

X-RAY IMAGE CATEGORIZATION AND RETRIEVAL USING PATCH-BASED VISUAL WORDS REPRESENTATION

Uri Avni, Hayit Greenspan*

Michal Sharon, Eli Konen

Jacob Goldberger

BioMedical Engineering
Tel-Aviv University

Diagnostic Imaging Department
Sheba Medical Center

School of Engineering
Bar-Ilan University

ABSTRACT

We present an efficient image categorization and retrieval system applied to medical image databases, in particular large radiograph archives. The methodology presented is based on local patch representation of the image content and a bag-of-features approach for defining image categories, with a kernel based SVM classifier. In a recent international competition the system was ranked as one of the top schemes in discriminating orientation and body regions in x-ray images, and in medical visual retrieval. A detailed description of the method (not previously published) is presented, along with its most recent results. In addition to organ-level discrimination, we show initial results of pathology-level categorization of chest x-ray data. On a set of 102 chest radiographs taken from routine hospital examination, the system detects pathology with sensitivity of 94% and specificity of 91%. We view this as a first step towards similarity-based categorization with clinical importance in computer-assisted diagnostics.

1. INTRODUCTION

The rapid growth of computerized medical imagery using picture archiving and communication systems (PACS) in hospitals throughout the world has generated a critical need for efficient and powerful search engines to classify and search the visual data. In addition, the growing workload on radiologists in recent years increases the need for computerized systems which could help the radiologist in prioritization and in the diagnosis of findings.

This work presents a classification and retrieval system that is based on the Visual Words (VW) paradigm which is a recently introduced concept that has been successfully applied to scenery image classification tasks (see e.g. [5, 3, 4]). The VW model is based on the idea that it is possible to transform the image into a set of visual words and to represent the image (and objects within the image) using the statistics of appearance of each word as feature vectors. In our system the visual words are image patches (small sub images) that are clustered to form a dictionary consisting of a small set

of representative patches. We utilize the VW approach while implementing modifications which are relevant for medical images. A main advantage of this approach is avoiding the need for explicit object detection features. Previous methods are based on explicitly predefined features that are locally extracted from the image (e.g. gradient orientation, edge, line length and orientation). The proposed approach, which is based on medium size image patches, avoids the need for explicitly specified medical object features. Instead, the features are implicitly found as part of the unsupervised learning step composed of building the visual dictionary.

In this study we present a patch-based classification system that has demonstrated very strong classification rates while also providing efficiency in the retrieval process. The system has been applied to several large radiograph archives. We have recently applied it within the ImageClef competition¹, and demonstrated strong results. A detailed description of the method is presented along with its most recent results.

The topic of retrieval becomes of value on the clinical front, once the content involves a diagnostic-level categorization, such as healthy vs pathology. In a collaborative effort with Sheba medical center, a large academic medical facility, we address this concept in the identification and categorization of x-ray lung disease. A description of the challenge and initial results are presented in the experiment section.

2. THE PATCH-BASED SYSTEM

Below we present our system for the task of large medical archive classification and retrieval. The system is composed of a feature extraction phase, a dictionary construction based on the training archive, an image representation phase and a classification phase. A schematic flow of the system is depicted in Figure 1.

We start by representing an input image as a collection of small patches. Several patch sampling strategies can be considered, including random sampling and grid sampling with several spacings. Following a comparative study, we select to extract a patch around every pixel, using a patch size of 9×9 pixels. Patches along the border of the image are considered

* H. Greenspan is currently on Sabbatical at IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95129, USA. hgreens@us.ibm.com

¹<http://imageclef.org>

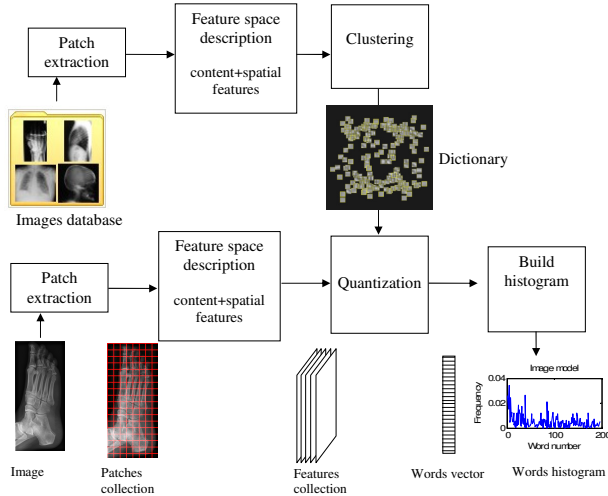


Fig. 1. System illustration

as noise and are ignored. The intensity values within a patch are normalized to have zero mean and unit variance. This provides local contrast enhancement and augments the information within a patch. Patches that have a single intensity value of black are ignored.

