IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TEXT RECOGNITION USING OCR

[1] RESHMA SONAR , [2] KIRTI RANDHE, [3] SUMAN MISHRA

[1,2] Department of Artificial Intelligence and Machine Learning, ISBM College of Engineering, Pune

[3] Applied Science Department , ISBM College of Engineering ,Pune

Abstract : Text recognition is a technique that recognizes text from the paper document in the desired format (such as .doc or .txt). The text recognition process involves several steps, including pre-processing, segmentation, feature extraction, classification, and post-processing. The pre-processing is performed as a binarized image to convert gray scale image, and noise is reduced on the input image of the basic operation performed by removing the noise of the image signal. The segmentation phase is used to segment the image given online and segment each character of the segmentation line. Feature extraction is to compute the characteristics of the image document. This document describes techniques for converting the textual content of a paper document into a machine-readable format. This paper analyses and compares the technical challenges, methods, and performance of text detection and recognition studies in color images. It summarizes the basic issues and lists the factors that should be considered when addressing them. The prior art is classified as step-by-step or integrated and highlights sub-problems including text localization, verification, segmentation and identification of text. This survey provides a basic comparison and analysis of the scope and challenges in the field of text recognition

Keywords: Classification, Data mining, Segmentation, Text recognition

I. INTRODUCTION

Text recognition is important for a lot of applications like automatic sign reading, navigation, language translation, license plate reading, content-based image search etc. So it is necessary to understand scene text than ever. Texts in images carry high-level semantic information of the scene. Images in the webs and database are increasing. Developing effective ways to manage and restore the content of these resources is an urgent task. With the rapid growth of digital technology and devices manufactured by megapixel cameras and other devices such as Personal Digital Assistants (PDA), mobile phones, etc., are responsible for increasing the attention for information retrieval and it leads to a new research task.

Text recognition is a tedious job as it involves recognizing text of different fonts, styles and with different background noise. Also recognizing handwritten text is even more complicated due to differences in letter size, orientation and spacing between letters which varies from one person to another. Thus, there is a need to develop an automated text recognition system which can identify the text component present in an image or scene and convert it into a machine recognizable format. The process of text recognition starts with capturing the image of the required document, pre-processing it to acquire the desired portion and then segmenting it to extract the text content present in it. This paper discusses different stages in the task of text recognition from images.[1]

II. METHODOLOGY

The text recognition module has to perform a number of tasks. The input to the module is the image containing text. The output of the module is the text information in machine readable form. The text recognition module has to perform the following tasks: pre-processing, segmentation, feature extraction and classification. Fig. 1 shows the various tasks involved in text recognition.[2]

