Text-Line Extraction and Word Spotting In a Handwritten Document

M. Jamuna [1], S. Haribabu [2]

Assistant Professor, M.E. Communication Systems, Maharaja Prithvi Engineering College

Abstract:-**Text-line extraction in unconstrained handwritten documents remains a challenging problem due to non-uniform character scale, spatially varying text orientation, and the interference between text lines. In order to address these problems, which propose a new cost function that considers the interactions between text lines and the curvilinearity of each text line. To detect text in natural scene images based on two machine classifiers. One is used to generate the candidate word regions and the other filter is used to filter the non text ones. In this the connected components in the images are extracted by using maximally stable external region algorithm and the Connected Components are then partitioned using classifier and cluster them by using their pair-wise relations. Text provides important information about images sequences in a documented image, but it always remains difficult to modify the static documented image. To carry out modification in any of the text matter the text must be segmented out from the documented image, which can be used for further analysis. In this paper we convert image in text file which enable editing options to search, modify etc. Also text in the images was converted into audio sound which may be useful for disabled peoples.**

I. INTRODUCTION

A very important step in the handwriting recognition process is that of text line extraction, it aims at extracting individual text lines from the text regions of the manuscript page. A novel proposes text line extraction algorithm for color manuscript pages without prior binarization. The algorithm is based on seam carving to compute separating seams between text lines. The unconstrained seam carving has the tendency to produce seams that move through gaps between multiple texts lines, if these are the lowest energy regions of the neighboring image space.

Due to the vast amounts of information contained in such collections, convenient access can only be achieved by entering some sort of index, very much like in the back of a book. Since the current approach, manual transcription and index generation from the transcript is extremely expensive and time-consuming, automatic approaches would be favorable.

A. Text Line Extraction

Text Line extraction in document images is an essential step for various document image processing tasks such as layout analysis and OCR (Optical Character Recognition). Therefore, there have been a lot of researches in this area, and a number of algorithms have been proposed for the extraction of text-lines in machine-printed document images. However, text-line extraction in handwritten documents is still considered a challenging problem, the scale and orientation of characters are spatially varying, inter-line distances are irregular and characters may touch across words and/or text-lines.

The most conventional work focused on specific character sets. The conventional algorithms address the variations caused by individual writers by exploiting language-specific features. A new text-line extraction method for handwritten documents is presented, by developing an effective CC (Connected Components) segmentation method, by partitioning under-segmented CCs into normalized ones, the estimate states reliably in a variety of documents. It introduces problems in the energy minimization, because the connectivity information is sometimes lost.

B. Word Spotting

The typically significant degradation present in historic documents traditional handwriting recognizers based on OCR and word spotting. The word spotting idea has been previously proposed as an alternative solution to this problem for single-author document collections. The approach is to segment pages into words, match the words as images, and use the match scores to cluster word images. Each word image cluster contains instances of the same word throughout the analyzed collection. By tagging a number of the resulting clusters, a partial index can be constructed for the collection. The problem of deciding whether two given words are the same is easier than the recognition of a degraded handwritten word. The results of the investigation into features which can be used for successful word image matching.

C. Connected Components

A connected component of an undirected graph is a sub graph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the super graph. The graph that is itself connected has exactly one connected component, consisting of the whole graph. Connected component labeling works by scanning an image, pixel-by-pixel in order to identify connected pixel regions, the regions of adjacent pixels which share the same set of intensity values Connected component labeling works on binary or gray level images and different measures of connectivity are possible. The connected components labeling operator scans the image by moving along a row until it comes to a point. Once region boundaries have been detected, it is often useful to extract regions which are not separated by a boundary.

- Any set of pixels which is not separated by a boundary is call connected.
- Each maximal region of connected pixels is called a connected component.
- The set of connected components partition an image into segments.
- Image segmentation is a useful operation in many image processing applications.

Detection of text, that is not salient in luminance channel images. A text detection algorithm based on machine learning techniques is used. To be precise, they developed two classifiers, one classifier was designed to generate candidates and the other classifier was for the filtering of non text candidates.

II. PROPOSED SYSTEM

Importance of segmentation technique accumulating periodically at its practical application base is expanding rapidly. It is the primary stage for numerous processes such as machine recognition of language script. Segmentation is also used to extract various useful features of a document. Segmenting accurately a script document to extract various features that the document contains is a very challenging work and a need concerted effort. Continuous research works are in field to make the segmenting process simple and efficient. A simple segmenting technique for a line and word segmentation of a script document has been proposed. In this space recognition technique the main objective is to recognize the spaces that separate two text lines and the similar procedure is followed for the word segmentation procedure. Three different scanned documents have been taken as input images for line and word segmentation experiment and result found were promising.

