# Text and Non Text Scene Image Classification for Visually Impaired Through Alexnet Transfer Learning Model

**Anilkumar B, Sreerama Murthy Velaga, A Aswani Devi**

***Abstract*****:** ***Natural Scene Image based text Recognition is a prevalent and exciting research field in computer vision in recent years. For visually impaired people there are some assistive devices which make them to sense the scene images through text extraction. The first and crucial task of the assistive device for text extraction is to detect the text in scene images. This paper proposes a transfer learning based approach with pre-trained CNN model to classify the text and non-text images. AlexNet is the pre-trained architecture that is used as binary classifier. The first 5 convolution layers of the AlexNet are freezed. The last 3 layers are fully connected layers, in which the final output layer is modified to size 2, as this is the binary classifier. The images in the dataset must be preprocessed either before training or testing. The preprocessing consists of Denoising and Augmentation. Denoising removes the noise in the input image using Denoising Convolution Neural Network (DnCNN). Data Augmentation includes image resizing, because AlexNet only accepts the RBG images of size 256x256. The proposed model has achieved the accuracy of 99% in classifying the test dataset.***

***Index Terms*****:** ***Scene Text detection, Transfer learning, Alexnet, Classification, DnCNN.***

## I. INTRODUCTION

According to recent statistics of World health organization (WHO) there are 1.3 billion people suffering from distance or near vision impairment. Scene text reading is attributable to many real-world applications including assisting persons with visual impairments, robotic vision, unmanned vehicle navigation, human–computer interaction, and geo-location [1]. Several research contributions are happening with visual substitutions and assistive technology devices are proposed to address these issues. Cameras are used to capturing pictures of the scene, but fail to process and classify the scene due to some limitations. There are two steps in the processing of natural scene text images, (a) text to be detected in the images and (b) text to be extracted and recognized in the images [2]. For visually impaired people the text present in the images should be available, which helps for their mobility in indoor and outdoor environment. Their navigation will become safer if they can understand the presence of sign boards or any kind of notice boards present in their surroundings.

**Revised Manuscript Received on 30 May 2019.**

***** Correspondence Author
**B. Anilkumar***, Department of ECE, GMRIT, Rajam, India.
**Sreerama Murthy Velaga**, Department of CSE, GMRIT, Rajam, India.
**A.Aswani Devi**, Department of CSE, GMRIT, Rajam, India.

The initial step to achieve this is to distinguish the scene images whether they contain text or not. After this classification, all the non-text images are ignored. Only the text images are taken into consideration and further the text detection and extraction can be performed. After text extraction this can be converted into speech, which will be fed to visually impaired people to help them in their independent mobility.

For this text and non-text image classification, Convolution Neural Networks (CNN) are widely used because of their ability and accuracy of image classification. There are 3 categories of using the CNNs for image classification.

1. Training the CNN from scratch to classify text and non-text images
2. Perform fine-tuning by modifying some layers of the feature extraction part of any pre-trained network
3. Perform transfer learning by modifying the output layer of pre-trained network and freezing the whole feature extraction part the network

There are many CNN, those were already trained well by experts for classifying the images. Train the network from scratch is very difficult and time consuming task. Fine tuning is preferable only in the situations like, our current task is not exactly fulfilled by the classification methodology of any existing pre-trained model and there is a need to re-train the feature extraction layers by modifying the weights. This fine tuning is the task which should be performed carefully to achieve the better performance of the resulted network for the required task. Transfer learning is the simple and effective mode of using the existing pre-trained network by simply replacing the size of output layer by the desired number of classes.

Many pre-trained models are available for image classification. Some of the popular Convolution Neural Network Architectures are:

a. AlexNet
b. VGGNet
c. GoogLeNet
d. ResNet
e. MobileNet

All of these models are popular according to the respective applications. In the proposed work, we are selecting AlexNet of text and non-text image classification.

AlexNet is the first deep neural network that gained the better ImageNet classification accuracy by a significant stride. It consists of 5 convolution layers followed by 3 fully connected layers. It resolved the vanishing gradient problem by replacing sigmoid activation function with ReLU activation. This network also solved the over-fitting problem by introducing Dropout layer after each Fully Connected layer of classification part.

