ISSN : 2231-0711

M.Malathi et al | International Journal of Computer Science Engineering and Technology(IJCSET) | March 2015 | Vol 5, Issue 3,62-65

Word Spotting in Scanned Tamil Land Documents using K-Nearest Neighbor

M.Malathi[1], M.Srividya[2]
Department of Computer Science,
Government Arts College (Autonomous), Salem, India [1, 2].

***Abstract*: word spotting is a technique which can extract the text from input image. Here, we implemented on scanned Tamil land documents. Using Gabor feature, we extract the feature values for the input image. The main goal is recognize the text from the document using K nearest neighbor classifier. The features were calculated and the features were combined. Using these features, we can classify and recognized a text using an unsupervised classifier. The Classifier needs training of a database. The database consists of some set of words (printed or scanned). For the classification of the feature vector the k nearest neighbor Classifiers were employed.**

Keywords: word spotting, recognition, text, feature, k nearest neighbor.

I.INTRODUCTION

Word spotting is a subfield of speech recognition which deals with the identification of word in utterances. It has several types. Specifically used some recognized types are keyword spotting in unconstrained speech and keyword spotting in isolated word recognition. Mostly used methods for word spotting are sliding window and garbage model, k-best hypothesis and iterative Viterbi decoding. Keyword spotting in isolated word recognition which deals the keywords is separated from other text by silence and the main problem is dynamic time wrapping. Word Recognition is the ability of a reader to recognize a written word correctly and virtually effortlessly. Word recognition has been recent research work on neuron functioning and the words can be segmented on horizontal and vertical lines characteristics. The words with character match with the visual representation of the observed word received from excitatory signals on neural network.

II.LITERATURE SURVEY

Word spotting and Recognition with Embedded Attributes was proposed by J.Almazan, A.Gordo, A.Fornes,E.valley[2] were the authors suggested the recognize the word in lexicon .With the help of four public datasets on handwritten images and natural image by using common subspace of label Embedding and attribute learning . In the subspace, images and strings that represent the same word are close together, allowing one to cast recognition and retrieval tasks as a nearest neighbor problem.

Word spotting in historical documents was proposed by Can E.F. and Duygulu P. [3] were the authors suggested to retrieve words in historical documents. Using washingpsilas handwriting data set was tested using image retrieval technique. With the help of word spotting idea they combine histogram of gradient features and correlation coefficient method were implemented.

Segmentation free word spotting in historical documents was proposed by Gatos B. and Pratikakis I. [4] were the authors suggested to retrieve the words using block based document descriptor for template matching method .It satisfying salient regions of image. It can be experimental on unconstraint layout and degrading documents. By using this method it can proved on applying matching process on salient regions of image and improvement on time consistency.

A novel word spotting method based on recurrent neural networks was proposed by Frinken V., Fisher A., Manmatha R. and Bunke H. [5] were experimented on keyword spotting technique used on handwritten documents. With the help of keyword, it would retrieve all instance of word from the document .It is a template free word spotting method. Using CTC token passing algorithm were modification can be done on keyword spotting and it conjunction to the neural networks.

Handwritten word spotting with correct attributes was proposed by J.Almazan, A.Gordo, A.Fornes, and E.valley[6]were discussed on query based method to recognize a word .A dataset were collected from public which one is multiwriter word spotting. Image is low dimensional which it is easy to compare by query –by-image method and fast to compute them. Here, query consider as a string and correct the attributes scores by canonical correlation Analysis.

Scene Text Localization and Recognition with oriented stroke detection was proposed by Neumann.L and Matas.J[7] were discussed on character detection and recognized using sliding window and connected component methods to recognize an end-to-end text. Using image region, strokes are detected from image gradient field. Characters can be classified on Euclidean nearest neighbor classifier on trained dataset of image region.

