Int. J. Advance Soft Compu. Appl, Vol. 9, No. 3, Nov 2017
ISSN 2074-8523

Optical Character Recognition on English Comic Digital Data for Automated Language Translation

Anggi Gustiningsih Hapsani[1], Fitri Utaminingrum[2], and Herman Tolle[3]

[1,2,3]Departement of Computer Science
Brawijaya University
e-mail: anggigustiningsih@ub.ac.id[1], f3_ningrum@ub.ac.id[2]
emang@ub.ac.id[3]

Abstract

Comic is a popular entertainment that tells about story line by using image as illustration. The distribution via internet makes comic market is not only local in one country but also international that across several countries. Many of comic lover did not understand foreign language. The current solutions that provided by comic community was translating comic from one language to another language. This solution done manually and caused a gap time between original comic and translate comic release time. This long difference release time give an addition problem. The problem of translation comic manually can be solved by using a system that translated the text of comic automatically. The challenge to built this system is how to detect and recognize the text of comic. We propose an approach for detect and recognize comic texts in digital English comics. Our method has a good accuracy to detect the frame and blob. Success rate while detect the frame is 90.54% and detect the balloon is 77.92%. We also have a good result where extraction text 98,60%, line numbering 98.91% and text recognition 85.95%.

Keywords: *comic, manga, text extraction, connected component labeling, detection text, recognition text, OCR*

1 Introduction

Comic is a popular entertainment that tells about story line by using image as illustration. Comic not only enjoyed by children, but also adults. Among with the increasing of technology and popularity of data digital, comic has been converted into digital form, usually called with digital comic. The term digital comic means

that the final form of comic is digital data and this comic created digitally or distributed via internet digitally.

The distribution via internet makes comic market is not only local in one country but also international that across several countries. Each country have local language and it is different from one country to another country. Unfortunately, many of comic lover did not understand foreign language. It would be a problem and caused the distribution of comic was constrained. The current solutions that provided by comic community was translating comic from one language to another language. This solution done manually and caused a gap of release time between original and translated comic. This gap time approximately are two or three weeks. The long difference release time give an addition problem, because the comic lovers need a fast service that can fulfilled their curiosity about the next story line of comic. However, if the comic lover feel satisfied, it will give a positive effect to the publisher of comic. The enthusiasts will be increased largely. It is actually also depends on the interest or not the storyline of the comic itself. The more popular a title of comic, the income that publisher get will be larger.

The problem of translation comic manually can be solved by using a system that translated the text of comic automatically. The challenge to built this system is how to detect and recognize the text of comic.

OCR (Optical Character Recognition) is a technology that developed to be able read text from an image containing image text in it. Many methods have been developed in to solved the OCR problem. The image of object used in OCR include image scene (car plate, traffic board, advertising, et al) and the scanned document (image of historic document, books, et al). Generally, OCR system has five components. They are optical scanning, location segmentation, preprocessing, feature extraction and recognition post-processing [1]. Overall, OCR system can divided into two main parts, text detection and recognition. The aim of text detection is to localize the prediction text in clutter image. The text recognition is step which recognize the image text that have been detected and the output of this step is string data. Many previous research have been conducted to detect the text of comic. The connected component used in text extraction for Japanese comic which text direction of this object comic is vertical [2]. A better result of extraction text requires an approach optimization by implementing a Median Filter in pre-processing [3]. Rigaud [4] make a better approach in frame and text segmentation comic by using connected component. The previous research project about extraction text in comic, only focus in how to detect the frame, balloon and extract text inside balloon correctly, do not include text recognition. However, Ranjini and Sundaresan [3] do the text recognition but did not explain clearly about how the step have been conducted.

Many methods had been used to solved the problem in OCR. Hidayatullah et al [1] use template matching to improve the optical character recognition for license plate in Indonesia. Hidayatullah convey that the template is a most important part

of OCR systems, especially when used template matching as method. The high of accuracy rate depend on dataset. The dataset should can accommodate the variance of font style. The better accuracy rate of recognition system can be gotten from capable dataset to accommodate the characteristic of character.

