Proceedings of the Fifth IIEEJ International Workshop
on Image Electronics and Visual Computing 2017
Da Nang, Vietnam, February 28 - March 3, 2017

RECOGNITION OF PANEL STRUCTURE IN COMIC IMAGES USING FASTER R-CNN

Hideaki Yanagisawa† Hiroshi Watanabe†
†Graduate School of Fundamental Science and Engineering, Waseda University

ABSTRACT

For efficient e-comics creation, automatic extracting technique for comic components such as panel layout, speech balloon, and characters is necessary. In the conventional methods, comic components are extracted using geometrical characteristics such as line drawings or connected pixels. However, it is difficult to extract all comic components by focusing on a particular geometric feature, since the components are drawn in various expressions. In this paper, we extract comic components using Faster R-CNN regardless of various comic expressions, and recognize panel structure. Experimental results show proposed method succeed to recognize 67.5% of panel structures on average.

1. INTRODUCTION

Current state of publishing industry has been shifting from the traditional paper based version to e-books. In the e-book market in Japan, e-comic dominates 80% of sales amount [1]. In order to improve convenience of e-comics, services using metadata of e-comics have been proposed. Such services are, e.g. comic search system using particular scene or dialogue in comics, or automatic digest generation system. However, most of e-comics are converted from scanned paper comics. Therefore, it is necessary to manually extract comic structure components such as panel layout, speech balloon, characters (in this paper, we use the word 'character' as actors in comics) and so on. To reduce a cost of metadata extraction, a technique which extracts comic components automatically is important. In this paper, we evaluate a system, which automatically obtains the number of speech balloons and characters in panels using Faster R-CNN from comics.

2. RERATED WORK

For detecting panel layout, Ishii et al. [2] proposed to identify panels by detecting dividing line using gradient concentration. Nonaka et al. [3] introduced panel layout recognition method by detecting lines and rectangles according to a characteristic that panels are often represented as rectangles. Next, for speech balloon extraction, Tanaka et al. [4] proposed a method that identify text areas using Ada-Boost and detect white areas in speech balloons. Moreover, in a study for structure recognition of comics, Arai et al. [5] proposed a detection method for panel, speech balloon and text area. That based on the image blob detection and extracting function using modified connected component labeling (CCL) method. For character detection, Ishii et al. [6] proposed an approach using machine learning with HOG features to detect character face areas. We applied Fast R-CNN in character face detection. [7] From its result, Fast R-CNN showed higher detection rate than HOG features.

Conventional methods extract comic components according to the geometric characteristics, e.g. line detection or extracting connected pixels. However, in some of comic images, panels and speech balloons are illustrated in special expressions. Therefore, it is difficult to detect such components as drawn in specific shapes or overlapped other objects.

3. FASTER R-CNN

Garshick et al. [8] proposed Regions with Convolutional Neural Network features (R-CNN) as a general object detection method using convolutional neural network (CNN). R-CNN detects objects in following process. First, objects' region proposals are extracted from input image by selective search [9]. Second, the region proposals are input to CNN and image feature values are calculated. Then, the output feature values are classified by support vector machine (SVM). Finally, the deviation of region proposals is corrected by bounding box regression. However, R-CNN is slow since it calculates convolutional network features for each object proposal. In order to improve this problem, Fast R-CNN is introduced. Fast R-CNN enables end-to-end detector training on shared convolutional features. Therefore, it shows compelling accuracy and speed [10].

Ren et al. [11] proposed Faster R-CNN as a faster improved object detection technique. Faster R-CNN is single network connected Fast R-CNN and Region Proposal Network (RPN) that share full-image convolutional features with the detection network. RPN is fully convolutional network that simultaneously predict object bounds and object scores at each position. In addition, RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. Therefore, Faster R-CNN can detect object more quickly and shows higher detection accuracy than state-of-the-art methods.

4. PROPOSED METHOD

We propose a method of panel structure recognition from comic images by detection of panels, speech balloons and character faces. We create annotations of comic images