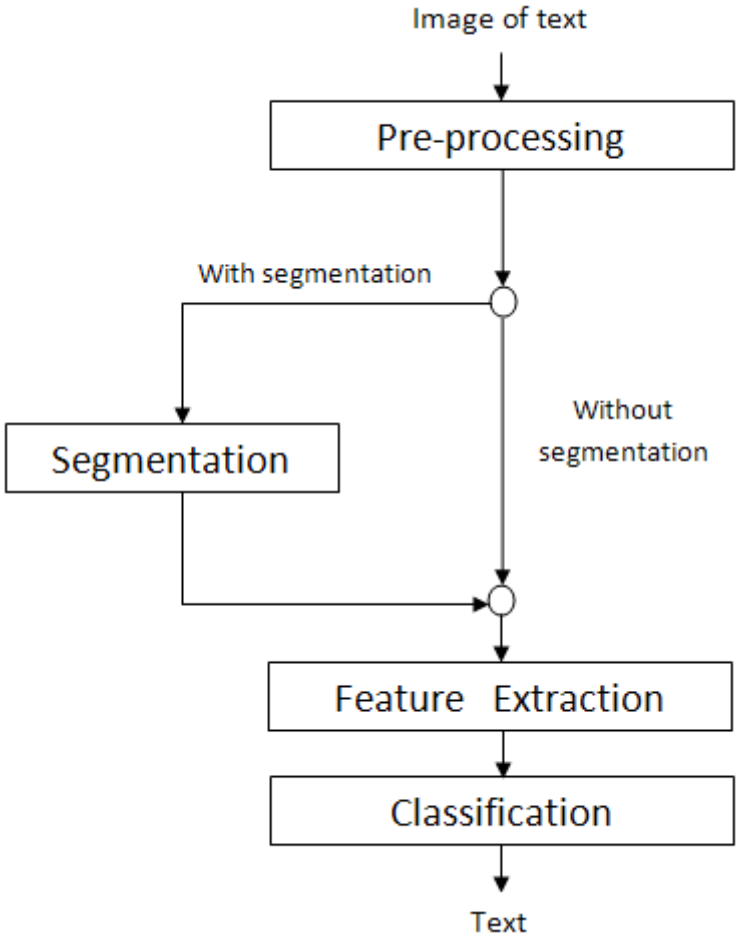

Fig 1: Various tasks involved in text recognition

The scanned document is usually in the form of an image. The first step is pre-processing which is to convert the image into a format suitable for further processing. The text image may contain noise or it may be skewed. In this step the image is enhanced by noise removal and then converted to binary. The noise present in the image has a major role to play in successful text recognition. Noise removal increases the probability of accurate text recognition and generates more accurate output. Various filters such as Gaussian filter, mean filter can be used for noise removal. Then normalization is done to ensure uniformity which is followed by binarization to convert the gray image into a binary image. After the pre-processing is done, the individual characters are separated using segmentation process. Then the vital data is retrieved from the raw data using the feature extraction step. Different techniques like Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), Histogram etc. can be used for segmentation purpose [6]. The next step is classification which involves recognizing each character and allocating it to the right character class thus converting the text into machine readable form. Different classifiers based on Artificial Neural Network (ANN); Support Vector Machine (SVM) could be used for this purpose. Post processing involves storing the recognized text in a format suitable for further processing[3]

TEXT RECOGNITION

Considering the type of text, machine-printed and handwritten OCR methods are two major parts of interest in the text recognition. The difficulties for fixed and multi-font OCR are solved with little constraints. The individual character separation of full text image is very important step in recognition. The documents generated with good quality paper yields test recognition accuracy as 99%. However, the commercially available products recognition rates much dependent on the quality of paper and ink, the age of the documents. The problem of the script could be split into two types as the cursive and normal script. In practice, it is difficult to draw a clear difference among them. A mixture of these two types could frequently be observed. On the writing style basis and the complexity of the segmentation stage, there are five phases of the problem in handwritten text recognition that is shown in Figure 7 [4]

III. MODELING AND ANALYSIS

The pre-processing is a fundamental stage that is proceeding to the stage of feature extraction; it regulates the appropriateness of the outcomes for the consecutive stages. The OCR success rate is contingent on the success percentage of each stage.

FACTOR AFFECTING THE TEXT RECOGNITION QUALITY

Many factors influence the precision of character recognized using OCR. The factors are scan resolution, scanned image quality, printed documents category either photocopied or laser printer, quality of the paper, and linguistic complexities. The uneven illumination and watermarks are few factors faced in OCR system that influence the accuracy of OCR

SIGNIFICANCE OF PREPROCESSING IN TEXT RECOGNITION

The pre-processing step is necessary to obtain better text recognition rate, using efficient algorithms of pre-processing creates the text recognition method robust using noise removal, image enhancing process, image threshold process, skewing correction, page and text segmentation, text normalization and morphological operations.

PREPROSSING METHOD

The majority of OCR application uses binary / Gray images. The images may have watermarks and/or non-uniform background that make recognition process difficult without performing the pre-processing stage. There are several steps needed to achieve this. The initial step is to adjust the contrast or to eliminate the noise from the image called as the image enhancement technique. The next step is to do thresholding for removing the watermarks and/or noise followed by the page segmentation for isolating the graphics from the text. The next step is text segmentation to individual character separation followed by morphological processing. The morphological processing is required to add pixels if the pre-processed image has eroded parts in the characters.