A. Block Diagram

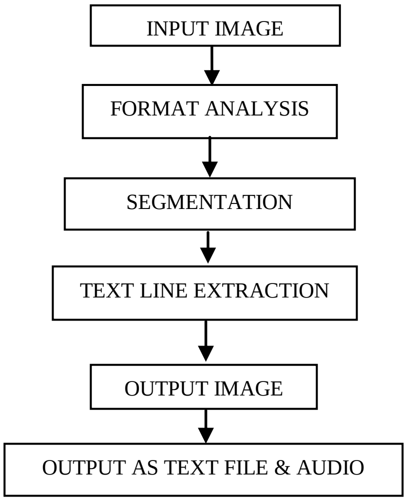

Fig. 1 Block Diagram of Classical OCR

B. Format Analysis

In the Fig.1, the Block Diagram of Classical OCR is shown. Multi-modality imaging usually requires handling of image formats with substantially different properties and capabilities: the amount of header information ranges from non-existent with Flat Image Format to overwhelming in the case of Image Format. To make matters worse, several image types come in different flavours, both Analyze images and data format can have Little Endian or Big Endian representation, and might contain integer or float values for one or more image frames. However, the real challenge is elsewhere and had seen it with almost all types of image data.

i) Data Format

The data format is native to several generations of scanners manufactured by Siemens Inc. It is a traditional binary format that uses struct-type data blocks for header information. The header and image data are contained in one file, the header data is split into one main header of size 512 Bytes located at the beginning of the file and a 512 Byte header prefixed to each image frame. Multi-frame files contain more than one image volume all images usually have the same dimensions and belong to one study, activation studies are saved in multi-frame files, each image frame has an entry in the directory block. The file format is not particularly suited as a general-purpose file format for data post-processing it uses a fixed-size header, different conventions exist for numbering of frames in different contexts, no information about coordinate systems is provided; space for free-text comments and patient data is very limited.

ii) Analyze Data Format

The Analyze data format offers even less in terms of header information but is important as this is a popular format for several packages for processing of data. It also has a fixed-size, binary header. It supports this data format for reading and writing, also for multi-frame files. This is easily the most complex data type available for medical imaging, and complexity is its main problem, even major vendors of medical imaging systems have histories of scanners which export DICOM data that violates the standard. Shadow Groups also pose a problem if manufacturers use this DICOM option to store important image parameters in an undocumented fashion.

iii) Flat Image Format

The term Flat Image usually refers to a binary file which contains raw vowel data for one image, without any embedded or external header information. Thus all meta-information must be supplied manually in the plugin, at the minimum, this would be the image file’s dimensions, data format and pixel sizes.

C. Segmentation

Segmentation is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols. It is one of the decision processes in a system for optical character recognition. Its decision, that a pattern isolated from the image is that of a character, can be right or wrong. It is wrong sufficiently often to make a major contribution to the error rate of the system. Segmentation is the initial step of the process. A character is a pattern that resembles one of the symbols the system is designed to recognize. But to determine such a resemblance the pattern must be segmented from the document image. Each stage depends on the other, and in complex cases it is paradoxical to seek a pattern that will match a member of the system's recognition alphabet of symbols without incorporating detailed knowledge of the structure of those symbols into the process.

D. Connected Components

The functionalities of Connected Components two ways as follows,

i) Under-Segmented CC

The states were estimated based on the distributions of super-pixels. The cost function that considers fitting errors of text-lines as well as distances between text-lines was developed based on the estimated states of CC. Then, the cost function was minimized by applying small variations to its coarse solution, and can extract text-lines. However, when the size of a CC is too large and there are only a small number of CC, the spatially varying states cannot be correctly estimated. In order to address these problems, a new CC extraction method is developed that partitions under-segmented CCs into normalized ones. This idea allows us to estimate the spatially-varying states even for documents having under-segmented CCs and can develop an algorithm that works for a range of languages and writing styles.

ii) CC Partitioning

This section presents a method to partition CCs into sub segments so that they have normalized sizes. Intuitively, the stroke length represents how far the pen moves to write a given CC. The random pixels were selected in and build a set of seed points where. The estimated stroke width is defined as the mean of the minimum distances: where N-E, W-E, NW-SE, NE-SW is a set of directions and is a width along the direction. Algorithm to partition CCs consists of two steps. First to select CCs that should be segmented and second to partition selected CCs into smaller ones.

