ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9379-9381

A Review of Extraction of Image, Text-line and Keywords from Document Image

Sneha L. Bagadkar[1], Dr. L. G. Malik [2]
[1] PG Student, G.H.R.C.E Nagpur-440016, India
snehabagadkar30@mail.com

[2]Professor, G.H.R.C.E Nagpur-440016, India
latesh.malik@raisoni.net

Abstract: ***In Document processing typed and handwritten text on paper-based & electronic documents converted into electronic information. To electronically process the contents of printed documents, information must be extracted from digitally scanned images. The manipulation of printed documents is largely in use like printed forms are delivered to end users for completion, storage and verification etc. In such situations these printed documents must return to digital form in order to participate in digitalized workflows. In printed documents, the contents of different regions and fields are highly heterogeneous. They have different layout, different printing quality and typing standards. The text line, keywords and image extraction from such complex printed document can be a difficult problem.***

Keywords: about Document image processing, Document image processing, image isolation, Text-line extraction, Keyword recognition.

1. Introduction

Traditionally paper documents are main form of transmission and storage for information. Data present in documents is very raw and not optimized for search, indexing etc. Document Processing is done to extract data from these documents using multiple stages. In Document Processing typed and handwritten text on paper and electronic documents is converted into electronic information. The working with printed documents is largely in use like business letters, forms, and maps, text books, technical manual etc. In order to participate in digitalized workflows wide variety of information which is stored on paper is converted into electronic form for better storage and intelligent processing. It deals with the textual and non-textual elements components of a document image.

Document processing use for extracting symbolic information like text, graphics, logical structures etc .There are various challenges associated with printed documents like distinguish text from graphics and sketches, Character and symbol recognition, line segmentation, identification of letters and words from text lines from document image etc. So to deal with such challenges document processing is necessary. When we are working with the documents having highly heterogeneous contents like multiple languages, images, printing quality, low resolution etc. the extraction of the contents of documents from a printed document image can be a difficult problem.

2. Review of Techniques

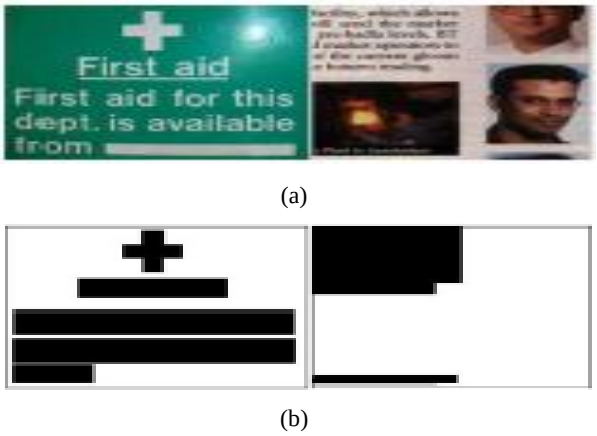
2.1 Title and authors

The Distribution characteristics of wavelet coefficients for segmenting document images [1] proposed by Jia Li et.al. Used for segmentation of document images into four classes - background, photograph, graph and text. Multiscale nature and fast classification are two important attributes of this algorithm. The multiscale structure provides context information from low resolutions to high resolutions. At start, classification is performed on large blocks to avoid over-localization. Blocks with extreme features are classified at higher resolutions to ensure that blocks of mixed classes are not yet classified. The unclassified blocks are divided into smaller blocks at the higher resolution. The smaller blocks are classified based on the context information received at the lower resolution.

Multiscale characteristics of wavelet coefficients for segmenting text and graphics part of document images [2] proposed by Mausumi Acharyya and Malay K. Kundu based on textural cues considers that the graphics part and the non-graphics (text) part have different textural properties. M-band wavelets use to decompose an image into M×M bandpass channels which represent the image at different scales and orientations in the frequency plane. Then feature maps are created to give a measure of the local energy around each pixel at different scales. A scale–space signature is derived from the feature maps which is then use for segmentation.

