Extraction and Recognition of Text From Digital English Comic Image Using Median Filter

S.Ranjini[1]
Research Scholar,Department of Information technology
Bharathiar University
Coimbatore,India
ranjinisengottaiyan@gmail.com

Dr.M.Sundaresan[2]
Associate professor, Department of Information Technology
Bharathiar University
Coimbatore,India
bu.sundaresan@gmail.com

***Abstract*— Text extraction from image is one of the complicated areas in digital image processing. Text characters entrenched in image represents a rich source of information for text retrieval application. It is a complex process to detect and recognize the text from comic image due to their various size, gray scale values, complex backgrounds and different styles of font. Text extraction process from comic image helps to preserve the text and formatting during conversion process and provide high quality of text from the printed document. Automatic text extraction from comic images receives a growing attention because of prospective application in image retrieval. In existing work, Japanese text is extracted vertically from Manga Comic Image using Blob extraction functions. At the same time, text is extracted from multiple constraints using optical character recognition (OCR) and make translation of Japanese language of Manga into some other languages in conventional way to share the enjoyment of reading Manga through the Internet. This paper talks about English text extraction from blob comic image using various methods.**

Keywords- Text extraction; Balloon Detection; Text Blob Extraction; Text recognition.

I. INTRODUCTION

A comic image is simply called as a comic image or funny image and it is a magazine made up of narrative artwork in the form of individual panels that represent the simple comic image or scenes or often comes with dialog, the dialog is usually in balloons or the emblematic scenes. The text extraction from English comic images is challenging where the text are arranged in horizontal wise in one or more blob. Text extraction from image should always meet three factors. First factor should preserve the text and data formatting during the conversion process. For example printed image should remain and look the same, like how it is looked in electronic databases. The second aspect of text extraction is aggregation. For example text extraction from image can easily be posted to another application like spreadsheet or document. And the third factor is image quality. For example text extracted from image should preserve full text quality from the printed document again [1].

The main intention of text extraction from comic image can be used for translation of language by the user using Google translator to any language as the user like. Main challenge in extracting the text from English comic image is how to detect the comic balloon, and extract the text in horizontal direction. Comic Balloon detection is done by connected component labeling algorithm. A Sample Comic image is shown in “Fig 1”.

Figure 1. Sample Comic image

Normally text extraction can be done in two methods such as texture based methods and region based methods. In this proposed method, we are using the region based text extraction technique for extracting the text from the English comic image.

II. LITERATURE REVIEW

According to Siddhartha Brahma, the text extraction from image is done by using the shape context matching [2].

According to Ruini Cao, Chew Lim Tan – the separation of overlapping text from graphics is a challenging problem in document image analysis. So they used a specific method for detecting and extracting characters that are touching graphics. It is based on the observation that the constituent strokes of characters are

usually short segment in comparison with those of graphics. It combines line continuation with the feature line width to decompose and reconstruct segments and improved the percentage of correctly detected text as well as the accuracy of character recognition significantly [3].

Q. Yuan, C. L. Tan presented a well designed method that makes use of edge information to extract textual blocks from the gray scale document images. It aims at detecting textual regions on heavy noise infected newspaper images and separate them from graphical regions. The algorithm traces the feature points in different entities and then groups those edge points of textual regions. By using the line approximation and layout categorization, it can successfully retrieve directional placed text blocks. Finally they used a connected component merging to gather homogeneous textual regions together within the scope of its bounding rectangles. They tested this method on a large group of newspaper images with multiple page layouts, promising results approved the effectiveness of their Method [4].

Kohei Arai and Herman Tolle stated that Reading digital comic on mobile phone is demanding now. Instead of creating new mobile comic contents, adaptation of the existing digital comic web portal is valuable. In this paper, they proposed an automatic e- comic mobile

content adaptation method for automatically creating mobile comic content from digital comic website portal. Automatic e-comic content adaptation is based on the comic frame extraction method combined with additional process to extract comic balloon and text from digital comic page. Their proposed method is an effective and efficient method for real time implementation of reading e-comic comparing to other methods. From their Experimental results they showed a 100% accuracy of flat comic frame extraction, 91.48% accuracy of non-flat comic frame extraction, and about 90% processing time faster than previous method [5].

According to Kohei Arai and Herman Tolle *Manga* is one of popular item in Japan and also in the rest of the world. Hundreds of *manga* printed everyday in Japan and some of printed *manga* book was digitized into web *manga*. People then make translation of Japanese language on *manga* into other language -in conventional way- to share the pleasure of reading *manga* through the internet. In their paper, they proposed an automatic method for detecting and extracting Japanese character within a *manga* comic page for online language translation process. Japanese character text extraction method is based on comic frame content extraction method using blob extraction function. Their experimental results from 15 comic pages showed that the proposed method has 100% accuracy of flat comic frame extraction and comic balloon detection, and 93.75% accuracy of Japanese character text extraction.96.45- 99.79% accuracy depending upon script [6].