CC partitioning is an important step in the proposed algorithm because to estimate the line spacing and orientations of CCs. To be precise, this method estimated the states of CCs from their distribution by assigning an equal weight to every CC.

E. Candidate Normalization

The process of clustering results the set of clusters. The clusters were then normalized to the corresponding region and undergone with reliable text/non-text classification. The clusters are first localized its corresponding region. Even though text boxes can experience perspective distortions, approximate the shape of text boxes with parallelograms whose left and right sides are parallel to *y*-axis.

F. Text-Line Extraction

The most important problems in optical character recognition and a number of algorithms have been proposed over the last decades. However, many methods have focused on the scanned images of machine-printed text, and they cannot handle complex cases such as camera-captured images and handwritten documents. The docstrum method was based on the assumptions that there is a global skew angle and the scale of text is almost constant in a given document, which is not true for many challenging cases. To be specific, text-line extraction in handwritten documents suffers from unclear spacing between text lines, an interline space may be smaller than an inter character distance or, even worse, characters in different text lines are connected Since it is not a simple task to address these problems based solely on local information, A new method that considers the interaction between text lines, as well as the local property of each text line is developed.

Conventional text-line extraction methods can be roughly classified into four categories,

- Projection-based methods
- Hough transform-based methods
- Bottom-up grouping methods
- Image segmentation-based methods

Projection-based methods allow the efficient extraction of text lines however, this approach does not work well for curved text lines. In, images are partitioned into several vertical strips, and lines are extracted in each strip by assuming that curved text lines are locally approximated with straight lines. Based on similar assumptions, the method in detects text lines by using Hough transform.

These methods are script independent and relatively robust to noise however, their computational complexity is high compared with other approaches. The regular structure of text lines is one of the most important clues in text-line extraction however, it is not straightforward to exploit this property with other information in conventional approaches. In order to address this problem, consider text-line extraction as a grouping problem of CCs and develop a cost function that encodes local and global observations

- A curvilinear cluster is desirable
- Text lines should not be too close.

In order to consider both terms simultaneously, the method estimates the local line spacing and normalizes those terms with the estimated spacing. Finally, can extract text lines by minimizing the cost function. The illustration the pixel-level labelling results with four colours. The method is robust to the interference between text lines, spatially varying skew, and irregular inter-character distance.

Text line patterns are found by building a fuzzy run-length matrix, at each pixel, the fuzzy run-length is a maximal extent of the background along the horizontal direction. ELS (Elementary Line Segments) are obtained by linking edge pixel and approximating them to piecewise straight line segment. These ELS are used as input to this approach. Adjacent line segments are grouped according to some grouping criteria and replaced by new line segment.

The threshold value CWR determines the spaces present in between text lines of the concerned image document. Special consideration has to be paid towards upper zone and lower zone of the text line as some spaces separates these zones from the middle zone, so that they are not abandoned by the processor as white spaces.

F. Text file & Audio sound

Here in the proposed system image was converted into a text file where user can search content easily and edit as user want. This prevents user for typing the content manually by seeing the image. Conversion into text advantageous only for able person, so content/text found in image are converted into audio sound which increase concentration of user and useful for disabled person.

III. EXPERIMENTAL RESULTS

A. Input Image

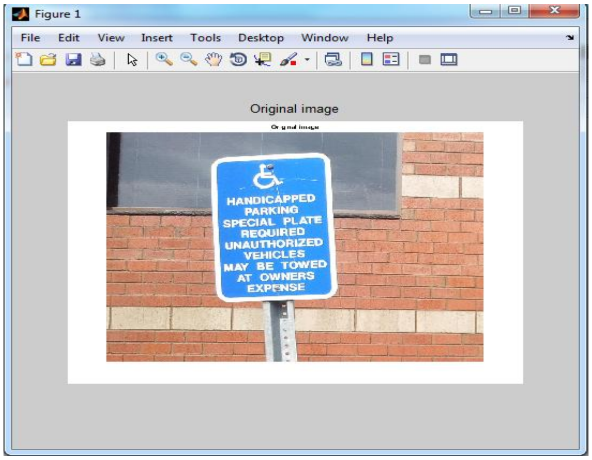
Figure 1
File Edit View Insert Tools Desktop Window Help
Original image
Original image
HANDICAPPED PARKING SPECIAL PLATE REQUIRED UNAUTHORIZED VEHICLES MAY BE TOWED AT OWNERS EXPENSE

Fig. 2 Input Image

The input image consists of text inside the image is given as an input to the MATLAB coding.