We consider that each character (letter) has a basic characteristic although in difference font style. Because of this basic characteristic, human can differentiate one character from another character although in different font style. We adopted these human behaviour to built the recognition system. As well as basic of OCR system, we divided the system into text detection and recognition. On the text detection, we adopted the connected component labelling that have been modified to detect and extract the text. Meanwhile, on the text recognition we use two type of point feature and built a system rule based on count of these feature for every character. We called our recognition method Hierarcycal Point Matching Classification.

Our research in this paper only focus on how to detect and recognize the text of digital english comic. The translation process will be explained in our next research.

The rest of the paper is organized as follows : In section 2 presents the explanation about digital english comic. Proposed method is reviewed in Section 3. Experiment result and discussion is given in Section 4. The conclusion in Section 5.

2 Comic Digital

Printed comic usually has been scanned to get digital comic. Comic tell the storyline by using image. Generally, comic divided into four main parts :

Panel or frame. This part contains one scene of the storyline in comic.

Balloon. This part contains text dialogue which said by actor to another actor.

Text. This part is the most important of comic image. Usually dialogue text located inside the balloon.

Character or actor. This part is illustration about actors in storyline. Usually like cartoon actors.

Detail about comic shown in Fig. 1.

3 Proposed Method

As well as the OCR system, our methods consist of three main part. It is text detection, feature extraction and text recognition.

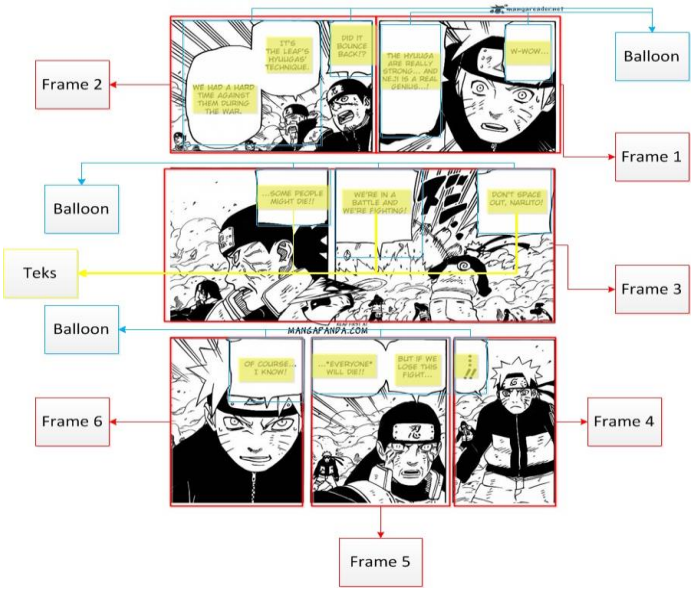


Fig. 1 Part of Comic Digital

3.1 Text Detection

The aim of this part is to detect the text location and extract it into new image text. In our study, we only focus to extract the text inside the balloon and ignore the other. We use connected component labeling to detect the text. The text detection process shown in Fig. 2.

3.1.1 Frame Extraction

The function of this step is to extract the frame of comic. The input can be a grayscale image or color image (RGB). First, the input image was converted into grayscale and we construct the binary image from it. The color of comic background usually is white and the object is black. The binary image then was inverted, this process caused the object to be white. The morphology reconstruction like dilation and erotion is applied on it. The holes inner frame was filled with white pixel based on morphological reconstruction [5]. The connected component labelling was used to detect the blobs of object. The area of each blob would be calculate and we used 700 pixels as threshold value of area that could detected as frame. We used 700 pixels as threshold because this value have a better accuration than another value.

3.1.2 Balloon Extraction

The frame images that have been cropped will be processed into balloon extraction process. First, the frame image is processed by converting into binary image. The background color is white and the object is black. After that, we fill the area with the dominant pixel. If the dominant is white, the area was filled with white and so do with black. The result of filling process was illustrated in Fig. 3.

