Accepted in IWSSIP 2009 Conference

Hierarchical discovery of patterns of infections in chest radiographs using non-negative matrix factorization

Dimitris K. Iakovidis, Spyros Tsevas
Technological Educational Institute of Lamia
Dept. of Informatics and Computer Technology
{dimitris.iakovidis, s.tsevas}@ieee.org

***Abstract*—Chest radiography is a screening tool for the detection and the primary assessment of abnormalities consistent with the infections of the lower respiratory tract. Most commonly, the radiographic manifestations of bacterial infection are areas of consolidation with a segmental or lobar distribution which appear as areas of increased pulmonary opacity. This paper addresses the problem of the discovery of consolidation patterns in chest radiographs, by a methodology for automatic segregation of lung consolidations from normal lung parenchyma. The proposed methodology is based on non-negative matrix factorization (NMF) of pulmonary radiographic patterns represented by intensity histograms and Gabor textural features. NMF is considered as a soft clustering algorithm which is iteratively applied according to an hierarchical cluster-merging scheme. This scheme reckons with the resulting NMF bases so as to overcome the limitations associated with the geometry of the clusters. The experimentation results validate the effectiveness of the proposed methodology and demonstrate its comparative advantage over the conventional hierarchical and NMF partitional clustering algorithms.**

Chest radiography, bacterial infections, image segmentation, clustering, cluster-merging, non-negative matrix factorization

I. INTRODUCTION

Respiratory tract infections represent a major cause of morbidity and mortality, and a leading cause of death worldwide [1]. Chest radiography is almost always the initial diagnostic test performed in patients who present with signs and symptoms suggesting pulmonary infection. The most common pattern in bacterial pneumonia is focal consolidation [2], which typically presents radiographically as single or multiple sites of focal consolidation, in either a segmental or lobar distribution. The radiographic appearance of focal consolidation is defined as an area of increased pulmonary opacity with obscuration of underlying bronchovascular structures and additional interference with the superimposed structures of the thoracic cavity such as the ribs and the mediastinum. The different types of superimposed structures that conflate to the final image contribute to both the diversity and complexity of its content which along with the quality variability induced by the parameters related to the radiation exposure, make its medical interpretation a challenging task. This task has motivated many researchers to develop various computational methods for automatic analysis of chest radiographs [3]. Such methods include detection of the lung fields [4], size measurements of structures of the thoracic cavity [5], detection of the ribs [6], lung nodule detection [7], whereas fewer methods, basically supervised, cope with the detection of abnormalities associated with the presence of pulmonary infections [8]-[10].

The major focus of this paper is on the discrimination of consolidation patterns associated with bacterial pulmonary infections, from patterns extracted from the normal lung parenchyma. For this reason, we propose a novel unsupervised approach that can be considered as a tool for exploratory analysis of the lung fields explicitly based on image features, and can be useful for the extraction of semantic-level features. Moreover, it avoids the need for feature normalization between images, which can be quite complicated and roughly approximative when it comes to the analysis of diverse sets of chest radiographs acquired with different settings.

The non-negative matrix factorization (NMF) of intensity and texture features extracted from radiographic patterns forms the basis of the proposed methodology. Intensity features are represented by grey-level histograms, whereas image texture is represented by Gabor energy features [11]. Since the radiographic opacities are first cues considered in the reading of a chest radiograph by the experts [12], clustering of the intensity feature space is performed firstly. This first step results in a cluster of possibly normal patterns and a cluster of ambiguous patterns, whose ambiguity will be resolved in a further step which involves clustering of the texture space. In order to extend the capabilities of the proposed methodology beyond the inflexible hyperellipsoidal cluster geometries posed by the conventional NMF approach an hierarchical cluster merging scheme is investigated.

The rest of this paper consists of three sections. Section 2 describes the proposed methodology, section 3 presents the results of its experimental evaluation, and section 4 summarizes the conclusions of this study.

II. METHODOLOGY

Non-negative Matrix Factorization (NMF) was introduced

as a dimensionality reduction method for pattern analysis [13] and gained popularity by the works of Lee and Seung [14], [15]. NMF makes use of non-negativity constraints on the data matrix so as to find a lower rank approximation of it. In contrast to other methods such as Principal Component Analysis (PCA), NMF allows only additive combinations of non-negative data, leading to a representation that is more intuitive and closer to the human perception.

