ISBN 978-93-84422-71-4
Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016)
London (UK), March 26-27, 2016 pp.37-43

Word Segmentation of Hand Written English Text for Improvement of Word Spotting Results

Muhammad Rashid Hussain, and Asif Masood
National University of Sciences and Technology, Islamabad

***Abstract*:** *Wealth of historical knowledge present in museums, libraries, homes etc is either scanned, typed, hand written or of very poor quality. Access to that knowledge is restricted to few individuals or is confined to a specific group of people. The linguistic barrier alongwith safe storage is another impediment that prevents this precious knowledge to be referenced by scholars around the globe. Recent advancement in science and technology pronounced this impediment and scholars / research organizations around the world are now transpiring into this very important and challenging research area. Content Based Image retrieval is an area of research wherein contents of an image are processed for desired output. Word spotting technique is used primarily for information retrieval systems in ancient / historical documents. Word Spotting Technique matches a input query word with all the targeted words contained in the document and outputs the matched words. Lot of scope and spectrum exists in this active research area. Within the scope of word spotting; processing hand written text / images are yet another challenging area, where style, content, quality of hand written text plays a pivotal role. In order to process / match a hand written word /text it is vital that those images are accurately segmented. Errors in segmentation will have ripple effect in the subsequent stages of feature extraction, indexing, matching etc. Removal of errors like varying writing styles, slanting lines, punctuations, commas, overlapping characters, erased words will enhance accuracy of desired output. An effort has been made in this paper to segmentise hand written documents using innovative techniques to make the foundation as accurate as possible. Five hand written text images(Images taken from IJDAR Competition-2013) written by five different authors having same textual content has been used as a dataset.*

Keywords: *Word spotting, Information retrieval systems, Handwritten Document Images, Databases, Segmentation, Indexing, Matching*

1. Introduction

History of printed documents reveals that ever since its inception, fonts and the layouts of the pages were almost similar to the handwritten books [1,2]. These printed documents were handcrafted and the technical constraints of the past introduced irregularities in book production like variations in spacing, margins, random alignments, etc. Defects were common in manufacturing process, conservation conditions, and absence of printing rules which came into practice at a much later stage. Johannes Gensfleisch Gutenberg was a German goldsmith and printer who is credited with being the first European to use movable type printing, in around 1439, and the global inventor of the mechanical printing press. His major work, the Gutenberg Bible (Fig 1) has been acclaimed for its high aesthetic and technical quality.

