MedIC/CISMeF at ImageCLEF 2006: Image Annotation and Retrieval Tasks

F.Florea[a,b], A.Rogozan[a], V.Cornea[b], A.Bensrhair[a] and S.Darmoni[a,b]
[a] LITIS Laboratory, INSA de Rouen, France
[b] CISMeF Team, Rouen University Hospital & GCSIS Medical School of Rouen, France
filip.florea@insa-rouen.fr

Abstract

In the 2006 ImageCLEF cross-language image retrieval track, the MedIC/CISMeF group participated at the two medical-related tasks: the automatic annotation task and the multilingual image retrieval task. For the first task we submitted four runs based on supervised classification of combined texture and statistical image representations, the best result being the fourth rank at only 1% of the winner. The architecture proposed for the second task is reposing on textual-retrieval using terms derived from the MeSH thesaurus, combined with ranking by visual similarity. Due to technical and practical difficulties, the only run we were able to submit was incomplete, resulting in a modest result in the pools. Therefore, the actual capacity of the proposed retrieval architecture could not be evaluated.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

Keywords

Content-based image retrieval, image categorization, visual/textual retrieval, classification, machine learning

1 Introduction

ImageCLEF cross-language image retrieval track was established in 2003 as part of the Cross Language Evaluation Forum (CLEF), a benchmarking multilingual information retrieval campaign held annually since 2000.

The CISMeF project[1] (French acronym for Catalog and Index of French-language health resources) [1] is a quality-controlled subject gateway initiated by the Rouen University Hospital. The objective is to describe and index the main French-language health resources (documents on the web) to assist the users (i.e. health professionals, students or general public) in their search for high quality medical information available on the Internet.

The CISMeF team currently is developing an image information extraction and annotation module, named MedIC (i.e. Medical Image Categorization), to allow direct access to images extracted from health-documents.

For the 2006 edition of ImageCLEF, we used the MedIC module to participate at the two tasks involving medical images: the medical image annotation task and the multilingual medical image

[1] http://www.cismef.org

retrieval task. We used different approaches for each of the two tasks, following their specificity and objectives. Thus, an approach based on supervised classification of image visual representations was used for the annotation task, and an approach reposing on bilingual (i.e. English-French) MeSH term text-retrieval combined with visual image similarity for the retrieval task.

2 Medical Image Annotation Task

When searching for images in non-annotated databases, medical image categorization and annotation can be very useful. Generally the first step is to define the categories (i.e. classes of images) you need to "recognize" and to select a set of image prototypes for each category. Then, each new/unknown image is projected in one of the categories (using some form of image similarity), and thus annotated with the identifier of the category (or the multilingual terms that can be associated to each category). One of the major disadvantages of this approach is the dependence on manually annotated prototypes for each of the categories, especially when treating domains as vast and rich as medicine. A second weakness is its difficulty to treat "unknown" images, an additional category being simply too big to create (basically it should contain everything else). Nevertheless, medical image annotation proved to be significantly accurate when an image sub-domain is considered, and image categories are well defined.

The image annotation task of ImageCLEF 2006 is using an image dataset provided by the IRMA project, consisting of 10000 fully annotated x-rays taken randomly from medical routine and 1000 x-rays for which classification labels are not available to the participants. The images are subdivided in 116 categories (classes), hierarchically organized according to image modality (x-ray, plain radiography), body orientation, body region, and biological system captured. The annotated dataset is intended to be used for *training* (9000 images) and *validation* - system tuning by parameter optimization (1000 images) and the 1000 un-annotated images for *tests*. Each approach/run is evaluated according to the annotation of the *test* set.

Our approach is based on the supervised classification of combined both local and global texture and statistic image representations. Because the resulted number of features can be significant compared to the available data samples (raising estimation problems for the classifiers) we added a dimensionality reduction phase.

2.1 Image Resizing, Global and Local Representations

For these experiments, all images are resized to a fixed dimension of 256×256 pixels. Even though this simplifies several aspects of the problem (e.g. simpler generation of texture filters, normalization of the resulted representation size), we want to point out that loosing the original image aspect ratio introduces some structural and textural deformations, which could result in poor performances when dealing, for example, with image fragments. However, from our observations (in general) and after the examination of the IRMA database we concluded that, most of the times, images of the same category have the same aspect ratio, and thus will finally be deformed in the same way.