Given a $m \times n$ non-negative matrix $\mathbf{V}$ and a reduced rank r $(r < \min(m,n))$, the NMF problem lies in finding two non-negative factors $\mathbf{W}$ and $\mathbf{H}$ of $\overline{\mathbf{V}}$ such that $\mathbf{V} \approx \overline{\mathbf{V}} = \mathbf{W} \times \mathbf{H}$, where $\mathbf{W} \in \Re^{m \times r}$ and $\mathbf{H} \in \Re^{r \times n}$

We may think of $\mathbf{W}$ as the matrix containing the NMF basis and $\mathbf{H}$ as the matrix containing the non-negative coefficients (or encodings) that exhibit a one-to-one correspondence with the data that consists $\mathbf{V}$. In order to quantify the similarity between the data matrix $\mathbf{V}$ and the model matrix $\overline{\mathbf{V}}$ we use as an objective function the Kullback-Leibler (KL) divergence measure $\mathbf{D}(\mathbf{V} \| \mathbf{W} \times \mathbf{H})$ [15] since it is better adapted to real applications where the data manifold is not always flat:

$$\mathbf{D}(\mathbf{V} \| \mathbf{W} \times \mathbf{H}) = \sum_{ij} \left[\mathbf{V}_{ij} \otimes \log \frac{\mathbf{V}_{ij}}{(\mathbf{W} \times \mathbf{H})_{ij}} - \mathbf{V}_{ij} + (\mathbf{W} \times \mathbf{H})_{ij} \right] \quad (1)$$

with $\mathbf{W}, \mathbf{H} \geq \mathbf{0}$. This optimization problem is solved by using the following multiplicative update rules:

$$\mathbf{W}_{ir} \leftarrow \mathbf{W}_{ir} \otimes \frac{\sum_j \frac{\mathbf{H}_{rj} \times \mathbf{V}_{ij}}{(\mathbf{W} \times \mathbf{H})_{ij}}}{\sum_j \mathbf{H}_{rj}}, \mathbf{H}_{rj} \leftarrow \mathbf{H}_{rj} \otimes \frac{\sum_i \frac{\mathbf{W}_{ir} \times \mathbf{V}_{ij}}{(\mathbf{W} \times \mathbf{H})_{ij}}}{\sum_j \mathbf{W}_{ir}} \quad (2)$$

where $\otimes$ denotes the Hadamard (element-wise) product and $\times$ denotes the matrix product.

NMF can be considered as an alternative clustering technique [17],[18]. However, due to its iterative nature, it is likely that the NMF converges to local minima of the objective function (Eq. 2). As a result, different initializations of the NMF algorithm may lead to different clustering results. Proper initialization schemes can improve the performance of the NMF either in terms of computational complexity or in its ability to analyze data. Fuzzy C-Means (FCM) has been proposed as a method to initialize NMF [19], [20], initializing $\mathbf{W}$ with the cluster centroids and $\mathbf{H}$ with the fuzzy membership values assigned to each data vector respectively. Since, $\mathbf{W}$ and $\mathbf{H}$ correspond to a clustering result after their initialization, NMF can be regarded as the method to improve this result, leading to a more visible cluster structure [20].

In this paper, we apply the described NMF clustering approach for the discrimination of consolidation from normal patterns in plain chest radiographs. It is assumed that the lung fields are isolated in regions of interest (ROIs) defined either manually or with a pre-processing lung field boundary detection algorithm [3], [21]. The proposed image analysis methodology is applied only on these ROIs.

A block diagram of the proposed methodology is given in Fig. 1. Considering that the radiographic opacities are evaluated first in the reading of a chest radiograph by the experts [12], as a first step in the proposed methodology, N local grey-level histogram signatures capturing image intensity information are extracted from non-overlapping square sub-images raster-sampled from the lung area. These signatures are subsequently clustered into $r_1 = \lfloor N/c \rfloor$ clusters, where $c \leq N/2$ is a properly chosen positive constant so that the cluster cardinalities remain small. By splitting the feature space into many small clusters, we expect that some of them will be formed from patterns of normal lung parenchyma, others from consolidation patterns, and others from both normal and consolidation patterns. The clusters comprising of normal patterns will be characterized by smaller intensity values than the rest ones, since the lungs are normally filled with air, which has the smallest radiographic density. Choosing directly $r_1=2$ ($c=N/2$) clusters, the cluster shapes would be limited to hyperellipsoids, and it would be a reasonable choice if the

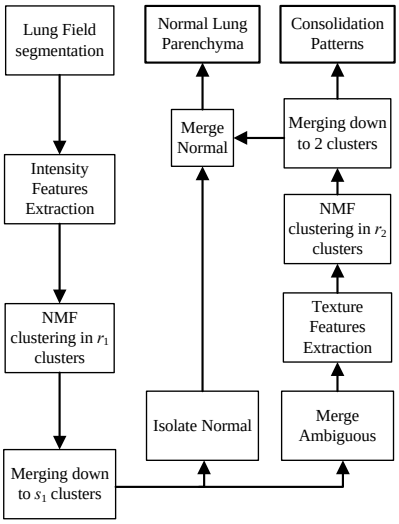

Figure 1. Block diagram of the proposed methodology.

target clusters were also hyperellipsoids.

