International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Proposed Approach for Layout & Handwritten Character Recognization in OCR

Manpreet Kaur[1], Navdeep Singh Randhawa[2], Vishal Garg[3]

[1,2,3]*Department of Electronics & Communication Engineering, Swami Vivekanand Institute of Engineering & Technology*

Abstract - *In this research work image analysis document segmentation is very important step. Document segmentation is the process in which first of all segment the document which contains the heterogeneous data means data like printed text, handwritten text, graph etc. In this we do the document segmentation because optical character recognition system is unable to recognize the whole document with multiple data type so before the recognition it is necessary to apply the document segmentation so to define the each region correctly. Document segmentation is the preprocessing step in document image analysis step. Document segmentation basically works on the document layout and segment the document into text and non-text component which contain the multiple type of component. Document segmentation gives the homogenous region to the optical character recognition system for the recognition. We would be using document segmentation on the handwritten bills which contain the heterogeneous content thereby segmenting the text and non- text region and the text into printed text and handwritten text and then we classify the text region into printed text and handwritten text. My main motive is to make this approach user specific so that the shopkeeper or the other person who is not so computer friendly store and analysis there bill. The work will start with design of the enhanced RSC (Radial Sector Coding) algorithm for detection of arbitrarily oriented text in an image. Information energy approach has been used to segment the text lines into rows that can be embedded into the word pad later which help to save the bill copy in e-format. It is very helpful to generate bill copy in e-format & it also saves so much manual work. It also saves so much time.*

Key Words: Segmentation, Recognition, Handwritten Bills, Information Energy, Radial Sector Coding

1. INTRODUCTION

Image processing is a method to convert an image into digital form. Some operations can be performed on image to get an enhanced image or to extract some important information from it. It is a type of signal dispensation in which input is image, like video frame or photograph and output may be image or characteristics associated with that image. Usually Image Processing system includes treating images as two dimensional signals while applying already set signal processing methods to them [1]. If one develop a method that extracts and recognizes those texts accurately in real time, then it can be applied to many important applications like document analysis, vehicle license plate extraction, text based image indexing etc. and many applications have become realities in recent years [2]. It is among rapidly growing technologies today, with its applications in various aspects of a business. Image processing technique gives better results than the original image.

In imaging science, image processing is a form of signal processing for which the input is scan by scanner in the form of image, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image, which are in the more enhanced form. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal processing techniques to it. Image processing refers to digital image processing, but optical and analog image processing also are possible. The acquisition of images is referred to as imaging in digital image processing [3]. Image processing is referred to processing of a 2D picture by a computer. An image is considered to be a function of two real variables, for example, a (x, y) with a as the amplitude (e.g. brightness) of the image at the real coordinate position (x, y). Localizing text in an image can be a computationally very expensive task as generally any of the 2N subsets can correspond to text (where N is the number of pixels). Modern digital technology has made it possible to manipulate multi-dimensional signals with systems that range from simple digital circuits to advanced parallel computers

Before processing an image, it is converted into a digital form. Digitization includes sampling of image and quantization of sampled values. After the digitization of an image, processing is performed. This processing technique may be Image enhancement, Image restoration, and Image compression.

Image Enhancement: It refers to accentuation, or sharpening, of image features such as boundaries, or contrast to make a graphic display more useful for display & analysis. This process does not increase the inherent information content in data. It includes gray level & contrast manipulation, noise reduction, edge crisping and sharpening, filtering, interpolation and magnification, pseudo coloring, and so on. Converting images into binary, the image has to remove the noise and trace the boundary of detected object. This process is done in image enhancement module.

Image Restoration: It is concerned with filtering the observed image to minimize the effect of degradations. Effectiveness of image restoration depends on the extent and accuracy of the knowledge of degradation process as well as on filter design. Image restoration differs from image enhancement in that the latter is concerned with more extraction or accentuation of image features.

Image Compression: It is concerned with minimizing the no. of bits required to represent an image. Application of compression are in broadcast TV, remote sensing via satellite, military communication via aircraft, radar, teleconferencing, facsimile transmission, for educational & business documents, medical images that arise in computer tomography, magnetic resonance imaging and digital radiology, motion, pictures, satellite images, weather maps, geological surveys and so on. Image processing basically includes the following three steps.