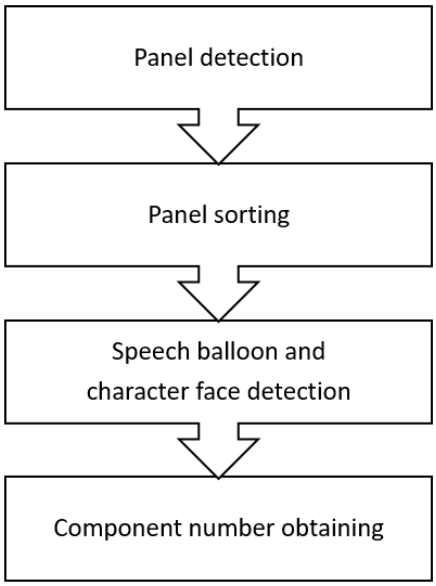


Fig.1 Flow diagram of panel structure recognition

by specifying peripheral regions of each component in rectangles, and 3 types of detectors are generated by training of Faster R-CNN. The flow diagram of panel structure obtaining is shown in **Fig.1**. First, panels are detected from an input image and sorted them. The sorting order is based on the height of detected areas. In addition, if the heights of areas are same, they are sorted from right side one. **Figure 2** shows example images of panel location and sorting orders. Then, there is a slight shift in the position of each panel detected by Faster R-CNN. Therefore, they are normalized per 50 pixels in y-axis direction. Next, speech balloon and character face are detected. They are belonged to the panel that overlapping more than 50% over the detected area. If there is a component which overlaps 50% or more on multiple panels as seen in **Fig.3**, the component is belonged to the panel sorted back side. Finally, the numbers of speech balloons and character faces that belong to each panel are obtained.

5. EXPERIMENT

In this section, we evaluate the detection accuracy of comic components using Faster R-CNN. Also, the recognition accuracy of panel structures is evaluated. In this experiment, we use an algorithm published in https://github.com/rbgirshick/py-faster-rcnn [11] for training and evaluation of Faster R-CNN, and set vgg_cnn_m_1024 [12] as architecture of CNN for training. Datasets for training and evaluation are made of comic images provided in Manga 109 database (http://www.manga109.org/) [13]. The training dataset consists of each 100 images in 20 titles of comics drawn by different authors. The test dataset consists of each 30 images in 5 titles of comic named as *Comic A*-to-*E* drawn by different authors from training images.

© Atsushi Sasaki

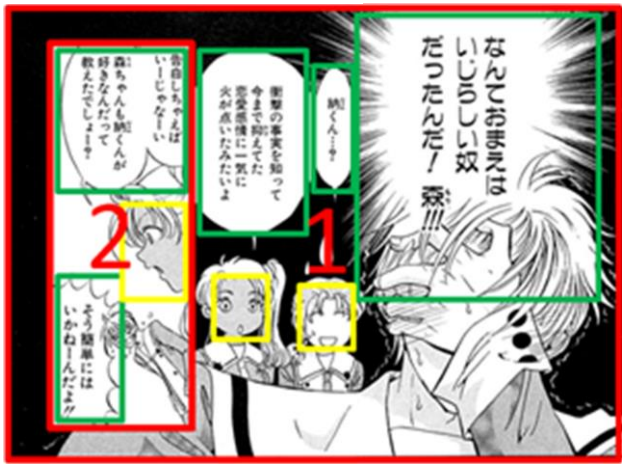
(a)

© Hishika Minamisawa

(b)

Fig.2 Examples of panel sorting

© Hishika Minamisawa



- Panel 1 has 2 characters and 3 balloons
- Panel 2 has 1 character and 2 balloons

Fig.3 Example of panel structure recognition

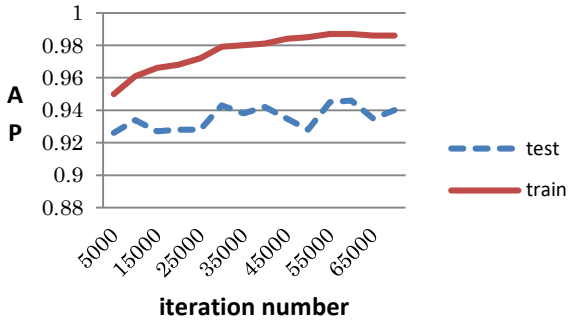
In this experiment, we define a true positive as the detected area overlapping the correct area more than 50%.

5.1. Iteration number

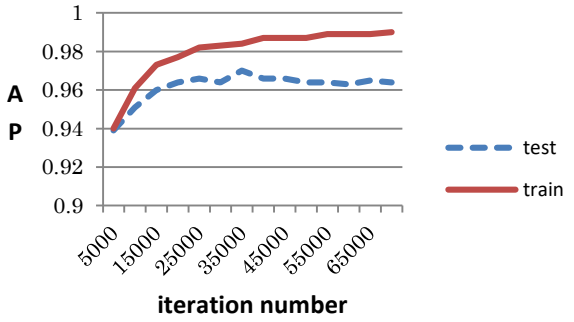
We verified relationship between iteration number in the training process of Faster R-CNN and average precisions (AP) for each comic component. AP means the average values of precisions at each level of recalls. In this experiment, AP is calculated for 2000 images in the training dataset and 150 images in the test dataset. Experimental results are shown in **Fig.4**. In this figure, x-axis indicates iteration number and y-axis indicates AP. From this result, the detection rates are increased by increasing of iteration number. In addition, when the iteration number is over 70000, the AP for training images is converged.

