Text Recognition from Natural Images by Using OCR

[1]Vaishali, [2] Dr. Shilpi Singh

[1]PG Scholar, CSE Department, Lingaya's Vidyapeeth, Faridabad, Haryana, India

[2]Assistant Professor, CSE Department, Lingaya's Vidyapeeth, Faridabad, Haryana, India.

Abstract :The aim of this paper is to extract and detect text from natural images and handwritten documents. This paper reviews how extraction of text is done by segmenting the image to individual characters. Characteristics are used as boundary to segment the image. The information from these image documents would give higher efficiency and ease of access if it is converted to text form. The process by which Image Text converted into plain text is Text Extraction. Text Extraction is useful in information retrieving, searching, editing, documenting, archiving or reporting of image text [6]. However, variation of these texts' font size, orientation style, and alignment, text is embedded in complex colored document images, degraded documents image, low quality image and complex background makes text extraction extremely difficult

***IndexTerms* - OCR, segmentation, classification.**

I. Introduction

In today's world all the major information is available on paper or in the image/ video format and most of the important data is stored in images. The current technology is restricted to extracting text from different natural images. Extraction of text from natural images and handwritten document is used in various applications such as digital libraries, multimedia systems, Information retrieval systems, and Geographical Information systems. The aim is to find and separate regions based on the text it contains, if a separated region consists of only text then it is highlighted or those regions are processed with OCR module. In this paper, the system takes colored background images that contain text and handwritten documents as input [7]. Text information extraction is divided into detection, enhancement and extraction, tracking, OCR and localization.

II. Phasesof OCR

The process of OCR is a composite activity comprises different phases. These phases are as follows:

2.1 Image acquisition

Image acquisition is the 1st step of OCR; it acquires or captures images from an external source such as camera, scanner etc.Image acquisition first obtains a digital image and then converts it into a format which can be easily processes by computer. It is the 1st step in OCR. Image acquisition involves of quantization and compression of image. Binarization is a subset of quantization or it can be also be referred as an exceptional case. In binarization image has only two levels that is black or white [13].

Binary image is used for image characterization as it can easily differentiate foreground apart from background. Compression used in this technique can be either lossy or loss less but in the most cases binarization results in lossy compression.

2.2 Pre-Processing

Different pre-processing steps can now be performed on the acquired image to enhance image quality. Denoise, segmentation, thresholding, morphology, Resize and extracting image base line are some of the basic pre-processing techniques.After acquiring image from the source, pre-processing starts which aims to enhance the quality of image.

Thresholding is performed it used binarization technique on image and resultant image is obtained based on the threshold value. Thresholding is one of the pre-processing methods. To achieve this threshold value is set at a local or global level and multiple types of filters are applied for e.g. Averaging, min and max filters [1].

There are many different morphological operations such as erosion, dilation; opening and closing are there which can also be used to achieve this purpose. There is a possibility of loss in image file and it is a very important to find it out in a document, another important role of pre-processing is Projection profiles, Hough transform, nearest neighborhood methods are few of the skew estimation techniques.

Thinning action is performed on an image in some cases; it is a step after which later phases are applied. Text lines of files are also detected in pre-processing phase. It works by projecting and clustering the pixels in order to create a suitable character recognition system.

2.3 Character Segmentation

Character segmentation separates the image by passing them through a recognition engine. Recognition engine can segment the images into separate parts or slices of image each segment contains text that can be easily detected as it is easier to work on small images than larger ones [12]. Projection profiles and connected component analysis are the simplest segmentation techniques that are used in preprocessing. Advance character segmentation is used in complex situations.Before the image is forwarded to classification phase the characters are segmented from the given input image. The segmentation can be performed explicitly or implicitly as a by the product of classification phase. Some of the other phases of OCR can also provide contextual information on the basis of which segmentation of an image is done [2].

2.4 Feature Extraction

Feature extraction extracts different features from the segmented characters. These features help in recognizing the characters. Moments are one of the different types of features that can be extracted and used from the image. The extracted features are only useful when it can minimize intra-class variations and maximizes inter-class variations and is efficiently computable.Feature Extraction As the name suggests all the features of characters are extracted in this step. These features are responsible to uniquely identify characters. Only the right features and the total number of features are useful. The image itself is also a feature and its geometrical features (loops, strokes) and statistical feature (moments) can be used for the purpose. Principal component analysis is used to reduce the dimensionality of the image.