1. Importing the image with optical scanner or by digital photography.
2. Analyzing and manipulating the image which includes data compression and image enhancement and spotting patterns that are not to human eyes like satellite photographs.
3. Output is the last stage in which result can be altered image or report that is based on image analysis.

1.1 Fundamental Steps of Digital Image Processing

The description of fundamental steps is as follow and also explains through fig 1: Problem Domain gives the input to the image acquisition.

- **Image Acquisition:** With the help of sensor image is captured and digitized it with the help of analog to digital convertor only when image is in analog form.
- **Preprocessing:** After image enhancement and restoration preprocessing is done before segmentation. For extracting the components tools are used for the representation and proper shape of the image.
- **Segmentation:** Segmentation divides the image into its constituent and objects. When the object inaccessible in interested applicant then segmentation stops. Segmentation of the text line in an un-constrained handwritten documents still a challenging task because handwritten text lines are often un-uniformly skewed and curved, and the space between lines is not obvious.
- **Representation and Description:** The output of the segmentation of the image is followed by this step. In representation decision is made which data should be used either boundary or complete.
- **Recognition and Interpretation:** It assigns the labels to the objects based upon some information according to its description.
- **Knowledge Base:** In the form of knowledge database, knowledge about problem domain is coded into image processing.

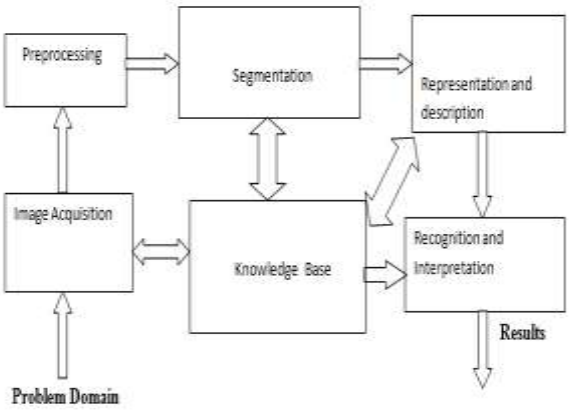

Figure 1: Fundamental Steps of Digital Image Processing

1.2 Purpose of Image Processing

The purpose of image processing is divided into 5 groups. They are:
1. Visualization - Observe the objects that are not visible.
2. Image sharpening and restoration - To create a better image.
3. Image retrieval - Seek for the image of interest.
4. Measurement of pattern – Measures various objects in an image.
5. Image Recognition – Distinguish the objects in an image.
Image Processing is a process to convert an image into digital form and perform some operations to get an enhanced image and extract useful information from it. It is a study of any algorithm that takes an image as input and returns an image as output. Image processing is referred to processing of a 2D picture by a computer. It is a form of signal privilege in which image is input similar to video frame or photograph and is image or characteristics associated with that image may be output. Image processing system treat images as two dimensional signals and set of signals processing methods are applied to them. It is latest technologies and its applications in various aspects of a business. The acquisition of images is referred to as imaging. Image processing is also known as digital image processing. Optical and analog image processing are also possible.

2. LITERATURE REVIEW

P. Barlas Et al., This paper present on document image analysis to extract the homogenous typed and handwritten text and successfully done the text/ non text segmentation and typed and handwritten text segmentation followed by the block segmentation to detect the white rectangle. This approach is applied on document of MOURDOUR COMPAIGN. In the previous method, we apply the segmentation one type of handwritten language with less variability in the document structure but when we apply this technique on MOURDOUR COMPAIGN dataset which contain the handwritten text of

Multilanguage like Arabic, English, French then this technique not give the good result.

In the present work we extract the homogenous block of handwritten and printed text with the help of connected component and block segmentation with white zone. In the first step we do the noise removal and remove the small and large connected component which are close to the border of the document and then the next step is to classification of text and non text with the help of predefined shape of connected component and then it further give to the multiple layer preceptor classifier to classify this and the next step is the layer separation in which we done the classification of printed text and handwritten text with the help of codebook in which we first of all construct the fragment and classify it by the multiple layer preceptor classifier .the next step is block segmentation in which we generate the block which contain the homogenous region with help of run length smearing algorithm and after this segment the document with white space [4].