## II. RELATED WORK

Many works have been proposed to classify text images from non-text images using various CNN architectures. Xiang Bai et.al. [6] have proposed the text and non-text image classification using multi scale spatial partition network (MSP-net). The input image is segmented as blocks and the text in the image is predicted by checking the each block of the image. In single forward propagation, this network can classify the images by predicting the all blocks simultaneously. This model achieved good accuracy with less time. But some times the model classifies wrongly if the scene image with low illumination conditions. Boris Epshtein et al.[7] defined a method of detecting text in scene images using Stroke Width Transform. This method is not using an CNN but uses an image operator to detect thepresence of text in the image pixels. It's performance is good and it can detect the texts of various fonts and languages.

Pengyuan Lyu et al [8] have used multi dimentiona recurrent neural network(MDRNN) to distinguish the text and non-text naturel images. This model gives the block level predictions by considering region pixels and dependencies among the local pixels. This performs well but much complex as it is using both CNN and MDRNN for text detection in the images.

Nikhil Damodaran et al [9] haveproposed the scene classification with transfer learning. A pre trained CNN called AlexNet is used only for feature extraction and Support Vector Machine(SVM) and Multi Layer Perceptron(MLP)is used for classification.

Zengxi Li et al [10] have proposed a compact CNN Transfer learning model for small scale image classification. In general, ImageNet CNNs have high complexity and requires GPU computing for faster performance. To overcome those drawbacks, transfer learning is decomposed into fine tuning and joint learning stages. It is much complicated process and it is not essential to implement this model for binary classification.

## III. PROPOSED METHOD

### A. System Block Diagram

Figure.1: **Block diagram of proposed method**

### B. Dataset

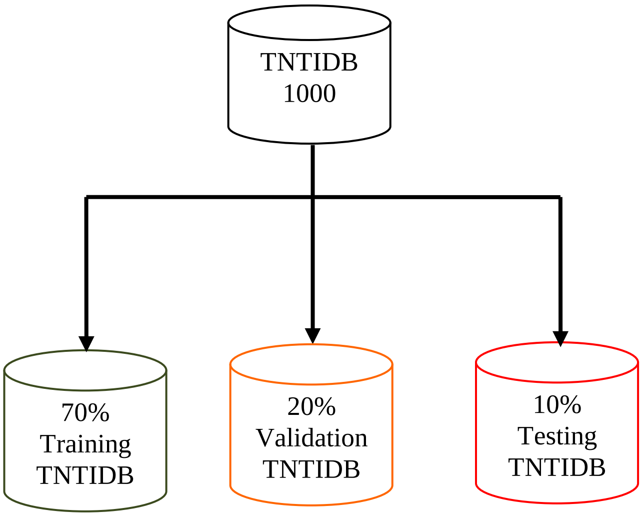


Figure.2: **Dataset split-up**

The entire data set is divided into 7:2:1 ratio for Training, Validation and testing data respectively.

**Table.1:** Training, Validation and Testing Data

| | TNTIDB | Training | Validation | Testing |
|---|---|---|---|---|
| No. of Images | 1000 | 700 | 200 | 100 |
| Text Images | 500 | 350 | 100 | 50 |
| Non-Text Images | 500 | 350 | 100 | 50 |