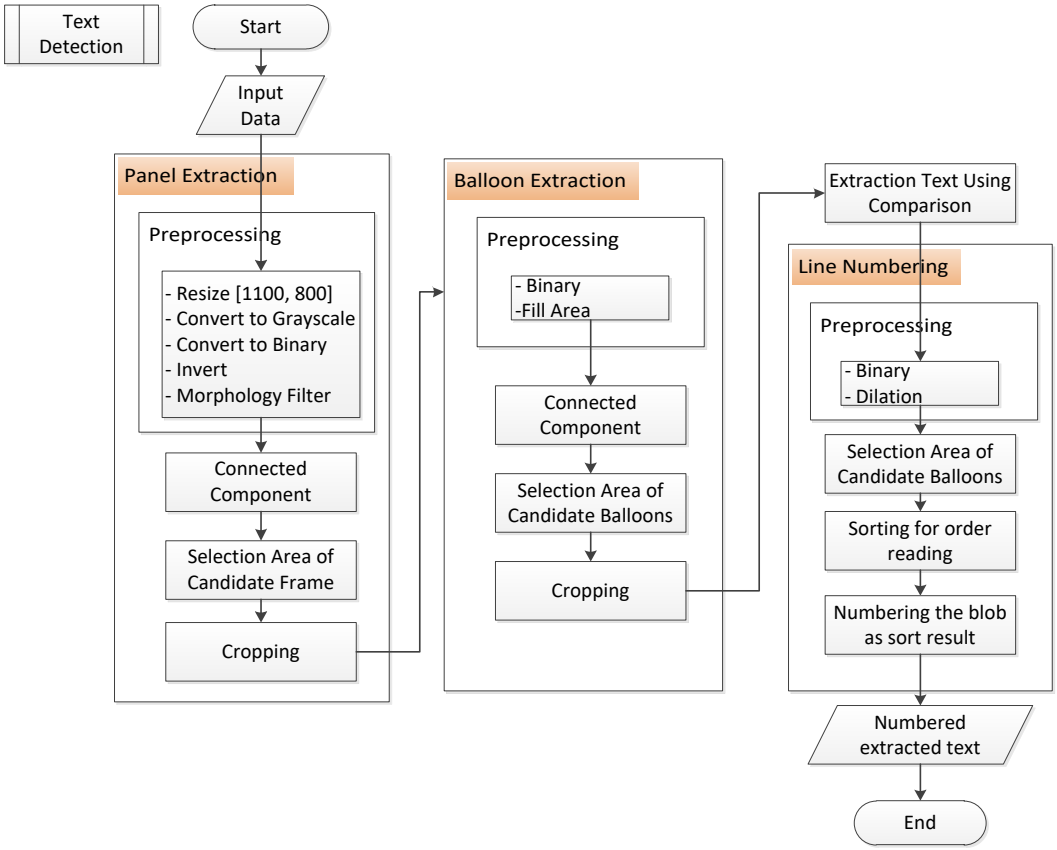


Fig. 2 Text Detection Process

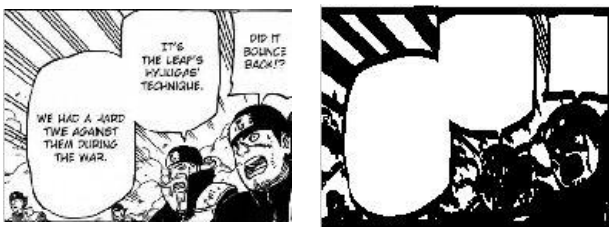
The balloon area is a big area of white pixels. This area was detected by using connected component labeling method and threshold value for selection area is 5000 pixels. The balloon extraction conducted by calculating the coordinate corner of balloon area using boundingbox. The detected balloon area was cropped into new image. The result of balloon extraction shown in Fig. 4.