TECHNIQUES INVOLVED IN IMAGE ENHANCEMENT

Image enhancement increases image quality for perception of humans by increasing contrast, minimizing blurring and removing noise (Nithyananda et al. 2016).

SPATIAL IMAGE FILTERING

The filters are applied to defeat the high or low frequency present in the image. Eliminating the high frequencies in the image is smoothing, and the low frequency elimination is enhancing or edge detection in the image. The following figure 2(a) shows the original image and 2 (b) & (c) shows the images applied with Prewitt and Canny edge detection methods. These filtering techniques may give effective text detection from images available in natural scene.

TECHNIQUES OF POINT PROCESSING

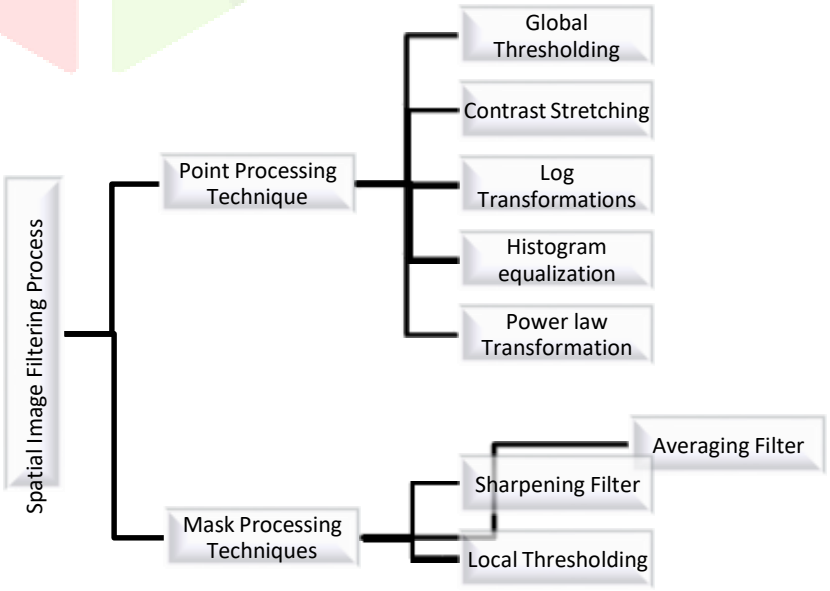

Fig 2: Techniques of point processing

a. GLOBAL THRESHOLDING

Image thresholding is the method of isolating the information from its background. Hence, this method is used normally to grey-level, or scanned colour images and itis categorized as global and local thresholding. Global method of thresholding chooses a value of threshold for the complete image from the intensity histogram (Mori 2010). Global thresholding automatically reduces a grey level image to a binary image. The local adaptive thresholding method for each pixel it uses different values based on the information of local area. The figure 4 shows the global threshold applied using Otsu's method.

b. CONTRAST STRETCHING METHOD

The image contrast level may change because of incorrect setting or the poor illumination in the acquisition process. The equation of linear mapping expressed as,

$$A(x, y) = A_1 + \left(\frac{A_2 - A_1}{I_2 - I_1}\right)\left[I(x, y) - I_1\right]$$

Where A1tends to 0 and A2denotes to desired level. I1 and I2are the min and max values of the input Gray range

a. LOG TRANSFORMATIONS

The common form of this method is expressed as

$$S = K \log (1+r)$$

b. HISTOGRAM EQUALIZATION

Histogram equalization is a global type method that gives the histogram of the whole pixels (0- 255) range and this technique can be applied in image understanding problems to normalize variations in illumination. For each level of brightness 'j' in the actual image, the level value (k) for the new pixel can be calculated as,

$$k = \sum_{i=0}^{j} N_i / T$$

Where T is the total pixels in an image (Russ 2007). The contrast enhancement of an image using histogram equalization

MASK PROCESSING TECHNIQUES

a. AVERAGING FILTERS:

