Automatic Detection of Body Parts in X-Ray Images

Vincent Jeanne and Devrim Unay
Philips Research Laboratories
Eindhoven, The Netherlands
vincent.jeanne@philips.com,unay@sabanciuniv.edu

Vincent Jacquet
ISEN
Brest, France
vincent.jacquet1@gmail.com

Abstract

The number of digital images that needs to be acquired, analyzed, classified, stored and retrieved in the medical centers is exponentially growing with the advances in medical imaging technology. Accordingly, medical image classification and retrieval has become a popular topic in the recent years. Despite many projects focusing on this problem, proposed solutions are still far from being sufficiently accurate for real-life implementations.

Interpreting medical image classification and retrieval as a multi-class classification task, in this work, we investigate the performance of five different feature types in a SVM-based learning framework for classification of human body X-Ray images into classes corresponding to body parts.

Our comprehensive experiments show that four conventional feature types provide performances comparable to the literature with low per-class accuracies, whereas local binary patterns produce not only very good global accuracy but also good class-specific accuracies with respect to the features used in the literature.

1. Introduction

Storing, archiving and sharing patient information among hospitals has become a very important topic for tomorrow's medical field. Companies but also governments are now looking forward to building patient centric IT systems like targeted by the Ratu e-Health Project of Northern Finland which is aimed at building a national electronic patient records archive that is estimated to manage 550 petabytes of data by 2025 [1].

According to the usual saying "To be seen by a doctor", digital images represent a huge part of the data that need to be stored in medical centers. Those images can, for example originate from standard radiography (X-Ray), magnetic resonance imaging (MRI) computer tomography (CT), endoscopy, ultrasonography, etc. In hospitals a dramatic number of those images have to be managed each year in order to be stored in an archiving system. Manual labeling of this data cannot be handled by a single person. Indeed, such a task is very time consuming and, moreover, it can easily lead to ambiguities due to different observations made by each individual. In order to realize an accurate classification we need to develop tools that allow high performance automatic image annotation, where, without any user input, a given image is automatically labeled with a describing text, or code.

Several attempts for automatic classification in the field of medical images have been performed in the past. For example, we can mention the WebMRIS system [6], aiming at retrieving cervical spinal X-Ray images, or the ASSERT system [11], aiming at retrieving CT images of lung. While these projects only aim at retrieving a specific body part, other initiatives have been proposed in order to retrieve multiple body parts.

The IRMA project [2](Image Retrieval in Medical Application), from the Aachen University of Technology (RWTH Aachen), is aiming at developing high-level methods for content-based image retrieval with prototypical application to medico-diagnostic tasks on a radiologic image archive. One of the goals of this project is the automated classification of X-Ray images with respect to the body region examined. To achieve this, the IRMA project has built a database containing more than 10,000 images that were arbitrarily chosen. The images represent body parts of different people from different ages, of different genders, under different viewing angles and with or without pathologies.

In order to build a powerful classification system, the image data has to be translated into a more compact and more manageable representation containing only the relevant features. Several feature representations have been investigated in the past for such a classification task. Among others, image features, such as average value over the complete image or blocks within this image, have been investigated by Rahman et. al in [10]. Or as proposed by Mueen et. al

[1] http://pre20090115.stm.fi/pr1105954774022/passthru.pdf

[2] http://libra.imib.rwth-aachen.de/irma

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

in [7] , one can also consider the usage of color features, such as color histograms. Because many medical images can easily be distinguished by their texture, texture features like gray-level co-occurrence matrices should also be considered as described Rahman et. al, which also mentions shape features to describe the images using the MPEG-7 edge histogram descriptor.

Motivated by the considerations above, we present in this paper the performance of four conventional feature types and the performance of the *Local Binary Patterns* (LBP) [9] applied to the IRMA images dataset.

The paper is organized as follows. In Section 2, we first introduce the image database on which the features extraction is performed. Then the five features representation for the generation of the classifier's input are discussed in details. Next, in Section 3, the classification technique we use is presented, namely *Support Vector Machines* [2]. In Section 4, the performed experiments and the obtained results are described. Finally, Section 5 presents our conclusions.

2. Method

2.1. Database

Since 2005, the automatic medical image annotation task of ImageCLEF -Cross Language Evaluation Forum- compares state-of-the-art approaches for automatic image annotation and classification.

