Pattern Recognition Letters 29 (2008) 2032–2038

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Effectiveness of global features for automatic medical image classification and retrieval – The experiences of OHSU at ImageCLEFmed

Jayashree Kalpathy-Cramer *, William Hersh

Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, United States

ARTICLE INFO

Article history:
Available online 11 July 2008

Keywords:
Automatic image annotation
Medical images
Benchmark
Image classification
Neural networks

ABSTRACT

In 2006 and 2007, Oregon Health and Science University (OHSU) participated in the automatic image annotation task for medical images at ImageCLEF, an annual international benchmarking event that is part of the cross language evaluation forum (CLEF). The goal of the automatic annotation task was to classify 1000 test images based on the image retrieval in medical applications (IRMA) code, given a set of 10,000 training images. There were 116 distinct classes in 2006 and 2007. We evaluated the efficacy of a variety of primarily global features for this classification task. These included features based on histograms, gray level correlation matrices and the gist technique. A multitude of classifiers including *k*-nearest neighbors, two-level neural networks, support vector machines, and maximum likelihood classifiers were evaluated. Our official error rates for the 1000 test images were 26% in 2006 using the flat classification structure. The error count in 2007 was 67.8 using the hierarchical classification error computation based on the IRMA code in 2007. Confusion matrices as well as clustering experiments were used to identify visually similar classes. The use of the IRMA code did not help us in the classification task as the semantic hierarchy of the IRMA classes did not correspond well with the hierarchy based on clustering of image features that we used. Our most frequent misclassification errors were along the view axis. Subsequent experiments based on a two-stage classification system decreased our error rate to 19.8% for the 2006 dataset and our error count to 55.4 for the 2007 data.

1. Introduction

Advances in digital imaging technologies and the increasing prevalence of picture archival and communication systems (PACS) have led to a substantial growth in the number of digital images generated and stored in hospitals and medical systems in recent years. On-line atlases of images have been created for many medical domains including dermatology, radiology and gastroenterology (Aisen and Broderick, 2003; Schmid-Saugeon and Guillod, 2003; http://mypacs.net; http://www.visualdx.com). As the number of images being generated and archived everyday increases, the ability to search and retrieve relevant images becomes a critical and challenging task. Medical and other image retrieval systems have historically relied on the annotation or captions associated with the images for indexing the retrieval system. The labor-intensive task of indexing and cataloging the images in these collections is typically performed manually, a process that can to be subjective and prone to errors.

Medical images typically store information including the imaging modality and anatomical location in the DICOM header (http://medical.nema.org/). This information can be useful for image retrieval when the query provides information about the desired imaging modality or anatomical location. However, this information is often lost when the images are compressed as JPEG images or stored in teaching and on-line collections. There have also been reported errors (Güld et al., 2002) that can occur in the DICOM headers.

Thus, the ability to automatically categorize and annotate medical images with information can be very useful for the purposes of image management and retrieval. Imaging modality, anatomical location, biological system and view are four possible dimensions along which medical images can be categorized (Deselaers et al., 2007b; Lehmann et al., 2003; Müller and Clough, 2005). Content-based image retrieval (CBIR), where the image itself is used in the query to find similar images, has been an area of interest to researchers in the last decade (Smeulders and Worring, 2000; Tagare and Jaffe, 1997). The algorithms developed in the field of computer vision and CBIR can also be applied to the task of image categorization and annotation. In recent years, researchers have been investigating methods to classify images based on visual appearance and annotate the image with the purported category (Florea et al., 2006; Müller et al., 2004). This information can be

* Corresponding author. Tel.: +1 5036576560; fax: +1 5034944551.
E-mail addresses: kalpathy@ohsu.edu, jayashreekalpathy@yahoo.com (J. Kalpathy-Cramer).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.05.013

used in addition to other textual annotations or captions to improve the performance of retrieval engines (Hersh et al., 2006; Hersh and Kalpathy-Cramer, in press; Kalpathy-Cramer and Hersh, 2007).

1.1. ImageCLEFmed and the automatic annotation task

The ImageCLEF medical image retrieval and automatic annotation tasks (ImageCLEFmed) (http://www.imageclef.org) provide test collections for the medical image retrieval community to benchmark their algorithms. Since 2003, the ImageCLEF campaign has been a part of the cross language evaluation forum (CLEF) (Braschler and Peters, 2004; Hersh and Muller, 2006; http://www.clef-campaign.org) which is an offshoot from the Text Retrieval Conference (TREC, trec.nist.gov).

