International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 22 No. 2 Apr. 2016, pp. 304-311

http://www.ijisr.issr-journals.org/

Degraded Document Image Binarization Using Optical Character Recognition

M. Manimaraboopathy[1], M. Anto Bennet[2], M. Kalpana[3], S. Premalatha[3], and G. Gayathri[3]

[1]Assistant Professor, Department of Electronics and Communication Engineering, VELTECH, Chennai-600062, India

[2]Professor, Department of Electronics and Communication Engineering, VELTECH, Chennai-600062, India

[3]UG Student, Department of Electronics and Communication Engineering, VELTECH, Chennai-600062, India

ABSTRACT: The proposed OCR algorithm to retrieve the text in the scanned document images. Here the text detection algorithm based on two machine learning classifiers: one allows generating candidate word regions and the other filters out non-text ones. The extract connected components (CCs) in images by using the maximally stable extremal region algorithm. In CC clustering adaboost classifiers are used to determine whether the region contains text or not. Then using binarization method, the gray image is converted into binary image. The binarization outcomes are subject to OCR and the corresponding result is evaluated with respect to character and word accuracy. As more and more text documents are scanned fast and accurate. Additional performance metrics of the percentage rates of broken and missed text, false alarms, background noise, character enlargement and merging. This effectiveness of the proposed method is also confirmed by tests carried on realistic document images. For proposed algorithm MATLAB version 13 software is used.

KEYWORDS: Maximally Stable Extremal Regions(MSER), optical character recognition (OCR).

INTRODUCTION

Document image binarization is an important pre-processing step to document image analysis and recognition. Also, it can be considered as a critical stage in OCR software systems since the result of the subsequent steps is highly dependent on its effectiveness. This is the reason why document image binarization has been a subject of extensive research during the last decades. The contrast enhancement methods, and then at spatial filtering methods that sharpen edges and remove much of the image blur. The image shown in the Fig.1 is the input image given for binarization. (Detector calibration is usually the first step of the image enhancement chain, but this was discussed earlier as part of the sensor modelling.) For simplicity, assume that the images have an eight-bit dynamic range; i.e., there are 28 = 256 possible gray levels, so the gray levels in the image will be in the range 0–255, with zero being black and 255 being white. Colour images have three arrays of numbers typically representing the red, green, and blue images that are combined to give the full spectrum of colours. The image shown in the Fig.2 is the binarization result. Focus on processing single-band images, i.e., black and white images.

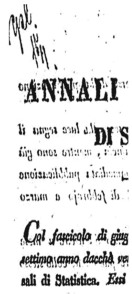


Fig.1 Input image

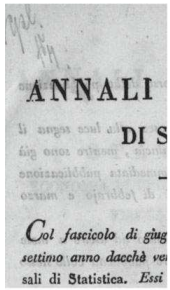


Fig.2 Binarization result

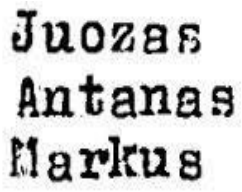


Fig.3 Sample for noise image

The image is divided into segments, denoted by (S), of fixed size. In each segment, the frequency of black pixels is calculated. The selected segments form areas by connecting neighboring segments in respect to their original position in the image. The row-by-row labeling algorithm is used for scanning the document by the window. The parameter k in the formula determines the sensitivity of the detection method. The higher the k, the less segments will be detected. The image shown in the Fig.3 is the sample for noise. Noise area can be having two type foreground and background.In the user was assisted by software to create the ground truth for machine-printed images, by merging and splitting clusters in the character clustering stage, as well as by adding and removing character models to degraded character instances in the character matching stage. The image shown in Fig.4 is the original image. However, the aforementioned procedure can be applied only on machine-printed document without many different font types.

