International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 15223-15227
© Research India Publications. http://www.ripublication.com

Kannada Script Recognitions from Scanned Book Cover Images

Preema P.Y
PG Scholar, Department of Computer Science
Christ University, Bangalore, Karnataka-560029, India.
Orcid Id: 0000-0002-9908-4982

Dr. Anita H B
Associate Professor, Department of Computer Science
Christ University, Bangalore, Karnataka-560029, India.
Orcid Id: 0000-0003-1608-0175

Abstract

Text extraction from the images plays a vital role in providing valuable information. Text extraction from images is still a challenging area specially extracting text from regional scripts of India like Kannada, Malayalam etc. Most of the times the images contain complex background then the cropping of text becomes even more challenging for extracting features. The input image is a scanned document images of Kannada book cover which is scanned with flatbed scanner of 400dpi resolution. The data sets are created by dividing the original images into number of varied size of blocks. Both spatial and frequency features are extracted for classifying images. This paper aims at recognizing the scanned images block which contains text or not by using multiple feature approach. The classification is analysed using Multilayer perceptron, Kstar and KNN. Experiments are performed on different sets of scanned documents of text cover images. Compare to all the classifiers KNN has given the encouraging results.

Keywords: Text extraction, Multilayer preceptor, Kstar,KNN, FFT, DCT

INTRODUCTION

The Text Processing is a fast thriving due to more surveillance cameras, usage of Robots and everything is automating for the human needs. In every fraction of second, it is being brought out something new, which surpasses our world of comprehension. People of every walk irrespective big and small, scholarly or non-scholarly, rich or poor everyone is making use of this technology. Information and applications with regard to computer enables man to accelerate his job more efficiently.

The text extraction from images and processing is useful in many cases like Banking, Visually impaired person, Industrialisation, Health care, Education, Researches, Document processing, Document analysis and retrieval, Library Automation etc. The extracted text will help to understand document in a proper way and the information that it contains. A document images contain graphics, pictures and texts in which text hold an important part to understand the image fully.

Text extraction from the scanned document images are really a challenging area because it may be of low contracts and resolution, complex background varies from size, colour, orientation and layout. Text extraction from the scanned images which has complex background document images is very difficult. Because text contain in images with complex background has many colour combination which makes the extraction very difficult and thought-provoking. Multicolour document images may contain different colour combination which increases the complexity in pre-processing, feature extraction and classification.

The proposed project can be updated in various ways to help the people in different scenarios. Once the text is extracted from the images it can be fed to the system which identifies the type of script and it can be converted to speech by text to speech converter. This converted speech can be used for many applications. In this way it can help the visually impaired persons. Similarly project can be modified to help travellers who don’t know the local script, people who don’t know to read, people who have eye sight problem etc. As per the survey many researchers worked on foreign type of scripts. Very few researchers worked on images which is having complex background and Indian type of script. Very less researchers worked for Kannada script which is using by Karnataka state of South India.

Vikas K et al have done the extraction of English script from video images, books, historical manuscripts, journals, scanned document, book covers, records, maps etc. [15] They have used variety of scanned images and acquired accuracy of 90.56%.

The proposed system work flow diagram Fig 1 depicts the step by step procedure. The spatial and frequency features are extracted from the images for classification.