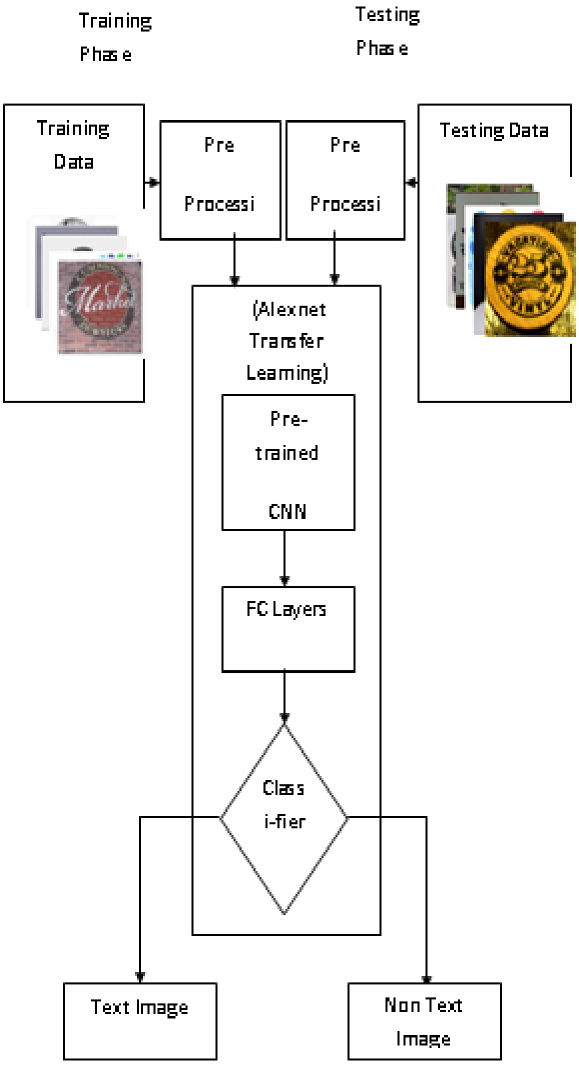
### C. Preprocessing

Pre-processing of Input image data is an essential step before the execution of algorithm on images. This step removes the complexity and the execution time of the entire algorithm by eliminating the unnecessary processing of image data at the time of execution.

### Denoising

For Denoising the input image, a feed forward Denoising Convolution Neural Network (DnCNN) is used. This DnCNN can handle Gaussian Denoising with unknown noise level. This can also detect the simple nosie level and other high frequency artifacts of images.

CNN can be trained to improve the image resolution and remove JPEG image compression artifacts. CNN can be trained as per the requirement for Denoising and that DnCNN is going to fit for the respective applications. DnCNN not only removes the noise from the images but also benefited by GPU computing.



## Data Augmentation

Data Augmentation is the major part of pre-processing. This involves many techniques as mentioned below:

i. Resizing
ii. Scaling
iii. Translation
iv. Rotation
v. Flipping
vi. Adding Salt and Pepper noise
vii. Lightening Conditions
viii. Perspective Transform

Among these data augmentation techniques, the particular operation can be chosen and applied to the images based on the requirement. AlexNet accepts RBG images of size 256x256. So all the training and testing images should be resized to 256x256. So, at present, Image resizing is the only augmentation technique that is going to be applied.

Image resizing is the primary part of data augmentation. Before feeding the image to a neural network, it must be resized according to the requirement of the respective network with which we are going to work on. All images must be resized to process them into batches.

### D. *Alexnet Transfer learning Architecture*

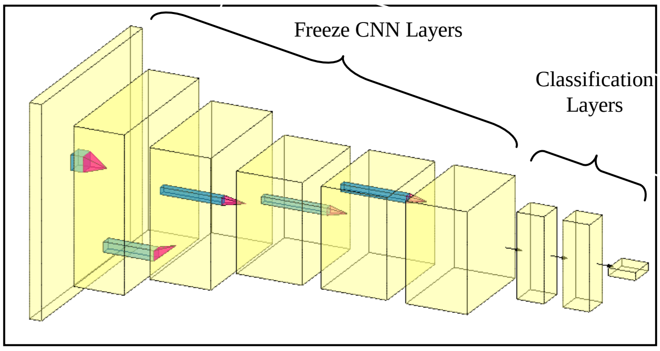


Figure.3: **CNN transfer learning Architecture (Alexnet)**

Many pre-trained models were used as mentioned in the Literature Survey to distinguish the text and non-text images. In this paper, we are using AlexNet to perform transfer learning in order to classify test and non-text images.

While performing transfer learning, we are freezing the feature extraction part of the AlexNet and only modifying the classification part of the network. AlexNet is trained on ImageNet dataset, which can classify 1000 types of images. So, the output layer is the fully connected layer of size 1000 with 'softmax' activation function. The proposed methodology replaces 1000 with 2 as the network is going to classify only 2 classes i.e. the input image contains text or not. After this modification, the network should be trained on the dataset of text and non-text images as mentioned in the respective ratio of TNTIDB1000.