Real Time Scene Text Localization and Recognition was proposed by Neumann and Matas.J [8] was discussed on detecting a character from sequential selection from the set of extremal regions. Extremal Region(ER) descriptor handles low contrast text. An Exhaustive Search method with feedback loops is applied to the group of ERs into words and selects the most probable character segmentation. Finally, a text is recognized with trained set of different synthetic fonts. The robustness of proposed system was implemented on low contrast character is demonstrated a false positives rate and it caused detected by watermark text on dataset

Unconstrained License plate and text localization and Recognition was proposed by Matas .J and Zimmermann .K [9] were suggested on recognize a license plate and traffic signs detection using neural network. It tested on database of several license plates with different view point, bad illumination and partial occlusion. Using Standard classifier (neural network) and locally separable threshold detector based on extremal region which attained the high detection accuracy 95% is achieved. The viewpoint is changed on invariant descriptor and modeling a shape variation using classifier.

An Efficient text extraction algorithm for complex image was proposed by Kumar .A [10] who suggested on extracting a text from extraction algorithm. It can be processed on three steps. First step, edge map generation using line edge detection mask. Second one is text area segmentation on horizontal projection and vertical projection with Heuristic Filtering method. At last, text can be recognized on image and video frames. It compared to existing method where achieved the precision rate is 96.20% and 99.61% for recall rate and the average computational time of 1.31 seconds per frames.

A simple approach for text segmentation in image based on curvelet transform was proposed by Doma, M.A.M, Faye .I and Jeoti.V [11] were suggested on segmenting an image on complex background and extracted text using curvelet transform method. Existing method were capture the horizontal and vertical direction. This proposed is recognizing for detecting an edges in multidirectional and taking advance of text segmentation. It was tested on different samples of images and would get an encouraging result with good accuracy.

III.METHODOLOGY

Our aim is to recognize a character wise classification in Tamil scanned land documents. First, words are extracted from the scanned document. And then characters are extracted based statistical texture features like area, centroid, eccentricity, equivalent diameter, extent and perimeter. These features are integrated to form a single feature vectors which are then used for handwritten and printed text separation via k-nearest neighbor. The following processes are required to do:

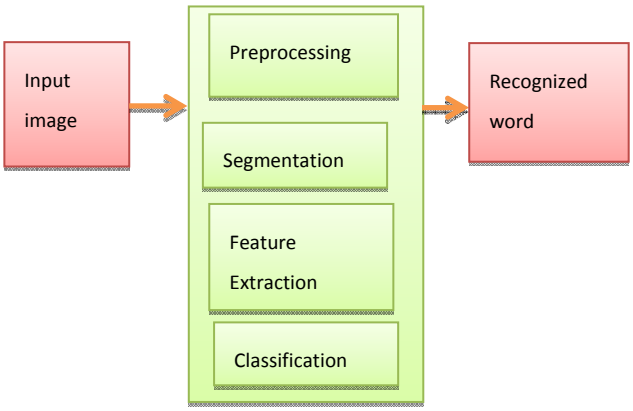


Figure .1 System Architecture of word spotting process

a. Dataset Preparation

We collect the Tamil land documents from friends and neighbor nearly up to 80 documents. The documents were real and accurate would not be faked one. We can manually cropped the each word for better accuracy .we can cropped the word from document it reaches up to 800 words .we kept the dataset separately for preprocessing stages.

b. Preprocessing

Preprocessing is a technique which would remove the noise from the input image using such technique like binarization, skew correction and slant correction etc., Image Enhancement is a method which improves the quality of images for different formats like .gif, .tiff., .jpg etc., Binarization is a method carries a logical value like zero (0) and ones (1).It takes a logical value ones (1) for foreground and zeros (0) for background. It stores on logical array. Skew correction and slant correction method are used for properly aligned the page layout for misaligned document.

c. Segmentation

Segmenting a page into paragraph, paragraph into lines and lines into word. In this work we will also assume that the location of the words in the images is provided, i.e., we have access to images of cropped words. If those were not available, text localization and segmentation techniques could be used .In the segmentation process we are cropping the words identically and show it in the bounding box.