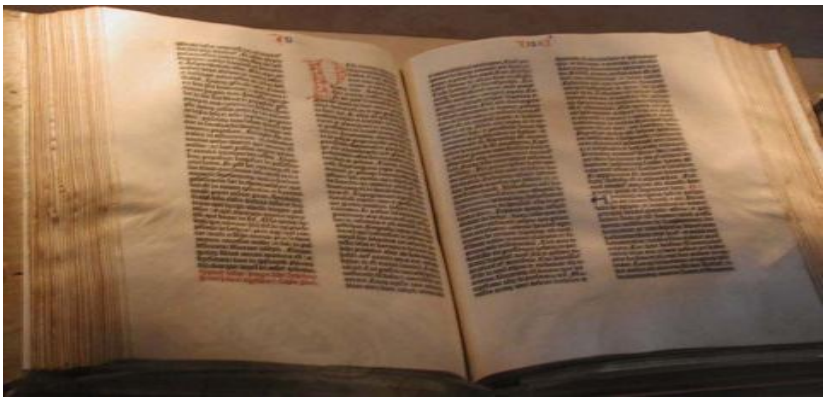


Fig 1. Marco Polo's "Le Livre des Merveilles", Latin edition of 14th century

An information retrieval sys is sub-divided into different stages as under:-

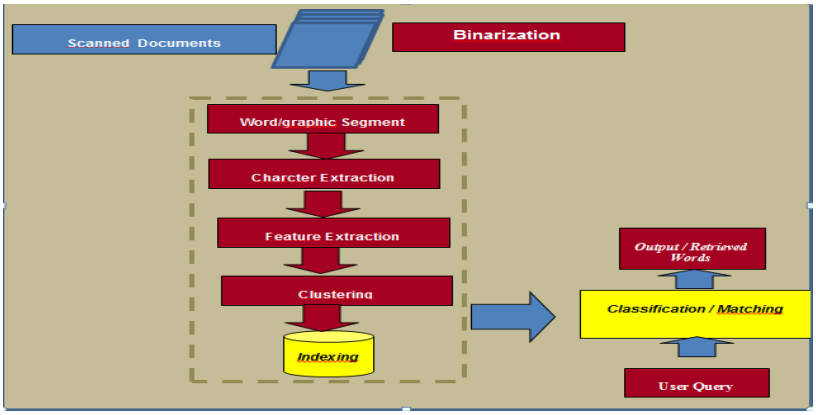


Fig 2. Stages of information retrieval through word spotting

2. Literature Review

Most Widely used implementation of feature extraction for word spotting is word profile [1] by T. Rath and R. Manmatha. An automatic segmentation scheme for cursive handwritten text lines using the transcriptions of the text lines and a hidden Markov model (HMM) based recognition system for word segmentation [2]. With reference to features position dynamic time wrapping [1] Hidden Markov models (HMM) [4]or neural networks [5] have been used for spotting purposes. Symbol/logo [6] spotting and text graphics segregation is presented [7]. Two-stage approach has been proposed towards word spotting in graphical documents, using rotation invariant features isolated components are recognized and matched with the characters of the query string. Word segmentation is the most critical pre-processing step for any handwritten document recognition/retrieval system [8]. Chen Huang and Sargur N. Srihar describes an approach to separate a line of unconstrained (written in a natural manner) handwritten text into words. Marti and Bunke[9] propose a full-page word segmentation algorithm and the evaluation is done by using the IAM database[10].

3. Problems in Segmentation

Hand written segmentation suffers from following errors / impediments:-

3.1 . Style of Writing

Same text written by different authors yield entirely different results. Following five images having same textual content proves this argument.

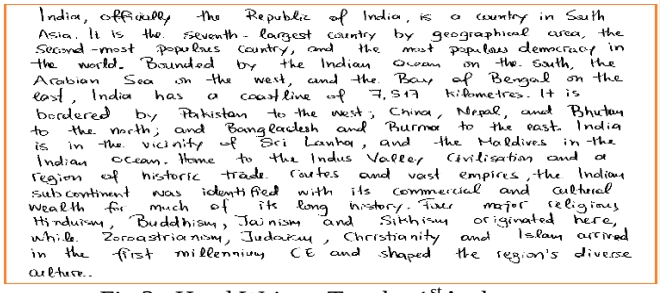


Fig 3. Hand Written Text by 1st Author

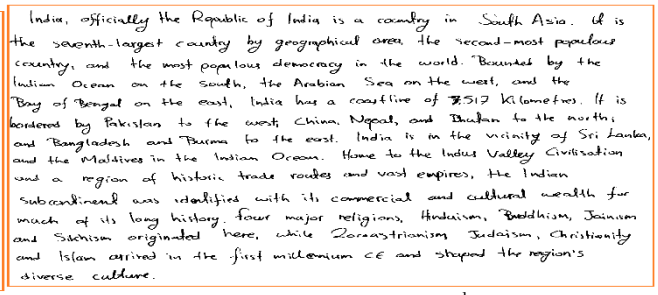


Fig 4. Hand Written Text by 2nd Author

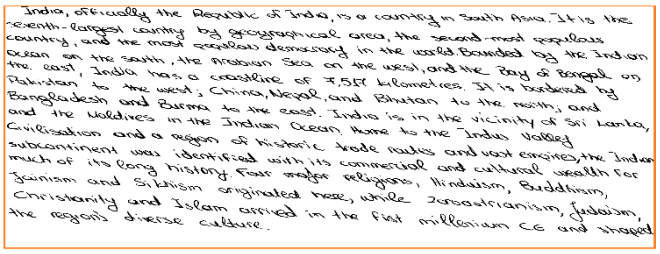


Fig 5. Hand Written Text by 3rd Author

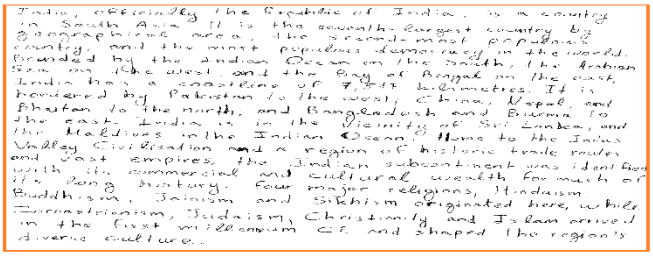


Fig 6. Hand Written Text by 4th Author

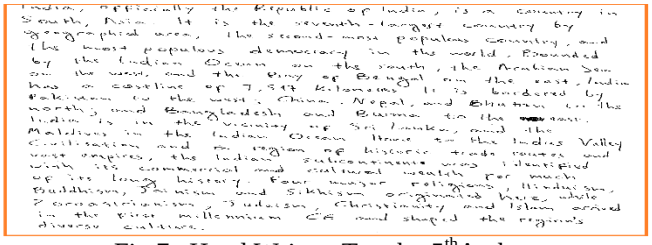


Fig 7. Hand Written Text by 5th Author

3.2. Slanting Lines

Hand written texts may contain slanting lines that may pose challenges like overlapping of characters between two different lines making the connected component analysis difficult.