Based on that observation, the r_1 clusters are dyadically merged down to $s_1 \geq 2$ clusters based on the similarity of their centroids, which are derived from factor $\mathbf{W}$. Since the signatures are intensity histograms, the similarity is evaluated by the histogram intersection metric [23]. Dyadic merging takes place by merging, in each iteration, the two clusters whose centroids exhibit the maximum similarity. Ideally the value of s_1 would be equal to two, since the patterns are expected to be either normal or abnormal. However, a perfect discrimination of the patterns in two clusters is usually infeasible due to limitations regarding the sampling policy or the superimposed structures, mainly posed by the feature extraction procedure. It can be noticed that the ambiguity is higher for higher intensity levels; however, the certainty of a histogram signature with very low intensity levels to belong to the normal lung parenchyma, is higher. Therefore, if the value of s_1 is greater than two, it is more likely that a cluster of normal patterns is formed and that the rest of the clusters contain ambiguous patterns. The exact value of s_1 is experimentally determined.

The ambiguous clusters are merged into a sole cluster of N_a ambiguous patterns and textural signatures extracted from the corresponding sub-images. The signatures are formed by the energies estimated from the outputs of two Gabor filter banks; one with symmetric and one with anti-symmetric Gabor kernels. Each bank comprises of 16 Gabor filters, resulting from the use of four values of orientation and four values of radial frequency [11], [22]. The textural signatures are subsequently clustered into $r_2 = [N_a / c_a]$, where $0<c_a \leq N_a/2$.

The r_2 clusters of textural signatures are dyadically merged down to two clusters based on the Euclidian similarity of their centroids. The resulting two clusters should correspond to the consolidation patterns on the one hand, and to the remainder normal patterns on the other. Finally, the image regions with the consolidations will be comprised of the sub-images corresponding to the discovered consolidation patterns, whereas the image regions of the normal lung parenchyma will be comprised of the normal patterns discovered in both the first and the second step of the proposed methodology.

III. RESULTS

For the evaluation of the proposed methodology, a collection of chest radiographs obtained from twenty four patients with diagnosed bacterial pulmonary infections hospitalized in an intensive care. In all radiographs the infections were manifested as foci of consolidations. The radiographic images were 8-bit grayscale with a size of 2K×2K pixels. The lung fields were isolated by manual delineation by an expert and were further sampled with 32×32-pixel sub-images. The initial number of clusters r_1 in the first step as well as the initial number of clusters in the second step of the proposed methodology were determined by setting the constants c and c_a to 1/10 of the number of signatures to be clustered.

Comprehensive experiments were conducted to investigate the performance of the proposed methodology using various combinations of configurations, feature sets and clustering approaches. These include clustering in two (s_1=2) vs three (s_1=3) target clusters in the first step, split-merge (SM) clustering vs split (S) using direct clustering. Moreover, experimentation took place using simple hierarchical clustering (H) for comparison purposes. For the evaluation of similarities in the simple hierarchical clustering, the histogram intersection metric [23] was used for the intensity histograms, whereas for the Gabor textural features Euclidean similarity of their centroids was considered. The combinations tested are illustrated in Fig. 2.

The performance measures considered in this study are: sensitivity, specificity and accuracy [24]. They were estimated from the number of pixels classified as true positive TP=GTP ∩ PCLA, true negative TN=GTN ∩ NCLA, false positive FP=GTN ∩ PCLA, and false negative FN=GTP ∩ NCLA, where PCLA (positive cluster lung area) is the area corresponding to the patterns considered as consolidations, NCLA (negative cluster lung area) is the area corresponding to the patterns considered as normal lung parenchyma, and GTP and GTN are the ground truth areas of consolidations and normal lung parenchyma, respectively.

The average results estimated from the application of the proposed approach on the whole dataset are illustrated in Fig. 3. It can be observed that the highest average accuracy (94.1%) is achieved in the case of SM(I)-S2(T), where we have hierarchical cluster merging into s_1=3 clusters followed by a second step where the ambiguous patterns were further clustered to r_2=2 clusters using their texture features.

