MIRACLE at ImageCLEFannot 2008: Nearest Neighbour Classification of Image Feature Vectors for Medical Image Annotation

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3]
José Carlos González-Cristóbal[1,3], and 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. During the last year, our own image analysis system was developed, based on MATLAB. 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 co-occurrence matrix statistics. A classifier based on the k-Nearest Neighbour algorithm is trained on the extracted image feature vectors to determine the IRMA code associated to a given image. The focus of our participation was mainly to test and evaluate this system in-depth and to compare among diverse configuration parameters such as number of images for the relevance feedback to use in the classification module.

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

The MIRACLE team is a research consortium formed by research groups of three different Spanish universities (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a private company founded as a spin-off of these groups and a leading company in the field of linguistic technologies in Spain.

This paper reviews our participation [1] at the Medical Image Annotation task of ImageCLEF 2008 [2]. While in previous participations [3] [4] we approached this task as a domain-independent machine learning problem, as our areas of expertise did not include image analysis research, a lot of effort was invested during the last year to develop our own image analysis system. Thus, the main purpose of our participation in this task was to test and evaluate this system in-depth and determine the optimum settings to use in the classification module. In the following sections, we will give an overview of our approach and describe and analyze the results achieved.

C. Peters et al. (Eds.): CLEF 2008, LNCS 5706, pp. 728–731, 2009.

2 Description of the System

Our system is composed of two different functional modules. First, the Feature Extraction module is in charge of the extraction of a variety of feature descriptors for a given image. It has been entirely developed during the last year, using MATLAB.

Each image is first converted to gray-scale and rescaled to 256x256 pixels. Then, the following feature descriptors are extracted: 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, Discrete Wavelet Transform, Tamura features (coarseness, contrast, directionality), and co-occurrence matrix statistics (energy, entropy, contrast, homogeneity, correlation). Both global features for the whole image and local features for 64x64 blocks are obtained. All features are linearly combined (weight=1), and no feature selection is carried out. The final feature vector contains 3,741 descriptors.

On the other hand, the Classification module determines the IRMA code associated to a given image, basically comparing 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.

3 Experimental Results and Discussion

Four runs were submitted. All of them are based on the classical k-Nearest Neighbour algorithm [5] with a specific adaptation to generate the output class. The IRMA code is generated from the combination of the codes of the first *k* images in the training set that are most similar to the image to classify. The combination consists of a simple "addition" of code strings characters in which, if both characters are different, the result is the wildcard "*" representing the ambiguity ("hesitation" to choose).

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

Results are shown in Table 1. The "Error score" is defined [2] so as to penalize wrong decisions in which there are few possible choices over wrong decisions with many possible choices. Furthermore, it also penalizes wrong decisions higher up in the IRMA code hierarchy over wrong decisions lower down in the hierarchy. The "Well Classified" column shows the images with complete correct predicted code.

Table 1. Results of experiments

Run[1]	Error Score	Well Classified
2I-0F	190.38	**219**
3I-0F	**187.90**	144
2I-2F	190.38	**219**
3I-2F	194.26	167

[1] Run identifiers hold *k* (neighbours) and *n* (images for relevance feedback)

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. Although it is not shown in the table, no image code was completely wrong, i.e., there was at least one valid code character for every annotated image.

One important issue related to the classification algorithm is the fact that, as the cost of making an incorrect decision is higher than the cost of not actually making a decision, the choice 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. As confirmed by results shown in Table 1, when the number of codes taken for generating the final IRMA code increases, so ambiguity does, thus the number of complete correct predictions decreases and also the error score.

Strategies for improving this decision criterion must be found for next years. One possible strategy is to assign a different relevance to each result according to its position P in the list, with different weighting factors, such as (1/P), (1-P), etc.

Comparing to other groups, we achieve average results and rank 4th (out of 6).

The analysis axis-by-axis shows interesting results. For each of the four axis of the IRMA code, the "Error Score" shown in Table 2 is 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 2. Axis-by-axis analysis

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

The Technical axis achieves the best error score. However, this is somehow misleading, as the whole image database holds only 4 possible values (codes) for this axis ("1121", "1123", "1124" and "112d"), thus making the decision easier than for the other axes. In fact, 93% of the images have the value "1121", i.e., it is possible to achieve a good error score even with a simple majority classifier (such as ZeroR [5]). Obviously, these different distribution and frequency of code values for each axis will be taken into account in future participations.

Moreover, the prediction for the Anatomical (A) axis shows a significant difference with respect to the Direction (D) and Biological (B) axes. We must conclude that the chosen features are not useful for modelling the image concepts that are intrinsic to those axes. Particularly, in the case of the D-axis the differences among images are very subtle and strongly rely on different brightness or contrast areas in the left or right side (or top or bottom) of the image. A possible approach for improving the classification of those axes is to give more weight to local features with respect to the global analysis, as local feature can model the differences among image regions.

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.

Another aspect is that 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. This fact will be taken into account for next years.

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

References

1. Lana-Serrano, S., Villena-Román, J., González-Cristóbal, J.C., Goñi-Menoyo, J.M.: MIRACLE at ImageCLEFannot 2008: Classification of Image Features for Medical Image Annotation. In: Working Notes of the 2008 CLEF Workshop, Aarhus, Denmark (2008)
2. Deselaers, T., Deserno, T.M.: Medical Image Annotation in ImageCLEF 2008. In: Peters, C., et al. (eds.) CLEF 2008. LNCS, vol. 5706, pp. 523–530. Springer, Heidelberg (2009)
3. Lana-Serrano, S., Villena-Román, J., González-Cristóbal, J.-C., Goñi-Menoyo, J.M.: MIRACLE at ImageCLEFanot 2007: Machine Learning Experiments on Medical Image Annotation. 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. 597–600. Springer, Heidelberg (2008)
4. Villena-Román, J., González-Cristóbal, J.C., Goñi-Menoyo, J.M., Martínez Fernández, J.L.: MIRACLE's Naive Approach to Medical Images Annotation. In: Working Notes for the CLEF 2005 Workshop, Vienna, Austria (2005)
5. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)