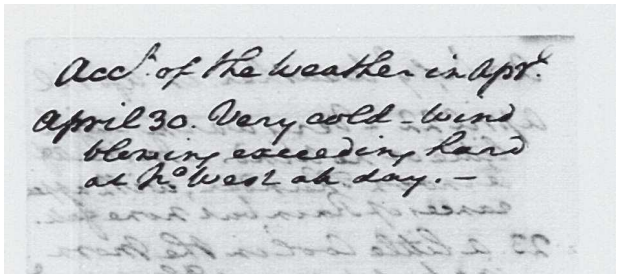


Fig.4 Original image

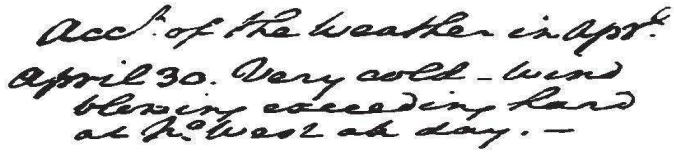


Fig.5 Ground truth

In the proposed methodology, the construction of ground truth plays an important role, since it aids towards the automation of the evaluation procedure. It consists of two distinct stages, namely Skeletonized Ground Truth (SG) stage and Estimated Ground Truth (EG) stage. The image shown in Fig.5 is the ground truth image.Transcript mapping (or text alignment) techniques are used in order to map the correct text information to a segmentation result produced automatically. Usually, these techniques are very useful in order to automatically create benchmarking data sets. The image shown in Fig.6 Sample for skeletonized ground truth image**.**They are mainly based on hidden Markov models (HMMs) and dynamic time warping (DTW) and mainly focus on the alignment of handwritten document images with the corresponding transcription on word level. An efficient transcript mapping technique to ease the construction of document image segmentation ground truth that includes text-image alignment in text line, word and character level.

Fig.6 Sample for skeletonized ground truth image

By concluding this noise in background and foreground can be quantitatively measured by using the garbour filter. Different type of noise can be detected and the ground truth images used for matching the input with grayscale image. Finally the degraded document can be detected.

