Segmentation And Analysis Of Interstitial Lung Disease Pattern In HRCT Images And Disease Classification By K- Means

Sonal P. Wasekar¹, Sneha Sahare², Sangharsh Lanjewar³
¹PG student CSE, DBACER/RTMNU, India
²,³Asst. Prof CSE, DBACER/RTMNU, India

Abstract— automated segmentation of pathological bearing region is the first step towards the development of lung CAD. Most of the work reported in the literature related to automated analysis of lung tissue aims towards classification of fixed sized block into one of the classes. This block level classification of lung tissues in the image never results in accurate or smooth boundaries between different regions. In this work, effort is taken to investigate the performance of the automated image segmentation or clustering algorithm that results in smooth boundaries among lung tissue patterns commonly encountered in microscopic images of the thorax. A public database that consists of HRTC images taken from patients affected with Interstitial Lung Diseases (ILDs) is used for the evaluation. The K-means algorithm is used here to find the clusters. The performance of K-means algorithm is evaluated for varied value of K (number of clusters) and its effect is studied.

Keyword—Lung disease, CBIR, Segmentation, k-means clustering, GMM-EM, HMRF-EM

I. INTRODUCTION

Segmentation of the lung tissue patterns into several regions is one of the most important steps for the computer aided detection of the abnormal region in High Resolution Computed Tomography (HRCT) images of the thorax. Most of the work reported in the literature related to lung tissue pattern classification aims to classify the fixed sized (either square or circular) region of interest (ROIs). These types of block level classification never result in the exact or smooth boundary of the pathological bearing region (PBR). So many boundary pixels are misclassified to one among the unintended neighboring regions. In practice, the spread of the disease is not symmetric but arbitrary. In addition at times, these predefined size ROIs are not representative of those diseases patterns. For example, an ROI of size 32 * 32 may represent an ILDs pattern called consolidation or emphysema well but may not represent patterns like fibrosis or micronodule. Therefore, the entire spread of the disease patterns needs to be considered to represent the pathology for computer-aided detection or analysis of the disease. It is also found that for most of the content-based image retrieval (CBIR) system for lung images reported in the literature delineation of these ROIs is a manual process which not only involves huge human interventions but also leaves chances of human error. Automated unsupervised clustering techniques could result in different regions with smooth boundaries between them which are helpful for representing the pathology better for both of these applications (classification and retrieval). Automated segmentation also substantially reduces the amount of human intervention for the image annotation task during the database creation for a CBIR system. The aim of this work is to evaluate the performance of three such algorithms in the context of automated clustering of lung tissues affected with Interstitial Lung Diseases (ILDs) into different regions of interest.

Interstitial Lung Disease (ILD):
Interstitial Lung Disease is a general category that includes many different lung conditions. All interstitial lung disease affects the interstitium, a part of the lungs anatomic structure. The
interstitium is a lace-like network of tissues that extends throughout both the lungs. The interstitium provides support to the lungs' microscopic air sacs (alveoli). Tiny blood vessels travel through the interstitium, allowing gas exchange between blood and the air in the lungs. Normally, the interstitium is so thin it can't be seen on chest X-rays or CT scans. Hence, it is very difficult to identify the diseases using conventional imaging techniques.

II. LITERATURE SURVEY

The research on the detection of the interstitial diseases by existing imaging methods is very limited to its application due to unclear boundaries and tiny nature of the infected region. However, several researchers have tried to propose a method for accurate detection of the critical interstitial diseases. The authors Pradipta Maji and Sankar K. Pal said that CLUSTER analysis is a technique in finding natural groups that are present in data. It divides a given data set into a set of clusters in such a way that two objects from the same cluster are as similar as possible and the objects from different clusters are as dissimilar as possible [1], [2]. Clustering techniques have effectively been applied to pattern recognition, machine learning, biology, medicine, computer vision, communications, remote sensing, etc. A number of clustering algorithms have been proposed to suit different requirements [3, 4]. Image segmentation methods are generally based on one of two fundamental properties of the intensity values of image pixels: similarity and discontinuity. In the first category, the concept is to partition the image into several different regions such that the image pixels belonging to a region are similar according to a set of predefined criteria’s. Whereas, in the second category, the concept of partition an image on the basis of abrupt changes in the intensity values is used. The performance of a clustering algorithm may be affected by the chosen value of K. Therefore, instead of using a single predefined K, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the specific characteristics of the data sets. At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering.

Image segmentation refers to the process of partitioning an image into mutually exclusive regions. It can be considered as the most essential and crucial process for facilitating the delineation, characterization, and visualization of regions of interest in any medical image. Despite intensive research, segmentation remains a challenging problem due to the diverse image content, cluttered objects, occlusion, image noise, non-uniform object texture, and other factors. There are many algorithms and techniques available for image segmentation but still, there needs to develop an efficient, fast technique of medical image segmentation [5]. The goal of segmentation is to change the representation of an image to be more meaningful and easier to analyze. It is used in order to locate objects and boundaries in images. The result of image segmentation occurs as a set of regions that collectively covers the entire image [6]. Therefore, medical image segmentation plays a significant role in clinical diagnosis. It can be considered as a difficult problem because medical images commonly have poor contrasts, different types of noise, and missing or diffusive boundaries of unlabeled data.