3.1.3 Text Extraction

The aim of text extraction process is to get the region of candidate text in image. The results of balloon extraction (shown in Fig. 4) still have many noises, which not only text but also scratch background. The solution to solve this problem, we just take the pixel value within the area that has been detected as balloon. We compare the filled area image of balloon and the binary image of original balloon image that have been detected. The text image was built into new image by using Definition 3.1

Definition 3.1.

$$I_{(x,y)} = \begin{cases} 0, & if \quad M_{(x,y)} = 0 \\ N_{(x,y)}, & if \quad M_{(x,y)} = 1 \end{cases} \tag{1}$$



(a) (b)

Fig. 3 The result of filling process (a) original frame image, (b) frame image after process

(a) (b)

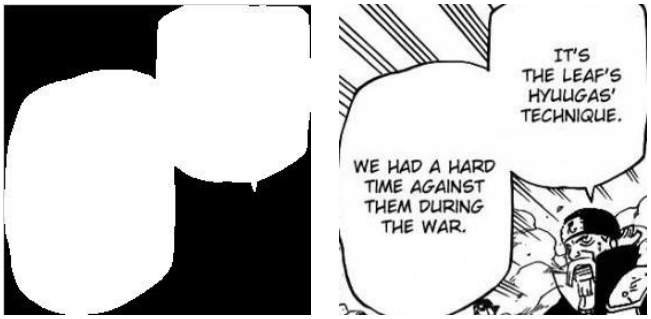
Fig. 4 The result of balloon extraction (a) first balloon, (b) second balloon

Denotes, I is the result text image, N is the binary image of balloon image, M is convex hull image of balloon image and (x,y) is pixel coordinate. The illustration of this text extraction process was illustrated in Fig. 5.

3.1.5 Line Numbering/Grouping

Line grouping text is a process that horizontally grouping each character that has been detected by using connected component algorithm into one line or not.

First, we convert the extracted text image into binary image. It is possible to make segmentation process easily. We apply dilation process to combine the near character into one line. This filter can help to segment the character into group based on lines. For reducing the noise, we omit the small blobs that have area less than 100. By using this statement, the blob that is detected as noise will be ignored. The way to read comic is different with the way while reading the usual book. In this study, we use comic that adapted from Japanese comic. The reading order to read the line of text (row) is from right to left and top to bottom. To complete this order, we sort the minimum coordinate-y of each blob boundingbox



(a) (b)

(c)

Fig. 5 Text extraction process (a) convexhull image, (b) binary image, (c) extracted text image

value ascending. The recognition text process based on this result sort order. The line that have been numbering was illustrated in Fig. 6.

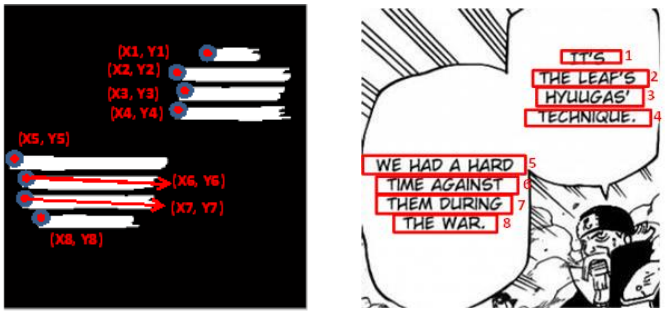
3.2 Feature Extraction

Every type of font in character always have two unique points. We used two kind of feature that can be a unique point of character, there are branchpoint and endpoint.

- The branchpoint is a point that be a branch or joint point of two or more lines which shaped the character of font. Fig. 7 illustrate about branchpoint.
- The endpoint is a point that be a corner or end of line character. The example of endpoint is shown in Fig. 8.

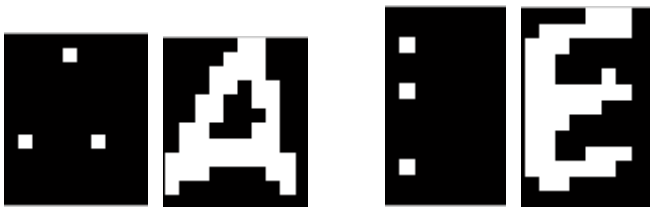
3.3 Text Recognition using Hierarchycal Point Matching

Text recognition is performed by using Hierarchycal Point Matching Classification. The design of this method illustrate in Fig. 9.