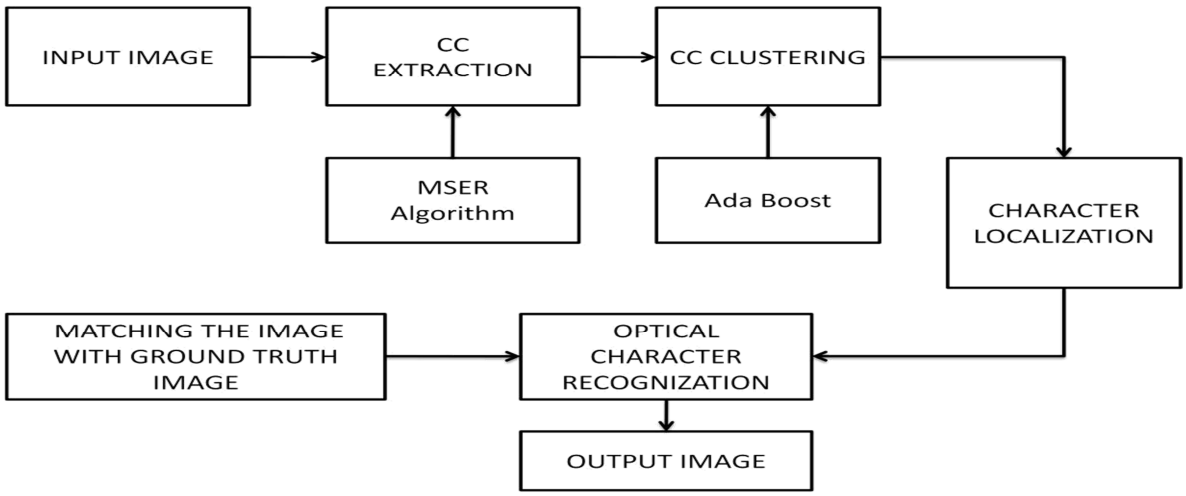
LITERATURE REVIEW

Lamiroy, B. et al., (2011) proposed techniques used here were Fuzzy logic, Self organized neural networks. The process done here was used the image colour values and additional local spatial features extracted in the neighbourhood of the pixels. Both image and local features values feed a kohonen self-organised feature map neural network classifier. An illuminated document easily modified but requires additional spatial information. Low, S.H. and Maxemchuk, N.F. (2010) proposed techniques used here were Bi-modal distribution, Optimization. The process done here was a strongly bi-modal image with smooth regions in both the foreground and background, while allowing for sharp discontinuities at the edges. Then bi-modal and average that measure desired properties in text image but it gives low-resolution image. Moghaddam, R.F. and Cheriet, M. (2010) proposed techniques used here were Sauvola‟s binarization method, Automation and Optimal selection of binarization method. The process done here was tried to compare binarization algorithms by using the precision and recall analysis of the resultant words in the foreground or by evaluating their effect on end-to-end character or word recognition performance in a complete archive document recognition system utilizing OCR. The process provides quality images but this model is not over parameterized. Matsui, E. et al., (2010) proposed techniques used here are adaptive filtering, restoration, scanning. The process done here was that the correction of show through of the printed documents. It offers rate improvement in thresholding and successfully eliminating show through from back side but the efficiency is not much better Anto Bennet, M. et al., (2012) proposed techniques are MLIR where it can find the information that expressed in any language. This paper measuring the effectiveness of MLIR system, this explaining the measures that are regularly used for document retrieval. Here the efficiency to evaluate the performance of the system is not more. Anto Bennet, M. et al., (2015) proposed techniques used here were Binarization algorithm, Hybrid algorithm. The process done here was binarization techniques focus either on finding an appropriate global threshold or adapting a local threshold for each area in order to remove smear, strains, uneven illumination etc. Here, a hybrid approach is presented that first applies a global thresholding technique and then, identifies the image areas that are more likely to still contain noise. The processing is efficient but algorithm used is Anto Bennet, M. et al., (2016) proposed classifying data using Boosting algorithm performs supervised learning which is known as machine learning meta-algorithm. Boosting methods are commonly used to detect objects or persons in videoconference, security system, etc. as an approximation of logistic regression, or enhanced with arithmetical improvements of calculation of weight coefficients. This paper provides a good survey of the literature on mining with rare classes and rare cases using Boosting techniques that shows original approach to classification and its variants. Different evaluation metrics on rarity mining are also discussed in this paper, but error reduction is not efficient.

PROPOSED METHOD

CC-based methods use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image. A geometrical analysis is needed to merge the text components using the spatial arrangement of the components so as to filter out non-text components and mark the boundaries of the text regions.A CC-based method could segment a character into multiple CCs, especially in the cases of polychrome text strings and low-resolution and noisy video images. Further, the performance of a CC-based method is severely affected by component grouping, such as a projection profile analysis or text line selection. In addition, several threshold values are needed to filter out the non-text components, and these threshold values are dependent on the image/video database.A block-matching algorithm using the mean absolute difference criterion is employed to estimate the motion. Blocks missed during tracking are discarded. Their primary focus is on caption text, such as pre-title sequences, credit titles, and closing sequences, which exhibit a higher contrast with the background. This makes it easy to use the contrast difference between the boundary of the detected components and their background in the filtering stage. Finally, a geometric analysis, including the width, height, and aspect ratio, is used to filter out any non-text components. Based on experiments using 2247 frames, their algorithm extracted 86% to 100% of all the caption text.A binarized character with missing foreground pixels(false negatives) from the contour that do not affect the character topology, compared to a binarized character for which the lack of the same amount of foreground pixels alters the character topology achieves: a) equal performance when the typical measures of Recall or PSNR are used, because of the same amount of false negative pixels, b) better performance when the distance-based measures MPM and DRD are used, because those measures apply lower penalization near the ground truth image.

BLOCK DIAGRAM OF PROPOSED METHOD



CANDIDATE GENERATION

For the generation of candidates, extract CCs in images and partition the extracted CCs into clusters, where the clustering algorithm is based on an adjacency relation classifier. In this section, first CC extraction method shown. Then (i) to build training samples (ii) to train the classifier, and (iii) to use that classifier in our CC clustering method.

MSER ALGORITHM

The core of the vision system are Maximally Stable Extremal Regions, or MSERs, introduced by Matas etc all for gray-scale images and later ex-tended to color as Maximally Stable Color Regions, or MSCR. Details about MSER and MSCR principles are given respectively. The main usage of MSER detection is for wide-baseline image matching mainly because of its covariance and high repeatability. To match two images of the same scene (taken from different viewpoints), MSERs are extracted from both images and then appropriately described using (usually in-variant) descriptor. Because MSER extraction is highly repeatable, the majority of the regions should be detected in both images.

CC EXTRACTION

For any two nodes s and t in a graph, their connected components are either identical or disjoint start BFS from some node s. This gives one component of the graph pick any currently unexplored node *u* start another BFS. This gives another component. Continue in this manner until all nodes are explored undirected graphs: s-t connectivity. Directed graphs: s-t mutual connectivity directed path from s to directed path from t to s.

