A Comparative Study for Chest Radiograph Image Retrieval using Binary, Texture and Deep Learning Classification

Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva and Hayit Greenspan

Abstract— **In this work various approaches are investigated for X-ray image retrieval and specifically chest pathology retrieval. Given a query image taken from a data set of 443 images, the objective is to rank images according to similarity. Different features, including binary features, texture features, and deep learning (CNN) features are examined. In addition, two approaches are investigated for the retrieval task. One approach is based on the distance of image descriptors using the above features (hereon termed the "descriptor"-based approach); the second approach ("classification"-based approach) is based on a probability descriptor, generated by a pair-wise classification of each two classes (pathologies) and their decision values using an SVM classifier. Best results are achieved using deep learning features in a classification scheme.**

I. 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.

Two main approaches can be used for retrieval in clinical application scenarios. In some cases the radiologist will be interested in obtaining similar cases in the sense of the appearance, and in other cases he might be interested in obtaining similar cases in the sense of the pathology. In this work we compare two possible distances schemes – corresponding with the two approaches mentioned above. The first is based on descriptor-based distances between images. The second scheme is based on the probabilistic output of the SVM classifier that defines the probabilities of each pathology class per image.

The dataset for this work consists of 443 frontal X-ray chest images, taken during routine examinations at Sheba Medical Center, a large community hospital. Each image was examined by radiologists and was assigned with a binary label for each of the following chest pathologies: enlarged heart shadow, enlarged mediastinum, left pleural effusion, right pleural effusion, left lung opacities, right lung opacities, congestion. An additional healthy/sick binary label was set, indicating the existence of any pathology. Figure 1 shows examples of the pathologies.

* Authors are from the Medical Image Processing Lab, Department of Biomedical Engineering, Faculty of Engineering, Tel Aviv University, Israel. Corresponding Author: Prof. Hayit Greenspan, hayit@eng.tau.ac.il

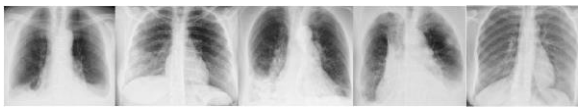


Fig. 1: Image Examples. Left to right: Left Pleural Effusion, Left Opacity, Enlarged heart, Enlarged Mediastinum and Healthy

Recent works in the domain of medical image retrieval use a bag of visual words model (BoVW). In the BoVW framework, the image patches are sampled densely by interest point detectors and represented by local patch feature vectors. Commonly used feature vectors are the intensity local patches [1], wavelets and LBP [2], where similarity and ranking is based on feature vectors distance. Euclidean, Histogram intersection, Manhattan and Hellinger distance are common metrics. In some works, metric learning approaches are used for the ranking task [3]. Another method which enhances performance is an initial classification and ranking according to the classification results. In [4], images are ranked according to the distance from the SVM decision boundary.

.

II. Methods

A. Descriptor Vs Classification approach

In a *descriptor–based* approach we generate a feature vector that takes into account the number of occurrences of the texture and binary patterns in the image and their location. The distance measure, which is used in this work is a histogram intersection-based, which was found to be slightly better than the Euclidean metric and is mathematically defined as follows:

$$\mathrm{Dist}_{\mathrm{Intersection}} = \sum_{i=1}^{n} \min\{(d_1)_i, (d_2)_i\} \quad (1)$$

where d_1 and d_2 are two feature vectors of the size n.

Based on the distance used, the images are ranked in order of similarity. When two feature vectors are compared to each other, a hidden assumption is that each feature vector bin has the same importance. However, usually the actual information regarding pathology will be "hidden" in a subset of the feature vector bins. This could be as a result of common areas or textures of the pathology. Therefore, when two images are compared only by their feature vectors and a predefined metric, their similarity will be based on a visual correspondence and not necessarily a pathology-related one.