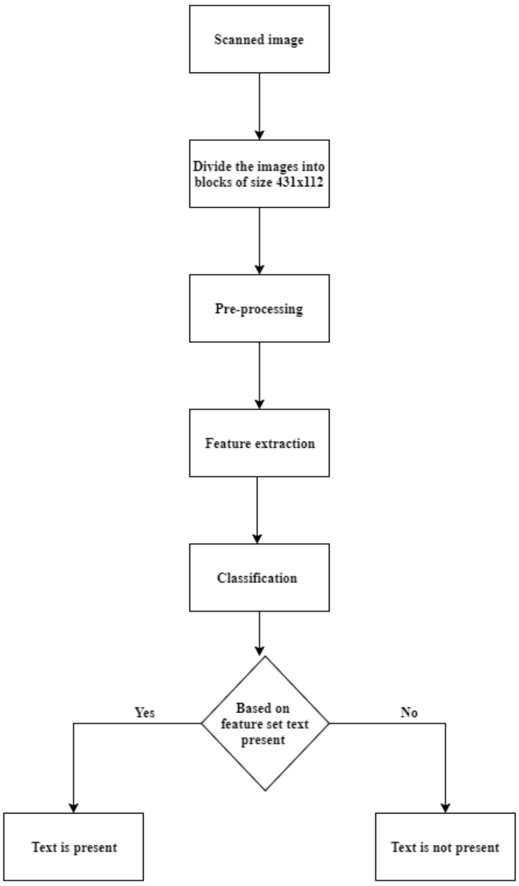

Figure 1: Flow of proposed work

LITERATURE REVIEW

Text extraction from images makes the world to get changed and improves the indexing of information. To create a database by editing the materials by human is hectic work. The text extraction helps in this kind of situation. Now a day's automatic indexing process reduces the work of humans. The morphological features are extracted [1] to crop text from Telugu regional language script images. Researcher has used multiple histogram projection. This algorithm is based on different types of segmentation especially on line as well as on word. [12] has designed the filter to process document images and full images for extracting text. Multi scale filtering is being used to cascade text region. To get aggregate centre points suppression on modified non-maximum of text region is done.

The edge detection method used to extract Kannada text is implemented [13]. The pre-processing include the conversion to gray scale and used medium filter for noise removal. Sobel edge detector is applied to detect edge as well as edges are being mapped to get the proper text.

Zaravi et al. [5] used wavelet transform method for extracting text from the book cover and journals. The images being decomposed into three ways means vertically, horizontally and diagonally. Edges were being detected using thresholding values of the three sub-bands. Morphological component analysis is done on the graphical document images. Morphological component analysis and Wavelet Transform are used for extracting features [4] to distinguish between text and graphics. [10] has used hybrid approach of connected components of their combination and texture analysis in complex background of document images. The problem faced due to the size, font, and orientation has been solved here. Koppula [14] et al proposed a method for text extraction from Telugu document images in which vertical as well as nearest neighbouring algorithm is used.

Automatic text identification and localization is being done on book cover and journals are described in [8]. The system uses mainly two ways to find hypothesis. One is based on top down analysis and other is bottom up analysis. In the case of top – down analysis work is done by splitting the image region. And in bottom up approaches uses region growing algorithm. Combining both methods result in understanding text elements and non-text elements. Binary text region is accepted as an input to OCR [16] module. The input images are being scanned with 200dpi resolution. The size of image can be nearly 1300x1800 which contain varieties of colours. Text extraction from Kannada images and videos is [9] focusing on three phases like text detection, localization and extraction. This algorithm includes the colour reduction technique, standard deviation for edge detection and connected components for text localization. The scanned images are accepted as input to the system and the output is computer file which contain information is able to edit. The Kannada text is being categorised into two groups such as vowels and consonants. The SVM classifier is being used. The texture based and a connected component analysis is explained in [11]. The background clearance from the noise is done by diffusion filtering. Some of the researchers used Gabor filter to identify the type of script [6] .The approach made by Angadi et al. [2] is different than to detect and extract text from low resolution of natural scene images by making the blocks of images. The features are extracted for each block of 50x50.

METHODOLOGY

Data collection

Kannada book covers are scanned and created database. Varieties of books are collected from different libraries for creating the database. The text of the cover pages contains different font, size and complex background. Flatbed scanning technique is used to capture image samples. Book cover images are scanned using an HP Scanjet 2610 flatbed scanner. This technique is used because it eliminates problems of non-uniform illumination and variation in the distance between camera and book.

Most of the book covers images size in the data base ranges from 21 cm height and 15cm width. All the book cover images scanned with the same size of 3500 x 2500. The book cover image contains pictures, more space, uneven spacing between lines, varied fonts and sizes of text. Therefore three levels of database are created by dividing the images into different sizes to extract features. The three different sizes are 431 X 335(data base 1), 431 X 168 and 431x112.This size is fixed empirically (Fig.2). The total number of books scanned is 35 and it is divided as mentioned above. The Database-1 contains total number of 1750 images of block size 431 x 335. Similarly database-2 contains 3500 images of size 431x168 and database-3 contains 5250 images of size 431x112.

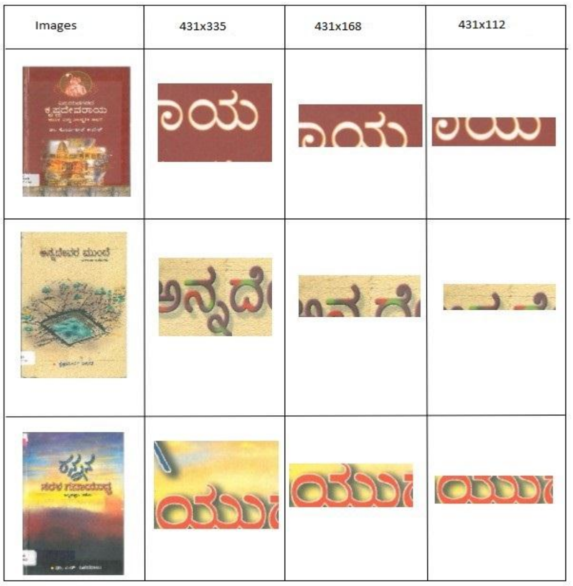

Figure 2: Different blocks of images

Pre-processing

Using the matlab software, images are converted to gray scale and black & white images to extract the features. Black and white image means binary images which contain mainly two pixel values that of zeros and ones.