The image features can be extracted both globally and from local representations. The global features are extracted at the image level and have the advantage of being less sensible of small local variations, noise and/or geometrical transformations (e.g. especially rotation). However due to the importance of details in medical imaging, local features are of great importance when representing the medical image content. To take in consideration the spatial disposition of features inside the images we choose to also use a local representation obtained by splitting the original image in 16 equal sub-blocks (of 64x64 pixels). This way each image is represented by a vector of 16 blocks, and from each block features are extracted to describe its content.

2.2 Feature Extraction

The x-rays used for the annotation task are by nature acquired in gray-levels, and electronically digitized using 8bbp (8 bit per pixel; $2^8 = 256$ gray levels). This renders some of the most successfully used (i.e. for image representation) features, like the color, inapplicable here. The texture based features combined with statistical gray-level measures proved to be a well suited global descriptor for medical images [3].

For describing image TEXTURE we employed several very different approaches:

- `(COOC)` In [5] the *co-occurrence matrix* is used to explore the gray-level spatial dependence of the texture. We compute 4 co-occurrence matrixes, one on each direction (horizontal, vertical and diagonals), after a 64 gray-level quantification. From each matrix, 4 features are extracted: energy, entropy, contrast and homogeneity, producing a **16 feature** vector.

- `(DF)` [16] made the assumption that textures are fractals for a certain range of magnifications. *Fractal dimension* is not an integer in contrast to the dimension in Euclidean geometry, but a number between 2 and 3. The more the texture is smooth (respectively rough), the more the fractal dimension is close to 2 (respectively 3). We used a modified box-counting texture analysis based on the probability density function described by [7]. The computing of the fractal dimension generates **a single** feature.

- `(GB)` The belief that simple cells in the human visual cortex can be modelled by Gabor functions [11] lead to the development of texture features extracted from response to *Gabor filters* [9]. The aim is to discriminate coarse textures which have spectral energy concentrated at low spatial frequency, from fine textures which have larger concentrations at high spatial frequency. The Gabor filters computed at $\lambda = 3$ scales and $\phi = 4$ orientations, produce a 12 level decomposition from which we extract the mean and standard deviation, resulting a **24 feature** vector.

- `(DCT)` The *discrete cosine transform* is popular in image coding due to good performance and fast implementation [17]. [13, 14, 15] suggest using a 3×3 DCT for texture feature extraction. They furthermore suggest excluding the low-frequency component of the DCT, thus yielding **8 features**.

- `(RL)` Galloway has proposed a run-length-based technique, which calculates characteristic textural features from *gray-level run lengths* in different image directions [4]. A total of **14 features** is derived.

- `(Laws)` Laws has suggested a set convolution masks for feature extraction [8]. Using the *Laws filter masks for textural energy* **28 features** are resulted.

- `(MSAR)` [10] proposed the classification of color textures using *Multispectral Simultaneous Autoregressive Model (MSAR)*. The basic idea of a simultaneous autoregressive(SAR) model is to express a gray level of a pixel as a function of the gray levels in its neighborhood. The related model parameters for one image are calculated using a least squares technique and are used as textural features. This approach is similar to the Markov random fields described in [6]. From this **24 features** are resulting.

In addition we used features derived from *gray-level statistical measures* `(STAT)`: different estimations of the first order (mean, median and mode), second order (variance and l2 norm), third and forth order (skewness and kurtosis) moments, thus obtaining a **7 feature** vector.

We choose to combine these descriptors because in previous experiments, using feature selection algorithms, we pointed out their complementarity [3].

2.3 Dimensionality Reduction

A significant obstacle in machine learning problems is learning from few data samples in a high-dimensional feature space. Unfortunately, most of the time, the number of data samples is given by the context of the application and thus it is difficult to change. Furthermore, with the increase of feature space dimensionality, it becomes impossible to estimate the probability density function (PDF) with a reasonable amount of training data and (very important) computational burden.

In previous experiments [3] we used various feature selection techniques, and the best ratio between (later) classification-accuracy and dimensionality-reduction are obtained with the Principal Component Analysis (PCA). PCA is a linear transformation that transforms the data into a new coordinate system, the first new coordinate (called the first principal component) containing the projection with the greatest variance of the data (from any projection possible), the second new coordinate containing the second greatest variance and so on. The dimensionality reduction is done retaining those characteristics of the dataset that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Generally, the low-order components often contain the "most important" aspects of the data.