In order to facilitate the comparison with [10], we also use the IRMA database. It consists of medical radiographs taken randomly from medical routine at the RWTH Aachen University Hospital and subdivided into 116 classes. Figure 1 displays example images from this database along with class numbers. Note that the classes in this database are not uniformly distributed (Figure 2): while the most populated class 111 holds 19.3% of the images, 6 classes have only 0.1% or less. This non-uniform distribution of images among the classes is actually done on purpose; indeed in the clinical settings one can expect that some radiographs are acquired more often than others (for e.g, more thorax X-Rays than single finger images).

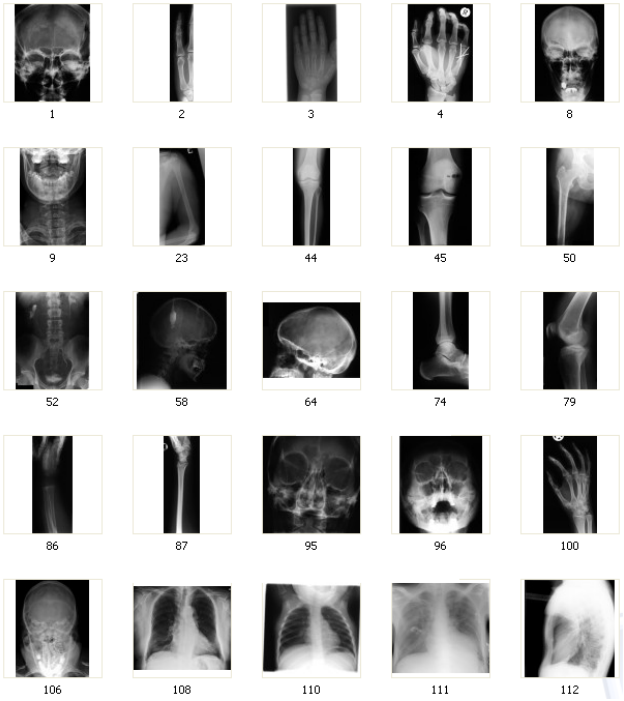

Figure 1. Exemplary images from the IRMA database and their corresponding class numbers.

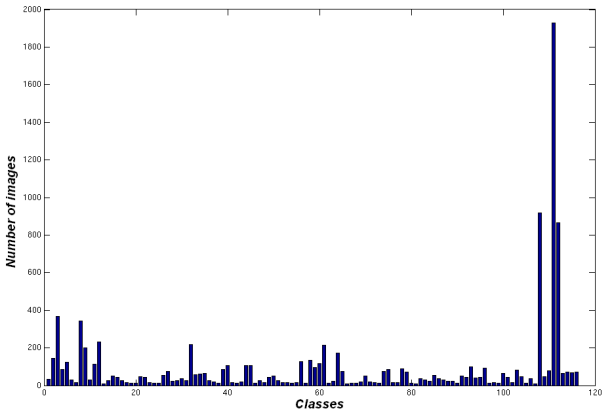

Figure 2. Distribution of images in the IRMA database.

2.2. Image Description

In this part we introduce the descriptors that characterize the content of the image relative to luminance, texture and shape. Please note that not all descriptors we present are invariant to translation and rotation. However as the database is nicely oriented and centered on the body part, this issue is not problematic for our application.

2.2.1 Average Gray Descriptor (AGD)

As gray level values are an indication of texture in radiographs, we use the *Average Gray Descriptor* (AGD) value over blocks to capture this information. As a measure of luminance, in AGD we resize images down to 64×64 and 128×128 pixels, equally divide them into non-overlapping blocks, compute the average gray value of each block and concatenate these values into a feature vector. Figure 3 shows the AGD performed using 4×4 block sizes.

2.2.2 Color Layout Descriptor (CLD)

The *Color Layout Descriptor* (CLD) [5] is a color feature description for images. This descriptor specifies the spatial

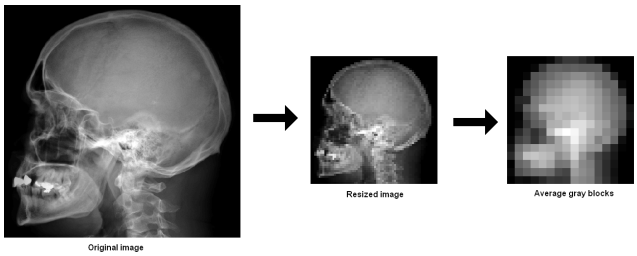

Figure 3. Example of average gray computation with 4×4 blocks from an image resized to 64×64 pixels.