2.5 Character Classification

Character classification maps the features which are extracted from the image with different categories and classes. There is a variety of character classification techniques which are used in today's text extraction systems such as Structural classification techniques and Statistical pattern classification; Structural classification techniques, it uses feature extraction from image structure to obtain knowledge by character classification, set of rules are used for classification.Statistical pattern classification methods are based on probabilistic models and other statistical methods to classify the characters.

The process of classification classifies the characters accordingly to category. In Structural classification of image classifies character into similar category based on the component relationship.

Structural approach for image classification used the image components relationships and discriminate functions are used in statistical approaches for classification. Bayesian classifier, decision tree classifier, neural network classifier, nearest neighborhood classifiers is few of the statistical classification. To compose an image from its sub-constituents there are different set of classifiers which is based on syntactic approach and uses grammatical approach to do the job.

2.6 Post Processing

The result we obtain after the character classification is not accurate. Post-processing techniques are used to increase the accuracy rate of the OCR system. Natural language processing and other context-based techniques are used for enhancing accuracy by minimizing errors [5]. For example, postprocessor operates by utilizing markov chains model for dictionary and spell check to ensure the characters detected are actually a genuine word or not. In most scenarios characters forms a word that has meaning and by using n-grams accuracy can be further improved.

Postprocessor main concern is not just increasing accuracy and correcting errors, other concern is that any application should not create a new error. Complexity of postprocessor should be less, if time and space complexity is not very high it can result in faster and better results as a simple approach is more usable and it is easily applicable.

To improve the accuracy of OCR results there are many approaches which can be used using different post processing techniques. In some special cases the post processing step uses more than one classifier for classification of image to get more accurate results. The classifier can be used in cascading, parallel or hierarchical fashion [4]. In order to improve OCR results the various approaches are combined together of different results of all the classifiers; contextual analysis can also be performed to get better results. The geometrical and document context of the image can help in reducing the chances of errors.

Lexical process used Markov models that also can be used with pca and lda based on the requirement. Dictionary based method also improves the detection accuracy. OCR aims to resolve the classical problem of text recognition; text can be present in natural image or any digital image [11]. It can categorize the pattern present according to alphabets training set. OCR comprises of segmentation, feature extraction & classification. Segmented texts are used by OCR to recognize texts. Given a closer and deeper attention on the properties of the candidate character regions in the segmented frames or image it is observed that most OCR software packages will not find critical and difficult to recognize the text.

Document images has a very low-resolution as it is those captured using small cameras of mobile phones, which makes it even more difficult to extract the complete layout structure (logical or physical) of the documents or to apply any standard OCR system. To extract the same signature of original electronic documents from different formats such as PDF or PPT, the file is first converted into a relatively high-resolution image such as TIFF, JPEG and then the signature is computed to perform the action.

Finally, the captured document's signature is compared to with all the original electronic documents' signatures in order to find a match. Text detection and recognition is an approach that is combination of steps that includes detection, tracking, localization, binarization and text recognition.

2.7 Text Detection

This phase takes image or video frame as input and decides it contains text or not. It also identifies the text regions in image.

2.8 TextLocalization

The work of text localization is to formulate the text objects by merging the text regions and it further defines the tight bounds around the text objects.

2.9 TextTracking

Text tracking is only applicable on video format data. Text which is embedded in the video shows up in more than 30 consecutive frames which improve the readability of the text. Text tracking can be done in continuous frames by the appearance of similar text. It is used for improving the text detection as many images of same scene can be capture and it leads to better results as mean of image set used and localization. Text tracking does not apply binarization and recognition step to every detected object which makes the process of text extraction fast.

2.10 TextBinarization

This step is used for separating the text segment apart from natural scene image. In binarization output image contains only two levels of color that is black and white. It creates a complete separation of foreground from background scene or if can also be useful in region-based segmentation or edge detection.

2.11 Character Recognition

Character recognition is the last step in text extraction. Character recognition converts the binary object into the ASCII text for the text extraction process [20]. Text detection, localization and tracking are most challenging module of extraction process. These modules are inter-related which makes it more difficult in the process of text extraction process.