(a) (b)

Fig. 6 (a) Minimum coordinate-x and coordinate-y, (b) and the result of line numbering process



(a) (b)

Fig. 7 (a) Branchpoint of A letter, (b) branchpoint of E letter

(a) (b)

Fig. 8 (a) Endpoint of A letter, (b) endpoint of E letter

The aim of Hierarchycal Point Matching Classification Method is to split characters into two groups based on the bounding box of the form letter, which is square or rectangular. Each of these groups will be divided into two groups based on the number of branchpoint dominant area, this author's new contribution to the solutions used in the process of solving problems in the field of optical character recognition. On the square group, if the number of branchpoint letter larger or dominant on the left of the area bounding box, the character will be

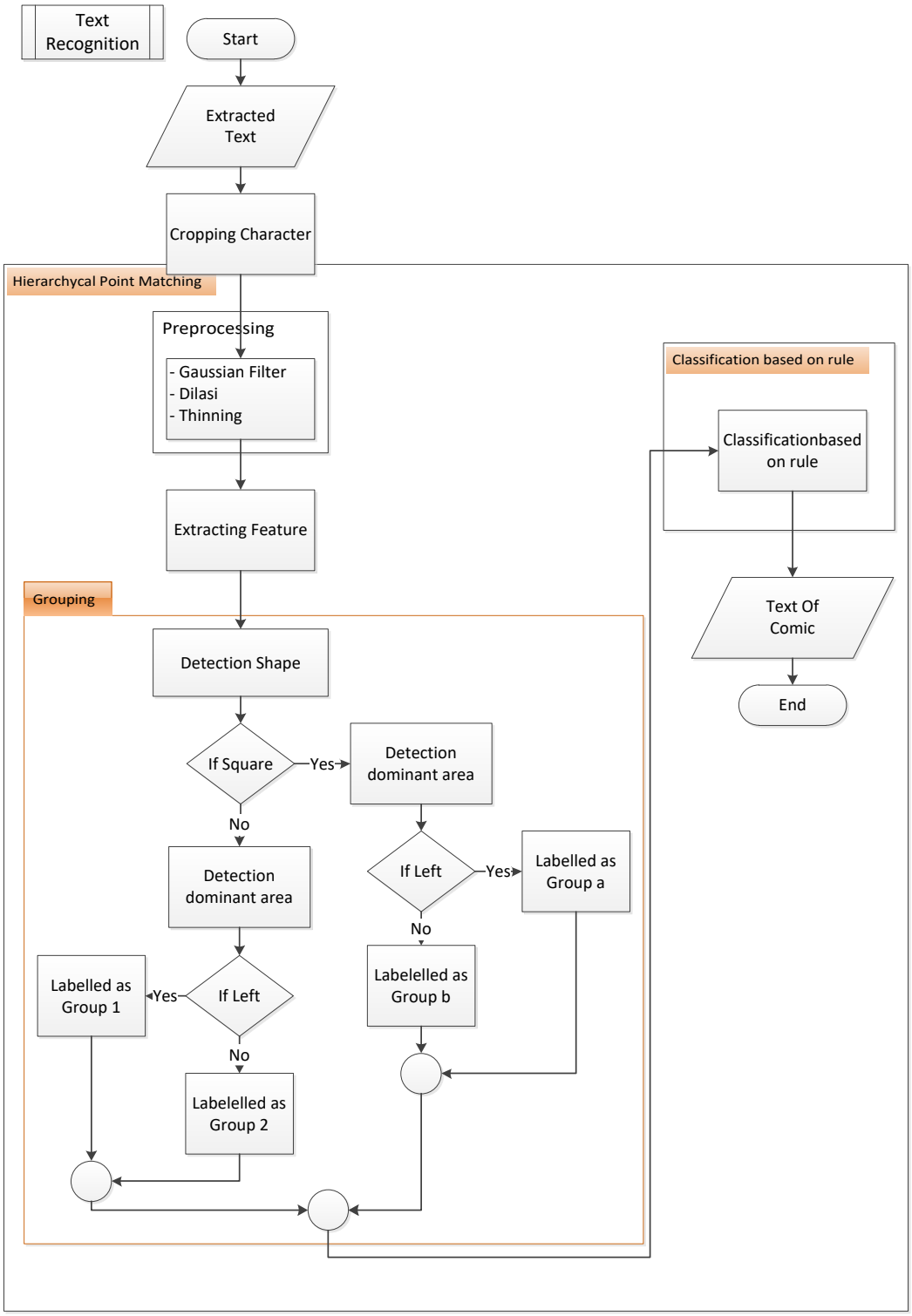


Fig. 9 Text recognition process

labeled as Group A, and then if dominant on the right, then will be labeled as Group B. On the other hand, the rectangle group, if the count of branchpoint letter larger or dominant on the left of the area bounding box will be labeled as Group 1 and vice versa if the dominant right will be labeled with Group 2.

In the classification character, we use rule that can classify the character based on the count of branchpoint and endpoint. First, the result branchpoint and endpoint image of character will divide into 9 part, according to Fig. 10. The rule generated from this two feature for each letter or character.

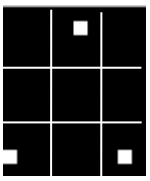


Fig. 10. The divided area of branchpoint image

4 Experiment