III. IMPLEMENTATION DETAILS

A. Experimental setup

The publicly available database12 is used to evaluate the performance of three segmentation algorithms. This database contains 110 ILD cases. All the images in the database are in DICOM format and are obtained with a slice thickness of 1 mm, and inter-slice distance of 10 mm. The images in the database are annotated with the 2-D arbitrary region of interests (AROIs) representing 17 image-findings related to ILDs by two experienced radiologists. From each lung tissue pattern, 25 AROIs are collected and the final database contains a total of 150 images each with a single arbitrary region of interest (AROI).
Three popular algorithms selected for this study are: Gaussian Mixture Model with Expectation Maximisation (GMM-EM), Hidden Markov Random Field with Expectation Maximisation (HMRF-EM) and k-means algorithm.

B. System architecture
The implementation of this system encompasses various modules in to a single architecture. The main purpose of this system is to detect the cancer by using content based microscopic image retrievals system and gives the information about the types of cancer and their subtype.

System Flow Diagram

C. Algorithms
1. Low level feature extraction:

   a) Color Features: - For color features extraction are using the histogram method. Gray-scale features are extracted using three different color spaces for a g image i.e red–green–blue (RGB) color space, CIELab (Lab) and hue–saturation–value (HSV) color spaces. Features extracted from the Lab space characterize the intensity and color information of images separately. The HSV color space can be separate the chromatic and achromatic components.

   b) Feature Extraction: - Co-occurrence histogram is used method for texture feature extraction. The color co-occurrence histogram (CH) keeps track of the number of pairs of certain colored pixels that occur at certain separation distances in image space. By adjusting the distances over which we check co-occurrences, we can adjust the sensitivity of the algorithm to geometric changes in the object’s appearance such as caused by viewpoint change or object flexing. The CH is also robust to partial occlusions because the project does not require that the image account for all the co-occurrences of the model. The project represents each model image as a color CH. The color CH holds the number of occurrences of pairs of color pixels. The purpose of the colors of each model image separately using a standard nearest neighbor k means algorithm. This quantization in color and distance means the CH is represented by CH (i, j, k), where I and j index the two colors from set C, and k indexes the distance range from set D. The intersection indicates how well the image CH accounts for the model CH. If the image accounts for all the entries in the model CH, then the intersection will be equal to the sum of all the entries in the model CH. The intersection is an attractive measure of similarity because if the image CH contains extra entries from the background, they will not reduce the quality of the match. The shapes identified will be stored in the database and will be used for later use. This will create a dictionary which will store all shapes like round, circular rectangular etc. It will help us in the initial search after first use.
2. GMM-EM Algorithm (Gaussian Mixture Model with Expectation Maximization (GMM-EM))

The image will be searched based on their image level similarities. Here purpose system uses the dictionary of objects. This will help us in reducing the search space, thus makes the algorithm efficient. This algorithm provides us the frequency of similar images per image in the dataset to a given query image set or a slide in terms of scores. Scores are computed by summing the number of occurrences of each image in the dataset for a $k$-nearest neighbor (KNN) search of that query image set. The output of this algorithm is the traditional image-level-based retrieving of most similar images from the given dataset and their image-level scores.

3. HMRF EM Algorithm (Hidden Markov Random Field with Expectation Maximization (HMRF-EM))

In our alternative approach to image-level retrieval, Propose to retrieve similar images from the database by keeping the slide-level semantic grade among the retrieve images For this purpose are introduced a slide-level retrieval methodology. The conventional way of ranking the similarity of slides to a given query image set is by sorting the similarity scores of the reference slides independent from their subtypes and retrieving the highest scored slides from the database, which means that subtypes of the slides are considered equally important. In our proposed approach, the first step is to scale the score of each slide by assigning different weight parameters based on subtype frequencies over the reference database.

4. Relevance:

Purpose system store the final output set of images with the given search criteria so that it will help us in making the algorithm intuitively learning by itself. it can also help us in increasing the efficiency.

5. Analysis

The efficiency of the present approach is compared with the already existing methods.

6. k-means algorithm

The K-means algorithm is a popular data-clustering algorithm. However, one of its drawbacks is the requirement for the number of clusters, K, to be specified before the algorithm is applied.

In order to perform image clustering, we first have to choose a representation space and then to use an appropriate distance measure (similarity measure), to match between images and cluster centers in the selected representation space. The image clustering is then performed in a supervised process, using human intervention or in an unsupervised process, relying on the similarity between the images and the various cluster centers.

The k-means algorithm is an iterative technique that is used to partition into $K$ clusters. The basic algorithm is

1. Pick $K$ cluster centers, either randomly or based on some heuristic.
2. Assign each pixel in the image to the cluster that minimizes the distance between the Pixel and the cluster center
3. Re-compute the cluster centers by averaging all of the pixels in the cluster
4. Repeat steps 2 and 3 until convergence is attained (i.e. no pixels change clusters).

IV. CONCLUSION

The Proposed system in the present study demonstrates the novel content-based microscopic image slide retrieval algorithm. The suggested weighting scheme is inspired by IR theory and reveals that the slide-level retrieval performance of the CBIR system is considerably better than the traditional image-level retrieval accuracy for all challenging diseases. The proposed system is CBIR based
which detects cancer very efficiently on the multiple slides and extracts more color features and texture. Application of the proposed weighting strategy, inspired by the IR theory, is not limited to microscopic images only and can be also useful for any type of multi-query search and content-based retrieval systems.

REFERENCES