X-ray Categorization and Spatial Localization of Chest Pathologies

Uri Avni[1], Hayit Greenspan[1⋆] and Jacob Goldberger[2]

[1]BioMedical Engineering Tel-Aviv University, Israel
[2]School of Engineering, Bar-Ilan University, Israel

Abstract. In this study we present an efficient image categorization system for medical image databases utilizing a local patch representation based on both content and location. The system discriminates between healthy and pathological cases and indicates the subregion in the image that is automatically found to be most relevant for the decision. We show an application to pathology-level categorization of chest x-ray data, the most popular examination in radiology. Experimental results are provided on chest radiographs taken from routine hospital examinations.

Keywords: Image categorization, x-ray, chest radiography, visual words, Computer-Aided Diagnosis (CAD), region-of-interest (ROI).

1 Introduction

In the last ten years the number of images that are acquired every day in any modern hospital has increased exponentially, due to the progress in digital medical imaging techniques and patient image-screening protocols. One immediate consequence of this trend is the enormous increase in the number of images that must be reviewed by radiologists. This has led to a concomitant demand for computerized tools to aid radiologists in the diagnostic process. Computerized systems assist the radiologist in the diagnostic process by categorizing the image content. This is done by learning from a large archive of image examples that are annotated by experts.

Image categorization is concerned with the labeling of images into predefined classes. The principal challenge of image categorization is the capture of the most significant features within the images that facilitate the desired classification. A single image can contain a large number of regions-of-interest (ROI), each of which may be the focus of attention for the medical expert, depending on the diagnostic task at hand. A single chest image for example, contains the lungs, heart, rib cage, diaphragm, clavicle, shoulder blade, spine and blood vessels, any of which may be the focus of attention. One way to enhance the image categorization process is to focus on a ROI within the image that is relevant to the presumed pathology. The advantage of the ROI approach is that the rest

⋆ Currently on Sabbatical at IBM Almaden Research Center, San Jose, CA

of the image can be ignored leading to computational advantages and increased accuracy in the classification.

Clinical decision support techniques based on automatic classification algorithms can produce a strong need to localize the area that is most relevant to the diagnostic task. A diagnostic system that, in addition to a decision on the existence of a pathology, can provide the image region that was used to make the decision can assist radiologists to better understand the decision and evaluate its reliability. Another advantage of an ROI based decision is that we can construct a detailed representation of the local region. We refer to this approach as ROI based image categorization. Of course it is not suited to every image categorization task; not every pathology includes a significant and identifiable ROI that appears across the data set. In such cases, a global, full-image categorization is appropriate.

The (bags of) visual words paradigm, which has recently been adapted from the text retrieval domain to the visual analysis domain (see e.g. [1][2]), provides an efficient way to address the medical imaging categorization challenge in large-size archives while maintaining solid classification rates [3][4]. The best categorization methods in recent ImageCLEF competitions are all based on variants of the visual words concept [5]. In this study we utilize a variant of the visual words framework, that combines content and location, to automatically localize the relevant region for a given diagnostic task. Besides localizing the decision based area, the proposed method yields improved results over a categorization system based on the entire image.

2 The Visual Words Framework for Classification

In this section we describe a state-of-the-art medical image categorization paradigm using the visual words framework, which is based on a large set of image patches, and their respective representation via a learned dictionary. This paradigm is the foundation for the proposed localized image classification system. The method was ranked first in the ImageCLEF 2009 medical annotation task [6].

Given a training labeled image dataset, patches are extracted from every pixel in the image. Each small patch shows a localized view of the image content. In the visual dictionary learning step, a large set of images is used (ignoring their labels). We extract patches using a regular grid, and normalize each patch by subtracting its mean gray level, and dividing it by its standard deviation. This step insures invariance to local changes in brightness, provides local contrast enhancement and augments the information within a patch. Patches that have a single intensity value are ignored in x-ray images (e.g. the brightness of the air surrounding the organ appears uniform especially in DICOM format). We are left with a large collection of several million vectors. To reduce both the computational complexity of the algorithm and the level of noise, we apply a Principal Component Analysis procedure (PCA) to this initial patch collection. The first few components of the PCA, which are the components with the largest eigenvalues, serve as a basis for the information description. In addition to the patch

