-2I-2F	Merge code of 2 first results + RF with 2 images
MIRACLE-3I-2F	Merge code of 3 first results + RF with 2 images

3. Results

Results are shown in Table 2. The "Error score" column contains the experiment score as computed by the task organizers [1]. This score is defined so as to penalize wrong decisions that are easy to take (i.e., there are few possible choices at that node) over wrong decisions difficult to take (i.e., there are many possible choices at that node). Furthermore, it also penalizes wrong decisions at an early stage in the code (higher up in the IRMA code hierarchy) over wrong decisions at a later stage (lower down in the hierarchy). The "Well Classified" column shows the actual number of images with complete correct predicted code. The "Bad-Classified" column shows the number of images with error score equal to 1.0 (wrong prediction of all code axis).

Table 2. Results of experiments

Run Identifier	Error Score	Well Classified	Bad Classified
MIRACLE-2I-0F	190.38	**219**	**0**
MIRACLE-3I-0F	**187.90**	144	**0**
MIRACLE-2I-2F	190.38	**219**	**0**
MIRACLE-3I-2F	194.26	167	**0**

The best score is achieved by the run that combines the codes of the first 3 images, with no relevance feedback. Moreover, runs using the codes of the first 2 images seem to get the same final score no matter if relevance feedback is considered or not. However, the analysis axis-by-axis shows interesting differences that will be described later.

Table 3 shows the average results from all groups. Comparing our scores to the scores of other participants in the task, we achieve average results and rank 4th out of 6 groups.

Table 3. Summary of results

Maximum error score	313.01
Minimum error score	74.92
Average error score	169.71
Mode error score	190.38
Number of runs	24

Table 4 shows an axis-by-axis analysis of the results. For each of the four axis of the IRMA code, this table shows the "Error Score", calculated as the sum of the errors made for each image, and the number of images in which the full prediction of the axis (i.e., no wildcards in the output) is correct.

Table 4. Axis-by-axis analysis

	T-Axis		D-Axis		B-Axis		A-Axis	
Run Identifier	**Error Score**	**Well Classified**	**Error Score**	**Well Classified**	**Error Score**	**Well Classified**	**Error Score**	**Well Classified**
MIRACLE-2I-0F	**5.24**	**852**	318.04	**381**	**362.56**	**283**	75.6	**789**
MIRACLE-3I-0F	6.32	808	**309.78**	293	367.82	206	**67.67**	735
MIRACLE-2I-2F	**5.24**	**852**	318.04	**381**	**362.56**	**283**	75.67	**789**
MIRACLE-3I-2F	6.12	818	322.09	318	374.74	235	74.07	744

Next figures allow to make a graphical comparison of the results obtained by each experiment, showing both the global evaluation of the experiment and the specific evaluation of each individual axis. Figure 1 shows the number of images for which the complete code (no wildcards) has been correctly predicted. Figure 2 shows the number of images for which the complete prediction of the axis is completely wrong.

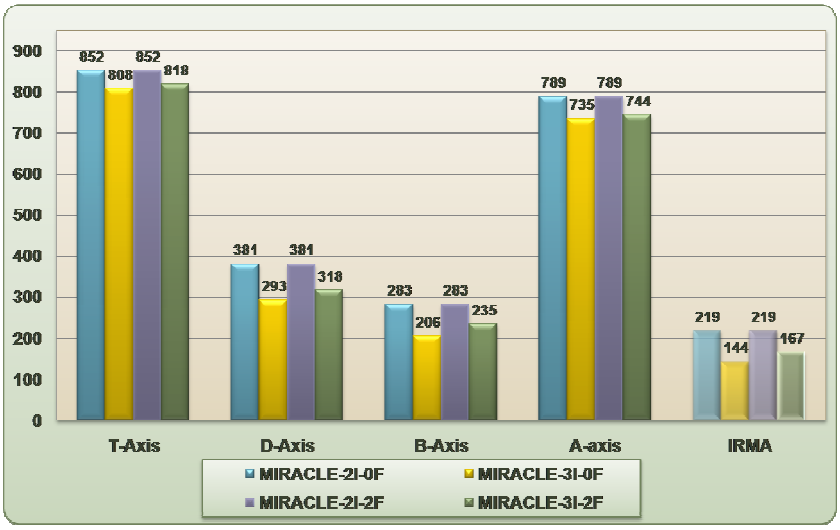

Figure 1. Correctly predicted axis

As observed in the previous figure, the Technical (T) and Anatomical (A) axis are the best predicted axis, with a significant difference with respect to the others. However this is misleading in the case of the Technical axis, as

the value of this axis for all images to classify is either “1121”, “1123”, “1124” or “112d”, thus, in practice, having to decide only among four codes – in fact, 93% of the images have “1121” and 4% have ”1124”.