In 2006, the goal of the automatic annotation task within ImageCLEFmed was to correctly classify 1000 radiographic medical images into 116 categories (Müller et al., 2006; http://www.imageclef.org). These anonymized images, selected from the image collection of the Department of Diagnostic Radiology, Aachen University of Technology (RWTH), Aachen, Germany, differed in the "modality, body orientation, body region, and biological system examined", according to the track Web site (http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef06/medicalaat.html). The task organizers provided a set of 9000 training images for which the class membership was provided.

In addition, a set of 1000 classified images was provided as a development set. The suggested procedure was to create a classifier based on the training images. The development set could then be used to test the effectiveness of the classifier. One could then combine the training and development tests to create a larger database to create the final classifier for the test images.

In 2007, although the 116 classes remained the same, the goal was to correctly classify a similar set of 1000 radiographic medical images using the hierarchical IRMA code (Deselaers et al.,2007a,b; Müller et al., submitted for publication). This code classifies the image along the modality, body orientation, body region, and biological system axes. The goal of the task was to classify the images to the most precise level possible, with a greater penalty applied for incorrect classification than for a less specific classification within the hierarchy. The task organizers provided a set of 10,000 training images and 1000 development images.

2. System description

Our overall goal in the long term is to annotate all images in our medical image retrieval system (Hersh and Kalpathy-Cramer, in press; Kalpathy-Cramer and Hersh, 2007) with the anatomical location and view of the images. The ImageCLEF automatic annotation task offered us the opportunity to create a fairly simple but extensible automatic image annotation system as it provided us with a large training dataset of well-annotated images. Since the images in our image retrieval system are extremely varied in imaging modality, anatomical location and quality, we were primarily interested in exploring the ability of low level, global features for automatic annotation and categorization.

We followed the steps described below in creating the classifier for the ImageCLEF med automatic annotation task using MATLAB (http://www.mathworks.com) as our primary development environment:

1. A variety of feature vectors were created for all images in the training, development and test images.
2. A 116-class classifier was developed using primarily open source toolboxes.
3. The classifier was optimized using the development set of images.
4. This optimal classifier was applied to the set of test images.
5. In 2007, the class code was converted to the IRMA code required for submission.

2.1. Features vectors

The images provided were resized such that at least one dimension was 512 pixels. We further padded the images to create a 512×512 image, with the original image centered within this new image. White (255) and black (0) pixels were tested for the padding. This was done since we had noted that the aspect ratio of the image could provide useful information for classification. For instance, the aspect ratios of most images of arm and leg bones are significantly different than the aspect ratio of lungs. We chose to pad the images instead of either extracting features directly from the resized image or from images resized to square dimensions as early experiments had shown that maintaining an indirect measure of the aspect ratio helped in the classification.

All images were then resized to 256×256 pixels using the bilinear interpolation algorithm provided in the MATLAB image processing toolbox (http://www.mathworks.com).

For our first efforts in the medical image automatic annotation domain in 2006, we started with low-level, commonly used, global, texture and histogram features. In addition, we tried to capture a sense of spatial differences between images classes by tiling the image and creating feature vectors for the overlapping tiles.

A variety of features described below were tested on the development set. These features were combined in different ways to try to improve the classification ability of the system, with the final submissions were based on the three best combinations of image features. The features included:

- *Icon*: A 16×16 pixel 'icon' of the image was created by resizing the image using bilinear interpolation. This resulted in a 256-dimensional feature vector.
- *GLCM*: Four gray level co-occurrence matrices (GLCM) (Haralick et al., 1973) matrices with offsets of 1 pixel, 0, 45, 90 and 135° were created for the image after rescaling the image to 16 levels of intensity. GLCM statistics of contrast, correlation, energy, homogeneity and entropy were calculated for each matrix. A 20-dimensional vector was created for each image by concatenating the 5-dimensional vector obtained by each of the four matrices.
- *GLCM2*: In order to capture the spatial variation of the images in a coarse manner, the resized image (256×256) was partitioned into 5 squares of size 128×128 pixels (top left, top right, bottom left, bottom right, centre). A gray level correlation matrix was created for each partition. A 20-dimensional vector was created for each partition. Subsequently, the 5 vectors from each of the partitions were concatenated to created feature vector of dimension 100. The difference between the padded and unpadded images is most apparent in this feature vector.
- *Hist*: A 32-bin intensity histogram was created for each image and counts in each bin were used as a 32-dimensional feature vector.
- *DCT*: A global discrete cosine transform was created for each image. The upper left (10×10) vectors were concatenated and used as a 100-dimensional feature vector.

Concatenating one or more of the above vectors created additional feature vectors.