MUTUAL CONNECTIVITY

If u and v are mutually reachable, and v and w are mutually reachable, then u and w are mutually reachable. If u and v are mutually reachable, and v and w are mutually reachable, then u and w are mutually reachable. To go from u to w, we can go via v. To go from w to u, we can again to via v.

STRONG CONNECTIVITY OF DIGRAPHS

A Directed graph is strongly connected if every two nodes „u and „v are mutually reachable from each other analogous to connectivity in undirected graph. The strong component containing *s* is the mutually reachable from each other. Analogous to connected component containing s in undirected graphs. For any node u in R, u and s are mutually reachable at any two nodes u and v in R are mutually reachable u and s are mutually reachable; s and v are mutually reachable. Hence, u and v are mutually reachable. The image shown in Fig.8 is the original image and text extracted from the original image. For any node u in R, u and s are mutually reachable any two nodes u and v in R are mutually reachable u and s is mutually reachable; s and v

are mutually reachable. Hence, u and v are mutually reachable for any two nodes u and v, the strong components of u and v are either identical or disjoint.

Fig.8 Original image and extraction of text from original image.

Text in images can exhibit many variations with respect to the following properties:

Size: Although the text size can vary a lot, assumptions can be made depending on the application domain.

Alignment: The characters in the caption text appear in clusters and usually lie horizontally, although sometimes they can appear as non-planar texts as a result of special effects. This does not apply to scene text, which can have various perspective distortions. Scene text can be aligned in any direction and can have geometric distortions.

Inter-character distance: characters in a text line have a uniform distance between them.

Color: The characters in a text line tend to have the same or similar colors. This property makes it possible to use a connected component-based approach for text detection. Most of the research reported till date has concentrated on finding text strings of a single color. However, video images and other complex color documents can contain „text strings with more than two colors for effective visualization, i.e., different colors within one word.

Motion: The same characters usually exist in consecutive frames in a video with or without movement. This property is used in text tracking and enhancement. Caption text usually moves in a uniform way: horizontally or vertically. Scene text can have arbitrary motion due to camera or object movement.

Edge: Most caption and scene text is designed to be easily read, thereby resulting in strong edges at the boundaries of text and background.

Compression: Many digital images are recorded, transferred, and processed compressed format.

CC grouping or Clustering

The main aim of CC grouping is to group adjacent characters detected in the previous steps into separated meaningful words and further reject false positives. Based on the observation that characters in the same word usually share some similar properties, such as intensity, size, stroke width etc., this valuable information can be utilized in CC grouping.For CC grouping Adaboost classifier is used which is use full in finding the adjacency relationship from CC. bounding box of c_i and denote its width and height as w_i and h_i respectively. Given a pair $(c_i, c_j) \in C \times C$ (i not equal to j), the horizontal distance, horizontal overlap, and vertical overlap between two boxes are denoted as d_{ij}, ho_{ij}, and vo_{ij} respectively.

ADABOOST ALGORITHM

This paper presents an algorithm for detecting and reading text in city scenes. This text includes stereotypical forms such as street signs, hospital signs, and bus numbers as well as more variable forms such as shop signs, house numbers, and billboards. The database of city images were taken in partly by normally sighted viewers and partly by blind volunteers who were accompanied by sighted guides using automatic camera settings and little practical knowledge of where the text was located in the image. The databases have been labeled to enable us to train part of our algorithm and to evaluate the algorithm performance. The negative examples were obtained by a bootstrap process similar to Drucker. First selected negative examples by randomly sampling from windows in the image dataset. After training with these samples, applied the AdaBoost algorithm to classify all windows in the training images (at a range of sizes).Those misclassified as text was then used as negative examples for retraining AdaBoost. The image regions most easily confused with text were vegetation, repetitive structures such as railings or building facades, and some chance patterns. The image shown in Fig.9 is the positive examples used for training Adaboost. The previous section described the weak classifiers used for training AdaBoost.

Fig.9 Positive examples used for training AdaBoost

TEXT READING

Then applied commercial OCR software to the extended text regions (produced by AdaBoost followed by extension and binarization). This was used both to read the text and to discard false positive text regions. Overall, the AdaBoost strong classifier (plus extension/ binarization) detected 97.2 % of the visible text in test dataset (text that could be detected by a normally sighted viewer).For typical examples of the text that AdaBoost fails to detect.