Ankush Gautam Et al., This paper employ the system wavelet and 2 mean classification for extract the text from image and in post processing step using the morphological operation like erosion and dilation. There are various approaches regarding the text segmentation from the image like region based and texture based approach but this approach is failed when we apply the segmentation in complex document. We start our approach to convert the RGB image into grey scale image

Grey (i, j) = 0.59 green (i, j) + 0.30 red (i, j) + 0.11 blue (i, j)

After this we apply the wavelet transformation which image decomposed into multi resolution frame in which every portion has distinct frequency and spatial properties. The image will be large so we apply the block processing which decompose the picture according to user specification and also resembles each block result into output image. After this we apply the k-mean clustering approach where k is for the input parameter in this approach we use the 2 mean clustering so we take the input parameter as a black pixel and white pixel. In post processing step we used morphological operation one is dilation to add the pixel to the object boundaries and erosion for remove the pixel from object boundaries the number of pixel add or remove from the image is depend upon the structuring element to process the image . This proposed method is very efficient for detecting the all kind of text and graphs from the real life document. The limitation of this method is failed when there is very complex layout structure. In the future work we improve the segmentation by segment the text from graphic image [5].

L. Neumann Et al., This paper concerned about the formation of text line. It kept multiple segmentations of each character till context of each element is not known. At last stage, it calculated various parameters like Region text line positioning, Character recognition confidence, Threshold interval overlap. Then directed graph constructed with corresponding scores. Output of this graph was a word or a sequence of word. To eliminate typographical artifact, a pre-processing technique is used with a Gaussian pyramid [6].

3. PROBLEM FORMULATION

A document image analysis technique to extract the handwritten text from the heterogeneous document by the block segmentation method based on white rectangles detection. This method applied on the document which contains the three type of language Arabic, English, French is in the form of printed text and handwritten text. The first step in this preprocessing step is used for the noise removal of small connected component (cc). In the second step the text & non text classification is done on based of learning based approach on the basis of cc and its neighborhood to the feed an mlp classifier. Third step is layer separation which is used to classify the typed text and handwritten on the behalf of code book method. The fourth step is block segmentation. In this first step include applying RLSA algorithm to connect close cc and the second step is segment the document by white space to filter the small rectangle. This approach is applied on the maurdour company which is of five types and has three different accuracies according to the classes. The algorithm is to identify the text line by using information energy technique. In this input image taken as the format of gray scale which is further taken as a binary or white and black pixel method by the previous defined threshold method. The pixel gray value highest from threshold value taken as black pixel and other one taken as white pixel but this method works well on printed document but in handwritten document the information energy technique is used. In this algorithm energy map show the amount of information associated with each pixel if the pixel contain the high information value it means it contain the text if the pixel information value is low than it mean it show the space between text line. First step of this algorithm is to calculate the energy maps the pixel value and for accurately segmentation the direction of each text line also taken as consideration. This algorithm shows good accuracy in printed as well as hand written document. The document image will be scanned and image segmentation technique will be applied. The image segmentation technique will detect the layout of the letter. The text extraction technique will extract the text from the letter. As explained previously, text extraction techniques will not extract the text properly. It reduces the efficiency of the algorithm. In this work, I will enhance the text extraction algorithm and increases the efficiency of algorithm.

4. OBJECTIVE OF THE WORK AND METHODOLOGY OF PROPOSED APPROACH

Document segmentation is the preprocessing step in document image analysis. Document segmentation basically works on the document layout and segment the document into text and non-text component which contain the multiple type of component. Document segmentation gives the homogenous region to the optical character recognition system for the recognition. Now these days it is very important to store the analysis of the document because it can store the very important information so for to store and

analysis the document we have to process the document and for the processing of document we need document segmentation. We enhance the document segmentation approach to segment the handwritten bill template which contains the heterogeneous component like image, printed text, handwritten text and the graphical image. My main motive is to make this approach user specific so that the shopkeeper or the other person who is not so computer friendly store and analysis their bills.

