Pattern Recognition Letters 29 (2008) 2011–2017

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Hierarchical classification using a frequency-based weighting and simple visual features

Xin Zhou [a,*], Adrien Depeursinge [a], Henning Müller [a,b]

[a] Service of Medical Informatics, Geneva University Hospitals and University of Geneva, 24, Rue Micheli-du-Crest, 1211 Geneva 11, Switzerland
[b] University of Applied Sciences, Sierre, Switzerland

ARTICLE INFO

Article hisotry:
Available online 22 April 2008

Keywords:
Image retrieval
Hierarchical image classification
Automatic annotation
Medical imaging

ABSTRACT

This article describes the use of a frequency-based weighting scheme using low level visual features developed for image retrieval to perform a hierarchical classification of medical images. The techniques are based on a classical *tf*/*idf* (term frequency, inverse document frequency) weighting scheme of the *GIFT* (GNU Image Finding Tool), and perform classification based on kNN (*k*-Nearest Neighbors) and voting-based approaches. The features used by the GIFT are very simple giving a global description of the images and local information on fixed regions both for colors and textures. We reused a similar technique as in previous years of ImageCLEF to have a baseline for the retrieval performance over the three years of the medical image annotation task. This allows showing the clear increase in quality of participating research systems over the years.

Subsequently, we optimized the retrieval results based on the simple technology used by varying the feature space, the classification method (varying number of neighbors, various voting schemes) and by adding new information such as aspect ratio, which has shown to work well in the past. The results show that the techniques we use have several problems that could not be fully solved through the applied optimizations. Still, optimizations improved results enormously from an error value of 228 to below 150. As a baseline to show the progress of techniques over the years it also works well. Aspect ratio shows to be an important factor to improve results. Performing classification on an axis level performs better than using the entire hierarchy code or not taking hierarchy into account at all. To further improve results, the use of more suitable visual features such as patch histograms or salient point features seems necessary. Small distortions of images of the same class have to be taken into account for very good results. Still, without using any learning technique and high level visual features, the approach performs reasonably well.

1. Introduction

Medical images are an extremely important part of the diagnosis process in medical institutions. As most hospitals now have computerized patient records and fully digitized image production, new possibilities arise for management of data and the extraction of information from the stored data (Müller et al., 2004a; Tagare et al., 1997; Vannier et al., 2002). At the same time of images becoming digital, the number of images produced and their complexity has increased strongly. The Geneva University Hospitals radiology department alone produced over 70,000 images per day in 2007 (Müller et al., 2007) and these numbers continue to rise.

In other domains, content-based image retrieval has been used for many years to manage the growing amount of visual data (Datta et al., 2008; Smeulders et al., 2000; Kato, 1992; Rui et al., 1999). While early approaches used fairly low level features such as global color distributions and texture characteristics (Niblack et al., 1993), more modern systems rather use local features either gained through segmentation (Winter and Nastar, 1999) or in the form of salient points and their relations (Fergus et al., 2004; Tommasi et al., 2007). The latter obtained the best result in ImageCLEF 2007.

Object recognition in images has been another active research area to extract important information from potentially non-annotated images (Everingham et al., 2006; Pinz, 2005). In the medical domain, similar approaches have been used for medical image classification to extract information from these images (Lehmann et al., 2005). The dataset of the IRMA project (Image Retrieval in Medical Applications) is also used in the ImageCLEF[1] benchmark, of which a participation is described in this article. Many of the techniques for image retrieval and for image classification are similar but

* Corresponding author. Fax: +41 22 372 8879.
E-mail addresses: Xin.Zhou@sim.hcuge.ch (X. Zhou), Adrien.Depeursinge@sim.hcuge.ch (A. Depeursinge), Henning.Mueller@sim.hcuge.ch (H. Müller).

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

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.04.004

whereas for classification, a finite number of classes is regarded and training data are often available, for information retrieval applications, the number of classes occurring in the dataset is often unknown and training data are rarely available.

Several steps can generally be tuned to optimize the final performance.

- Image pre-processing such as segmentation (Antani et al., 2004), normalization of gray levels, or background removal (Müller et al., 2005).
- Extraction of domain-specific visual features (Müller et al., 2004b).
- Optimization of the distance measure or weighting scheme to determine distances between elements.
- Application of a learning strategy (such as Support Vector Machines) (Qiu, 2006).