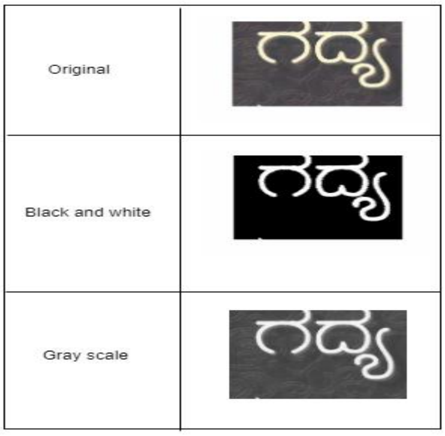

Figure 3: Different form of original image

Feature extraction

Features are measures or values which use for classification. The features are extracted using Fast Fourier transform and Discrete Cosine Transform.

Fast Fourier Transform

It used to find the frequency component of signal. It mainly compute the discrete Fourier transform. The frequency value represents the variation in the color of the image blocks.

The Fourier Transform is a significant mathematical method used in image processing to convert an image into its sine and cosine components. [7] The output image is represented in sine and cosine terms which is frequency domain. FFT images are produced using matlab function. Fig 4 represents the Fourier Transformed image means frequency content of the spatial domain image.

Discrete Cosine Transform

It is expressed as the sum of cosine function of data points of different frequencies that are being padded in images. DCT image consists of frequency values which are calculated from the images. The functionality of DCT is to transform the image from spatial to frequency domain. Discrete Fourier transformed of the text block is depicted in Fig 4.

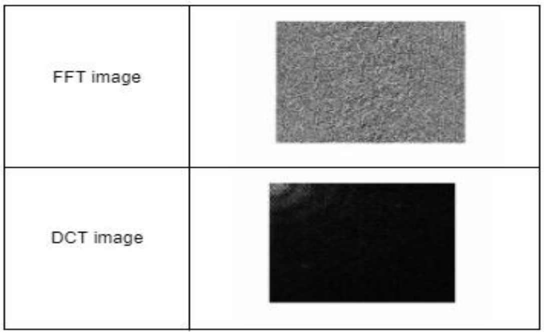

Figure 4 : Feature extraction of FFT and DCT

Features extracted for classification are mentioned below.

Algorithm

Input: Kannada Scanned Images of book cover

Output: 9 Features are extracted from the input image

Method:

a) Fast Fourier Transform is applied on the image and standard deviation and mean of the spectrum is considered as one feature.

b) Discrete Cosine Transform is applied and DC component of the frequency spectrum is considered as one feature.

c) Mean and standard of DCT image is considered. This form 2 features.

d) Row wise, column wise and diagonal wise standard deviation is considered as 3 features.

e) Standard deviation of gray scale and binary image is considered. This form 2 features.

CLASSIFICATION

Weka is a tool used for classification. The classification technique used in proposed system are Multilayer perceptron, Kstar and KNN. Different levels of folds for cross validation and percentage split used in classification. The different classification techniques which are tested are explained below.

Multilayer perceptron

Multilayer perceptron (MLP) is method in which data moves in one direction. It is merely neural network with one or more layers between input and output layer. MLPs are widely used for pattern classification, recognition, prediction and approximation. Here Single Neurons are called perceptron. This classification has mainly three phases. The three phases are input layer, hidden layer, output layer. The repeated and mathematical function is performed in hidden layer for better classification.

K star

The K star algorithm based on clustering concept. It perform based on ‘n’ partition with ‘k’ clusters. The observation is done based on cluster with the nearest mean. It uses distance metric which mainly compare each element with all elements of the metrics. It is an instance based learner and functioning with entropic distance which used to find out the similar instances from data sets. The other benefits of this method is to find out real valued attributes and missing values.

K Nearest Neighbour

K-nearest-neighbour classifier uses the distance metric concept with the number of nearest neighbours. It mainly used in for prediction related projects. As per the distance between trained data and testing data, this algorithm decides the sample belongs to which type of class. Proposed method has got encouraging results with this method.

EXPERIMENTAL RESULT

The performance of the algorithm is evaluated using different classifiers. The result obtained by using different classifier is mentioned in tabular form (Table 1).

