IJMTARC – VOLUME – IV – ISSUE - 15 - SEPT 2016

ISSN: 2320-1363

Adaptive Robust Document Image Retrieval Technique for Degraded Document Images and Text

[1]R.Vedashri, [2] N.Sujata Gupta

[1]M.Tech Student, Department of CSE, Sridevi Womens Engineering College, Vatinagullapally(v), Rajendranager(m), Ranga Reddy(d), Telangana state, India.

[2]Assistant Professor, Department of CSE, Sridevi Womens Engineering College, Vatinagullapally(v), Rajendranager(m), Ranga Reddy(d), Telangana state, India.

ABSTRACT:

The Jewelers of the technology always take the credit to the extent of the Human Society, in order to make life easier. In the Context of Information Technology; it has its vast area of coverage towards the automation. Today's IT Industry is completely competitive and always Process oriented rather than data oriented. Optimization and best easy way to avail the service is the main theme behind the technological advancement. In this Paper, We put forward the layer Based Granular pixel methodology where the yeas of years text which is typed or hand written is not perfectly or partially readable will read based on the binary technology. The Process of text and back ground is typically involved the robust approach where accuracy makes a difference, in the approach we have consider each letter as some expected pixel based on the carbon dating process of the Chemical engineering and comparing the Pixel which gives the amazing statically best of the best process to maintain the robustness and artifact. In this aspect if we compare with the classical methodology of the contrast flow algorithm will be different and the sensitized camera is required specially designed to read the carbon dating process to give accuracy of near around 99% , which is most amazing percentage still yet.

INDEX TERMS: Adaptive image contrast, Bit pattern feature, color co-occurrence feature, Content-based image retrieval, ordered dither.

INTRODUCTION:

Historical collections are of interest to a number of people, like historians, students and scholars who need to study the historical originals. Unfortunately, digitization alone is not enough to render historical document collections useful for such research purposes. Having the information available in an electronic image format makes it possible to share it with many people across large distances via the Internet, Digital Versatile Discs (DVDs) or other digital media. However, the size of a collection is often substantial and the content is generally unstructured, which makes it hard to quickly find particular documents or passages of interest. There is no universal or generic text line extraction method that can be robustly applied to a diverse collection of document images.

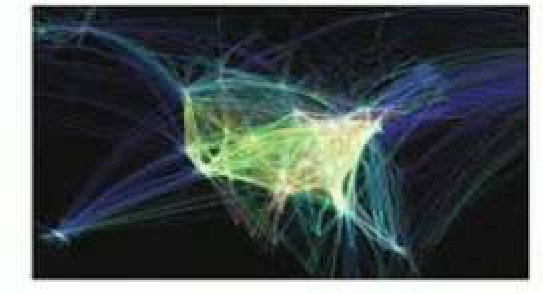

Fig.1.1. Illustration of the Image Pixel

However, a generic text line extraction method can solve a variety of document image processing problems/tasks such as it can be used to overcome the requirement of specific text line extraction methods for different categories of document images, to improve the performance of existing OCR software for complex typed-text document layouts to remove geometric and perspective distortions of warped camera-captured document images, so that de-warped document images can be directly processed by traditional/existing scanner based OCR software, to solve layout analysis problems for document images containing complex script (like Urdu, Telugu, and Kannada), to upgrades/promotes the document image processing pipeline for handwritten/historical document images to page level, which traditionally focuses on the recognition at line, word, or character levels because of complex (irregular) page layouts.

II.RELATED WORK

In data set A, we have document images taken from 5 different books. The image sizes vary from 1256 x 1939 to 1708 x 2721 for different books. From each book, four different pages are selected making the size of the data set to 20 document images. A total of 6,632 words exist in these pages. For query, five different words are selected from the images of each book thus giving a total of 25 different query words having 175 instances in total. Instances of a query word are the different occurrences of the same word in similar or different sizes and styles. For example, the instances of a word “query” could be “query”, “query”, “query”, “Query”, etc. (The word “QUERY” is not counted as the instance of word “query”). The coupled snake-lets based text line extraction method, is designed for extracting curled text lines from typed-text camera-captured document images. It performs better than previously reported curled text line extraction methods for Latin script document images. In addition to Latin script, the coupled snake-lets based text line extraction method can be applied to different scripts like Chinese, Devanagari, and Arabic with specialized script-specific adaptations. In contrast to that, the matched filtering and ridge detection based text line extraction method, which is described in is a domain-independent, generic text line extraction algorithm that can be equally applied to a large variety of document image classes.