In our approach, we do not take into account any pre-processing and neither any learning strategy. Efforts are concentrated on the optimization of the feature space and particularly on a classification strategy with our simple features to test the limits of our retrieval engine, the GIFT.[2] This cannot rival in performance with more modern approaches particularly for learning/classification such as the use of Support Vector Machines (SVMs) (Chapelle et al., 2002) or salient point-based visual features (Tommasi et al., 2007).

More on the ImageCLEFmed benchmark, the corresponding classification setup, error calculation, and the other participating techniques can be read in (Deselaers et al., 2008).

In Section 2, the methods of our approach are explained in detail. Section 3 presents the results obtained with these methods. In the last section, we critically interpret our results and present the conclusions of this article.

2. Methods

This section describes the data used and the techniques employed.

2.1. Database and task description

We use the dataset of the ImageCLEFmed 2007 automatic classification task containing in total 10,000 training images, 1,000 validation images and 1000 test images. The 1000 test images had to be classified according to the full IRMA code (Lehmann et al., 2003), which is a mono-hierarchical code with four distinct axes (image modality, anatomic region, biosystem under examination, and the body orientation all have their own hierarchy). Classification was allowed to stop at any level of the hierarchy within any of the axes. Non-classified hierarchy levels were regarded as better than incorrectly classified parts to force participants to think about measures of confidence in the classification strategy. A single image can be classified completely incorrectly (error value equal to 1), completely correctly (error value equal to 0) or partly incorrectly (error value between 0 and 1). The maximum error value can be obtained when all the 1000 test images are incorrectly classified, equaling 1000. If all the images are classified as "unknown" the total error value equals 500. A short explanation of this error value calculation is detailed in.[3] More information about the system setup and the error scoring methodology can be found in (Deselaers et al., 2008).

[2] http://www.gnu.org/software/gift/.

[3] http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef07/hierarchical.pdf.

2.2. Technical description

The techniques used for visual similarity calculation are mainly those used in the *GIFT* system (Squire et al., 2000). This tool is open source and can be used by other participants of ImageCLEF as well, so all results are reproducible. The image classification is processed in four steps:

(1) indexation of the entire image database with visual features (including the images to be classified);
(2) execution of queries with images to be classified to get similar images with known classification;
(3) re-ordering of the similar images with additional features;
(4) classification of the query image based on the list of similar images and their classes.

Varying parameters were used in steps 1, 3, and 4 to obtain improvement. Several gray level quantizations were used in the indexation step. Varying weights were attributed to the additional features (mainly aspect ratio). These two parts were already studied for a similar task in 2006 (Gass et al., 2007), so this paper investigates rather the effect of varying classification strategies.

2.2.1. Visual features

The four distinct visual feature sets used by *GIFT* are:

- Local color features at different scales by partitioning the images successively into four equally sized regions (four times) and taking the mode color of each region as a binary descriptor.
- Global color features in the form of a color histogram, compared by a simple histogram intersection.
- Local texture features by partitioning the image as before and applying Gabor filters in various scales and directions, quantized into 10 strengths (where the lowest band can be discarded).
- Global texture features represented as a simple histogram of responses of the local Gabor filters in various directions and scales.

The color histogram is originally based on the HSV (Hue, Saturation, Value) color space. Gray levels are added in a varying number as the entire database contains no color images. The texture feature space is based on two parameters: the number of directions and the scale of the Gabor filters. A more detailed description of the *GIFT* feature set can be found in (Squire et al., 1999).

Based on the results from 2006, a varying number of gray levels (4,8,16,32) were tested in this paper. Together with HSV values of (9,3,3), this results in a total of 60,833 possible features descriptors, most of them of binary nature. A large part of this feature space is unpopulated as the database contains only gray scale images and no color features are thus possible. A normal image contains around 1000 of these features but the numbers can vary depending on the amount of texture and the number of gray levels present.

2.2.2. Feature weighting

A particularity of *GIFT* is that it uses many techniques well–known from text retrieval. Visual features are quantized and the distributions of the features are fairly similar to those of words in texts (sparsely populated spaces). A simple *tf*/*idf* weighting is used and the query weights are normalized by the results of the query itself. The features using histograms are compared based on a simple histogram intersection (Swain and Ballard, 1991). The four feature groups are combined in normalized form with an equal weight. Feature groups can also be used directly without separate normalization leading to significantly worse results. This