distribution of colors with a few nonlinear quantized *Discrete cosine transform*(DCT) coefficients of grid-based average colors. We can use this feature for a gray level image in order to describe the luminance. We compute a two dimensional DCT. The two-dimensional transform is equivalent to a one-dimensional DCT performed along a single dimension followed by a one-dimensional DCT in the other dimension. In the following formula, A represents the input image.

$$DCT_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N},$$

where :

$$\alpha_p = \begin{cases} 1/\sqrt{M}, & p = 0 \\ \sqrt{2/M}, & 1 \leq p \leq M-1 \end{cases}$$

$$\alpha_q = \begin{cases} 1/\sqrt{N}, & q = 0 \\ \sqrt{2/N}, & 1 \leq q \leq N-1 \end{cases}$$

To obtain more than one feature per image, we compute the DCT on sub-areas. For example the image can be divided in non-overlapping blocks. Figure 4 represents the low frequency values of CLD captured with different block sizes.

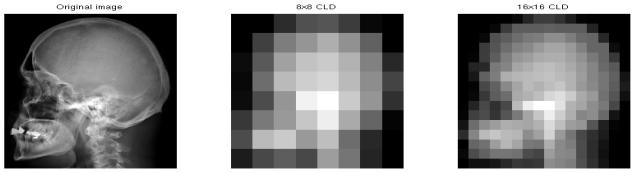

Figure 4. Example of CLD : Input image and the output image for 8×8 and 16×16 sized blocks.

2.2.3 Edge Histogram Descriptor (EHD)

The essential idea of the *Edge Histogram Descriptor* (EHD) descriptors is that local object appearance and shape can be described rather well by the distribution of local intensity gradients or edge directions. The method uses local histograms of image gradient orientations as image features. In practice the EHD features are extracted by dividing the image into subparts and for each subpart a local 1-D histogram of gradient directions over the pixels is computed. Typically, the contributions to the histogram are weighted by the gradient magnitude. The combined histograms then form the object representation as a feature vector.

2.2.4 Gray-level Co-occurrence Matrix (GLCM)

Gray-level Co-occurrence Matrix (GLCM) is defined as a sample of the joint probability density of the gray levels of two pixels separated by a given displacement. This co-occurrence of values is computed using two distance levels between pixels and in four orientations (Figure 5). For each distance level, results from all orientations are averaged to form the averaged GLCM matrix. We then extract the eleven features proposed by Haralick et. al in [3] from the averaged GLCM matrix and use their statistical properties as descriptors.

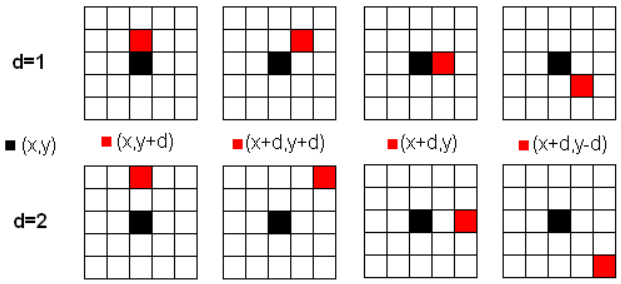

Figure 5. Distance and orientation for GLCM computation.

2.2.5 Local Binary Patterns (LBP)

LBP [9] is a gray-scale invariant local texture descriptor with low computational complexity. The LBP operator labels image pixels by thresholding a neighborhood of each pixel with the center value and considering the results as a binary number. The neighborhood is formed by a symmetric neighbor set of *P* pixels on a circle of radius *R*. Formally, given a pixel at (x_c,y_c), the resulting LBP code can be expressed in the decimal form as follows,

$$LBP_{P,R}(x_c, y_c) = \sum_{n=0}^{P-1} s(i_n - i_c) 2^n$$

where *n* runs over the *P* neighbors of the central pixel, i_c and i_n are the gray-level values of the central pixel and the neighbor pixel, and *s(x)* is 1 if $x \geq 0$ and 0 otherwise.