5.2. Threshold of confidence

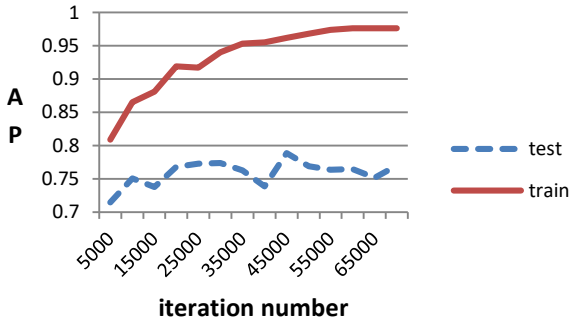
We evaluate the detailed results of comic component detection for 150 images in test dataset using the detectors trained with 70000 iterations. Faster R-CNN calculates a confidence of object in the region proposals, and detects a region when its confidence is larger than a threshold. In this experiment, the threshold of confidence is set to 0.6 at panel detection, and those are set to 0.8 at speech balloon and character face detection. The thresholds are heuristic values. Experimental results are shown in **Table 1**. In this table, "*Total*" means total numbers of comic components in test images, "*TP*" means true positive, "*FN*" means false negative and "*FP*" means false positive. We also measure parameters of recall (*R*) and precision (*P*). **Table 2** shows the detection results of panels and speech balloons by the method of [5] for same test set.



(a) Panel detection



(b) Speech balloon detection



(c) Character face detection

Fig.4 Relationship elation between average precision and iteration number increasing

Experimental results show that the precision rates of Faster R-CNN are more than 90%, and this method exceeds the conventional method at panel and speech balloon detection. Examples of detection results are shown in **Fig.5**. From this figure, it is shown that blob extraction is hard to separate panels when a panel overlapping another panels. On the other hand, R-CNN can detect panels independently of those layouts.

5.3. Recognition rate of panel construction

We evaluate a recognition accuracy of panel structures for each 30 pages in 5 comics. The recognition accuracy

© Atsushi Sasaki

(a) Examples of panel detection by [5]

(b) Examples of panel detection by Faster R-CNN

Fig.5 Examples of panel detection for flat panels and connected panels

Table 1 Results of comic component extraction for 5 comic sources by Faster R-CNN

	Total	*TP*	*FN*	*FP*	*R* (%)	*P* (%)
Panel	859	770	90	40	89.5	95.1
Balloon	1190	1161	29	42	97.6	96.5
Character	937	803	134	50	85.7	94.1

Table 2 Results of comic component extraction for 5 comic sources by [5]

	Total	*TP*	*FN*	*FP*	*R* (%)	*P* (%)
Panel	859	481	378	183	56.0	72.4
Balloon	1190	790	400	650	66.4	54.9

Table 3 Results of panel structure recognition for 5 comic sources

	B (%)	*C* (%)	*B* + *C* (%)
Comic A	83.0	74.5	68.1
Comic B	91.4	89.8	84.9
Comic C	81.7	72.8	66.3
Comic D	94.6	69.0	65.2
Comic E	62.3	62.9	52.8

is defined as follows: "*B*" means the panels which speech balloon numbers correctly extracted, "*C*" means the panels which character face numbers correctly extracted and "*B* + *C*" means the panels which both numbers of speech balloon and character face correctly extracted. An experimental result is shown in **Table 3**. From this result, the highest value of *B* + *C* is 84.9% in comic B and the lowest value is 52.8% in comic E.

An example case of failure to panel structure recognition is the detection failure caused by deformed faces as shown in **Fig.6**. In addition, the reason of low recognition rate in *Comic E* is that it contains fuzzy panel layout as shown in **Fig.7**. In Fig.6 and Fig.7, red rectangle shows the detected area as comic component.

6. CONCLUSION & FUTURE WORK

In this paper, we evaluated panel structure recognition using Faster R-CNN. Experimental results show our proposed method success to recognizing 67.5% of panel structures on average.

For future works, there are some possible improvements in detection for panels and character faces those are hard to detected in this method. As a specific technique, it is considerable to combine image processing such as highlighting division lines of panels with Faster R-CNN detection. In addition, for obtaining metadata to be used for automatic generation of comic summaries, we need to consider a technique for classifying main characters from detected character faces.

7. REFERENCES

[1] Internet Media Research Institute: "eComic Marketing Report 2012", Impress R&D, pp.14 (2012).

[2] D. Ishii, K. Kawamura, H. Watanabe: "A Study on Frame Decomposition of Comic Images", IEICE Transactions, Vol. J90-D, No.7, pp. 1667−1670 (2007).