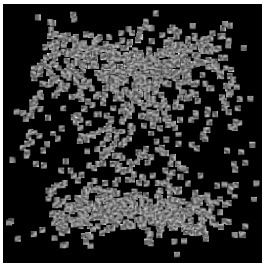


Fig. 1. An example of a region based visual dictionary. Each visual word is placed according to its (x, y) coordinates.

content information represented, we add the patch center coordinates to the feature vector. This introduces spatial information into the image representation, without the need to explicitly model the spatial dependency between patches. The final step of the visual-words model is to convert vector-represented patches into *visual words* and generate a representative *dictionary*. A visual word can be considered to be a representative of several similar patches. The vectors are clustered into k groups in the feature space using the k-means algorithm. The resultant cluster centers serve as a vocabulary of k visual words. The location of the cluster center is the average location of all the patches in the cluster. Due to the fact that we included spatial coordinates as part of the feature space, the visual words have a localization component in them, which is reflected as a spatial spread of the words in the image plane. Words are denser in areas with greater variability across images in the database. Fig. 1 shows a region based visual dictionary. Each visual word is placed according to its (x, y) coordinates.

A given (training or testing) image can now be represented by a unique distribution over the generated dictionary of words. In our implementation, patches are extracted from every pixel in the image. For a 512×512 image, there are several hundred thousand patches. The patches are projected into the selected feature space, and translated (quantized) to indices by looking up the most similar word in the generated dictionary. Note that as a result of including spatial features, both the local content and spatial layout of the image are preserved in the discrete histogram representation. Image classification of a test image is based on the ground truth of manually categorized images that are used to train an SVM classifier which is applied to the image histogram representation.

3 Localizing Image Categorization

When radiologists look for a specific pathology in an image, they usually focus on a certain area. For example right pleural effusion is diagnosed using the lower bottom and the peripheral lateral zones of the right lung, while ignoring the rest of the chest x-ray image. In this section we describe how to automatically find a relevant area, without prior knowledge about the organ structure or the characteristics of the pathology. This step is designed to improve the classification accuracy of the global approach. It can also provide a useful visualization of the area used in the automatic classification.

Assume we already learned a visual dictionary as described in the previous section and we are now concentrating on a specified pathology. We are given a training image set in which each image is manually labeled as either healthy or pathological. The visual-words representation of an image x is a histogram vector $(x_1, ..., x_k)$ such that x_i is the relative number of image patches in x that were mapped to the i-th visual word based on content and location similarity. These k numbers are the features extracted from the image to be used in the classification algorithm, where each feature corresponds to a visual word. The first step toward localization of the pathology decision is finding the relevance of each feature. Feature relevance is often characterized in terms of mutual information between the feature and the class label. It is reasonable to expect that for a feature (visual word) that is located far from the pathology area, the class label and the feature values random variables should be independent. We compute the mutual information between the image label and each of the features in the following way. Suppose we are given n images with binary (healthy/pathological) labels $c_1, ..., c_n$ and the feature vector of the t-th image is denoted by $(x_{t1}, ..., x_{tk})$. To obtain a reliable estimation of the mutual information for each feature i, we first quantize the i-th feature values across all the images $x_{1i}, ..., x_{ni}$ into L levels (in our implementation we sort the n values and divide them into four groups of equal size). Denote the quantized version of x_{ti} by $y_{ti} \in \{1, ..., L\}$. Denote the joint probability of the (quantized) i-th feature and the image class by:

$$p_i(v, c) = \frac{1}{n} |\{t | y_{ti} = v, c_t = c\}| \tag{1}$$

where c is the class label, v is the quantized bin level and $|\cdot|$ is the set cardinality. The mutual information between the class label variable C and the quantized feature Y_i is [7]:

$$I(Y_i; C) = \sum_{v=1}^{L} \sum_{c} p_i(v, c) \log \frac{p_i(v, c)}{p_i(v) p(c)} \tag{2}$$

where $p_i(v)$ and $p(c)$ are the marginal distributions of the i-th feature and the class label respectively. C is a binary random variable and therefore $0 \leq I(Y_i; C) \leq 1$.