To reduce both the algorithm’s computational complexity and the level of noise, we apply a principal component analysis procedure (PCA) and reduce the data dimensionality from 81 to 7. Due to the normalization step, the resultant PCA components do not contain information regarding the patch average intensity. This average value contains information that discriminates between the dark background and the bright tissue, and may distinguish between tissue types. We therefore take the patch mean gray level as an additional feature. We note in passing that we also examined SIFT descriptors of 128 features which turn to be less effective for our medical retrieval task. Finally, we include the patch center (x, y) coordinates as two additional features, for an overall ten-dimensional patch representation. The addition of the spatial coordinates to the feature vector introduces spatial information into the image representation. Special care should be taken when combining features of different units, such as coordinates, gray level, and PCA coefficients. The relative feature weights in the proposed system were tuned experimentally on a cross-validation set (see the experiment section).

The next step of our system is to learn a dictionary of visual words based on a representative set on images. We represent each image as a collection of patches. To accelerate the learning process we randomly take a subset of all the patches. The main step in the dictionary building procedure is clustering the patches, using the k-means algorithm, to form a small-size dictionary of visual words. Note that this dictionary development step is done in an unsupervised mode without any reference to the image categories.

Using the feature extraction parameters that were learned (PCA, feature weights) and the generated dictionary, each image is represented as a histogram of visual words. In this step images are sampled with a dense grid. Note that as a result of the feature selection, which includes spatial features, both the image local content and spatial layout are preserved within the discrete histogram representation.

The classification is based on a set of manually categorized images. A k-nearest neighbor classifier is a reasonable choice. We have found that we gain much better results using a multi-class SVM classifier. The multi-class SVM is implemented as a series of one-vs-one binary SVMs with a RBF kernel, based on the LIBSVM library².

Given a query image the retrieval is based on finding the nearest image in the labeled training set. Several distance measures can be considered. We have found the using the L_1 norm between the word histograms of the two images yields the best results.

3. EXPERIMENTS AND RESULTS

In this section we evaluate the proposed system. We first investigate the sensitivity to various parameters that define the system. We then show classification and retrieval experiments on large radiograph archives. We conclude with classification results in a lung pathology detection application.

3.1. System parameters validation

The system validation was conducted using a database of 12,000 categorized radiographs. This dataset is the basis for the ImageClef 2007 medical image classification competition [2]. A set of 11,000 images are used for training, and 1000 serve for testing. There are 116 different categories within the archive, differing in either the examined region, the image orientation with respect to the body or the biological system under evaluation. Several of these images are presented in figure 2(a). The distribution of the images across the categories is non-uniform, the most frequent category contains over 19% of the images in the database, while many categories are represented by less than 0.1% of the images.

We optimized the system parameters using the training portion of this set, by running 20 cross-validation experiments trained on 10,000 images and verified on 1000 randomly drawn test images. Each parameter was optimized independently. As Figure 2(b) shows, increasing the number of dictionary words proved useful up to 700 words. Beyond this value the running time increased significantly, with no evident improvement in the classification rate. Figure 2 also demonstrates that using an SVM classifier provides results that are more than 3% higher than the best K-NN classifier (k=3). The effect of the number of PCA components was examined next. Figure 2(c) shows similar classification results