III. Methodology

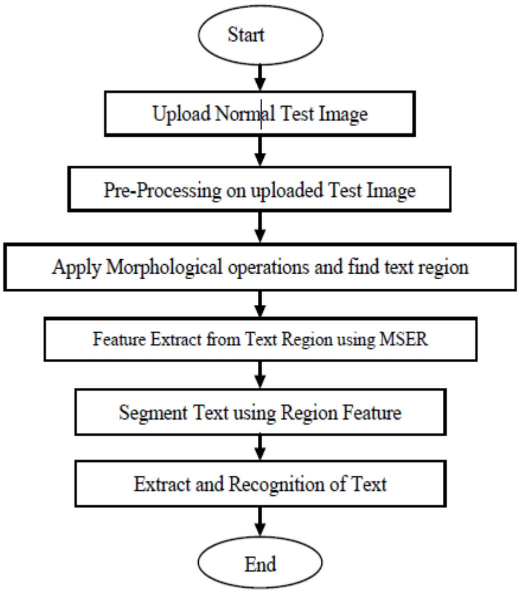
Start
Upload Normal Test Image
Pre-Processing on uploaded Test Image
Apply Morphological operations and find text region
Feature Extract from Text Region using MSER
Segment Text using Region Feature
Extract and Recognition of Text
End

Fig.1- DFD

- Design and develop a proper GUI of proposed text extraction and recognition from the normal images.
- Develop a code to upload test normal image for the text extraction and recognition. Apply pre-processing on uploaded image for testing.
- In pre-processing step, we apply some basic process like Binarization, filtering, resizing,conversion etc. to make the uploaded image useful in simulation.
- Develop a code for the region detection using the morphological operations and find only text region.
- Develop a code for the feature extraction from the extracted region of pre-processed image using MSER feature extraction techniques.

The MSER extraction implements the following steps

- Remove threshold of intensity from black to white from the binarized image.
- Find out the connected components which are Extremely Regions of image.
- Set a threshold when an extremely region of binarized image is Maximally Stable and suitable for the region extraction.
- Approximate an extracted region.
- Save those regions as feature sets of image.
- After that, develop the code for the image segmentation on the basis of MSER feature region and segment text from the image and match with the dataset and recognized the letters [16].

IV. Proposed Work

A text recognition system receives an input in the form of image which contains some text information. The output of this system is in electronic format i.e. text information in image are stored in computer readable form. In the proposed work, we train the OCR with different samples of handwritten texts of alphabets in the English language, so that the text extraction system can extract various kinds of text documents or the natural image containing text regions. The text recognition system can be divided in following modules:

(4.1) Pre-processing
(4.2) Text recognition
(4.3) Post-processing.

Each module is further described in detail as below:

4.1 Pre-Processing

Scanning of a document is usually done by using optical scanner and retrieved output is saved as an image format.An image is a combination of pixels, pixel is the single unit or element of picture [14]. At this stage we have the data in the form of image and this image can be further analyzed so that's the important information can be retrieved. Image quality can be enhance using operations like histogram processing, noise reduction or removal and normalization etc.

i. Noise Removal

Noise removal is one of the most important processes. Due to this quality of the image will increase and it will affect recognition process for better text recognition in images [17]. And it results in generation of more accurate output at the end of text recognition processing. There are many methods for image noise removal such as mean filter, min-max filter, Gaussian filter etc.

ii. Normalization

Normalization is one of the important pre-processing operations for text recognition. Normalization is necessary to get the result that is similar in nature i.e. size and uniformity [18].

iii. Binarization

Binarization is one of the important pre-processing operations for text recognition. Binarization technique is used to convert the gray scale images into binary images [15]. This separation of text from background that is required for some operations such as segmentation.

4.2 Text Recognition Module

Text recognition module is applied on output image of pre-processing model and gives computer understandable output data. These are the few techniques which are used in this module [8].

i. Segmentation

The segmentation is the most important process among the entire text recognition module. Segmentation separates the individual characters of an image.

ii. Feature Extraction

Feature extraction filters the raw data to retrieve the most important data. The characters of the data that can be represented accurately are considered as important data.

iii. Classification

Classification identifies each character and assigns correct character class which makes it easier to convert the text from the images in a computer understandable format [19]. This process used extracted feature of text image for classification i.e. input to this stage is output of the feature extraction process. Classifiers are able to find the similarity in input features based on pattern and can give output based on similar cases. There are many techniques used for classification such as Artificial Neural Network (ANN), Template Matching, Support Vector Matching (SVM) etc. [9]

4.3 Post-processingModule

Output of this modules results in text or character form that is readable by human or a computer and this result is saved in a document file for farther use such as editing or searching in that data [3]. To accomplish additional tasks such as optical character recognition (OCR) on image segmenting text from an unstructured scene can be very useful.The automated text detection algorithm can detect a large number of text region candidates and can also progressively remove those regions which are less likely to contain text [10].