After labeling an image with a LBP operator, a histogram of the labeled image $f_l(x, y)$ can be defined as

$$H_i = \sum_{x,y} I(f_l(x, y) = i), \qquad i = 0, \ldots, L-1 \qquad (1)$$

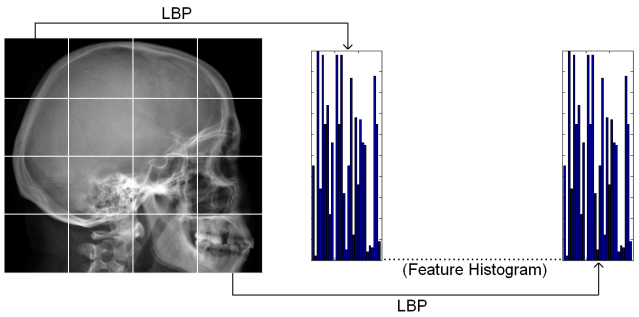

Figure 6. The image is divided into small regions from which LBP histograms are extracted and concatenated into a single, spatially enhanced histogram.

where L is the number of different labels produced by the LBP operator and

$$I(A) = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 0, & \text{if } A \text{ is false} \end{cases} \quad (2)$$

The derived LBP histogram contains information about the distribution of local micro-patterns, such as edges, spots and flat areas, over the image, so can be used to statistically describe image characteristics. As mentioned by Ojala et. al in [9], not all LBP codes are informative, therefore we use the uniform version of LBP and reduce the number of informative codes from 256 to 59 (58 informative bins + 1 bin for noisy patterns). Finally, similar to the face recognition work of Ahonen et. al [1], to obtain a more local description of the image we divide the images into small equal-sized non-overlapping regions and concatenate the LBP histograms extracted from each region into a single (Figure 6).

3. Learning Algorithm: Support Vector Machines

SVM is a popular machine learning algorithm that provides good results for general classification tasks in the computer vision and medical domains: e.g. nine of the ten best models in ImageCLEFmed 2006 competition were based on SVM [8].

In a nutshell, SVM maps data to a higher-dimensional space using kernel functions and performs linear discrimination in that space by simultaneously minimizing the classification error and maximizing the geometric margin between the classes. Figure 7 illustrates this classification approach for a two-class case.

Among all available kernel functions for data mapping in SVM, Gaussian radial basis function (RBF) is the most popular choice, and therefore it is used here.

$$RBF : K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2), \gamma > 0$$

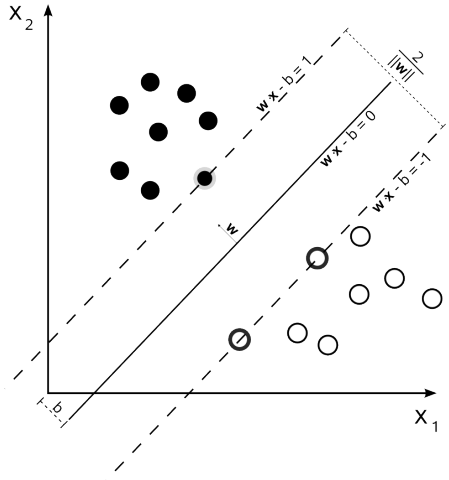

Figure 7. SVM trained on two classes. Samples on the margin are called the support vectors.

where γ is a parameter defined by the user. Besides γ, there exists an error cost C that controls the trade-off between allowing training errors and forcing rigid margins. An optimum C value,that should also be explored by the user, creates a soft margin while permitting some misclassifications. In this work we used a Matlab implementation of SVM, namely the Spider toolbox (version 1.71) [3]. In our framework we use a RBF kernel and optimum values C and γ that we found empirically on the IRMA dataset.

4. Evaluation

4.1. Experiments

According to Hsu et. al in [4], the extracted features need to be normalized before being introduced to SVM in order to avoid domination of attributes with greater numeric ranges over small ones. Accordingly we linearly scale each feature to [-1,+1] range.

The evaluation of the SVM-based learning is performed using cross validation, which is the statistical practice of partitioning a sample into subsets such that the analysis is initially performed on a single subset, while the other subsets are retained for subsequent use in confirming and validating the initial analysis. In our case, we use five-fold cross validation, where the database is partitioned into five subsets, each subset is used once for testing while the rest are used for training, and the final result is assigned as the average of the five validations. Note that for each validation all classes were equally divided among the folds.

The measures we will use to compare the image classification results are the True Positive (TP) and False Positive (FP) rates, as well as the F-Measure defined as follows :

[3] http://www.kyb.tuebingen.mpg.de/bs/people/spider/

$$\textbf{F-measure} = \frac{2 \times Precision * Recall}{Precision + Recall}$$

Note that higher F-measure scores depict more accurate results.

4.2. Optimal Features Parameters

Based on experiments we conducted during the feature investigation process, we observed that the following parameters allowed the best performance per feature type:

AGD : non-overlapping blocks of 4×4 pixels on a resized version of the input image (128×128 pixels).

CLD : computed on the original image using non-overlapping blocks of 32×32 pixels.