### E. Sample data from training

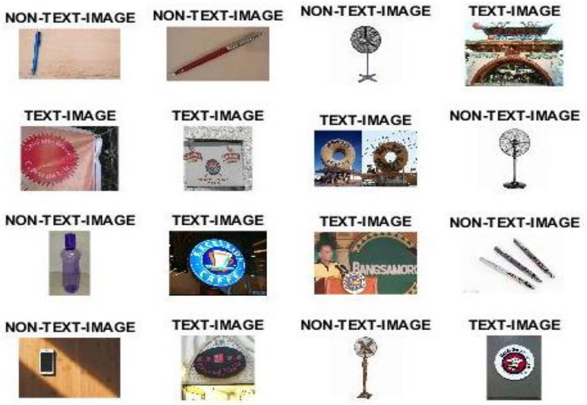


Figure.4: **Sample training Data**

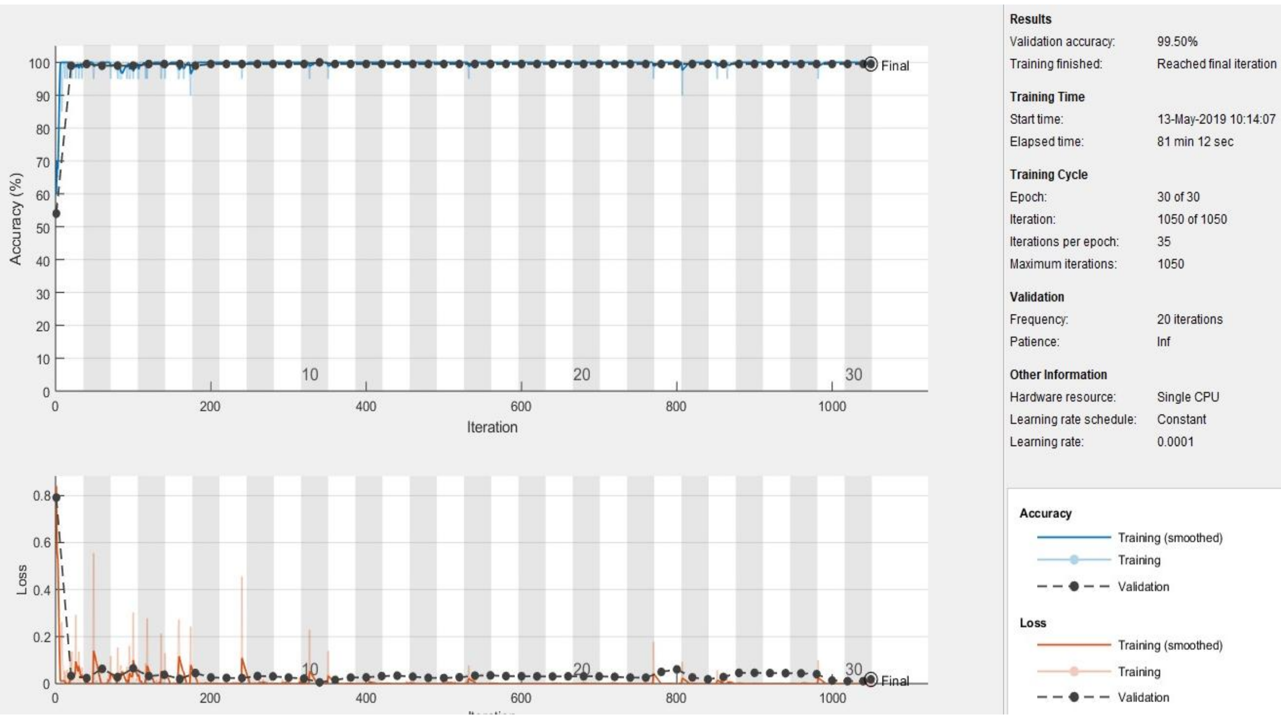
This is the set of training data of 16 images that has been chosen randomly with their corresponding labels.

The implementation of the proposed system is done using MATLAB. Training and the validation process of this binary classification model is plotted in Results section, with desirable level of accuracy.

MATLAB provides the easy and clear path to produce the result regarding the performance of the model.