The globally matched wavelet filters [3] are used for segmenting text and graphics part of document images proposed by Sunil Kumar based on clustering technique used for estimating globally matched wavelet filters using a database of images. Multiple two-class Fisher classifiers are used for segmenting text and non-text part in document image

Refinement of the segmentation results is done by using Markov random field model. This document image segmentation algorithm separates the document image independent of its content into three classes – text part, image part and background as shown in Figure 1.



(a)

(b)

Figure 1: Output of globally matched wavelet filters (a) Original Image (b) classified image [3]

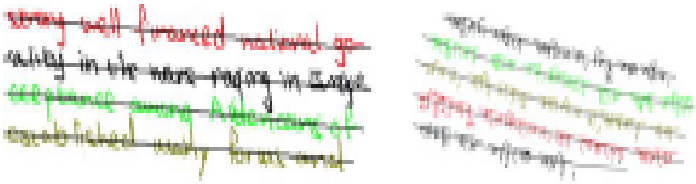
2.2 Text-Line Extraction

Text-line extraction technique [4] proposed by Jisheng Liang et. al. states the document structures extraction as problem of optimal partitioning. It is used to find an optimal solution by partitioning the set of glyphs of a given document image into a hierarchical tree structure. There entities within the hierarchy are associated with their physical properties and semantic labels. The Bayesian framework is used to assign and update the probabilities during structures extraction process. A text line extraction algorithm implemented to demonstrate the usage of this framework.

A local linearity based technique [5] proposed by U. Pal and Partha Pratim Roy used for identifying Multi-oriented or curved text lines from English and Chinese documents. The line extraction technique used here in this is performing component labelling to get individual components and then analyzing the reservoirs obtained in a component. The concept of water reservoir is used for this purpose. This method is efficient as it is independent of font, size, and style of the text lines and it provides flexibility.

A Text-line extraction technique [6] proposed by Hyung Il Koo and Nam Ik Cho consider that textline extraction as a grouping problem of CCs. The cost function was developed which considers the fitting error and the distances between text-lines. A cost function is based on local and global observations like the fitting error of each text line should be small and text lines should not be too close. For that, this method estimates the local line spacing and normalizes those terms with the estimated spacing. Text-lines are then extracted by minimizing the cost function.

A language-independent Text-line extraction algorithm [7] proposed by Jewoong Ryu, Hyung Il Koo based on connected components (CCs) is used to extract text-line in handwritten documents. In this technique strokes and partition under-segmented CCs are analysed into normalized ones. Because of this normalization, this technique is able to estimate the states of CCs for a range of different languages and writing styles. A cost function is built from this estimated state. Minimization of cost function gives required text-lines as shown in Figure 2.



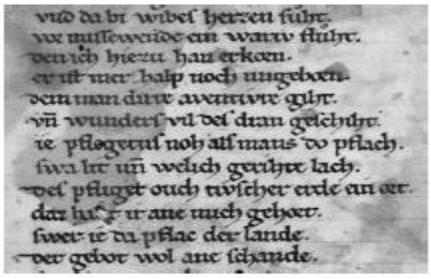
(a) (b)

Figure 2: Result of text-line extraction (a) Original Text (b) Extracted text-lines [7]

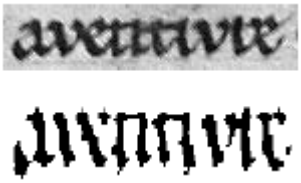
2.3 Keyword Extraction

Multi-font HMM-based recognition model [8] proposed by Aria Pezeshk et. al. used for isolation of text from graphics and then recognizes it. Preprocessing on word images used to reduce the noise. Then the input image is normalizing instead of sending it for recognition. The letter transition probabilities utilizes by system to improve its performance. This method is unable to solve the problem of extraction of linear features from text printed in the same color.