The proposed method for extracted text from digital English comic image has been evaluated using various naruto comic image pages. We use 20 pages image with total number of frame is 121 and 157 of balloon. We implement the proposed method using Matlab on desktop computer with Pentium Dual Core processor and 2 Gigabyte of RAM. Fig. 11 show the snapshot of our application. The experiment is conducting through 20 comic pages to evaluate the success rate (accuracy) of text extraction. We calculate the value of accuration by using simple equation (Definition 4.1).

Definition 4.1 TP is true positive, TD is count of all data.

$$\%\ accuration = \frac{TP}{TD} x\ 100\% \quad (2)$$

Based on Table 1, we can see that our method has a good accuracy to detect the frame and blob. Success rate while detect the frame is 90.54% and detect the balloon is 77.92%. The result of this experiment is shown in Table 1.

Regarding to Table 1, our method have good result to detect the individual frame, but has limitation to detect the overlap frame. The overlap frame will be detected as one frame by our method. The illustration about overlap frame is shown in Fig. 12. However, the result of the balloon extraction was not good enough because we take the data comic directly from website for purpose at online processing time on our future work. We do not recondition the comic data, on the other word we use raw data. Usually, the owner of the website adds some label which gives their identity into digital comic data that their upload on their web. This issue makes

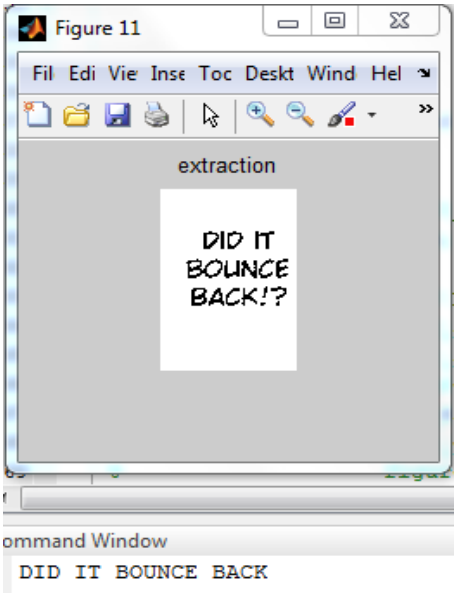


Fig. 11. The screenshot of application

Table 1: Result of Detection Text

	Accuracy (%)
Frame Extraction	90.54
Balloon Extraction	77.92
Extraction Text	98.60
Line Numbering	98.81

our data have a noise, and give a bad effect to our result of frame and balloon extraction. In text extraction and line numbering, we have a good result where extraction text 98,60% and line numbering 98.91%.