²<http://www.csie.ntu.edu.tw/~cjlin/libsvm>

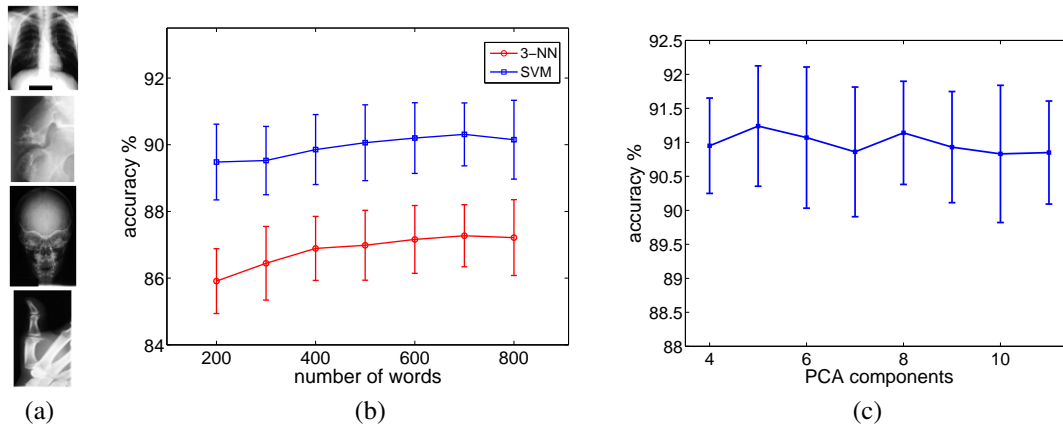


Fig. 2. (a): Sample images from ImageClef medical annotation challenge (b): Effect of dictionary size, for K-NN and SVM classifiers. (c): Effect of the number of PCA components in a patch. The y-axis shows the classification accuracy. Bars show mean and standard deviation of 20 random experiments.

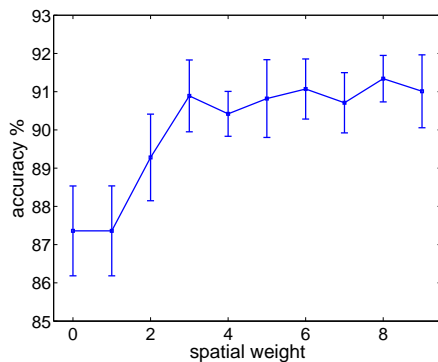


Fig. 3. Effect of spatial features: Weight of spatial features (x-axis); Classification accuracy (y-axis). Bars show mean and standard deviation of 20 experiments.

in the range of 5 to 8 components, with an average classification rate of approx. 90% using the SVM classifier. Based on the above experiments, a dictionary size of 700 visual words was selected, where each word contains 7 PCA coefficients.

Incorporating spatial coordinates of the patch as additional features improves the classification performance noticeably, as seen in Figure 3. The optimal range for the x,y coordinates is $[-3, 3]$. The patch variance normalization step improves the classification rate as well: with no normalization, the average classification rate is 88.19, while with normalization it climbs to 90.9. Using SIFT features with the SVM classifier increased significantly the feature extraction time, and achieved an average of 85.4% classification accuracy; well below the classification rate of the raw patch based classifiers.

Using the parameter set defined above, classification of previously unseen 1000 test images was conducted. The overall classification rate achieved is 89.1%. The total running

time for the whole system, training and classification, was approximately 40 minutes on the full resolution images, and 3 minutes on the 1/4 scaled down images. Times were measured on dual quad-core Intel Xeon 2.33 GHz.

3.2. Large scale image retrieval

In ImageClef 2008 a large-scale medical image retrieval competition was conducted. A database of over 66,000 images was used with 30 query topics. Each topic is composed of one or more example images and a short textual description in several languages. The objective is to return a ranked set of 1000 images from the complete database, sorted by their relevance to the presented queries. A sample query from this challenge and the first few returned images are seen in Figure 4. The retrieved results were manually judged for relevance by medical experts.

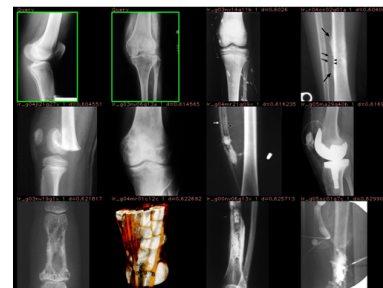


Fig. 4. Retrieval example: First two framed images are the query images; the following images (left to right, top to bottom) are retrieval results.