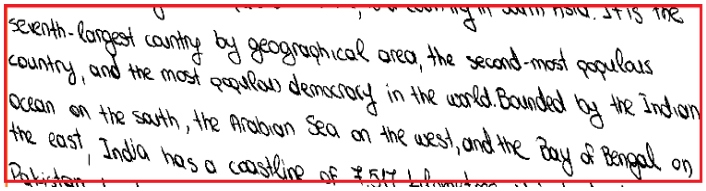


Fig 8.Slanting Lines have their own challenges

3.3. Irregularity In Inter Character / Inter Word Distances

As oppose to typed textual contents where gaps are maintained in uniformity by the software, hand written texts are generally left at the discretion of writer. The irregularity in this aspect poses enormous challenges for word spotting and a common threshold is not possible for a variety of different layouts. Impact of Commas, punctuations becomes more pronounced.

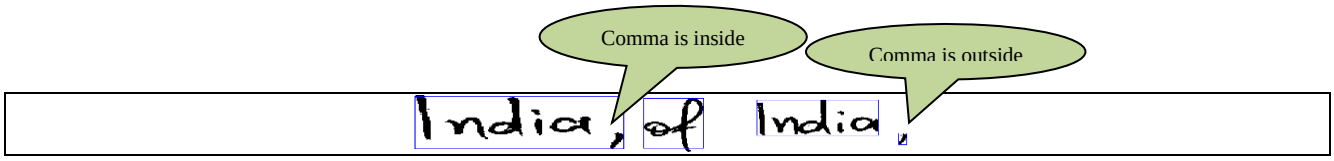


Fig 9.Challenges faced due to Irregularity in spaces

3.4. Misc Errors

Apart from above mentioned errors, hand written texts have variety of errors like inter-spacing, Erasing / cutting, over-writing etc.

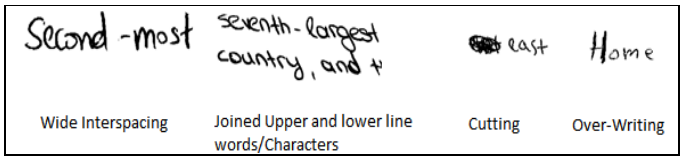


Fig 10. Punctuations, cutting, Over-writing

4. Techniques Employed for Segmentation

Segment selected for display Original Imageis as follows :-

4.1. Processing Steps and Screen Shots

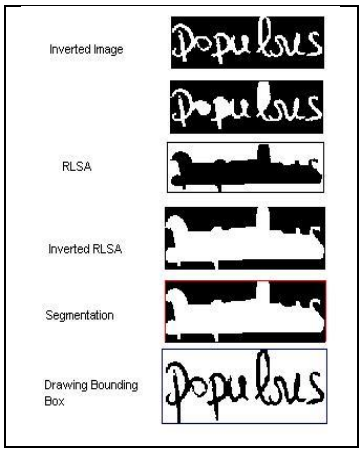


Fig 11. Step Wise Approach in Segmentation

4.2. Explanation of Employed Techniques

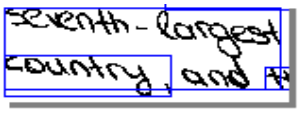
- Stage-I - The image is inverted and holes are filled
- Stage-II - Each line is processed separately and this is done innovatively by counting the sum of number of pixels for each line separately and also the number of pixels (if any) in the gap that exists between two adjacent lines. Threshold is adjusted to equate the pixels in the gap to zero thereby separating each line in turn.
- Run Length Smoothing Algorithm (RLSA) is than applied that uses the concept of connected components and their aspect ratios.
- RLSA image is than inverted and segmentation is done. Rectangular box is drawn by joining end points of each component horizontally and vertically.

- Once each component is bounded in a rectangular box, it is extracted for post processing steps. Errors like punctuations, commas, erasers etc may also form part of segmented word.
- The words can be further subdivided into characters within the bounding box by adjusting the threshold, if character level processing is required.

4.3. Techniques used for Improvement in Post Processing Stage

4.3.1. Detecting Overlapping Components Between Two Adjacent Rows

Hand written text suffer from this unique problem which is not be possible in ordinary printed / typed texts i.e. two characters may overlap each other thus de-shaping each component.