The results obtained in the case of SM(I)-S(I) are ranked second with an accuracy of 89.3%. This makes clear the importance of texture in the discrimination of the consolidation from the normal patterns. Moreover, it can be noticed that the advantage of the hierarchical cluster merging scheme is prevalent in the first step, whereas it has a marginal effect in the second step that can be attributed to the geometry of the Gabor space.

The results regarding the rest of the cases that evolve in two steps, either if they use a merging scheme (S2-SM, S3-SM) or not (S2-S2, S3-S2), exhibit inferior performance in contrast to the SM-S2 case for both intensity and textural features. It should also be noted that clustering in three target clusters in the first step has a major advantage over clustering in two, since in the first case, at least one of the resulting clusters contains solely normal patterns demonstrating 100% purity, in contrast to the second case where no-one strictly normal cluster can be defined (purity about 75%).

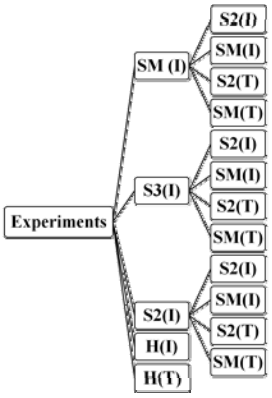

Figure 2. Combinations of methodologies and features considered for the evaluation of the clustering performance of the proposed methodology. SM stands for Split-and-Merge into 3, S3 stands for Split into 3 clusters and S2 stands for Split into 2 clusters, whereas H stands for the simple hierarchical clustering. I and T represent the type of features that were used in each step. I stands for Intensity and T stands for textural features.

Finally, the accuracies obtained in the cases that involve one step clustering only (SM(I), S(I) and simple hierarchical clustering, H(I) and H(T)), are poor. In the case of SM(I) and S(I) the accuracy achieved is less than 70%, whereas in the case of hierarchical clustering, though accuracy is relatively high (about 85%) the exhibited sensitivity rates are particularly low. The low sensitivity values in H(I) and H(T) can be attributed to the fact that hierarchical clustering is able to distinguish only a small fraction of the consolidation patterns classifying the rest with the normal ones. As a result, the inability of the one-step process to conclude to a well defined consolidation cluster underlines the need for a second

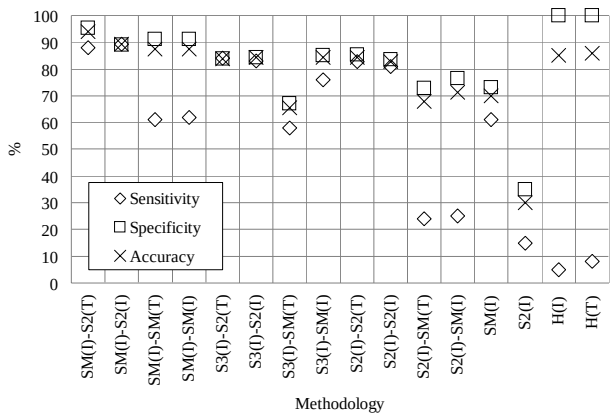

Figure 3. Results of the experiments outlined in Fig. 2.

refinement step. An example segmentation of a chest radiograph using the proposed methodology (SM(I)-S(G)) is illustrated in Fig. 4.

Figure 4. Images resulting from each clustering step: (a) lung fields indicated by the expert, and the discovered (b) consolidation and (c) normal patterns.

IV. CONCLUSIONS

This study presented a novel methodology to the discovery of patterns of bacterial pulmonary infections from plain chest radiographs, based on hierarchical merging of clusters derived from NMF. The proposed methodology is capable of automatically discriminating the normal lung parenchyma from consolidations caused by bacterial pulmonary infections and as its experimental evaluation demonstrates, it is more discriminative than the conventional hierarchical and NMF-based partitional clustering algorithms. Moreover, considering that NMF is used to improve the output of the FCM [20], it will consequently outperform also FCM. The advantage of the proposed methodology in the clustering performance comes with only a small cost in the computational complexity.

Another important conclusion is that texture is an important feature that should be co-evaluated with the image intensity for the discrimination of the radiographic patterns of interest. Using only intensity features and three target clusters instead of two in the first step leads to a cluster of normal patterns of high purity. The textural features are necessary to resolve the ambiguity among the remainder patterns in the second step of the methodology.