Up to now we have computed the relevance of each feature (visual word) separately. However, since each visual word has a location, we can consider the relevance of an entire region as the relevance of all the features located in that region. The proposed method can be viewed as a filter-based feature selection. Unlike general feature selection problems, here the features are located in the image plane. Hence, instead of feature selection we apply region selection. Since the visual words have a spatial location, the relevance information can be represented in the image plane. We create a relevance map $R(x, y) = \sum_i I(Y_i, C)$ such that i goes over all the visual words that are located at (x, y). $R(x, y)$ is a matrix with mutual information values at the positions of the visual-word feature centers, and zero at other locations. In this representation, areas that contain highly relevant features are likely to be the regions of interest for the classification task.

We define the relevance of the rectangular $[x_1, x_2] \times [y_1, y_2]$ sub-region in the image plane to the pathology as

$$\text{score}(x_1, x_2, y_1, y_2) = \sum_{x_1 \leq x \leq x_2} \sum_{y_1 \leq y \leq y_2} R(x, y) = \sum_i I(Y_i, C) \tag{3}$$

such that i goes over all the visual words that are located in the rectangle $[x_1, x_2] \times [y_1, y_2]$. We look for a rectangular region that contains the maximum amount of relevant features. For a given region size, we examine all the rectangles in the image and look for the rectangle with the highest score. This search can be efficiently done using the integral image algorithm which is part of the seminal Viola-Jones object detection framework [8]. The integral image algorithm efficiently generates the sum of values in rectangular subsets of a grid. It can be computed in one pass over the image using the following recurrence relation:

$$\begin{aligned} \text{sI}(x, y) &= \text{sI}(x, y-1) + R(x, y) \\ \text{II}(x, y) &= \text{II}(x-1, y) + \text{sI}(x, y) \end{aligned} \tag{4}$$

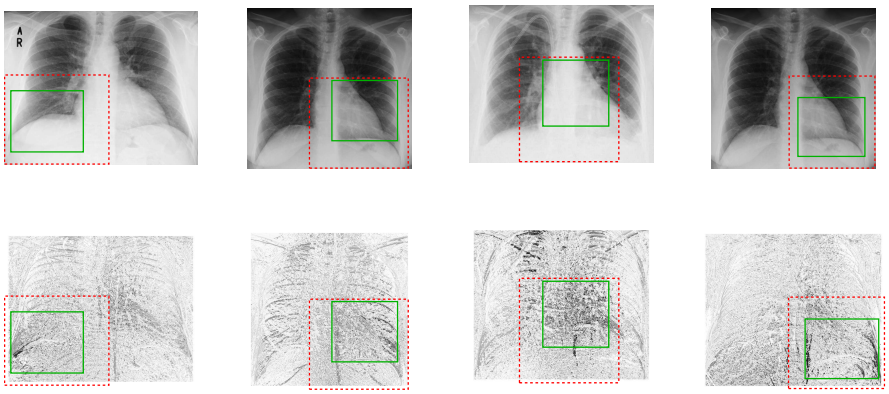
The region score is thus:

$$\text{score}(x_1, x_2, y_1, y_2) = II(x_2, y_2) - II(x_1, y_2) - II(x_2, y_1) + II(x_1, y_1) \tag{5}$$

The relevance map $R(x, y)$ takes into account all of the training images and therefore the ROI is little affected by noisy images. However, since the ROI is calculated globally, this procedure finds a rough estimation of the region of interest. The exact region might vary between images. The ROI can be refined individually for each image by examining the mutual information map in the image space instead of the visual dictionary space. Every patch in the image is translated into a visual word. We can create a relevance map per image $R_{image}(x, y)$ by placing in each image patch center the mutual information of the visual word it is assigned to. In other words, $R_{image}(x, y) = I(Y_i, C)$ such that i is the visual word that the image patch centered at (x, y) is assigned to. We can then repeat the integral-image process on the map R_{image} to find a smaller rectangle with the maximal amount of relevant patches in the image relevance map. The search is confined to the rough ROI found in the first step. A two-step process is required because if the relevance map of the images is noisy, it may generate erroneous ROI, especially if its relevant area is relatively small.