d. Feature extraction

It is the heart of a pattern recognition application. Feature extraction techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Gradient based Feature Zoning and Euclidean distance. Histogram might be applied to extract the feature of individual characters. This stage is represented as a feature vector which it is identity one. The main goal is extract a set of features from recognized character. It improves the recognition rate and reduces the misclassification. Here, we used Gabor feature who calculated the statistical region properties for each word. Gabor filter is a linear filter and it widely used in pattern recognition for extracting the feature of an image. The Gabor space is used in image processing application on such as optical character recognition, iris recognition and fingerprint recognition. It has a specific spatial location are very distinctive between objects in an image. It can be extracted from the Gabor space in order to create a sparse object representation.

e. Classification

Classification is a technique which classifies a data from training dataset which compared to unknown Labels and finally created a new one. Classification has two process are model usage and model construction method. Model usage method deals with classifying the previously unseen objects. The test sample is compared with classified data for getting accuracy rate of classifier. Test set is independent of trained set which avoid pruning. Model

construction method deals with previously known objects. A K nearest neighborhood method follows some steps are.
Steps:

- Determine the parameter k= number of nearest neighborhood
- Calculate the distance between query-instance and all the training set
- Sort the distance and determine nearest neighbor based on k-th minimum distance
- Gather the category Y nearest neighbors
- Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

We recognized the character from matching the target dataset of character and the cropped word from Tamil land document where implemented on preprocessing and post processing stages.

IV. RESULT AND CONCLUSION

Word spotting is a technique which implemented a preprocessing technique like noise removal and post-processing technique like classification and recognition. These two stages were implemented on scanned Tamil land documents successfully. We get an accurate recognized character from the input image using MATLAB and hence, we processed few images.

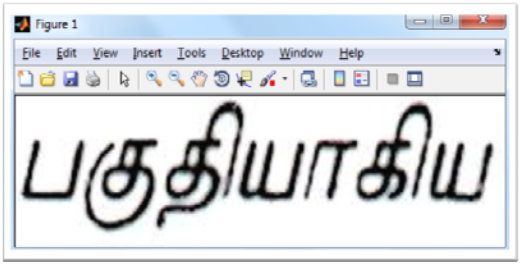


Figure .2 Input image

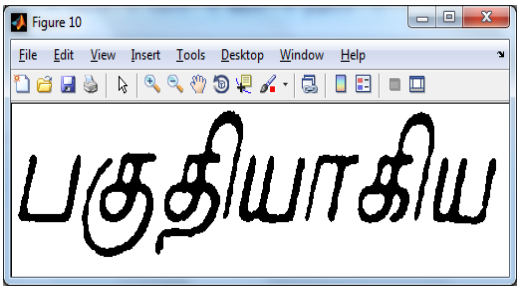


Figure .3 Binary image

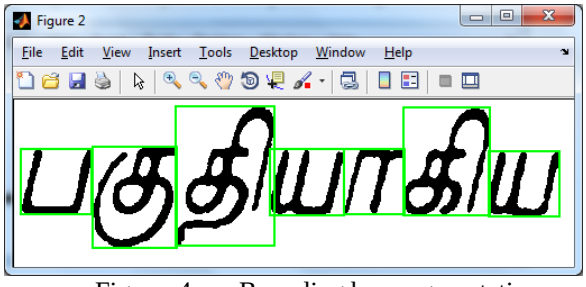
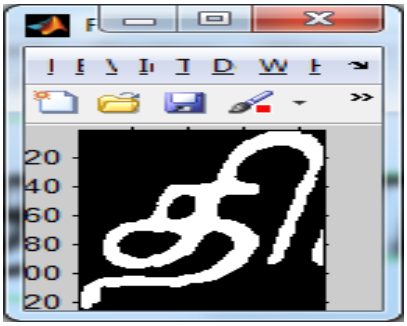