Most of these errors correspond to text which is blurred or badly shadowed. Others occur because do not train AdaBoost to detection vertical text or individual letters. (The training examples were horizontal segments usually containing two or three letters/digits). For the 286 extended text regions correctly detected by the AdaBoost strong classifier (plus extension/binarization), then obtained a correct reading rate of 93.0 % (proportion of words correctly read). This required a preprocessing stage to scale the text region. The 7 % errors are caused by small text areas. For text that can read successfully and for text that cannot read.

CANDIDATE NORMALIZATION

After CC Clustering, we have a set of cluster, normalizing these clusters corresponding regions for the reliable text/non-text classification. Also there are important differences between text and face stimuli because the spatial variation per pixel of text images is far greater than for faces. Facial features, such as eyes, are in approximately the same spatial position for any face and have similar appearance.

GEOMETRIC NORMALIZATION

Given $w_i \in W$, first localize its corresponding region. Even though text boxes can experience perspective distortions, approximating the shape of text boxes with parallelograms whose left and right sides are parallel to y-axis. This approximation alleviates difficulties in estimating text boxes having a high degree of freedom (DOF): only have to find a skew and four boundary supporting points. To estimate the skew of a given word candidate w_k , build two sets:

$$T_k = \{t\ (c_i)\ |c_i \in w_i\ \}$$

$$B_k = \{b\ (c_i\ )|c_i \in w_i\ \}$$

Where t (ci) and b (c_i) are the top-center point and the bottom center point of a bounding box of ci, respectively illustration of B_k .For every pair in B_k and T_k, the slope of a line connecting the pair is discredited into one of 32 levels in [−π/8, π8], and each pair votes for the skew angle. After voting, the most common angle is considered as a skew. Then, perform geometric normalization by applying an affine mapping that transforms the corresponding region to a rectangle.

BINARIZATION

Given geometrically normalized images, binary images are to be considered. However, performing the binarization separately by estimating text and background colors. It is because (i) the MSER results may miss some character components and/or yield noisy regions (mainly due to the blur) and (ii) have to store the point information of all CC for the MSER based

binarization consider the average color of CC as the text color and consider the average color of an entire block as the background color.Then, obtain a binary value of each pixel by comparing the distances to the estimated text color and the estimated background color.12 norms in RGB space are used.

TEXT/NON-TEXT CLASSIFICATION

Developing a text/non-text classifier that rejects non-text blocks among normalized images. In this classification, do not adopt sophisticated techniques such as cascade structures, since the number of samples to be classified is usually small. However, one challenge in this approach is the variable aspect ratio.

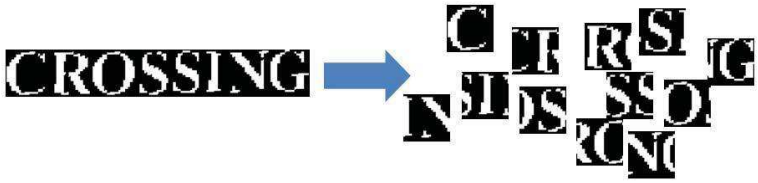


Fig .10 Example for Text/Non-text classification.

One possible approach to this problem is to split the normalized images into patches covering one of the letters and develop a character/non-character classifier. The image shown in Fig.10 is example for Text/Non-text classification. However, character segmentation is not an easy problem. Rather, split a normalized block into overlapping squares and develop a classifier that assigns a textness value to each square block. Finally, decision results for all square blocks are integrated so that the original block is classified.

FEATURE EXTRACTION FROM A SQUARE BLOCK

For the feature extraction, divide a square block into four horizontal and four vertical blocks and extract the features. The image shown in Fig.11 for the feature extraction, by horizontal and vertical blocks.

For a horizontal block Hi (i = 1, 2, 3, 4), consider

1) The number of white pixels,

2) The number of vertical white-black transitions,

3) The number of vertical black-white transitions As features, and features for a vertical block is similarly defined.

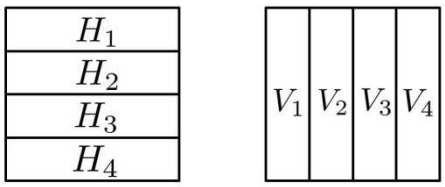


Fig.11 For the feature extraction, by horizontal and vertical blocks.