After finding an ROI and refining it for each image in the healthy/pathological labeled training set, we run a second training stage, where sub images are cropped to the region of interest of the pathology. A new dictionary is generated for each pathology, and the SVM classifiers are trained using the words histograms from the cropped regions.

In the test phase, a new image is converted to a word histogram using the dictionary learned in the first step, and an image relevance map is calculated. For each pathology we crop the image using the rough global ROI that was found in the training phase. Next we find a refined ROI by applying the integral image algorithm to the test image relevance map R_{image}. The refined ROI is then



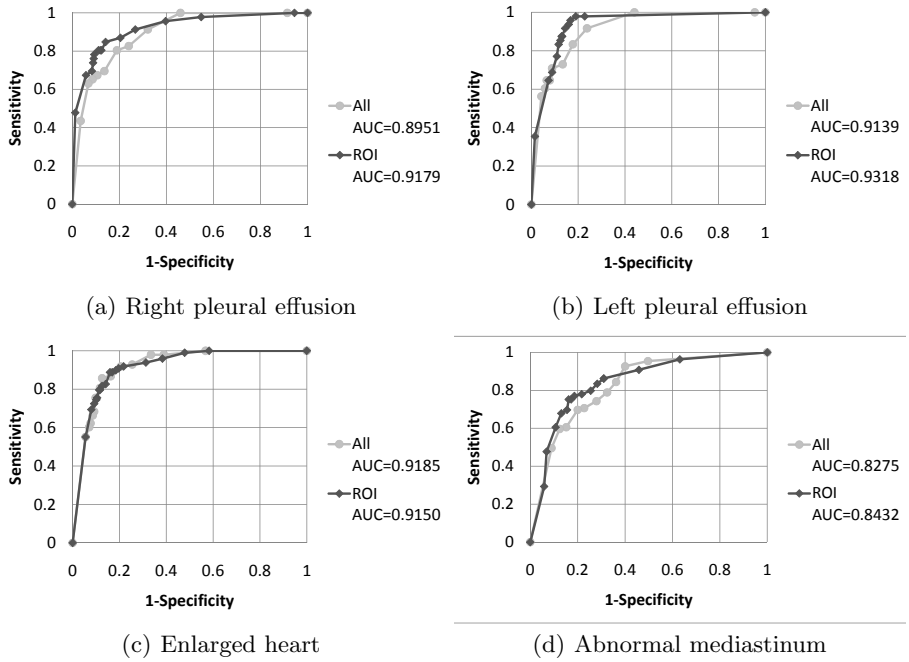
(a) (b) (c) (d)

Fig. 2. Frontal chest x-ray images, top row shows the ROI detection overlayed on the original image and bottom row shows the corresponding relevance map. Global ROI is shown in red and the refined ROI in green. The pathologies are (a) right pleural effusion, (b) enlarged heart shadow, (c) enlarged mediastinum and (d) left pleural effusion.

converted to a words histogram using the dictionary from the second training step, and passed to the region based healthy/pathology classifier that was also trained in the second training step. The reported result is the binary decision and the image subregion that was used to obtain the decision. Fig. 2 shows examples of relevance maps and ROI detection (both global for a pathology and image refined) of chest x-ray images. The entire processing of a test image including the translation to visual words and the integral image computation takes less than a second (time was measured on a dual quad-core Intel Xeon 2.33 GHz.)