We also test the result from recognition text. We compare our result with another method about recognition text. The comparison of results show in Table 2. Our method has the best result than another experiment. It was concluded that the text recognition using a rule that based on the features of endpoint and branchpoint have good results and being able to distinguish one character with another character.

5 Conclusion

The urgency of automatic translation for comic reader is to help them read the comic by giving a translated text from English comic text into another language. The automatic translation in the digital comic requires an extraction and recognition text inside a comic. The challenge is how to extract the comic text

Fig. 12. The overlap frame

Table 2: Result of Recognition Text

	Match Value (%)	Match Value with correlation (%)	Bayes (%)	Our Method (%)
Text Recognition	68	85.57	82.67	85.95

inside the balloon and recognize the text. We were used connected component labelling algorithm to detect the frame and blob also to extract the text, and Hierarcy Point Matching Classification to recognize the text. Our study has good enough results which success rate of frame detection is 90.54% while balloon detection is 77.92%, text extraction is 98,60% , line numbering is 98.91% and text recognition 85.95%. We used simple method that based on connected component labelling which modified with a natural human way to solved the problem in comic. The results show that our method are good enough. In the text extraction, we just do compare between filled area image of balloon and binary image of balloon. This method success to extract the text inside balloon even though the binary image of balloon still have scattered background. The urgency about how the comic was read, we just sort the coordinate of boundingbox each blob text line based on y coordinate first and then continue with x coordinate by ascending. There are many additional attributes from web owner on digital comic that cause our method detect the blob area which not a frame and balloon. The solution if we want good result of both process, frame and balloon detection, the comic data must be clear from unimportant attributes.

The text recognition result show that our method has better accuracy than another method. The branchpoint and endpoint are enough to distinguish the difference character or letter. Most of the false classified caused by two or more letter that connected, so the method detect it as one letter. This result will improve if the method combine with the method for solved the occlusion problem.

For next our research, we will propose the translation text in comic. So by getting information about text in comic, we can use the result of extraction to enable an automated translation from English comic into Indonesian.

References

[1] Hidayatullah, P., Syakrani, N., Suhartini,I., Muhlis,W. (2012, November). Optical Character Recognition Improvement for License Plate Recognition in Indonesia. In *UKSim-AMSS 6th European Modelling Symposium* (pp. 249-254). IEEE.

[2] Arai, K. and Tolle, H. (2010). Method for automatic e-comic scene frame extraction for reading comic on mobile devices. In *2010 Seventh International Conference on Information Technology: New Generations* (pp. 370–375).

[3] Ranjini, S. and Sundaresan, M. (2013). Extraction and recognition of text from digital English comic image using median filter. *International Journal on Computer Science and Engineering (IJCSE)* ,5, 238–244.

[4] Rigaud, C., Tsopze, N., Burie, J.-C. and Ogier, J.-M. (2013). Robust frame and text extraction from comic books, Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 7423: 129138.

[5] Soille, P. (2004). Morphological Image Analysis: Principles and Applications, Vol. 2, Springer-Verlag, New York.

[6] Ngo ho, A. K., Burie, J. C. and Ogier, J. M. (2012). Panel and speech balloon extraction from comic books. In *IAPR International Workshop on Document Analysis Systems* (pp. 424–428).

[7] Su, C. Y., Shoji, Chang, R. I. and Liu, J. C. (2011). Recognizing text elements for svg comic compression and its novel applications. In *International Conference on Document Analysis and Recognition* (pp. 1330–1333).

[8] Tanaka, T., Shoji, K., Toyama, F. and Miyamichi, J. (2007). Layout analysis of tree-structured sceneframes in comic images. *IJCAI* , 2885–2890.

[9] Utaminingrum, F., Uchimura, K. and Koutaki, G. (2013). High density impulse noise removal based on linear mean-median filter. In *The 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision* (pp. 11–17).

[10] Arai, K. and Tolle, H. (2011). Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)* ,4, 669–676.