Table 1: Experimental Results

Classifier	% split		Cross validation fold		
	70%	80%	4	7	10
Multilayer perceptron	91.77	92.66	91.39	91.92	90.86
K star	90.88	89.33	90.46	91.92	91.92
KNN	93.75	93.00	94.13	94.26	94.63

The encouraging result is produced using KNN classifier. To experiment the system, 70 percent of the data is considered for training and 30 percent of data is for testing. Similarly 80 percent of data as training and 20 percent for testing is also executed.

The result acquired by the KNN classifier with 93% in both cases. KNN classifier has performed well when compare to other two classifiers.

CONCLUSION

The proposed system objective is to find the position of text in scanned book cover images. This work can be used for applications like library, government offices etc. The proposed work is basically extracted the features from the scanned document images for classification. The scanned images are cropped into same size of blocks to retrieve the features from it. Features are extracted using DCT, FFT and spatial domain features. The total number of 9 features is being extracted from each block of images. The Multilayer perceptron, Kstar and KNN classifiers are used for experimenting. Based on the experiment we got 94 percent accuracy using KNN classification. It has given better accuracy compare to other classifiers. Frequency based features are observed to be a good feature set for classification. Once the text is extracted from image, it can feed to the system which converts text to speech and this helps vision impaired, illiterate people etc.

Further this project may be enhanced by recognizing the area of text in photographic images camera images, natural scenery images, signboards etc. for classification. In the proposed method, we have used regional script ie Kannada and with the slight modification this can be applied to other regional scripts.

REFERENCES

[1] Anupama N, C. Rupa, E.S. Reddy, "Character Segmentation for Telugu Image Document using Multiple Histogram Projections", Global Journal of Computer Science and Technology Graphics and Vision, 2013.

[2] Angadi S.A ,M.M Kodabagi, "A texture based methodology for text region extraction from low resolution natural scene images", International journal of image processing.

[3] E.A. Dayana C. Tejera Hernández, "An Experimental Study of K* Algorithm", I.J. Information Engineering and Electronic Business, 2015.

[4] Deepika Ghai, Neelu Jain, "Text Extraction from Document Images- A Review", international journal of computer application, 2013.

[5] D. Zaravi, H. Rostami, A. Malahzaheh, S.S Mortazavi, Journals Subheadlines "Text Extraction Using Wavelet Thresholding and New Projection Profile", World Academy of Science, Engineering and Technology, 2011.

[6] G G Rajput and Anita H.B., "Handwritten Script Recognition using DCT and Wavelet Features at Block Level", IJCA, Special Issue on RTIPPR (3), 2010.

[7] G. G. Rajput, Anita H. B., "Kannada, English, and Hindi Handwritten Script Recognition using multiple features", Proc. of National Seminar on Recent Trends in Image Processing and Pattern Recognition, 2010.

[8] Karin Sobottka, Horst Bunke and Heino Kronenberg , "Identification of Text on Colored Book and Journal Covers",international journal

[9] Keshava Prasanna, Ramakhanth Kumar P, Thungamani.M, Manohar Koli, "Kannada Text Extraction From Images And Videos Forvision Impaired Persons", International Journal of Advances in Engineering & Technology, 2011.

[10] P.Nagabhushan,S.Nirmala, ,"Text extraction in complex colour document images for enhanced readability", Intelligent Information Management, 2010.

[11] S. Malakar, S. Halder, R. Sarker, N. Das, S. Basu, M. Nasipuri, "Text line Extraction from Handwritten Document pages using spiral run length smearing algorithm", International Conference on communications, Devices and Intelligent Systems, Kolkata, 2012.

[12] Seong Jong Ha,Bora Jin,and Nam lk Cho, "Fast Text Line Extraction In Document Images",IEEE, 2012.

[13] S.V. Seeri, S. Giraddi, Prashant B.M, "A Novel Approach for Kannada Text Extraction, Proceedings of the International Conference on Pattern Recognition", Informatics and Medical Engineering, Tamil Naidu, 2012.

[14] V.K.Koppula, N.Atul,U.Garain,"Robust Text Line, Word And character Extraction From Telugu Document Image",2nd International Conference on Emerging Trends in Engineering and technology, 2009.

[15] Vikas K. Yeotikar, Manish T. Wanjari, Dr. Mahendra P. Dhore, "A Framework for Text Extraction from Document Images" ,International Journal of Computing Science and Information

[16] Z. Li, J. Luo, "Resolution Enhancement from Document Images for Text Extraction", 5th International Conference on Multimedia and Ubiquitous Engineering, Loutraki, 2011.