## IV. RESULTS AND DISCUSSIONS

### A. Training and Validation



Results
Validation accuracy: 99.50%
Training finished: Reached final iteration

Training Time
Start time: 13-May-2019 10:14:07
Elapsed time: 81 min 12 sec

Training Cycle
Epoch: 30 of 30
Iteration: 1050 of 1050
Iterations per epoch: 35
Maximum iterations: 1050

Validation
Frequency: 20 iterations
Patience: Inf

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Figure.5: **Training and Validation Process**

This is the training and validation graph which contains the smooth transition in accuracy curve which gains 99% . The loss curve is depicting significantly less loss in the training process

### B. Confusion Matrix

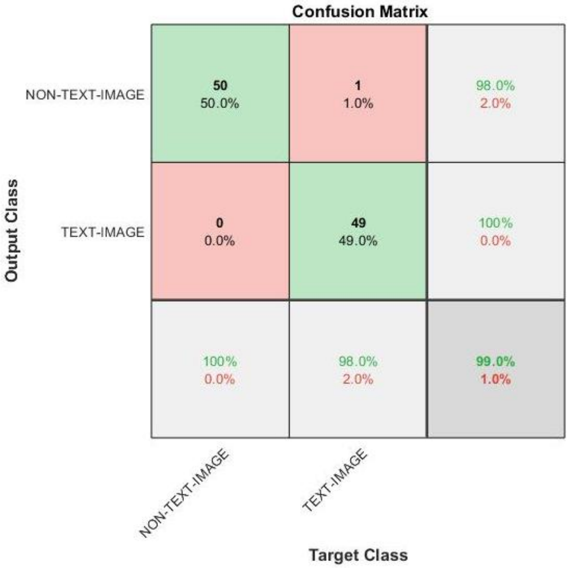
The performance of the classification model is often described by a table called 'confusion matrix'. The accuracy of the binary classifier can be calculated from the below formula:

True Positives: No.of Text images those are classified as they are

True Negatives : No.of Non-Text images those are classified as they are

$$Accuracy = \frac{TruePositives + TrueNegatives}{Total\ \mathrm{Im}\,ages} *100 \qquad (1)$$

$$= \frac{50+49}{100} *100 = 99\%$$



| | NON-TEXT-IMAGE | TEXT-IMAGE | |
|---|---|---|---|
| NON-TEXT-IMAGE | 50<br>50.0% | 1<br>1.0% | 98.0%<br>2.0% |
| TEXT-IMAGE | 0<br>0.0% | 49<br>49.0% | 100%<br>0.0% |
| | 100%<br>0.0% | 98.0%<br>2.0% | 99.0%<br>1.0% |

Figure.6: **Confusion Matrix**

### C. Predicted Outputs

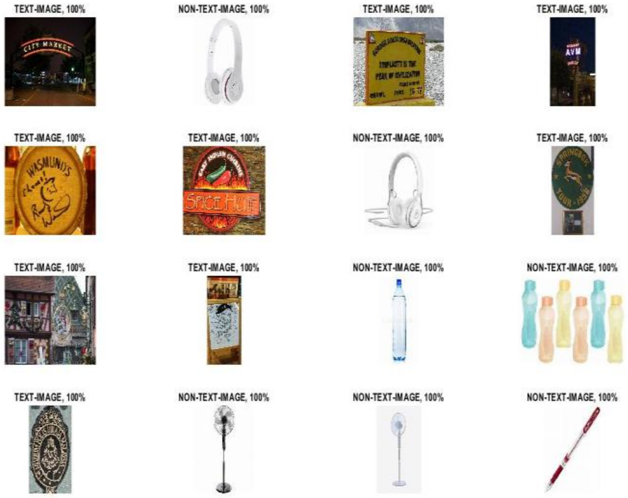


Figure.7: **Predicted Classes testing Images**

Classified. The accuracy is calculated for each test image using confusion matrix.

## V. CONCLUSION

The presented work includes the simple implementation of transfer learning using the classical AlexNet for binary classification of scene images based on whether it contains text or not. The experimental results have shown that the accuracy of the classification model is 99%, which is too impressive. The suggested application of the present work is that, this can be used as the preliminary step for scene text extraction for visually impaired people. In future, this work can be further extended by using a CNN architecture for text detection, text recognition and text extraction which can be further converted into speech for visually impaired people. This can be modeled as a prototype device to assist the navigation of blind people.

## ACKNOWLEDGMENT

This research work is fully funded by Department of Science and Technology, Science for Equity Empowerment and Development Division, Government of India with the wide Number SEED/TIDE/158/2016. The authors would like to thank DST, particularly TIDE division for supporting funding for this project.