V.CONCLUSION

This segment describes the result and simulation of proposed work for text extraction and recognition from the normal images and handwritten documents using the MSER feature extraction technique.

In proposed work, text region-based segmentation is used to segment the text region from the original image and on the basis of text region; feature extraction is applied to extract the feature from the image. In the proposed work, text extraction and recognition are based on combining efficient segmentation and connected component techniques within the image and text recognition is based on template matching techniques based on OCR system.

The proposed technique uses different templates of multiple handwritings to detect text from the given image; we have trained the OCR with different samples of handwritten texts in English language. The proposed work has been extensively tested on different types of normal images.

REFERENCES

[1] Liu, Ying, and Sargur N. Srihari. "Document image binarization based on texture features." IEEE Transactions on Pattern Analysis and Machine Intelligence 19.5 (1997): 540-544.

[2] Casey, Richard G., and Eric Lecolinet. "A survey of methods and strategies in character segmentation." IEEE transactions on pattern analysis and machine intelligence 18.7 (1996): 690-706.

[3] Vincent III, Winchel Todd. "System for creating and editing mark up language forms and documents." U.S. Patent No. 8,127,224. 28 Feb. 2012.

[4] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons, 2012.

[5] Bissacco, Alessandro, et al. "Photoocr: Reading text in uncontrolled conditions." Proceedings of the IEEE International Conference on Computer Vision. 2013.

[6] Fleizach, Christopher Brian, and Reginald Dean Hudson. "Intelligent text-to-speech conversion." U.S. Patent No. 8,996,376. 31 Mar. 2015.

[7] Starostenko, Oleg, et al. "Breaking text-based CAPTCHAs with variable word and character orientation." Pattern Recognition 48.4 (2015): 1101-1112.

[8] Sonka, Milan, Vaclav Hlavac, and Roger Boyle. Image processing, analysis, and machine vision. Cengage Learning, 2014.

[9] Nahin, AFM Nazmul Haque, et al. "Identifying emotion by keystroke dynamics and text pattern analysis." Behaviour& Information Technology 33.9 (2014): 987-996.

[10] Liang, Jian, David Doermann, and Huiping Li. "Camera-based analysis of text and documents: a survey." International Journal of Document Analysis and Recognition (IJDAR) 7.2-3 (2005): 84-104.

[11] Fink, Gernot A. "Handwriting Recognition." Markov Models for Pattern Recognition. Springer, London, 2014. 237-248.

[12] King, Martin T., et al. "Automatically providing content associated with captured information, such asinformation captured in real-time." U.S. Patent No. 8,990,235. 24 Mar. 2015.

[13] Jung, Keechul, Kwang In Kim, and Anil K. Jain. "Text information extraction in images and video: a survey." Pattern recognition 37.5 (2004): 977-997.

[14] Li, Jingquan, and Stephen Patrick Deloge. "Method and system operative to process color image data." U.S. Patent No. 8,849,019. 30 Sep. 2014.

[15] Pratikakis, Ioannis, et al. "ICDAR2017 competition on document image binarization (DIBCO 2017)." 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 1. IEEE, 2017.

[16] Islam, Md Rabiul, et al. "Text detection and recognition using enhanced MSER detection and a novel OCR technique." 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV). IEEE, 2016.

[17] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[18] Burger, Wilhelm, and Mark J. Burge. Digital image processing: an algorithmic introduction using Java. Springer, 2016.

[19] Kaushik, Deepti, and Vivek Singh Verma. "Review on Text Recognition in Natural Scene Images." Innovations in Computational Intelligence. Springer, Singapore, 2018. 29-43.

[20] Shahab, Asif, Faisal Shafait, and Andreas Dengel. "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images." 2011 international conference on document analysis and recognition. IEEE, 2011.

JETIR