Figure 5 shows the scores of our submitted runs, marked with (*), along with visual retrieval algorithms submitted by additional groups [1]. In this Figure, the run labeled 'Proposed System' uses patch normalization and the run labeled

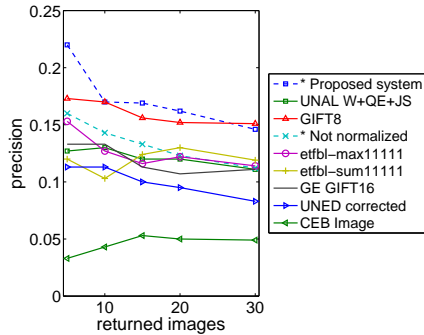


Fig. 5. Precision vs Recall graph of visual retrieval systems; ImageClef 2008 medical database. Precision shown for first 5, 10, 15, 20 and 30 returned images. Alg. results marked with dashed lines.

‘Not Normalized’ uses the patch original gray levels. The normalized patch approach in the proposed system is shown to rank first among the automatic purely visual retrieval systems. The retrieval system is computationally efficient, with an average retrieval time of less than 400ms per query.

3.3. Categorization on the pathology level

Image similarity-based categorization and retrieval becomes of clinical value once the task involves a diagnostic-level categorization, such as healthy vs pathology. We examined our system on chest x-rays obtained for various clinical indications in the emergency room of Sheba medical center. We used 102 frontal chest images, 26 of which are diagnosed as normal, and 76 of which have one or more pathologies, including lung infiltrates, left or right pleural effusion or an enlarged heart shadow. X-ray interpretations, made by two radiologists, served as the referral gold standard. Inconclusive results were not included in this set. Four sample images from this data are presented in Figure 6. The patch-based classifier was applied using an SVM classifier with two classes, the classification was conducted for each pathology type, and for healthy vs. any pathology. In order to preserve the generalization ability of the classifiers, system parameters were tuned using the general ImageClef 2007 database and were not specifically tuned to the lung pathology task. A leave-one-out classification was performed (results averaged over 102 trials). Table 1 summarizes the classification results. The software identified correctly 74 out of 76 abnormal and 22 out of 26 normal x-rays with 4 false positives and 2 false negatives cases, resulting in a sensitivity of 94.87% and specificity of 91.67%. In the task of between-pathology discrimination, the performance depends on the pathology type: it is highly accurate in detecting enlarged hearts, with a sensitivity of 95.24% and specificity of 93.48%. It is less accurate in detecting lung infiltrates and effusions.

To conclude, in this study we applied a patch-based classi-



Fig. 6. Frontal chest x-ray images. Left to right: Healthy, enlarged heart, filtrate, left+right effusion.

Table 1. Classification of chest images. Two left columns: (Software diagnosis)/(Radiologist’s interpretation)

	Normal images	Abnormal images	Sensitivity	Specificity
Any Pathology	22/26	74/76	94.8	91.7
Enlarged heart	20/23	43/44	95.3	93.5
Lung Infiltrates	23/33	27/34	76.7	73.0
Right pleural effusion	12/23	42/51	57.1	79.2
Left pleural effusion	15/27	38/47	62.5	76.0

fication system to a variety of medical image archives, in categorization and retrieval tasks. The proposed system was tuned to achieve high accuracy, with an average of over 90% correct classification on a publicly available database of 12,000 medical radiographs. In the ImageClef 2008 medical annotation challenge it ranked second. It is a highly efficient, with less than 200 milliseconds training and classification time per image. Using the same methods, we have developed an image retrieval utility, which was ranked first in ImageClef 2008 among the visual retrieval systems. Extending the system to pathology-level discrimination, we showed initial results for lung disease categorization. In ongoing collaborative efforts with Sheba medical center, we have demonstrated the ability to identify lung diseases on a chest image dataset. Future work involves extending the system to handle larger collection of chest images and pathology types.

4. REFERENCES

- [1] H. Muller et al. Overview of the imageclefmed 2008 medical image retrieval task. In *CLEF working notes*. http://www.clef-campaign.org/2008/working_notes/CLEF2008WN-Contents.html. 3
- [2] T. Deselaers et al. Overview of the imageclef 2007 object retrieval task. in workshop of the cross language evaluation forum 2007. volume 5152, 2008. 2
- [3] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR 05*, volume 2, pages 524–531, 2005. 1
- [4] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, pages 490–503. Springer, 2006. 1
- [5] M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *CVPR 03*, volume 2, pages 691–8, 2003. 1