In the *classification-based* approach, a one-versus-one classification is performed. For each image, a trained model using a set of the rest 442 images is generated. The output of the SVM classification is a normalized vector of size 8,

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

which contains the probabilities of the images to be in each of the 8 classes. The probabilities are calculated by using $\frac{k(k-1)}{2}$ binary classifiers, which is the total number of possible classes pairs, where k=8 is the number of classes. Using the LIBSVM library [5], the probability estimates $\{p_i\}_{i=1}^{8}$ are calculated using a minimization function, which takes into account the decision values of each binary classifiers.

$$\left\{\left((p_1,p_2,\ldots.p_8)\middle|p_i \stackrel{def}{=} P(\text{Image} \in \text{Class i}),\ \sum_i^8 p_i=1\right)\right\} \quad (2)$$

Given two probabilities feature vectors $Q=\{q_i\}_{i=1}^{8}$ and $M=\{m_i\}_{i=1}^{8}$, their distance is calculated using a histogram intersection-based metric:

$$Dist(Q,M) = \sum_{i=1}^{8} \min(q_i,m_i) \quad (3)$$

This metric was found to be the optimal one in this case since it favors similarity of high probabilities in contrast to differences metrics, as the Euclidean one for instance, which is indifferent to high values similarity and is affected only by difference. In addition, multiple pathologies similarity is also addressed using the presented metric. If for instance, two images contain enlarged heart, and mediastinum pathologies, then p_i, p_j and q_i, q_j are expected to be high, where i,j match the mediastinum and enlarged heart pathologies

B. Features examined

B.1 Pyramid dense SIFT-BoVW feature vector

The dense SIFT feature vector is a dense version of the SIFT feature vector (Scale Invariant Feature Transform). First, a set of SIFT feature vectors is generated using 16X16 pixels patches computed over a grid with spacing of 8 pixels for each image in the data set. Then, the union of all SIFT features is extracted from all the images in the database and used in the BoVW model, where the PCA algorithm is first performed in order to accelerate the convergence of the K-means algorithm and remove noise. The result of the K-means algorithm is a set of 300 indexes (of the size 20 after using PCA). Each image is then represented using these indexes and a spatial pyramid.

The spatial pyramid [6] is a concatenation of image representations in sub-regions and computation of histograms of local features found inside each sub-region. In this work the number of levels was set to be 3 for the dense SIFT-BoVW pyramid. *Level 0:* This level contains the histogram of the indexed features of the image (of size 300).

Level 1: The image is partitioned to 4 quarters. For each quarter and index k (k=1...300) the number of features that were indexed as k are summed for an overall size of 4X300.

Level 2: Similar to level 1 with 16 blocks, the size of this level is 16X300. Therefore, the total size of the feature vector is 300 + 1200 + 4800 = 6300. Images are ranked using the above feature vector using both descriptor and classification-based approaches and an intersection distance\kernel.

B.2 Binary images with region of interest

X-ray images are relatively noisy due to the acquisition conditions. A feature vector that may be less affected by noise is a descriptor-based on a corresponding binary image. A binary image BW is defined as follows:

$$BW(x,y) = \begin{cases} 1 & I(x,y)>\text{Threshold} \\ 0 & I(x,y)<\text{Threshold} \end{cases} \quad (4)$$

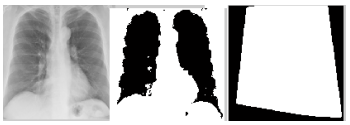
where $I(x,y)$ is the intensity value of the pixel (x,y). The optimum threshold is extracted using otsu's clustering method [7].

The binary feature vector can be used for the extraction of the lungs from the image and removal of regions, such as image corners and other which we select not to include in the chest representation. We extract a region of interest as the convex hull of the following set:

$$S = \cup_{(x,y)}\{BW(x,y) = 0\}, \quad (5)$$

mathematically defined as the following set:

$$\text{Convex Hull} \stackrel{def}{=} \left\{\sum_{i=1}^{|S|} a_i x_i \ ,\ x_i \in S\ ,\ (a_i \geq 0) \cap \sum_{i=1}^{|S|} a_i = 1\right\} \quad (6)$$



(a) (b) (c)

Fig. 2: (a) X-ray image (b) Binary Image (c) Convex hull (White)

When using the descriptor-based approach in this case, the distance between two images is calculated as follows:

$$BW_{Dist} = \sum_{x,y \in \text{Convex Hull}} |BW_1(x,y) - BW_2(x,y)| \quad (7)$$

where $BW_1(x,y)$ and $BW_2(x,y)$ are the binary images values at the (x,y) location.