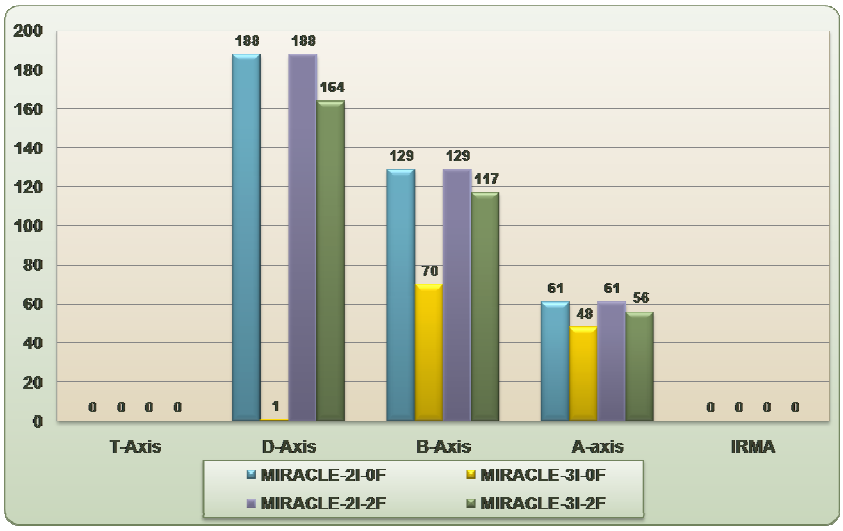

Figure 2. Incorrectly predicted axis

4. Conclusions and Future Work

Based on the analysis performed over each axis, the first conclusion to be drawn is that the first weak point of our experiments is the prediction of the Direction (D) and Biological (B) axis. Some extra effort must be invested on determining which image features could be most useful to predict those axis.

In addition, it can be observed that although the number of incorrect predictions is relatively low, this does not correspond to a high number of correct predictions (which in fact is also relatively low), as it would be expected applying a binary “correct” vs. “incorrect” logic. This is due to the fact that, as the cost of making an incorrect decision is higher than the cost of not actually making a decision, the design criteria of the system is biased for “hesitation”, i.e., the system is very cautious and assigns a wildcard “*” if there is any kind of ambiguity. This explanation also confirmed by the result of the run that takes 3 codes for generating the final IRMA code: when the number of codes increases, so ambiguity does, thus the number of complete correct predictions decreases and also the error score.

Finally, in all runs, the calculation of the distance among vectors assigns the same weight to every dimension of the vectors, regardless of the nature of the feature to which this component belongs and/or the number of components belonging to that feature. This was actually our mistake when carrying out the experiments and the feature matrix should have been divided into the different feature sub-matrixes that employ different distances for calculating similarity and are combined to each other using different weight strategies. For sure this will be taken into account for future participations.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project BRAVO (Multilingual and Multimodal Answers Advanced Search – Information Retrieval), TIN2007-67407-C03-03 and by Madrid R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

[1] Deselaers, Thomas; Kalpathy-Cramer, Jayashree; Müller, Henning; Deserno, Thomas . Hierarchical classification for ImageCLEF 2008 Medical Image Annotation. On line http://www.imageclef.org/system/files/hierarchical2008.pdf [Visited 14/08/2008].

[2] FIRE: Flexible Image Retrieval System. On line http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html [Visited 14/08/2008].

[3] Goodrum, A.A. Image Information Retrieval: An Overview of Current Research. Informing Science, Vol 3(2), pp 63-66, 2000.

[4] IRMA project: Image Retrieval in Medical Applications. On line http://www.irma-project.org/ [Visited 10/08/2008].

[5] Lana-Serrano, Sara; Villena-Román, Julio; González-Cristóbal, José Carlos; Goñi-Menoyo, José Miguel. MIRACLE at ImageCLEFannot 2007: Machine Learning Experiments on Medical Image Annotation. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.

[6] ImageCLEF Medical Automatic Image Annotation Task. On line http://www.imageclef.org/2008/medaat [Visited 14/08/2008].

[7] Villena-Román, J.; González-Cristóbal, J.C.; Goñi-Menoyo, J.M.; and Martínez Fernández, J.L. MIRACLE's Naive Approach to Medical Images Annotation. Working Notes for the CLEF 2005 Workshop. Vienna, Austria, 2005.

[8] Witten, Ian H.; Frank, Eibe. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.