B. Detecting MSER Region

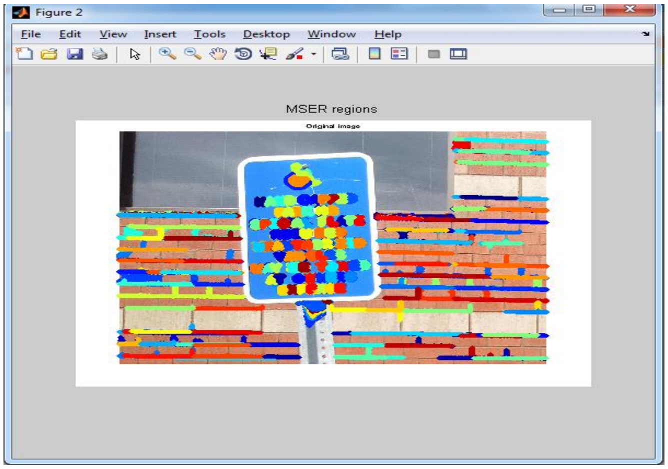

Fig. 3 MSER Extraction

Since text characters usually have consistent colour, thus begin by finding regions of similar intensities in the image using the MSER region detector. Fig. 3 shows the MSER Region.

C. Canny Edge Detector For Segmentation

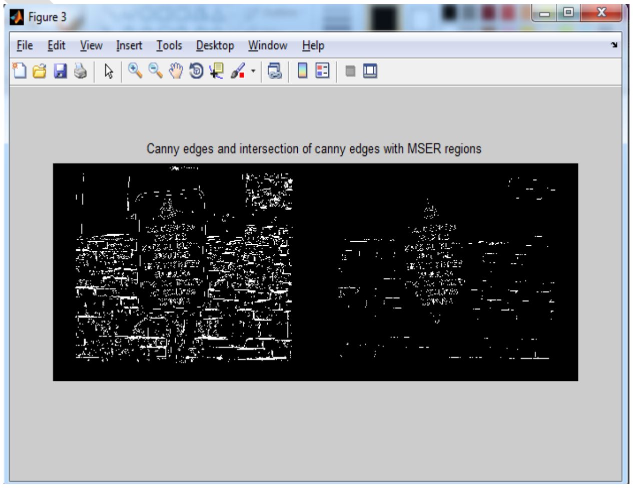
Figure 3
File Edit View Insert Tools Desktop Window Help
Canny edges and intersection of canny edges with MSER regions

Fig.4 Canny Edge Detector

The written text is typically placed on clear background; it tends to produce high response to edge detection. Furthermore, an intersection of MSER regions with the edges is going to produce regions that are even more likely to belong to text, for the need of segmentation Canny edge Detection is done over her it is shown in Fig. 4. In gradient direction the growing of edge is shown in Fig. 5.

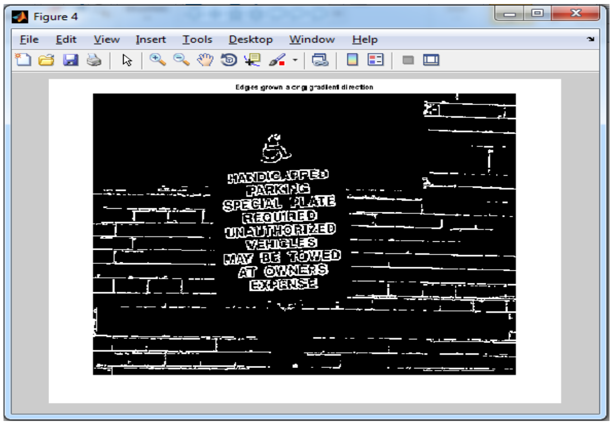

Fig. 5 Edge Grown in Gradient Direction

D. Filter Character Candidates Using Connected Component Analysis

Some of the remaining connected components can now be removed by using their region properties. The thresholds used below may vary for different fonts, image sizes, or languages. The process of Filtering Character candidates using Connected Component analysis is shown in Fig. 6.