For the ImageCLEF Annotation submissions, the experiments are conducted choosing enough eigenvectors to account for either 95% or 97% of the variance in the original data (on the training set).

We show in [3] that some of the texture features (cooccurence, fractal dimension and Gabor wavelets) as well as the statistic features are complementary, all the ten feature selection methods used, selecting subsets of each feature set. The other texture features were considered and implemented after the experiments described in [3], but they all behaved similarly, adding a small amount of useful information to the image representation. However most of the information is redundant, imposing some form of dimensionality reduction when using several descriptors.

2.4 Classification

For the projection of the test instances (i.e. the feature representation of each test image) in corresponding categories the MedIC module uses several well known supervised classification approaches based on neural networks, decision trees, support vector methods and nearest-neighbor architectures. In previous experiments [3, 2] we noted the best performances obtained by the Support Vector Machine (SVM) classifier and the good performances/classification_time of the k-Nearest Neighbor approach. Given that in most application the classification task (especially the learning phase) are expected to be conducted off-line (therefore making time a less important issue), for the experiments of ImageCLEF 2006 we only submitted runs classified with SVM, as they are expected to be better.

The SVM classification is conducted using different parameters for: kernel = polynomial, radial basis function (RBF) and sigmoid, C = the cost parameter ($1 \rightarrow 10^4$), γ = the gamma parameter of each kernel ($10^{-1} \rightarrow 10^{-4}$) and d = the degree of the polynomial kernel ($1 \rightarrow 5$). From these, using the 1000 images validation dataset, the best parameters are selected: RBF kernel, $\gamma = 10^{-2}$, C=10^2.

2.5 Results

Twelve (12) groups participated to the Image Annotation Task of ImageCLEF 2006 and submitted a total of 27 runs. The MedIC/CISMeF group submitted four runs. The parameters used for each run are presented in Table 1. For all the four runs a combined descriptor `COOC, DF, GB, DCT, RL, Laws, MSAR, STAT` is used. This represents (16+1+24+8+14+28+24+7) = 122 features for each extracting window (according to section 2.2). The runs `local+global_PCA450` and `local+global_PCA450` are considering one global extraction window and 16 local windows, and thus have an original number of features equal to 122*17 = 2074. From these `PCA450` uses a 97% variance PCA to select 450 features, while using 95% produces 335 features. When considering

Run label	(a) Parameters					(b)		
	local	global	orig. no.feat.	PCA var	final no.feat	error valid.	error test	rank
`local+global_PCA450`	√	√	2074	97%	450	**12.3%**	17.9%	7
`local+global_PCA335`	√	√	2074	95%	335	12.7%	**17.2%**	**4**
`local_PCA333`	√	×	1952	97%	333	12.6%	17.2%	5
`local_PCA150`	√	×	1952	95%	150	15.9%	20.2%	10

Table 1: Run details

only the 16 local extraction windows, the original number of features becomes 122*16 = 1952 and the PCA 97% and 95% are producing 333, respectively 150 features.

We obtained the forth rank but we situated third in the hierarchy of groups (only `RWTHi6`[2] and `UFR`[3] obtained better scores) and also we obtained the third score (the second and third rank having the same error rate: 16.7%). The best score was obtained by the `RWTHi6` with an error rate of 16.2%. That represents an improvement of 1% compared to our best score (`local+global_PCA335`), meaning that, compared to our 828 correctly annotated test images, 10 more images (not necessarily from those we missed) are correctly annotated.

However it is interesting to note that in all our experiments conducted on the validation set, we obtained better results with on average 4.75%, reaching up to 87.7% correctly annotated validation images (12.3% error rate) with (`local+global_PCA450`). This indicates different difficulties for the validation and test sets, and it will be interesting to compare if different systems reacted in the same way.

Also, we note that equal test error rates are obtained for two runs: `local+global_PCA335` and `local_PCA333`, with reduced representations of comparable sizes. This could indicate that adding the globally extracted features is redundant and the first information to be discarded, in the case of `local+global_PCA335`, by a more compacting 95% PCA transform (the same information seems to be captured locally only and preserved with PCA 97%). An inspection of the list of miss-annotated images for each run could show exactly if this assumption is true. Using the same parameters, we obtain similar performances on the validation dataset.