III. METHODOLOGY

Extraction of the text from digital English comic image is based on the assumption that only text is situated within a comic balloon (blob). That is the reason comic balloon detection has to be done clearly before text extraction as shown in the flow diagram of comic text extraction method in “Fig 2”. Many researchers proposed method for text extraction from images but not specific for extraction from comic image. There are two base methods for text extraction, one is region based method and other is texture based method. In the existing work, they proposed a method for extraction of text in online way and automatically make Translation using online translation feature on internet like Google language translation. Morphological filter is used for pre- processing.

In the proposed method, new method for automatically extracting text inside comic balloon from digital English comic (e-comic) image is proposed. Comic contents such us balloon and text inside balloon is extracted for further purpose, for example language translation, multimedia indexing or data mining. Region based text extraction method is applied for English character text extraction. These methods can be divided further into two sub- approaches: Connected Component (CC) and edge- based. In binary image analysis objects are usually extracted by means of the connected components labelling methods, which consist of assigning a unique label to each maximal connected region of foreground pixels. CC-based methods apply a bottom-up

approach by grouping small components into successively larger ones until all regions are identified in the image.

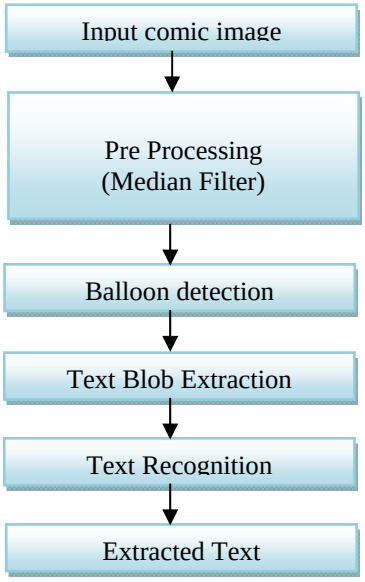

Figure 2. Flow Diagram for Text Extraction

IV. EXPERIMENTS

A. *Input Comic Image*

Comics representing individual scenes, often come with dialog, usually in balloons in art form, as well as including brief descriptive style. In this research work colour digital English comic image is taken as input image. RGB band values are applied to input image for band selection which is shown in "Fig 3".

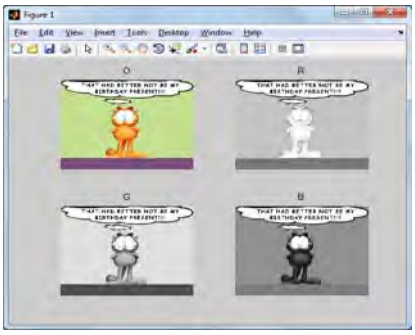

Figure 3. Sample RGB Image

B. *Pre-Processing Digital Comic Image*

In the pre-processing step the RGB images are converted into a binary image by applying the threshold values between 0 to 1. Threshold value T is obtained from average pixel value of comic image μ. Threshold value 0.9 above from the average empirically is set. This is depicted in "Fig 4".

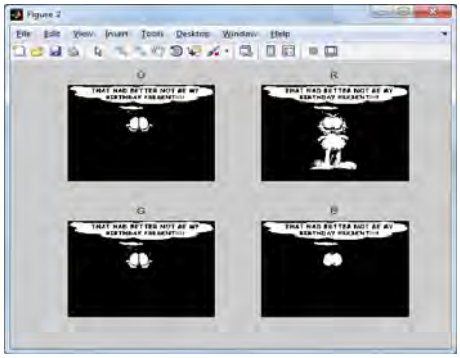

Figure 4: Binary Images

As noise images (RGB) are taken as input image, pre processing steps are done to remove the noise which helps to improve the efficiency of text extraction. In this research work Median Filter is applied for Pre-processing process to remove the noise. Median Filter averages the neighboring pixels for smoothening the image. “Fig 5” shows the noise reduced image.

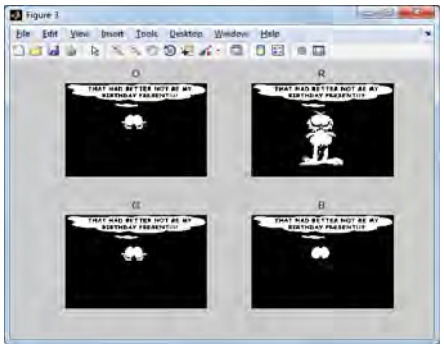

Figure 5. Sample Noise reduced image by median filter

C. ic Balloon Detection

Comic balloon (blob) detection is one of the important processes in comic text extraction. The accuracy of balloon detection is correlated with successful text extraction. CCL algorithm is applied to the noise removed RGB images for detecting the connected components in the image. Connected-component labeling (alternatively called as connected-component analysis, blob extraction, region labeling, blob discovery, or region extraction) is an algorithmic application of graph theory, where subsets of connected components are uniquely labeled based on a given heuristics.