EHD : The edge descriptors is computed over 16 non-overlapping blocks of identical size. The orientation is quantized over 8 orientations on a $0°$-$360°$ range.

GLCM : The Gray Level Co-Occurrence Matrix is computed over five equal-sized sub-parts of the image using a distance of 6 pixels.

LBP : The LBP histograms are computed on 16 equally sized image sub-part using the uniform version of this feature.

5. Results

Here, we present our results at a global level, meaning the performance on the complete database, but also at a class-specific level, where we average the classification rates of all classes evaluated separately. We follow this procedure, because as mentioned earlier the IRMA database is very unbalanced, and therefore a very good performance on the major classes can hide a lack of accuracy on the less predominant ones.

As presented in Table 1, we can see that using the conventional features (EHD, GLCM, CLD and Average gray) we achieve results comparable to those reported in the literature. It is also interesting to notice that the Local Binary Patterns clearly outperform the other types of features. In average, they provide an improvement of at least 10% in the true positive rate without hampering the false positive rate.

This big difference in performance comes from the fact that LBP operator is invariant to illumination change, and therefore is more robust on an image set where exposure varies noticeably. Moreover, the uniform version of this operator is less sensitive to noise in the image because by default it discards non-informative patterns from the feature vector.

	global accuracy %		
	TP	FP	F-measure
AGD	80.7	0.2	89.3
CLD	79.0	0.2	88.2
EHD	77.9	0.2	87.5
GLCM	79.6	0.2	88.5
LBP	89.1	0.1	94.2

Table 1. Global accuracies on the IRMA database.

At the class level, we can see in Table 2 that even if the performance at the global level is acceptable for the four conventional features, the classification is not accurate enough in terms of the true positive rate per class. Indeed, while at the global level those features present a true positive rate above 75% they drop to slightly above 50% when considering their performances at the class level. As a consequence many small classes have an error rate of 100%, but due to their small amount of images their impact on the overall performance is not visible.

On the contrary, we can clearly see that the LBP operator provides significant improvement compared to the previously mentioned features. Indeed, while we observe a decrease of at least 30% in the true positive rates of other features when comparing global vs. class-specific accuracies, this decrease is only 15% for the LBP operator. As a consequence, LBP exhibits clearly high F-measure figures with respect to the others both at the global and class-specific levels.

	class-wise accuracy %		
	TP	FP	F-measure
AGD	54.8	0.2	70.8
CLD	52.9	0.2	69.1
EHD	47.6	0.2	64.4
GLCM	53.5	0.2	69.6
LBP	73.5	0.1	84.7

Table 2. Class-specific accuracies on the IRMA database.

6. Conclusion

Due to the high number of digital medical images routinely acquired in the medical centers, automated classification and retrieval of these images has become a popular research area. In this paper we introduced a classification work where the aim is to automatically detect body parts

from X-Ray images. For this purpose, we employed four conventional feature types and the recently popular local binary patterns within a SVM-based learning framework. Our comprehensive experiments on the IRMA database revealed that local binary patterns not only outperform other feature types in global and class-specific accuracies, but also provide performance values better than the features proposed in the literature.

7. Acknowledgement

The IRMA database was provide with the courtesy of the RWTH Aachen University, Germany.

References

[1] T. Ahonen, A. Hadid, and M. Pietikãinen. *Face Recognition with Local Binary Patterns*, pages 469–481. 2004.

[2] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[3] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 3(6):610–621, 1973.

[4] C. W. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification, 2000.

[5] E. Kasutani and A. Yamada. The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 674–677 vol.1, 2001.

[6] L. R. Long, S. R. Pillemer, R. C. Lawrence, G.-H. Goh, L. Neve, and G. R. Thoma. WebMIRS: web-based medical information retrieval system. In I. K. Sethi and R. C. Jain, editors, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 3312 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 392–403, Dec. 1997.

[7] A. Mueen, M. Sapiyan Baba, and R. Zainuddin. Multilevel feature extraction and x-ray image classification, 2007. Journal of Applied Sciences 7 (8):1224-1229.

[8] H. Müller, T. Deselaers, T. Deserno, P. Clough, E. Kim, and W. Hersh. Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. pages 595–608. 2007.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[10] M. Rahman, B. C. Desai, and P. Bhattacharya. Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. *Computerized Medical Imaging and Graphics*, 32(2):95–108, March 2008.

[11] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick. Assert: a physician-in-the-loop content-based retrieval system for hrct image databases. *Comput. Vis. Image Underst.*, 75(1-2):111–132, 1999.