3 Medical Retrieval Task

The multilingual medical image retrieval task uses an image dataset containing 50026 images from four image collections: Casimage, MIR, PEIR, and PathoPIC. Each collection contains textual annotation and case descriptions in XML format and various languages: Casimage (Fr-En), MIR (En), PEIR (En), and PathoPIC (Ge-En).

There are 30 topics for ImageCLEFmed 2006, organized in three categories (with 10 topics for each): Visual, Textual and Mixed. The categories are defined according to the type of approach the participants are expected to use on each.

For the medical image retrieval task we submitted a single run using an approach based on bilingual (i.e. English-French) MeSH term text-retrieval and visual image similarity.

3.1 MeSH dictionaries

The first step is the extraction of terms, from the textual annotations of the image collection. The MedIC terms are originally based on the French version of the MeSH thesaurus (Medical Subject Headings [4]) and they are reorganized in several image-dependent categories: image modality,

[2]Human Language Technology and Pattern Recognition Group of the RWTH Aachen University
[3]Chair of Pattern Recognition and Image Processing of the Albert-Ludwigs University of Freiburg
[4]http://www.nlm.nih.gov/mesh/

anatomical region, disease, technical acquisition parameters (i.e. view angle), image formats (i.e. JPEG, PPT). Each category has its own dictionary, and contains for each MeSH term declinations like inflected (plural) MeSH terms, synonyms of MeSH terms, inflected synonyms of MeSH terms, abbreviations, initials and others. We can observe an extract of the modalities dictionary, containing the ultrasound declinations (in French):

```
echographie,echographie.N+MeSH+TR+QMesh:fs
echographies,echographie.N+MeSH+TR+QMeSH:fp
ECHO,echographie.N+MeSH+TR+QMeSH
us,echographie.N+MeSH+TR+QMeSH
ultrasonographie,echographie.N+MeSH+TR+QMeSH:fs
ultrasonographies,echographie.N+MeSH+TR+QMeSH:fp
```

The French dictionaries were created by the CISMeF team for experiments on automatic textual indexation, of health-resources (i.e. medical documents) in French. For the experiments presented at ImageCLEF 2006 we constructed corresponding English dictionaries to be able to treat all the textual annotations (only the Casimage collection has French textual annotations). However, due to the size and complexity of this task, the English dictionaries are containing only a small part of the French terms, and all the results from non-French collections are thus seriously influenced.

Once the terms are extracted form the textual annotations, a second step is the extract the search terms from each topic. This is performed similarly as for the extraction of annotation terms. Of course that using the same incomplete English dictionaries, the extraction of search terms from non-French annotations is as well negatively influenced.

The extraction of terms is performed using the linguistic INTEX/NOOJ environment [18]. This methodology is derived from the automatic text indexing approach that CISMeF is developing [12].

3.2 Visual similarity

Once the textual annotations containing all the search terms of each topic are obtained, the relevance of each retrieved image is evaluated according to the mean similarity between each retrieved image and the two (or three) query images (of each topic). The visual similarity between two images is estimated as the L2 distance between feature representations of images. We employed some of the features presented at section 2.2: `COOC, RL, DCT` and `STAT` as well as additional color features: mean color, mean saturation and color histograms. The features are extracted from 64x64 image sub-blocks of 256×256 resized images, as for the image annotation experiments.

3.3 Results

A total of 10 groups participated at the multilingual medical image retrieval task of ImageCLEF 2006, with a number of 100 runs. The MedIC/CISMeF group submitted a single run, and was placed on 69-th position, with an 0,0531 Mean Average Precision (MAP). Comparatively, the best score was obtained by the `Image Processing & Application Lab (IPAL)`[5], with 0,3095 MAP. We expected this modest score due to several unexpected problems we experienced during the preparation of our run. The most significant problem was that due to last moment technical problems we were able to treat only ∼30% of the 50026 images. Furthermore, the dictionaries we normally use for the extraction of medical terms are in French, and their translation in English (with all the derived forms: plurals, synonyms) was very limited. Knowing that ∼82% of the images have non-French annotations, we actually expected even poorer results. The treatment of less then a third of the collection and with incomplete dictionaries is shown also by the small number of answers our run proposed (for all the 30 topics), 1114, the smallest of all the runs.