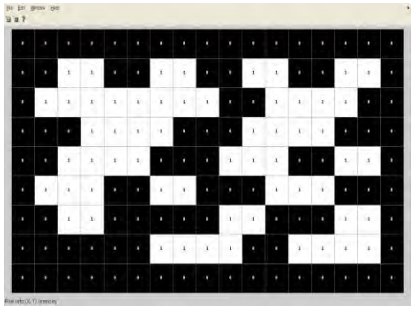

Figure 6 . Sample CCL image

By applying CCL, region boundaries have been detected, it is often useful to extract regions which are not isolated by a boundary. A set of pixels which are not separated by a boundary is called as connected components [7]. Each maximal region of connected pixels is called a connected component. Balloon detection process produces text blobs and also non text blobs which is represented in “Fig 6”. And here B-Band for next input text blob extraction is taken as, this Band provides better result when compared to R and G image. Because a set of connected components is less in B-Band Image when compared to other Bands(R and G) which is shown in “Fig 7”.

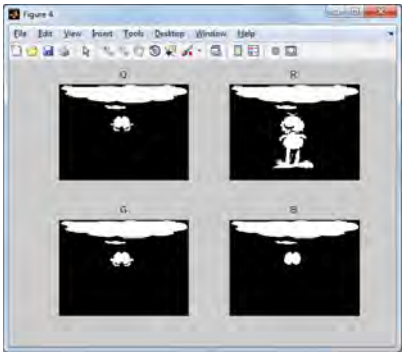

Figure 7. Connected components labeling

D. Text Blob Extraction

Blob Extraction is needed to identify text blobs from non- text blobs. After applying CCL, the number of CCL objects are obtained. Figure 7 shows text blobs and non text blobs. To avoid the false detection and to reduce the complexity the text blobs are to be identified exactly. The identification is done, based on the features of blob size [8]. For that the Area of the blobs are calculated using the following equation (1).

$$A.TB[i]=TB[i].Width*TB[i].Height \quad (1)$$

If the area of the blobs is 10% (or 8%) from the original image then it is classified as a text blob and all others are classified as non text blob, which is shown in "Fig 8".

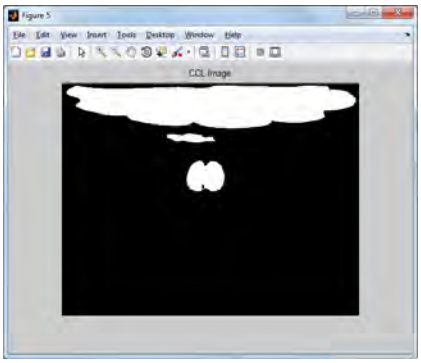

Figure 8. Sample Text blob extraction

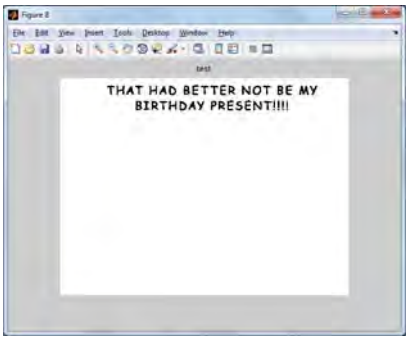

Figure 9. Text Extraction from Blob.

E. Text Recognition

During the Text Blob Detection, there is a possibility of the occurrence of false detection but that is not a serious problem, if text is recognized using Optical Character Recognition. The extracted text from the blob is recognized by the OCR which is shown in the "Fig 10". The process of OCR is segmentation, correlation, and classification. During segmentation, OCR crops each character in the text blob and in the correlation phase OCR matches the cropped characters with the datasets. During the classification process, it recognizes the text in text blob, if the crop characters matches with the datasets. Finally the extracted text is stored in text file for user convenience.

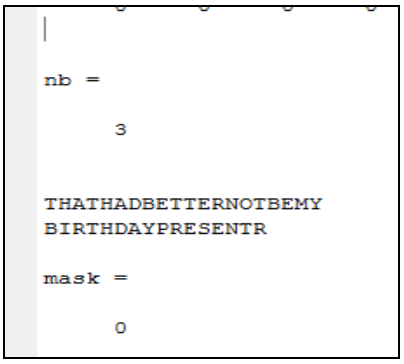

Figure 10: Extracted Text

V. RESULTS AND DISCUSSION

The proposed methodology for text extraction from digital English comic image character is implemented using MATLAB version 10. As MATLAB is widely used in image processing application, Comic text extraction was developed using MATLAB, that acts as a tool and facilitate text extraction from comic images. In this research work text is extracted in Horizontal manner. But in the conventional method they used a Japanese comic image for text extraction in vertical manner using Morphological filter. Experiment result is discussed below using the median and morphological filter for the comic image. Time is the main issue for noise reduction. In this paper the processing time of the proposed method using the median filter is evaluated and compared with morphological filter process timing. From that, the Median filter gives a better noise reduction timing which is shown in the "Fig 11".