In the classification-based approach the binary image (flattened to a vector) is used as the input to the SVM classifier with an intersection kernel. An advantage of the presented feature vector is that it can be used in real-time systems, as binary feature vectors are highly effective and efficient memory-wise.

B.3 LBP (Local binary patterns)

Local Binary Pattern (LBP) is an efficient texture operator, which labels image pixels by thresholding the neighborhood of each pixel and generating a binary number that contains information regarding the gradients of the local neighborhood. Due to its computational simplicity it is used in various applications [2]. Similar to the dense SIFT-BoVW pyramid presented above, a spatial pyramid with the LBP feature vector is generated. In this work a uniform rotation-invariant LBP [8] feature vector was applied in each block in the spatial pyramid. Similar to the dense SIFT-BoVW pyramid the number of patterns are summed in each sub-region of the image. In this case a pyramid consisting of levels 2 and 3 only, obtains the best performance. The

advantage in using the LBP feature vector is that a clustering algorithm, which is a very time consuming (K-means for instance) is avoided with the predefined 40 binary indexes of the rotation-invariant LBP. The total size of this feature vector is 3200 (40X(16+64)).

In the descriptor-based approach the feature vector which is generated by the spatial pyramid is used for the ranking with the intersection metric. In the classification approach a one-versus-one model is used to generate the probability vector using the SVM classifier. Using the probability vector and the histogram-intersection distance, images are then compared and ranked.

B.4 Deep Learning

The final set of features tested is extracted from a convolutional neural network (CNN) [9]. In earlier work we have shown strong performance in the classification of X-ray images, using a network trained on the Imagenet general image dataset, which contains one million images and 1000 classes [10]. Using a model which was trained on large dataset of non-medical images is beneficial due to the lack of sufficient amount of labeled medical imagery. In the case of chest X-rays images some of the visual patterns are similar in both medical and non-medical domains and we find that the filters that are learned can contribute to the medical task.

We use a similar approach here. The trained neural network consists of 60 million parameters, 650,000 neurons and five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax and activation function for all weight layers (except for full8) is the REctification Linear Unit (RELU). The deep learning (DL) package used is the Matconvnet framework [11], which is a MATLAB implementation of CNNs. The CNN was trained on the Imagenet dataset with the "Imagenet-vgg-m-2048" architecture [12] as shown in Figure 3.

In the descriptor-based approach a vector containing the activations values of the Full7 layer, which is the last fully connected layer before the output layer is used. The size of the feature vector is 2048. In addition, the concatenation of the last convolutional layer(Conv5) and the following two fully connected layers (Full6-Full7) is examined as a feature vector as well. The total size of this feature vector is 24,576.This was found to be slightly better than the feature vector extracted from the Full7 layer. In the classification-based approach the same feature vector is used with the intersection kernel in order to generate the probability vector which is used to rank the images.

III. EXPERIMENTS

We ran an extensive set of experiments to evaluate the performance of the Dense SIFT-BoVW, LBP, Binary and Deep learning (DL)_features. The dataset used consists of 443 images as listed in Table I. Each of the 443 images was used as a query image and the remaining images were ranked using two approaches according to the similarity to

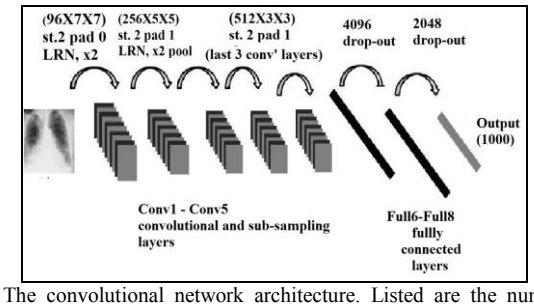


Fig.3 The convolutional network architecture. Listed are the number of convolution filters and their receptive field size, convolution stride ("st."),spatial padding ("pad") and if Local Response Normalization (LRN)

the query image. For the comparison, we use Recall estimates at the K=30 first images, denoted as r@K=30, where:

$$r@K = \frac{\#\text{relevant images in top K images retrieved}}{\#\text{relevant images in the database}}. \quad (8)$$

Another commonly used estimate is the Precision measure. We found this measure not to be relevant in the current evaluation, as it is highly sensitive to unbalanced data, which is the case here.