The system is conceived to be automatic, but at the last moment we chosed to manually intervene at the search term extraction phase in four of the topics: 1.1, 1.5, 1.9, 3.3. Therefore we chose to declare the whole run as manual.

[5] http://ipal.imag.fr/

4 Conclusion

In this paper we present the methods we used for the ImageCLEF 2006 evaluation. We participated in the medical tasks: the automatic annotation task, where we obtained the fourth rank (also the third score and the third placed group), and the multilingual medical image retrieval task, where our run was significantly less competitive, due technical an practical problems we were not able to overcome until the last moment.

The results obtained in the annotation task shows that the approach we propose is capable of obtaining very competitive results. A comparison between the correctly annotated images of different systems could be interesting, and could indicate how to combine different architectures to further improve the classification/annotation results.

References

[1] S.J. Darmoni, J.P. Leroy, B. Thirion, F. Baudic, M. Douyére, and J. Piot. Cismef: a structured health resource guide. *Meth Inf Med*, 39(1):30–35, 2000.

[2] F Florea, H Müller, A Rogozan, A Geissbuhler, and S Darmoni. Medical image categorization with medic and medgift. In *submitted to Medical Inforamtics Europe (MIE)*, 2006.

[3] F Florea, A Rogozan, A Bensrhair, and SJ Darmoni. Comparison of feature-selection and classification techniques for medical image modality categorization. In Faculty of Electrical Engineering The Transilvania University of Brasov and Computer Science, editors, *accepted at 10th IEEE International Conference on Optimization of Electrical and Electronic Equipment (OPTIM2006), Special Session on Image Processing - Technical and Medical Applications*, volume 4, pages 161–168, Brasov, Romania, May 18-19 2006.

[4] M. M. Galloway. Texture analysis using graylevel runlengths. *Computer, Graphics and Image Processing*, 4:172–179, 1975.

[5] R. M. Haralick, K. Shanmugam, and I. Dinstein. Texture features for image classification. *IEEE Trans. Systems, Mans and Cybernetics*, SMC-3:610–621, 1973.

[6] R. L. Kashyap, R. Chellappa, and A Khotanzad. Texture classification using features derived from random field models. *Pattern Recognition Letters*, 1:43–50, 1982.

[7] J.M. Keller, S. Chen, and R.M. Crownover. Texture description and segmentation through fractal geometry. *CVGIP*, 45:150–166, 1989.

[8] K.I. Laws. *Textured Image Segmentation*. PhD thesis, University of Southern Califomia School of Engineering, 1980.

[9] T.S. Lee. Image representation using 2d gabor wavelets. *IEEE Trans. on PAMI*, 18(10):1–13, october 1996.

[10] J Mao and A.K Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 5(2):173–188, February 1992. ISSN:0031-3203.

[11] S. Marcelja. Mathematical description of the response of simple cortical cells. *J. Optical Soc. Am.*, 70:1297–1300, 1980.

[12] A. Nèvèol, A. Rogozan, and S.J. Darmoni. Automatic indexing of health resources in french with a controlled vocabulary for the cismef catalogue: a preliminary study. *Medinfo*, 2004.

[13] I. Ng, T. Tan, and J. V. Kittler. On local linear transform and gabor filter representation of texture. In *International Conference on Pattern Recognition*, pages 627–631, 1992.

[14] C. W. Ngo. Exploiting image indexing techniques in DCT domain. In *APR International Workshop on Multimedia Information Analysis and Retrieval*, pages 196–206, juin 1998.

[15] Chong-Wah Ngo, Ting-Chuen Pong, and Roland T. Chin. Exploiting image indexing techniques in DCT domain. *Pattern Recognition*, 34(9):1841–1851, 2001.

[16] A.P. Pentland. Fractal-based descriptors of natural scenes. *IEEE Trans on PAMI*, 6(6):661–674, 1984.

[17] Tor Audun Ramstad, Sven Ole Aase, and John Håkon Husøy. *Subband Compression of Images – Principles and Examples*. ELSEVIER Science Publishers BV, North Holland, 1995.

[18] M Silberztein. *Dictionnaires électroniques et analyse automatique de textes: le syst ème INTEX*. Masson, Paris, 1993.