## REFERENCES

1. LeenaMaryFrancis, N.Sreenath, “Live detection of text in the natural environment using Convolutional Neural Network”, Systems Volume, September 2019, pp. 444-455.
2. Ghulam Jillani Ansari, Jamal Hussain Shah, Mussarat Yasmin, Muhammad Sharif, Steven Lawrence Fernandes, “A novel machine learning approach for scene text extraction” Future Generation Computer Systems, Volume 87, October 2018, pp. 328-340.
3. Dinh NguyenVan a , e , ∗, Shijian Lu b , Shangxuan Tian c Nizar Ouarti a , e , Mounir Mokhtari d , e, “A pooling based scene text proposal technique for scene text reading in the wild”, Pattern Recognition 87 (2019) 118–129
4. Yuanwang Weia,b, Zhijiang Zhanga, Wei Shena, Dan Zenga,∗, Mei Fanga,b, Shifu Zhoua, “Text detection in scene images based on exhaustive segmentation”, Signal Processing: Image Communication 50 (2017) 1–8
5. Yuanwang Wei a,b, Wei Shen a, Dan Zeng a, Lihua Ye b,c, Zhijiang Zhang a,*, “Multi-oriented text detection from natural scene images based on a CNN and pruning non-adjacent graph edges”, Signal Processing: Image Communication 64 (2018) 89–98
6. Xiang Baia, Baoguang Shia, Chengquan Zhanga, Xuan Caib, Li Qib,∗, “Text/non-text image classification in the wild with convolutional neural networks”, Pattern Recognition 66 (2017) 437–446
7. Boris Epshtein, Eyal Ofek, Yonatan Wexler,” Detecting Text in Natural Scenes with Stroke Width Transform”, Microsoft Corporation
8. Pengyuan Lyu, Baoguang Shi, Chengquan Zhang, Xiang Bai, “Distinguishing Text/Non-Text Natural Images with Multi-Dimensional Recurrent Neural Networks”, (ICPR), 2016, December 4-8, 2016
9. Nikhil Damodaran, V. Sowmya, D.Govind and K. P. Soman, “Scene Classification Using Transfer Learning”
10. Zengxi Li_ Yan Song_ Ian Mcloughlin† Lirong Dai_, “COMPACT CONVOLUTIONAL NEURAL NETWORK TRANSFER LEARNING FORSMALL-SCALE IMAGE CLASSIFICATION”, ICASSP 2106
11. Tong He, Weilin Huang, Member, IEEE, Yu Qiao, Senior Member, IEEE,” Text-Attentional Convolutional Neural Network for Scene Text Detection”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 25, NO. 6, JUNE 2016
12. Youbao Tang and Xiangqian Wu, Member, IEEE, “Scene Text Detection and Segmentation Based on Cascaded Convolution Neural Networks”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 26, NO. 3, MARCH 2017
13. Youbao Tang and Xiangqian Wu, “Scene Text Detection Using Superpixel-Based Stroke Feature Transform and Deep Learning Based Region Classification”, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 20, NO. 9, SEPTEMBER 2018
14. Text Image Dataset from GitHub [Online] Available: https://github.com/cs-chan/Total-Text-Dataset/tree/master/Dataset

## AUTHORS PROFILE

**B. Anil kumar** has completed Bachelor of Technology from JNTU, Hyderabad in 2005 and completed Master of Technology with Embedded systems specialization from JNTU, Hyderabad in 2008. He is working as an Assistant Professor in the department of ECE, GMRIT-Rajam, India since 2009 and pursuing PhD(Part-time) in Medical Image analysis from Andhra University, Visakhapatnam, India.

**Sreeram Murthy Velaga** has completed Bachelor of Engineering in Computer Science and Engineering from Madurai Kamaraj University, Tamil Nadu, India in 2000 and completed Ph. D. from Birla Institute of Technology, Ranchi, India, 2011. Currently he is working as a Professor in the department of CSE, GMRIT, Rajam, India.

**Aswani Devi A** has completed her Bachelor of Technology from IIIT, Basara in 2014 and completed her Master of Technology from GMRIT, Rajam in 2018. She has published 2 research papers in the areas of Image processing and Cryptography. She is currently working as JRF under DST-TIDE project at GMRIT.