MULTILAYER PERCEPTRON LEARNING

For the training, normalized images. For this goal, the algorithm is used (i.e., candidate generation and normalization algorithms) to the training images. Shows the multilayer perceptron with two hidden layers. Then manually classified them into text and non-text. Discarded some images showing poor binarization results, and collected 676 text block images and 863 non-text block images. However, it has been found that more negative samples are needed for the reliable rejection of non-text components and collected more negative samples by applying the same procedure to images that do not contain any text 3,568 text images. Multi-layer perceptron is trained for the classification of square patches. One hidden layer is used consisting of 20 nodes and set the output value to +1 for text samples and 0 otherwise. To help the learning, input features are normalized. Finally, text is alone detected by filtering the non text areas.

IMAGE ENHANCEMENT

Feature vectors are extracted to measure useful information from the decomposed sub images. Many feature vectors have been use document image binarization. Most of them were applied to printed documents with clean (white) backgrounds but did not work well for degraded images. Three feature vectors are proposed in this paper, which focus on handwritten document image with messy background and faded writing.

OCR

OCR addresses the problem of reading optically processed characters and has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Most optical character recognition (OCR) systems are designed to transform text images to readable text codes, but perform poorly when text is embedded into complex background because of background interferences and low frequency of occurrence of text.An Adaboost text detection algorithm based on machine learning techniques is proposed. To be precise this method two classifiers are used: one classifier was designed to generate candidates and the other classifier was for the filtering of non-text candidates. MSER method to exploit multi-channel information. Optical character recognition which is used to match the image with the ground truth image. Finally original images are retrieved from the degraded image. In this method yields the state-of-the-art performance both in speed and accuracy.

CONCLUSION

In this paper an Adaboost scene text detection algorithm based on machine learning techniques is proposed. To be precise this method two classifiers are used: one classifier was designed to generate candidates and the other classifier was for the filtering of non-text candidates. MSER method to exploit multi-channel information.Optical character recognition which is used to match the image with the ground truth image. Finally original images are retrieved from the degraded image. In this method yields the state-of-the-art performance both in speed and accuracy. In this method is designed to address the text detection problem in document images, where English alphabets are placed horizontally. This method should be changed in order to detect Asian scripts and texts of arbitrary orientations. The general framework should be extended.

REFERENCES

[1] Lamiroy, B. Lopresti, B. and Sun, T. (2011) „Document analysis algorithm contributions in end-to-end applications: Report on the ICDAR 2011 contest‟, in Proc. Int. Conf.Document Anal. Recognit, Beijing, China. Vol. 5, No. 11, pp. 165-177.

[2] Low, S.H. and Maxemchuk, N.F. (2010) „Performance comparison of two text marking and detection methods‟, IEEE J. Select. Areas Commun., to be published. Vol. 4, No. 9, pp. 165-177.

[3] Moghaddam, R.F. and Cheriet, M. (2010) „A multi-scale framework for adaptive binarization of degraded document images‟,Pattern Recognit., Vol. 43,No. 6, pp. 2186-2198.

[4] Matsui, E. Lins, R.H. and Paredes, R. (2010) „ICFHR 2010 contest:Quantitative evaluation of binarization algorithms‟, in Proc. Int. Conf. Frontiers Handwrit. Recognit, Kolkata, India.Vol.9, No.2, pp.23-36.

[5] AntoBennet, M & JacobRaglend, "Performance Analysis Of Filtering Schedule Using Deblocking Filter For The Reduction Of Block Artifacts From MPEQ Compressed Document Images", Journal of Computer Science, vol. 8, no. 9, pp. 1447-1454, 2012.

[6] Anto Bennet, M, Mohan babu, G, Rajasekar, C & Prakash, P, "Performance and Analysis of Hybrid Algorithm for Blocking and Ringing Artifact Reduction", Journal of Computational and Theoretical nanoscience vol.12,no.1,pp.141-149,2015

[7] Dr. AntoBennet, M , Resmi R. Nair, Mahalakshmi V, Janakiraman G "Performance and Analysis of Ground-Glass Pattern Detection in Lung Disease based on High-Resolution Computed Tomography", Indian Journal of Science and Technology, Volume09 (Issue02):01-07, January 2016.