4 Experiments

Chest radiographs are the most common examination in radiology. They are essential for the management of various diseases associated with high mortality and morbidity and display a wide range of potential information, many of which are subtle. According to a recent survey [9], most of research in computer-aided detection and diagnosis in chest radiography has focused on lung nodule detection. However, lung nodules are a relatively rare finding in the lungs. The most common findings in chest x-rays include lung infiltrates, catheters and abnormalities of the size or contour of the heart [9]. Distinguishing the various chest pathologies is a difficult task even for human observers. Research is still needed to develop an appropriate set of computational tools to support this task. We used 443 frontal chest images in DICOM format from the Sheba medical center hospital PACS, taken during routine examinations. X-ray interpretations,



(a) Right pleural effusion (b) Left pleural effusion

(c) Enlarged heart (d) Abnormal mediastinum

Fig. 3. ROC curves for pathology detection in frontal chest x-rays, using the entire image and automatically detected ROI; the areas under the curves (AUC) were calculated using trapezoidal approximation.

made by two radiologists, served as the reference gold standard. The radiologists examined all of the images independently; they then discussed and reached a consensus regarding the label of every image. The pathology data include 98 images with enlarged heart shadow, 109 images with enlarged mediastinum, 46 images with right pleural effusion and 48 images with left pleural effusion. Some patients had multiple pathologies.

The original high-resolution DICOM images were initially resized to a maximal image dimension of 1024 pixels, with aspect-ratio maintained. We followed the method described in Section 2 to extract features, build a visual dictionary, and represent an image as a histogram of visual words. To avoid overfitting and to preserve the generalization ability of the classifiers, model parameters were chosen following the experiments on the ImageCLEF database, described in [6, 10]. For each pathology we found the relevant ROI and utilized the cropped images (both healthy and pathological) to learn a visual dictionary. The rough ROI size was selected to be 36% of the image area; it was cropped to a smaller rectangle if it passed the image border. The fine ROI was set to 15% of the image area. We then detected each of the four pathologies using a binary SVM classifier with a χ^2 kernel, trained on words histogram built from the fine ROI. The sensitivity and specificity were calculated using leave-one-out cross validation.

Modifying the relative cost of false negative errors in the SVM cost minimization problem determines the tradeoff point between sensitivity and specificity. This technique was used to produce the receiver operating characteristic (ROC) curves for several pathologies, shown in Fig. 3. The figure clearly indicates that in three out of the four pathologies our localized categorization method outperformed the global categorization that utilizes patches from the entire image. In the right pleural effusion case, the AUC is improved from 0.895 to 0.92; In the left pleural effusion case, the AUC is improved from 0.91 to 0.93, and in the abnormal mediastinum case, an improvement in the AUC is from 0.827 to 0.84.

To conclude, in this study we showed how visual word information, based on both content and location, can be used to automatically localize the decision region for pathology detection. The method is based on choosing the visual words that are most correlated with the diagnoses task. We have shown that the proposed method outperforms methods that are based on the entire image, both in terms of classification performance and in enabling human interpretation of the decision. The method proposed is general and can be applied to additional Chest x-ray pathologies, currently being explored, as well as to additional medical domains.

Acknowledgments The authors wish to thank Dr. Eli Konen and Dr. Michal Sharon from the Diagnostic Imaging Department of Sheba Medical Center, Tel Hashomer, Israel - for the data collection, guidance and helpful discussions.

References

1. M. Varma and A. Zisserman. Texture classification: are filter banks necessary? In *CVPR*, volume 2, pages 691–8, 2003.
2. L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005.
3. T. Deselaers, A. Hegerath, D. Keysers, and H. Ney. Sparse patch-histograms for object classification in cluttered images. In *DAGM*, pages 202–211, 2006.
4. T. Tommasi, F. Orabona, and B. Caputo. Discriminative cue integration for medical image annotation. *Pattern Recogn. Lett.*, 29(15):1996–2002, 2008.
5. H. Muller, P. Clough, and T. Deselaers. Imageclef: Experimental evaluation in visual information retrieval. *Springer*, 2010.
6. U. Avni, J. Goldberger, and H. Greenspan. Addressing the imageclef 2009 challenge using a patch-based visual words. *The Cross Language Evaluation Forum Workshop (CLEF)*, 2009.
7. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.
8. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *IEEE Computer Vision and Pattern Recognition*, 2001.
9. B. van Ginneken, L. Hogeweg, and M. Prokop. Computer-aided diagnosis in chest radiography: Beyond nodules. *European Journal of Radiology*, 72(2):226–230, 2009.
10. U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. *IEEE Trans. Medical Imaging*, pages 733–746, 2011.