A word spotting method [9] proposed by Volkmar Frinken, Andreas Fischer et. al. based on recurrent neural networks used to extract words from text. For pre-processing IAM offline DB, PARZIWAL DB databases are used in which all documents are already segmented into individual text lines. Then sequence of feature vectors is extracted from each line. These features are then given to the neural network for further processing. They BLSTM Neural Networks used for the handwriting recognition task. Then CTC Token Passing algorithm takes sequence of letter probabilities and dictionary and language model as its input. Then it computes a likely sequence of words as shown in Figure 3.



(a)



(b)

Figure 3: Result of keyword extraction (a) Sample Text (b) Extracted keyword [9**]**

3. Conclusion

Extraction of image, text-line and keyword from document image proves to be a troublesome in document image analysis. We have reviewed various techniques for Image and Text-line isolation, Text-line extraction and keyword recognition in brief. Globally matched wavelet filter's method is one efficient method for content independent image and text isolation. Connected-Component technique for text-line extraction is an efficient method extracting text-lines from documents independent of its language. A word spotting method based on

recurrent neural network is efficiently used for keyword extraction from document image.

References

[1] Jia Li and Robert M. Gray, "Context-Based Multiscale Classification of Document Images Using Wavelet Coefficient Distributions", IEEE transactions on image processing, September 2000.

[2] Mausumi Acharyya and Malay K. Kundu, "Document Image Segmentation Using Wavelet Scale-Space Features", IEEE transactions on circuits and systems for video technology, December 2002.

[3] Sunil Kumar, Rajat Gupta, et.al, "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model", IEEE transactions on Image Processing, August 2007.

[4] Jisheng Liang, Ihsin T. Phillips, and Robert M. Haralick**, "**An Optimization Methodology for Document Structure Extraction on Latin Character Documents", IEEE transactions on pattern analysis and machine intelligence, July 2001.

[5] U. Pal and Partha Pratim Roy, "Multi-oriented and Curved Text Lines Extraction from Indian Documents", IEEE transactions on Systems, Man, and Cybernetic, August 2004.

[6] H. I. Koo and N. I. Cho, "Text-line extraction in handwritten Chinese documents based on an energy minimization framework", IEEE Trans. Image Process., March 2012.

[7] Jewoong Ryu, Hyung Il Koo, "Language-Independent Text-Line Extraction Algorithm for Handwritten Documents", IEEE Signal processing letters, September 2014.

[8] Aria Pezeshk and Richard L. Tutwiler, "Automatic Feature Extraction and Text Recognition from Scanned Topographic Maps", IEEE transactions on Geoscience and Remote sensing, December 2011.

[9] Volkmar Frinken, Andreas Fischer, R. Manmatha, Bunke, "A Novel Word Spotting Method Based on Recurrent Neural Networks", IEEE transactions on Pattern Analysis and Machine Intelligence, February 2012.

Author Profile

1) Sneha L. Bagadkar

She has completed Bachelor of Engineering (Information Technology) from S.R.P.C.E., Nagpur in 2011. She is currently pursuing final year of M.Tech (Computer Science & Engineering) from G.H. Raisoni College of Engineering, Nagpur.

2) Dr. L.G. Malik

She has completed Ph.D. (Computer Science & Engineering) from Vishveshraya National Institute of Technology in 2010, M.Tech. (Computer Science & Engineering) from Banasthali Vidyapith, Rajasthan, India and B.E. (Computer Engineering) from University of Rajasthan, India. She is gold medalist in B.E. and M.Tech.

She is currently working as Professor and Head of Department in Department of Computer Science & Engineering at G.H. Raisoni College of Engineering, Nagpur University, Nagpur, MH, India. She has teaching experience of 16 years. Dr. L. G. Malik is life member of ISTE, CSI, and ACM and presented 22 papers in international journal and 39 papers in international conference. She is recipient of one RPS and 1 MODROBs by AICTE.

She guided 26 PG projects and 8 Ph.D. students are registered under RTM Nagpur University.