In order to solve these problems, following technique has been employed:-

- Detecting overlapping characters, following pseudo code is used
- Step-I - Find the average height of each row
- Step-II - Find the average gap between rows
- xx = Heights of two consecutive rows + gap threshold =10
- If any component is in between xx - threshold to xx + threshold then it is a candidate for a connected component between two adjacent row or in other words any character of upper row has some connection with a character in bottom row.

4.3.2. Solution

- Step-I - Find Height of Rows of Connected Components.
- Step-II - Find the difference in the gap between the two adjacent lines.
- Step-III - Add row heights and gap.
- Step-IV - Divide the component from center.

4.3.3. Segment Filtration

Following steps have been performed in this technique:-

4.3.3.1. Step-I

Removal of small objects with the pixels containing n pixels (thresholding). This will remove commas and full stops from initial segmented words.

4.3.3.2. Step-II

For Commas, it is a known fact that they are on extreme right bottom corner of the box containing a small fraction of pixel value as compared to words. Using this idea, right bottom corner of the original image is replaced with the processed image. To make it dynamic, cropping is done based on percentage value of height and width of bounding box. Here width is kept similar to height for accuracy. Pseudo code is as follows:-

- Make a copy of image1 named Image1.
- Find the font average size by adding the height of each segment and dividing it by number of segments.
- Apply Bwareaopen to Image2.

- Crop Image1 from the bottom. The right bottom corner of 30% height and the width should be same as height.
- Crop Image2 from the bottom and concatenate this part with Image1.

4.3.4. Moving Bounding Boxes

- It is important that bounding boxes should be placed accurately for further processing of the image in case errors like ‘Commas’ or ‘Full Stops’ are encountered. Pseudo code is as follows:
- Step-I, Start from the right of the image and keep seeking the first white pixel.
- Step-II ,As soon as the pixel is found. Replace the bounding box original width with the new one received.
- Step-III . Repeat the same step from the bottom and replace the height of bounding box with the new one.
- Screen Shot of Post Processing steps

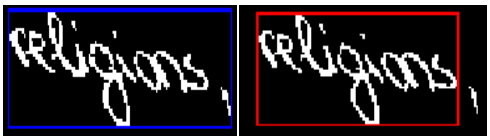


Fig 12. Pre and Post Processing Results

5. Results

Results of our employed techniques are displayed below. Characters in blue bounding box are the initial segmented characters which may contain some errors. Characters in red bounding box are the results of post processing steps employed on segmented characters.