Figure .4 Bounding box segmentation



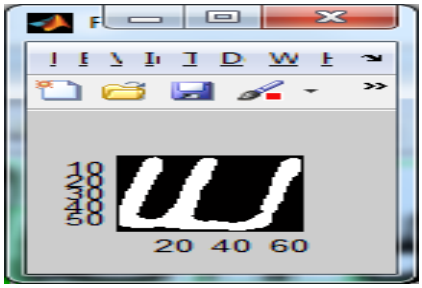


Figure .5 Segmenting Letters

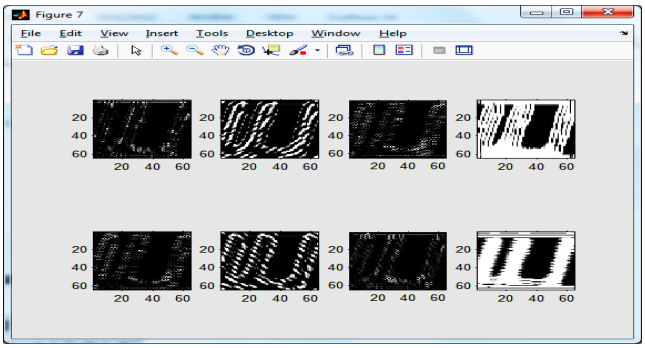


Figure .6 Feature Extraction using Gabor Filter

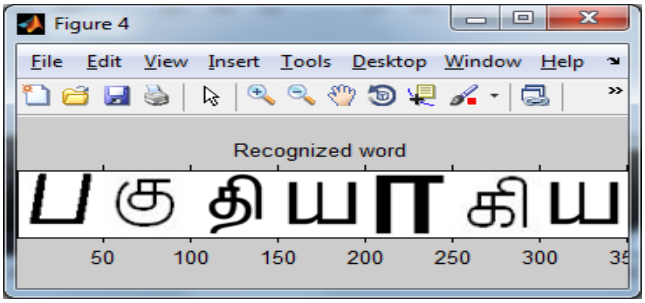


Figure .7 Recognized word using classifier

REFERENCES

[1] http://en.wikipedia.org/wiki/word spotting.
[2] J.Almazan, A.Gordo, A.Forne's and E.Valveny.,"Word spotting and Recognition with Embedded Attributes", Patten Analysis and Machine Intelligence,IEEE Transaction on Volume 36,Issue 12, pp. 2552-2566.
[3] Can E.F., Duygulu P, "Word spotting in historical documents", Signal Processing, Communication and Applications Conference, 2008. IEEE 16th Conference on 20-22 April 2008, Pp.1 – 4, E-ISBN: 978-1-4244-1999-9.
[4] Gatos B, Pratikaki I,"Segmentation-free Word Spotting in Historical Printed Documents", Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on 26-29 July 2009, pp.271 – 275, ISSN : 1520-5363
[5] Frinken V, Fischer A ,Manmatha R and Bunke H, "A Novel Word Spotting Method Based on Recurrent Neural Networks", Pattern

Analysis and Machine Intelligence, IEEE Transactions on 09 June 2011, volume :34,Issue:2 ,pp. 211 – 224 ISSN : 0162-8828.

[6] J.Almazan, A.Gordo, A.Forne's and E.Valveny ,"Handwritten Word Spotting with Corrected Attributes", Published on Computer Vision (ICCV), 2013 IEEE International Conference on 1-8 Dec. 2013,pp. 1017 – 1024, ISSN : 1550-5499.

[7] Neumann L and Matas J, "Scene Text Localization and Recognition with Oriented Stroke Detection" ,Published in Computer Vision (ICCV), 2013 IEEE International Conference on 1-8 Dec. 2013,pp. 97 – 104, ISSN : 1550-5499.

[8] Neumann L and Matas J ,"Real time scene text localization and recognition" , Published in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on 16-21 June 2012,pp. 3538 – 3545, ISSN 1063-6919.

[9] Matas, J. And Zimmermann, K., "Unconstrained license plate and text localization and recognition", Published in Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE Conference on13-15 Sept. 2005, pp.225 – 230, ISBN 0-7803-9215-9.

[10] "An efficient text extraction algorithm in complex images", Published in Contemporary Computing (IC3), 2013 Sixth International Conference on 8-10 Aug. 2013,pp. 6 – 12,ISBN: 978-1-4799-0190-6.

[11] Doma M.A.M, Faye I. and Jeoti V," A simple approach for text segmentation in images based on curvelet transform", Published in Intelligent and Advanced Systems (ICIAS), 2010 International Conference on 15-17 June 2010, pp. 1 – 5, ISBN: 978-1-4244-6623-8.

[12] www.mathwork. com