Figure 4 shows average Recall curves over all pathologies for each evaluated feature vector and ranking approach. The x-axis consists of the number of images retrieved (K), and the y-axis is the recall estimate r@K. The plot shows that the recall performance achieved using the classification approach is significantly better than the descriptor-based approach for all features used. We also see that the deep learning feature obtains the best results. In Table II recall rates are presented per pathology. Results are presented in both descriptor-based and classification approaches. We note the higher overall results of the classification-based approach. We also see that the DL features give best results in this approach and that the combined Conv5 and Full6-Full7 layers obtained better results than the Full7 layer.

When the descriptor-based approach is used, better results are achieved using dense SIFT-BoVW pyramid. Example retrieval results are shown in Figure. 5. In rows 1 and 2 examples of the dense SIFT-BoVW features, with descriptor-based and classification-based retrieval schemes are presented using the same query image. We see that the results obtained using the classification-based approach are better, where the top four images have the same pathology (Enlarged Heart) in contrast to three images using the descriptor-based approach. In row 5 an example is shown for a healthy pathology query image using the deep learning feature descriptor-based query, where the top four retrieved images contain the same pathology as the query image. In rows 3 and 4 images are ranked using the binary feature vector in a classification-based scheme. Both the query image and the top 4 retrieved images contain the enlarged heart pathology.

TABLE I. THE RATIO OF IMAGES IN THE DATABASE FOR EACH PATHOLOGY (443 IMAGES IN TOTAL, SOME OF THE IMAGES CONTAIN MULTIPLE PATHOLOGIES). LARGEST 5 CLASSES ARE LISTED.

	Effusion Left'	Opacity Left'	Enlarged Heart'	Enlarged Mediastinum'	Healthy'
Ratio	0.11	0.08	0.21	0.25	0.50

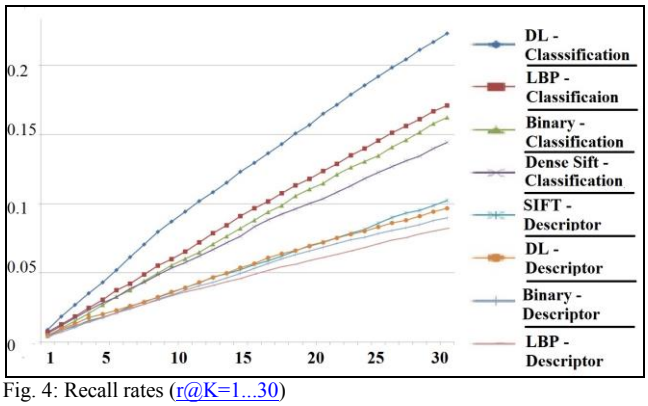


Fig. 4: Recall rates (r@K=1...30)

TABLE II. RECALL OF TOP 30 IMAGES

	r@K =30 , Classification based				
Feature	Effusion left	Opacity left	Enlarged Heart	Enlarged Mediastinum'	Healthy
Dense SIFT-BoVW Pyramid	0.159	0.072	0.192	0.125	0.103
LBP Pyramid	0.232	0.074	**0.236**	**0.165**	0.101
Binary	0.199	0.086	0.208	0.163	0.102
DL (Full7)	0.3023	0.1797	0.1811	0.1552	0.0986
DL (Conv5 + Full6 -Full7)	**0.310**	**0.182**	0.231	0.164	**0.103**
	r@K =30 , Descriptor based				
Feature	Effusion left	Opacity left	Enlarged Heart	Enlarged Mediastinum'	Healthy
Dense SIFT-BoVW Pyramid	**0.138**	**0.085**	**0.107**	**0.104**	**0.102**
LBP Pyramid	0.089	0.068	0.083	0.078	0.099
Binary	0.086	0.063	0.110	0.087	0.101
DL (Full 7)	0.1098	0.0812	0.0991	0.0832	0.0998
DL (Conv5 + full6-full7)	0.086	0.071	0.099	0.085	0.085