III.PROPOSED METHODOLOGY

There are libraries all across the world that are working on the preservation and digitization of ancient historical books and document images. The promises and challenges of digital libraries have been discussed in detail in. The digitization process of the document images, which are made available for research purposes by these digital libraries, is not simple scanning. It is because the crude scanned document images cannot be used directly for information retrieval. Some post processing (like cropping, rotation, etc.) is done for each document to get them in shape for use in information retrieval. An example of this processing is shown in Figure 1.1. This sort of trivial post processing is usually done by the digital libraries before making the digitized document images available for public reading and research purposes. A document retrieval system holds great promise for providing access to historical printed documents containing immense knowledge for a large set of audience. Given a word query, the document retrieval system would find all document images containing relevant "answers" to the query, which saves the user the tedious work of browsing or reading through an entire collection of documents while looking for a particular document image or ROI. This work provides a thorough examination of several retrieval techniques for historical document images that will allow queries in the form of a word image or ASCII text.

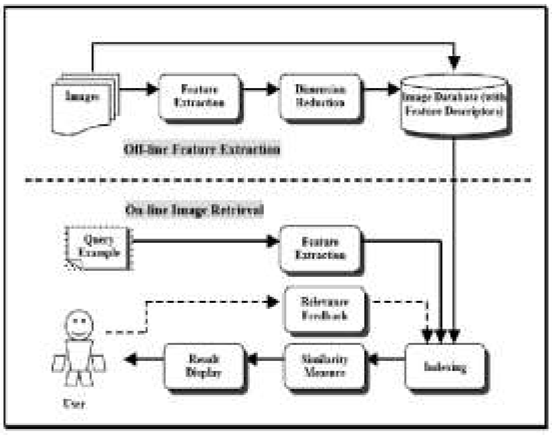

Fig.3.1. Architecture Model of the Content-based image retrieval

It is particularly appealing that the queries can be textual, a fact that makes this system very practical. The task of retrieval is to retrieve document images, that is, the pages of books in the collection at hand according to their relevance to the given query. All document images containing the given query instance are retrieved and are presented to the user in a chronological order. Retrieval is not limited to document images; it can often be easily extended to other retrieval units, such as paragraphs or lines. For example, lines are taken as unit to allow figure captions to be retrieved from the document images to create figure's index. As we are working on printed documents, where a high level of granularity is possible, it prompted us to work at character image level thus allowing more flexibility in the matching phase. For that, the third indexing step revolves around the connected component analysis of words for the extraction of characters. We call these extracted connected components as segmented characters or 'S-characters' while the actual (perfectly segmented alphabetical) characters are termed as true characters or 'T-characters'. The S-characters may or may not be same as T-characters. To get better S-characters (which relate closely to T-characters), S-characters of a word are post-processed using a 3-step process to get better S-characters, which in most cases are the true characters. But still some character segmentation errors remain (see Figure 1.12) which are handled in the matching part. In the last step, multidimensional features are defined for each S-character of a word. To take into account the bad quality of the document images that causes character segmentation problems, a new Merge-Split Edit distance has also been proposed to match two words irrespective of the fact that their characters have been segmented correctly or not. Apart from that, we have discussed different applications of our system such as an automatic figure caption retrieval system by fusion of spatial and perceptual information of the document image. An easy to use graphical user interface (GUI) has also been developed which facilitates us in easily searching the required information by giving ASCII or word image queries and can be used with any kind and number of document base for the purpose of document image indexing and retrieval.

IV.EVALUATION AND ANALYSIS

Overall, we can see that every type of methods has some pros and cons and the final choice of a method depends solely on the type of application for which the segmentation is required.