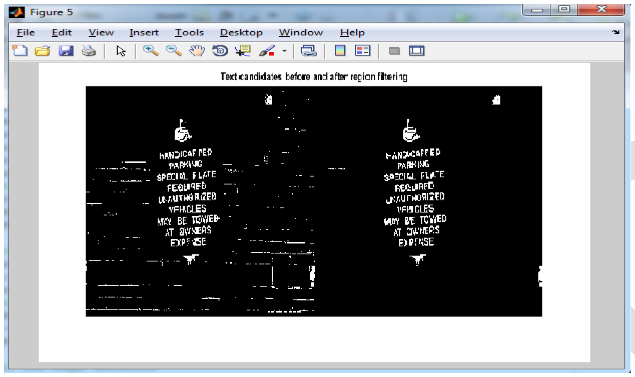
Figure 5
File Edit View Insert Tools Desktop Window Help
Text candidates before and after region filtering
HANDICAPPED
PARKING
SPECIAL PLATE
REQUIRED
UNAUTHORIZED
VEHICLES
MAY BE TOWED
AT OWNERS
EXPENSE

Fig. 6 Connected Component Text filtering

E. Determine Bounding Boxes Enclosing Text Regions

The process of computing bounding box of the text region, merging the individual characters into a single connected component. This can be accomplished using morphological closing followed by opening to clean up any outliers. The bounding box text region is shown in Fig. 7.

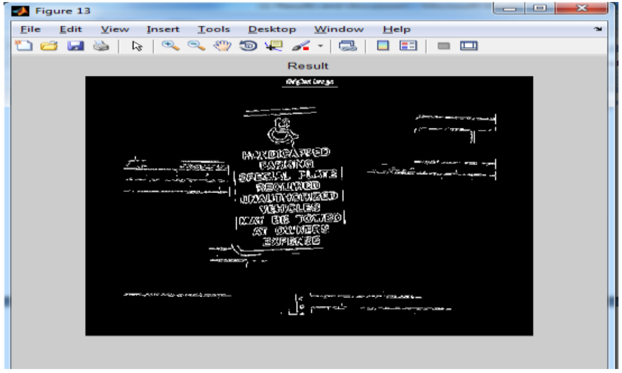

Fig. 7 Text Region

IV. CONCLUSION

The text-line extraction algorithm for the processing of handwritten document images has been proposed. The energy minimization technique is used for the handwriting databases. A machine learning technique algorithm is used for text detection. To be precise, the developed two classifiers: one classifier was designed to generate candidates and the other classifier was for the filtering of non text candidates. The method to exploit multi-channel information has been used, that yielded the state of the art performance in both new and traditional evaluation protocols. In the proposed system the text has to be extracted from the handwritten images. Text-line extraction can be considered a CC segmentation problem and the method is based on this observation to extract CCs and group them into text-lines. The CC extraction method and text-line segmentation algorithm based on an energy minimization framework. Image is converted into text file and audio file.

REFERENCES

[1] Jewoong Ryu, Hyung Il Koo, Nam Ik Cho, (2014) “Language-Independent Text-Line Extraction Algorithm for Handwritten Documents” IEEE Signal Processing Letters, Vol. 21, No. 9.

[2] Boris Epshtein, Eyal Ofek and Yonatan Wexler, (2010) “Detecting Text in Natural Scenes with Stroke Width Transform” IEEE Transactions on Image Processing, vol. 12, no. 7, pp 2963-2709.

[3] Hyung Il Koo, and Duck Hoon Kim, (2013) “Scene Text Detection via Connected Component Clustering and Non-text Filtering”, IEEE Transactions on Image Processing, vol. 22, no. 6, pp 2269 – 2305.

[4] Hyung Il Koo and Nam Ik Cho, (2012) “Text-Line Extraction in Handwritten Chinese Documents Based on an Energy Minimization Framework”, vol.21,no. 3, pp 1169-75.

[5] Jaime S. Cardoso , Artur Capela, Ana Rebelo and Carlos Guedes, (2008) “A Connected Path Approach for Staff Detection on a Music Score”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 3, pp.1005 – 1008.

[6] Lawrence O’Gorman, (1993) “The Document Spectrum for Page Layout Analysis”, IEEE Transactions on Pattern analysis and Machine Intelligence, vol. 15, no. 11, pp 1162 – 1173.

[7] Louloudis.G, Gatos. B, Pratikakis.I, and Halatsis.C, (2008) “Text line detection in handwritten documents,” Pattern. Recognition, vol. 41, no. 12, pp. 3758–3772.

[8] Richard G. Casey and Eric Lecolinet, (1996) “A Survey of Methods and Strategies in Character Segmentation”, IEEE Transactions on Pattern Analysis, vol. 18, no. 7, pp. 690-706.

[9] SatoA, Kise.K, and M. Iwata, (1998), “Segmentation of page images using the area Voronoi diagram” Comput. Vis. Image Understand., vol. 70, no. 3, pp. 370–382.

[10] Yin.F and C.-L. Liu, (2009) “Handwritten Chinese text line segmentation by clustering with distance metric learning,” Pattern Recognition, vol. 42, no.12, pp. 3146–3157.