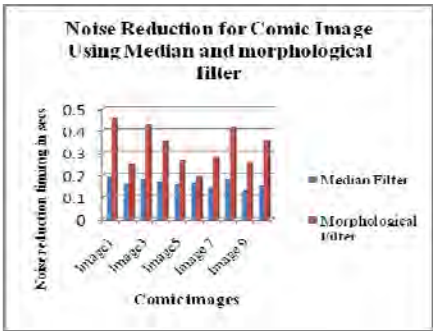

Figure 11. Noise Reduction Timing

After the Blob detection from comic image using connected component, text extraction process are implemented. During the text extraction process only blobs are considered, because balloons always represent non text blob. Text Extraction results are classified into two groups such as text extracted and text not extracted. OCR is applied for text extraction. When the OCR is applied to comic image, the image which is processed by Morphological filter and the results for text extraction count is low and text not extraction count is high which is recorded in Table 1. When the OCR is applied to comic image, the image which is processed by Median filter and the results for text extraction count is high and text not extracted count is low which is recorded in Table 1. Therefore from the above discussion, comic image which is processed by Median Filter provide better result for text extraction than the Morphological Filter.Where the CD denotes the Correctly detected and MC denotes missed characters.

TABLE I. Text Extraction using Median and Morphological filter

Comic Image	Total characters	Median Filter		Morphological Filter	
		CD	*MC*	*CD*	*MC*
Image 1	***35***	***35***	***0***	***13***	***22***
Image 2	24	21	3	16	18
Image 3	27	27	0	25	2
Image 4	14	14	0	3	11
Image 5	15	13	3	15	0
Image 6	32	32	0	25	7
Image 7	19	17	2	11	8
Image 8	29	28	1	20	9
Image 9	23	23	0	12	11
Image 10	17	15	2	6	11

VI. CONCLUSION

A new automatic method for text extraction from English comic image is proposed in this work. Median Filter and Black White threshold improves the performance of text extraction. This method can automatically detect the comic balloon using the median filter. After the blob detection, the system will extract the English text characters text inside the blob. So from the results the text extraction phase achieved 95.35 % successively.

REFERENCES

[1] White paper. "Extraction of text from image".
[2] Siddhartha Brahma , "Text Extraction Using Shape Context Matching". COS429: Computer Vision. Vol.1, Jan 12, 2006.
[3] Ruini Cao, Chew Lim Tan, "Separation of overlapping text from graphics," vol.29,no.1, pp.20-31, Jan/Feb 2009.
[4] Q. Yuan, C. L. Tan ,"Text Extraction from Gray Scale Document Images Using Edge Information," proceedings of sixth international conference on document analysis and recognition, pp.302-306, 2001.
[5] Kohei Arai and Herman Tolle, "Automatic E-Comic Content Adaptation," International Journal of Ubiquitous Computing (IJUC) vol.1, Issue(1), pp1-11, 2010.
[6] Kohei Arai and Herman Tolle, "Method of real time text extraction from digital manga comic image," International Journal of Image Processing (IJIP), vol.4, Issue (6), pp 669-676, 2010.
[7] C. A. Bouman "Connected Component Analysis," Digital Image Processing , pp 1-19, Jan 10, 2011.
[8] Sundaresan.M, Ranjini.S. "Text extraction from digital English comic image using the two blobs extraction method" Proceedings on the international conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), pp 467-471, 978-1-4673-1039-0/12/$31.00 ©2012 IEEE, Mar 2012.

AUTHORS PROFILE

S.Ranjini did her B.Sc (Computer Science) Degree from Periyar University in 2008, M.sc (Computer Science) Degree from Periyar University in 2010, M.Phil(Computer Science) from Bharathiar University – Coimbatore in 2012. Her area of interest is Digital Image Processing. She has presented her research paper in International Conference and National Conference.

Dr.M.Sundaresan did his B.Sc (Applied sciences) Degree at Madurai Kamaraj University, MCA at Bharathidasan University, M.Phil and Ph.D at Bharathiar University. He has presented more than 20 research papers in International and National Conferences. He has published in 15 Journals. He is Senior Life member of Professional bodies like CSI, ISCA and ISTE. Currently, he is working as an Associate Professor and Head i/c of the Department of Information Technology at Bharathiar University. His areas of interest are Image Processing and Data Compression.