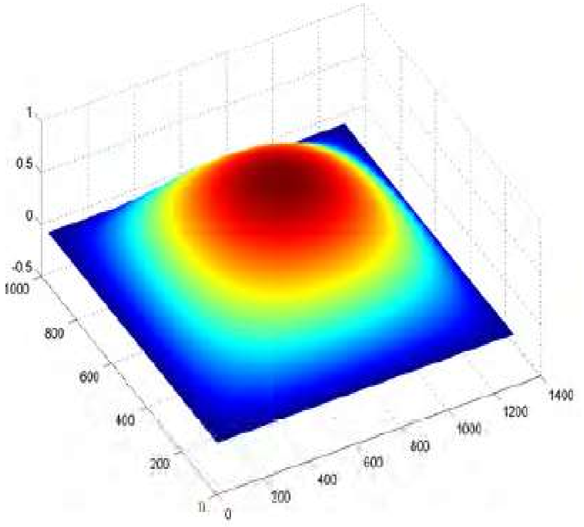

Fig.3.1.1. Comparison of the Contrast with Image PIXEL

If the requirement is limited to text/graphic segmentation on the images for which we don't have any prior knowledge, then hybrid methods are more efficient for that. If, however, we have the prior knowledge about the class of the documents that we are working on, then top-down methods are more efficient and precise while if we want to segment on a smaller/finer level (e.g. characters, etc.), then bottom-up methods seem to do the job better. In the present work, a multi-step bottom up method has been used by taking use of the classic Run Length Smoothing Algorithm (RLSA) in horizontal direction followed by an analysis of the connected components in the image.

V.CONCLUSION

Hand-held camera-captured document images usually contain warped/curled text lines because of geometric and/or perspective distortions. Unlike existing approaches, the presented algorithm performs text lines segmentation and their x-line and baseline pair's estimation simultaneously that results in improved segmentation with better estimation of x-lines and baseline than other approaches. It also yields the lowest over segmentation and missed text lines errors, and a small number of under segmentation errors. The presented method contains six free/tunable parameters. Most of these parameters are non-sensitive with respect to the performance the presented method.

VI.REFERENCES

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in *Proc. Int. Conf. Document Anal. Recognit.*, Jul. 2009, pp. 1375–1382.

[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1506–1510.

[3] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in *Proc. Int. Conf. Frontiers Handwrit. Recognit.*, Nov. 2010, pp. 727–732.

[4] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *Int. J. Document Anal. Recognit.*, vol. 13, no. 4, pp. 303–314, Dec. 2010.

[5] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," in *Proc. Int. Workshop Document Anal. Syst.*, Jun. 2010, pp. 159–166.

[6] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images," in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 13. 2003, pp. 859–864.

[7] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imag.*, vol. 13, no. 1, pp. 146–165, Jan. 2004.

[8] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12, pp. 1191–1201, Dec. 1995.

[9] O. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 312–315, Mar. 1995.

[10] A. Brink, "Thresholding of digital images using two-dimensional entropies," *Pattern Recognit.*, vol. 25, no. 8, pp. 803–808, 1992.

[11] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria,"

IEEE Trans. Syst., Man, Cybern., vol. 15, no. 5, pp. 652–655, Sep.–Oct. 1985.

[12] N. Otsu, “A threshold selection method from gray level histogram,” *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 62–66, Jan. 1979.

[13] N. Papamarkos and B. Gatos, “A new approach for multithreshold selection,” *Comput. Vis. Graph. Image Process.*, vol. 56, no. 5, pp. 357–370, 1994.

[14] J. Bernsen, “Dynamic thresholding of gray-level images,” in *Proc. Int. Conf. Pattern Recognit.*, Oct. 1986, pp. 1251–1255.

[15] L. Eikvil, T. Taxt, and K. Moen, “A fast adaptive method for binarization of document images,” in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 1991, pp. 435–443.

[16] I.-K. Kim, D.-W. Jung, and R.-H. Park, “Document image binarization based on topographic analysis using a water flow model,” *Pattern Recognit.*, vol. 35, no. 1, pp. 265–277, 2002.

[17] J. Parker, C. Jennings, and A. Salkauskas, “Thresholding using an illumination model,” in *Proc. Int. Conf. Doc. Anal. Recognit.*, Oct. 1993, pp. 270–273.

[18] J. Sauvola and M. Pietikainen, “Adaptive document image binarization,” *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000.