Document segmentation is the preprocessing step of document image analysis. It play very important role because without this we can never recognize the document image. But there are lots of works to be done in document segmentation and there is some objective to work on document segmentation.

1. Segment the documents which contain the heterogeneous component.
2. To analyze the various document segmentation technique.
3. To propose the enhancement in document segmentation technique to improve the accuracy.
4. Provide the user specific application.

Research Methodology: The proposed algorithm in this research work detects the arbitrarily oriented text in an image. The work will start with the design of the enhanced RSC (Radial Sector Coding) algorithm for detection of arbitrarily oriented text in an image. It will include the definition of the constraints and structure of the proposed algorithm. I will use 26 uppercase English characters for performance evaluation of RSC. Different rotated and scaled versions of Arial and Tahoma fonts for each character were used. We used two sets of characters. One set for training of artificial neural network and other set is used for performance test. Further proposed algorithm is implemented in suitable environment of MATLAB.

Radial Sector coding: RSC uses simple and new method to extract invariant topological features. At first we have found Center of Mass (CoM) which locates the character within the image independent of translation. CoM is also a rotation invariant feature. Scaling invariance is easily obtained by normalization of features. The most challenging part was achieving rotation invariance. I achieved it by finding Axis of Reference which is a rotation invariant feature for all characters. Then we found Line of Reference from Axis of Reference which is considered as 0 lines for feature generation and thus we generated Translation, Rotation and Scale invariant features.

5. CONCLUSION

In this work, the RSC algorithm is applied on the shopkeeper bills to extract the handwritten character and layout of the bill. In future, others can propose a technique which saves handwritten character in different and printed characters in other file. Future scope also includes taking more complex images and applying enhanced techniques of character recognition on it. They will apply advance techniques on the input images to detect the characters i.e. handwritten and printed characters and also can detect different types of images like graphs, tables etc. More enhanced techniques will be applied on the images for better analysis of the heterogeneous images.

REFERENCES

[1] A. Iqbal, B. M. Musa, A. Tahsin, A. Sattar, M. Islam, and K. Murase , "A Novel Algorithm for Translation, Rotation and Scale Invariant Character Recognition," in Proceedings of SCIS & ISIS Nagoya, Japan, 2008, pp. 1367–1372.

[2] [2] A. J. Jadhav, "Text Extraction from Images : A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no.3, pp. 333–337, 2013.

[3] [3] A. Mishra, K. Alahari and C. V. Jawahar, "Top-down and Bottom-up Cues for Scene Text Recognition," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2687–2694, 2012.P.Barlas, S.Adam, C Chatelaine and T Paquet, "A Typed and Handwritten Text Block Segmentation System for Heterogeneous And Complex Document", publish in document analysis system, France in 2014.

[4] Ankush Gautam "Segmentation of Text from Image Document", publish in international journal of computer science and information, Vol. 4(3), 2013, pp. 538-540.

[5] L. Neumann, "On Combining Multiple Segmentations in Scene Text Recognition," in International Conference on Document Analysis and Recognition, 2013, pp. 1020-1024

[6] C.A Boiangiu, R.Laonitescu, M. CTanase, "Handwritten Document Text Line Segmentation Based on Information Energy", publish in International journal computer comm. ISSN 1841-9836.

[7] [10] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting Texts of Arbitrary Orientations in Natural Images," in IEEE Conference on Computer Vision and Pattern Recognition, 2012, vol. 8, pp. 1083–1090.

[8] [11] D Sasirrekha, Dr. Chandra "Enhanced Technique for PDF Image Segmentation and Text Extraction", International journal of computer science and information security, Vol.10, No.9, 2012.

[9] [12] Fattahc Zirari, Driss Mammes, Abdellatiifi Ennaji and Stephane Nicolas, "A Graph Based Approach for Heterogeneous Document Segmentation", Springer velag berlin Heidelberg, 424-431, 2012.

[10] [13] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Edge-Enhanced Maximally Stable Extremely Region," in 18th IEEE International Conference on Image processing, 2011, pp. 2609–2612