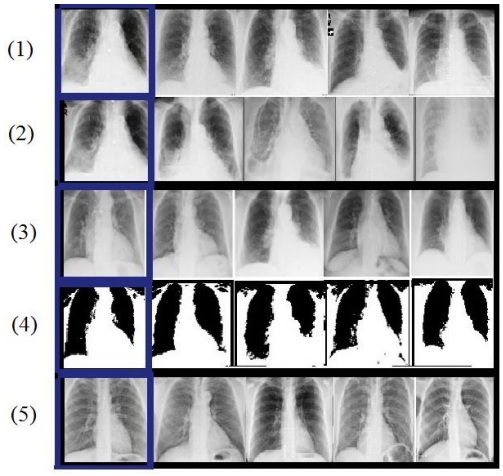


Fig 5. Images are ranked using a sift-BoVW descriptor-based approach (1) and using a classification approach (2). From left to right: (1,a) Query image: Enlarged heart, (1,b) Mediastinum, (1,c) – (1-e) Enlarged Heart , (2,a) Same Query image: Enlarged Heart, (2,b) –(2,e) enlarged Heart.
(3-4) Images are ranked using the binary vector feature vector. Left to right: (3,a) Query image: enlarged heart, (3,b) – (3,e) Enlarged heart.
(5) Images are ranked using a deep learning descriptor-based approach. From left to right: (a) Query image, Healthy (b) - (e) Healthy

IV. DISCUSSION AND CONCLUSION

In this work we evaluated a system for chest X-ray image retrieval using various features and two ranking approaches (a descriptor and classification-based approaches). Earlier work (e.g. [4]) has shown that using an initial classification followed by ranking according to the classification, significantly improves precision and recall performance. Similar conclusions are shown here. Our results show that a retrieval system which is based on a CNN using a classification-based scheme, obtains the best results.

As shown in this work, the final layers of the CNN are used to generate a quality image descriptor. The unique descriptor contains the probabilities to be in each class (pathology) and addresses well the multiple pathology labels challenge. In future work, other features from the convolutional layers should be examined as well, where metric learning and feature selection techniques can be applied. In addition, fine-tuning of the CNN using medical data is a future goal.

REFERENCES

[1] Yang, W., Lu, Z., Yu, M., Huang, M., Feng, Q., & Chen, W. "Content-based retrieval of focal liver lesions using bag-of-visual-words representations of single-and multiphase contrast-enhanced CT images". *Journal of digital imaging*, 2012, *25*(6), 708-719.

[2] Takala, V., Ahonen, T., & Pietikäinen, M. "Block-based methods for image retrieval using local binary patterns". H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, 2005, pp. 882–891

[3] Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., & Li, J. "Deep Learning for Content-Based Image Retrieval: A Comprehensive Study". In *Proceedings of the ACM International Conference on Multimedia,* November 2014, (pp. 157-166).

[4] Arandjelovic, R., & Zisserman, A. "Three things everyone should know to improve object retrieval". In *IEEE -CVPR,* June 2012. (pp. 2911-2918).

[5] Chang, C. C., & Lin, C. J. "LIBSVM: a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2011, *2*(3), 27.

[6] Lazebnik, S., Schmid, C., & Ponce, J. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". In *IEEE -CVPR,* June *2006,* (Vol. 2, pp. 2169-2178)

[7] Otsu, N. "A threshold selection method from gray-level histograms".*Automatica*, 1975, *11*(285-296), 23-27.

[8] Ojala, T., Pietikainen, M., & Maenpaa, T "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". *IEEE-PAMI* (2002). *24*(7), 971-987.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. "Imagenet classification with deep convolutional neural networks". In *Advances in neural information processing systems,* 2012 (pp. 1097-1105).

[10] Bar, Y., Diamant, I., Wolf, L., & Greenspan, H. "Deep learning with non-medical training used for chest pathology identification". In *SPIE Medical Imaging* Feb. 2015

[11] Vedaldi, A., & Lenc, K. "MatConvNet-Convolutional Neural Networks for MATLAB". 2014, *arXiv preprint arXiv:1412.4564*.

[12] Chatfield K., , Simonyan, K. , Vedaldi, A. & Zisserman A. *"Return of the Devil in the Details: Delving Deep into Convolutional Networks', BMVC 2014*