[3] S. Nonaka, T. Sawano, N. Haneda: "Development of "GT-Scan", the Technology for Automatic Detection of Frames in Scanned Comic", FUJIFILM RESEARCH & DEVELOPMENT, No.57, pp.46−49 (2012).

[4] T. Tanaka, F. Toyama, J. Miyamichi, K. Shoji: "Detection and Classification of Speech Balloons in Comic Images", Journal of the Institute of Image Information and Television Engineers, Vol.64, No.12, pp.1933−1939 (2010).

© Satoshi Arai

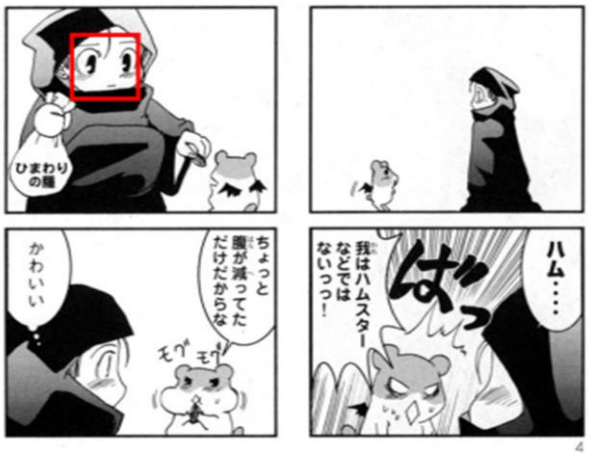


Fig.6 Example of failure to detect character faces

[5] Arai K, Tolle Herman: "Method for Real Time Text Extraction from Digital Manga Comic", International Journal of Image Processing Vol.4, No.6, pp.669−676 (2011).

[6] D. Ishii, H. Watanabe: "A Study on Automatic Character Detection and Recognition from Comics", The Journal of the Institute of Image Electronics Engineers of Japan, Vol.42, No.4 (2013)

[7] H. Yanagisawa, H. Watanabe: "A study of Multi-view Face Detection for Characters in Comic Images", Proceedings of the 2016 IEICE General Conference, D−12−12 (2016).

[8] R. Girshick, J. Donahue, T. Darrell, J. Malik: "Rich feature hierarches for accurate object detection and semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, (2014).

[9] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders: "Selective Search for Object Recognition", International Journal of Computer Vision, Vol.102, No.2 pp.154−171, (2013).

[10] R. Girshick: "Fast R-CNN", arXiv:1504.08083, (2015).

[11] S. Ren, K. He, R. Girshick, J. Sun: "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Advances in Neural Information Processing Systems (NIPS), (2015).

[12] S. Farfade, M. Saberian: "Multi-view Face Detection Using Deep Convolutional Neural Networks", arXiv:1502.02766, (2015).

[13] Y.Matsui, K.Ito, Y. Aramaki, T.Yamasaki, K. Aizawa: "Sketch-based Manga Retrieval using Manga109 Dataset", arXiv:1510.04389, (2015).

© Saya Miyauchi

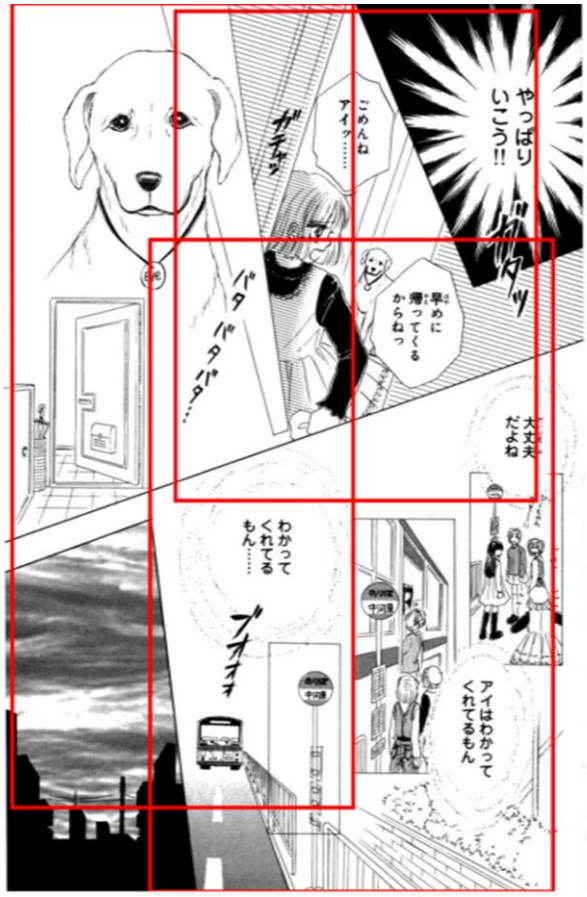


Fig.7 Example of failure to detect panels in *Comic E*