Future work includes further experimentation with the proposed and alternative cluster merging schemes on a larger dataset, investigation of combined supervised-unsupervised approaches and integration into a multimodal data mining system for adverse events detection, which will be capable of co-evaluating radiographic findings of patients with bacterial infections.

ACKNOWLEDGMENT

Great thanks to G. Papamichalis, M.D. who generously offered his help and advice on the medical aspects of this study. This work was supported in part by the European Commission's Seventh Framework Information Society Technologies (IST) Programme, Unit ICT for Health, project DEBUGIT (no. 217139).

REFERENCES

[1] World Health Organization, The top ten causes of death, Fact sheet no. 310, Nov. 2008

[2] N.L. Müller, T. Franquet, K.S. Lee, C. Isabela, and S. Silva, Imaging of Pulmonary Infections, Lippincott Williams & Wilkins, 2006.

[3] B.V. Ginneken, B.T.H. Romeny, and M.A. Viergever, "Computer-Aided Diagnosis in Chest Radiography: A Survey," IEEE Tr. Med. Im., 20(12).

[4] B.V. Ginneken, M.B. Stegmann, M. Loog, "Segmentation of Anatomical Structures in Chest Radiographs using Supervised Methods: A Comparative Study on a Public Database," Med.Im.Anal., vol. 10, 2006.

[5] I.C. Mehta, Z.J. Khan, and R.R. Khotpa, Volumetric Measurement of Heart Using PA and Lateral View of Chest Radiograph, AACC 2004, LNCS 3285, pp. 34–40, 2004.

[6] M. Loog, B.van Ginneken: Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. IEEE Trans. Med. Imaging 25(5), pp. 602-611, 2006.

[7] Giuseppe Coppini, Stefano Diciotti, Massimo Falchini, N. Villari, Guido Valli: Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms. IEEE Tr.Inf.Tech.Biomed. 7(4), 2003.

[8] B.V. Ginneken, S. Katsuragawa, B.T.H. Romeny, K. Doi, and M.A. Viergever, "Automatic Detection of Abnormalities in Chest Radiographs Using Local Texture Analysis," IEEE Trans. Med. Im. vol. 21, no. 2.

[9] X. Xie, X. Li, S. Wan, and Y. Gong, Mining X-Ray Images of SARS Patients, G.J. Data Mining, LNAI 3755, pp. 282–294, 2006.

[10] L.L.G. Oliveiraa, S. Almeida e Silvaa, L.H. Vilela Ribeirob, R. Maurício de Oliveiraa, C. J. Coelhoc and A.L.S.S. Andrade, "Computer-Aided Diagnosis in Chest Radiography for Detection of Childhood Pneumonia," Int. J. of Med. Inf., vol. 77, no. 8, pp. 555-564, 2007.

[11] Anil K. Jain and Farshid Farrokhnia, Unsupervised Texture Segmentation Using Gabor Filters, Patt. Recog., 24(12), 1991.

[12] Novelline, R.A., Squires's Fundamentals of Radiology, Harvard University Press, 1997.

[13] C. Bezdek, J. Keller, R. Krisnapuram, and N. R. Pal. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, 1999.

[14] Paatero and U. Tapper. Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics, 5(1), pp. 111–126, 1994.

[15] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Adv. Neural Inf. Process. Systems, 13, 2000.

[16] C. Ding, X. He, H.D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," Proc. of the SIAM Int'l Conf. on Data Mining, pp. 21–23, April 2005.

[17] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization, Proc.ACM Conf.Res. Dev. in IR, pp.267–273, 2003

[18] Ding, C., Li, T., Peng, W. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing, Comp. Stat. & Data Anal., 52, 2008.

[19] Z. Zheng, J. Yang, Y. Zhu, Initialization enhancer for nonnegative matrix factorization, Eng. Appl. Artif. Intell. 20 (1), pp. 101–110, 2007.

[20] Okun, O., Priisalu, H. Unsupervised data reduction Sign. Proc., 87 (9).

[21] D.K. Iakovidis, and G. Papamichalis, Automatic Segmentation of the Lung Fields in Portable Chest Radiographs Based on Bézier Interpolation of Salient Control Points, in Proc. IEEE Int'l Conf. on Imaging Syst. and Techniques, Chania, Greece,pp. 82-87, 2008.

[22] Grigorescu, S.E., Petkov, N., Kruizinga, P., Comparison of texture features based on Gabor filters, IEEE Tr.Im.Proc., 11 (10), 2002.

[23] M.J. Swain, D.H. Ballard, Color Indexing. Int. J. Comp. Vision, 7(1).

[24] Han, J., Kamber, M., Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.