Average filter also known as mean filter is an uncomplicated method for smoothing images. It minimizes the intensity variation amount between adjacent pixels also used to decrease or remove the noise. The averaging filter will be acting as a low-pass frequency filter that reduces the intensity of spatial derivatives in the image. Mean filter depends on a kernel that represents the size and shape of the neighbourhood pixels to be sampled during mean value computation. The 3×3 square kernel or mask is shown in Figure 6 (Mori 2010); other large masks like 5×5, 7x7, and 9x9 could be used for smoothing in severe case

b. LOCAL THRESHOLDING:

Local thresholding techniques can be applied to the document images that are having complex background like water marked images or images having non-uniform illumination in its background. The global thresholding method fails to separate the foreground from the background because the histogram of images may have more than two peaks.

SEGMENTATION

The two main segmentation types are external and internal segmentation. External type segmentation splits the page layout into the logical units. It is the decisive component of the text analysis, as it is an essential step before the off-line text recognition. It extracts the paragraph, sentence or words. The internal segmentation decomposes an image of a series of texts into sub-images. It gives the extraction of letters, especially, in cursive written words.

MORPHOLOGICAL PROCESSING

The morphological filtering technique comprises opening and closing, erosion and dilation, thinning and skeletonization. This practice suitable only on binary images (Phillips 2000). Dilation and erosion are the two morphological masking and threshold techniques that increase or decrease the size of the object. Erosion process

erodes or removes the pixels in edges and makes an object smaller; however, dilation process adds pixels around the object edges and creates an object larger.

FEATURE EXTRACTION

Feature extraction is the stage to eliminate redundancy from the data. The classification accuracy can be enhanced by selecting or searching most relevant features (Mohamad et al. 2015). Feature set should have a sufficient discriminating power to enable the accurate classification even among very similar symbols. Thus, features which are beneficial for classification ensure that objects from different classes have different values; objects from the identical class have similar feature values. A good feature set should be efficient about computation time.

IV. RESULT AND DISCUSSION

In this paper, we were successfully able to develop a robust and modular web application for image text extraction and multilingual translation using the Pytesseract based OCR Engine. The system was able to extract text from handwritten and printed documents with high accuracy which further strengthens the fact that OCR based applications can bring a lot of convenience to our daily activities and streamline a lot of workflows that can result in the efficient storage, retrieval, sharing and back up of the information. Although the results look promising, the OCR systems need to be made customizable so that they can be trained so as to efficiently recognize the characters from all sorts of handwritten data sources.

V. CONCLUSION

The computer vision and digital image processing are fast growing fields that are essential in many aspects of other areas like multimedia, artificial intelligence, robotics and much more. Image analysis involves the study of segmentation, feature extraction, and classification techniques. Humans interact quite naturally with each other over writing and speech, similarly human – computer interaction would make things exciting and easier to the user from the study it is found that optimal results can be obtained with less computation time as well as multilingual character segmentation and recognition also possible with better rate. It is to be noticed that the segmentation free approach using DNN is also possible in OCR. Our work may bridge the knowledge on automatic interaction between the human-system and system – system interaction.

VI. REFERENCES

[1] M.S. Akopyan, O.V. Belyaeva, T.P. Plechov and D.Y. Turdakov, “Text recognition on images from social media”, 2019, Ivannikov Memorial Workshop (IVMEM).

[2] Matteo Brisinello, Ratko Grbić, Dejan Stefanović and Robert PečkaiKovač, “Optical Character Recognition on images with colorful background”, 2018, IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)

[3] Wang,Y. DingX. and Liu,C. "Topic Language Model Adaption for Recognition of Homologous Offline Handwritten Chinese Text Image," in IEEE Signal Processing Letters, vol. 21, no. 5, pp. 550- 553, May 2014.

[4] SahareP. and Dhok,S. B. "Multilingual Character Segmentation and Recognition Schemes for Indian Document Images," in IEEE Access, vol. 6, pp. 10603-10617, 2018.

[5] https://www.irjet.net/archives/V4/i6/IRJET-V4I629.pdf