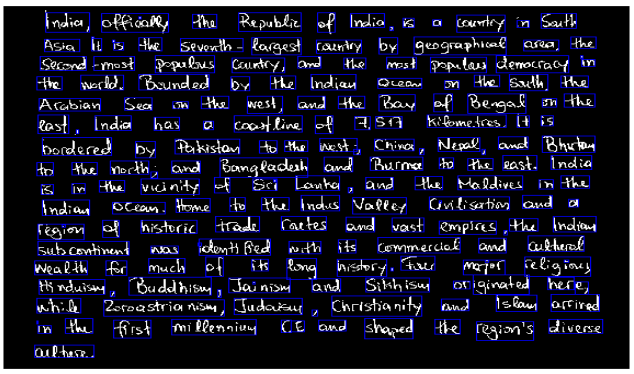


Fig 13. Initial Segmented Characters – Image 1

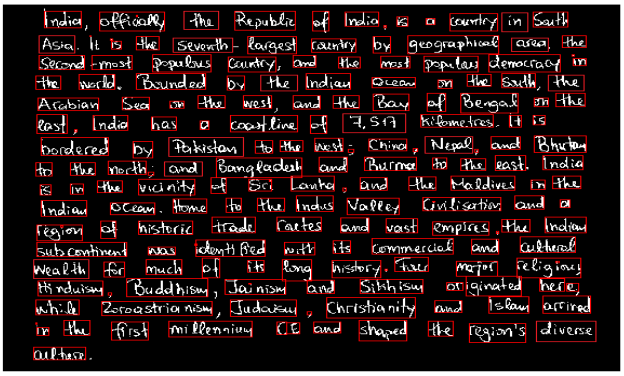


Fig 14. Post Processing of Segmented characters – Image 1

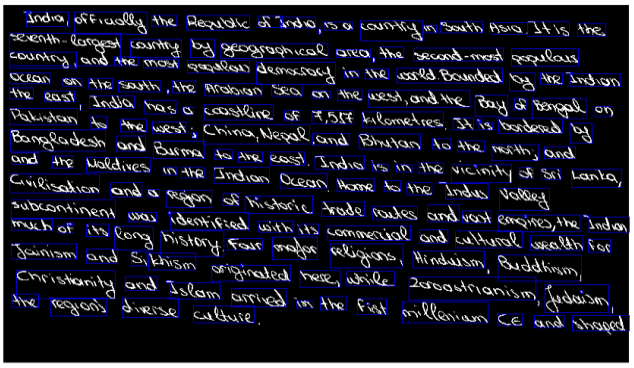


Fig 15. Initial Segmented Characters – Image 2

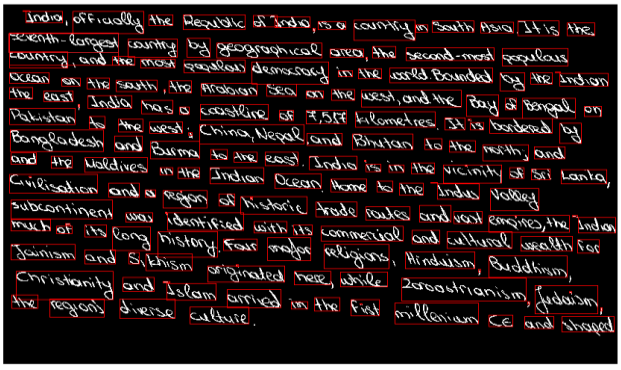


Fig 16. Post Processing of Segmented characters – Image 2

TABLE-1: Summary of Results

Image No	Total Words	Hits	Misses	Accuracy
1.	161	158	3	98.1%
2.	161	160	1	99.3%
3.	161	159	2	98.75%
4.	161	157	4	97.5%
5.	161	160	1	99.3%

6. Future Work

Binnarization and Segmentation are the initial but most critical steps of information retrieval through word spotting. Accuracy in segmentation will have profound impact on later stages like feature extraction, indexing, matching etc. Hand written textual documents are very difficult to segment, and with accuracy of approx 99% results, our research will now focus on matching process. Query word will act as an input and desired output will be targeted. Feature extraction will be explored further and innovative techniques will be employed to get accurate results.

7. Conclusion

Hand written textual documents suffer from various issues like different styles of writing, slanting lines, inter and intra word spaces, irregular punctuations like Commas, full stops etc, overlapping characters, erasers, overwriting etc. In word spotting technique, feature selection plays a pivotal role and in turn feature selection is pre-dominantly dependent on accurate segmentation. In this paper we have used simple yet innovative combination of techniques to achieve 99% accurateresults that are free from errors. Accurate segmentation will act as a strong foundation for further steps in information retrieval systems like feature selection, indexing and matching.

8. References

[1] T. Rath and R. Manmatha, "Word spotting for historical documents," Int. Journal on Document Analysis and Recognition, vol. 9, no. 2–4, pp. 139–152, April 2007.

http://dx.doi.org/10.1007/s10032-006-0027-8

[2] Matthias Zimmermann, Horst Bunke. Automatic Segmentation of the IAM Off-line Database for Handwritten English Text. .1051:465:l/02 Q 2002 IEEE

[3] J. Rodr´ıguez-Serrano and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," Pattern Recognition, vol. 42, no. 9, pp. 2106–2116, September 2009.

http://dx.doi.org/10.1016/j.patcog.2009.02.005

[4] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," Pattern Recognition Letters, vol. 33, no. 7, pp. 934–942, May 2012.

http://dx.doi.org/10.1016/j.patrec.2011.09.009

[5] R. Cao, and C. Tan, "Text/graphics separation in maps." In, Blostein, D.,Kwon,Y.-B. (eds.) GREC 2001. LNCS, vol. 2390, p. 167. Springer, Heidelberg (2002)

http://dx.doi.org/10.1007/3-540-45868-9_14

[6] Arundhati Tarafdar, Umapada Pal, Partha Pratim Roy1, Nicolas Ragot and cJean-Yves Ramel, "A Two-Stage Approach for Word Spotting in Graphical Documents," 2013 12th International Conference on Document Analysis and Recognition

[7] Chen Huang and Sargur N. Srihari," Word Segmentation of Off-line Handwritten Documents,"

[8] U. V. Marti and H. Bunke, "Text line segmentation and word recognition in a system for general writer independent handwriting recognition," Proc. of the 6th Int. Conference on Document Analysis and Pattern Recognition, pp. 159–163, 2001.

http://dx.doi.org/10.1109/ICDAR.2001.953775