A Study on Object Detection Method from Manga Images using CNN

Hideaki Yanagisawa, Takuro Yamashita, Hiroshi Watanabe
Waseda University
Graduate School of Fundamental Science and Engineering
Tokyo, Japan
bule-cosmo@ruri.waseda.jp

***Abstract*—Japanese comics (manga) are popular content worldwide. In order to acquire metadata from manga images, techniques automatic recognition of manga content have been studied. Recently, Convolutional Neural Network (CNN) has been applied to object detection in manga images. R-CNN and Fast R-CNN generate region proposals by Selective Search. Faster R-CNN generates them using CNN layers called Region Proposal Network (RPN). Single Shot MultiBox Detector (SSD), the latest detection method, performs object classification and box adjustment for small regions in an image. These methods are effective to natural images. However, it is unclear whether such methods work properly to manga images or not, since those image features are different from natural images. In this paper, we examine the effectiveness of manga object detection by comparing Fast R-CNN, Faster R-CNN, and SSD. Here, manga objects are panel layout, speech balloon, character face, and text. Experimental results show that Fast R-CNN is effective for panel layout and speech balloon, whereas Faster R-CNN is effective for character face and text.**

Keywords—Object Detection; Manga; CNN; Fast R-CNN; Faster R-CNN; SSD

I. INTRODUCTION

Electronic comic is an important content which accounts for about 80% of sales in e-book market in Japan. Many of existing manga images are obtained by scanning paper mediums. By recognizing content in such manga images, it is possible to obtain useful metadata for services such as searching and image processing. Since handwritten manga objects have big change in shape than general objects, it is necessary to construct a dedicated detection method. Attempts to detect objects from manga images have been studied. Arai et al. proposed panel layout and speech balloon extraction using blob detection [1]. Ho et al. proposed to extract them based on region merging and mathematical morphology [2].

Particularly in recent years, detection methods using Convolutional Neural Network (CNN) have shown high accuracy for manga object detection. Chu et al. proposed a detection method for character faces using Selective Search and CNN [3]. Iyyer et al. detected panel layouts by Faster R-CNN, which was trained using manga images and created annotation data [4]. Ogawa et al. applied Single Shot MultiBox Detector (SSD) to manga object detection. And they proposed an anchor allocation for collective detection of manga objects [5]. Methods [3]-[5] are differ in region proposals for CNN feature computation. In general object detection, SSD shows the best accuracy. However, manga objects have different characteristics. Thus, it is not clear whether similar results can be obtained or not. In this paper, we verify an effectiveness of region proposals in manga images by comparing different object detection methods.

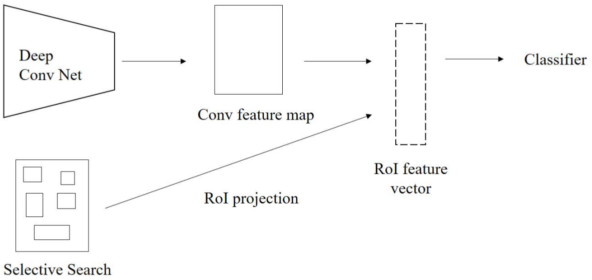


Fig. 1. Outline of object detection by Fast R-CNN

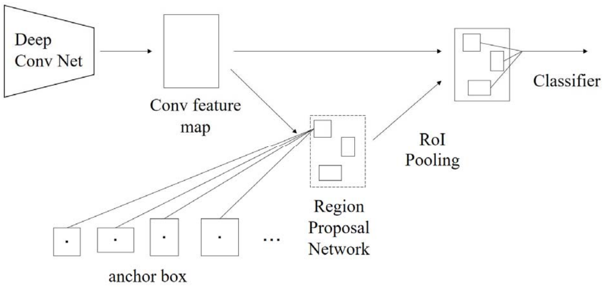


Fig. 2. Outline of object detection by Faster R-CNN

II. OBJECT DETECTION FROM MANGA IMAGES

3.1. R-CNN, Fast R-CNN, Faster R-CNN

A. R-CNN

Girshick et al. proposed Regions of CNN features (R-CNN) that is an object detection method using CNN features [6]. The object detection procedure of R-CNN is as follows. First, input images are segmented by Selective Search [7] and region proposals are generated. Next, region proposals are

978-1-5386-2615-3/18/$31.00 ©2018 IEEE.

normalized, and input to CNN. Then, image features that output from CNN are classified. Finally, the region proposals are determined whether include target objects or not.

B. Fast R-CNN

One of the problems of R-CNN is increasing of processing time, since generating of CNN features are processed for all region proposals. To solve this problem, Fast R-CNN [8] is proposed. The outline of Fast R-CNN is shown in **Fig. 1**. First, image feature map is generated from all over input image. Next, region proposals of Selective Search are projected on the feature map by RoI Pooling. Finally, feature vectors extracted from region proposals are classified. This method succeeded to detect object faster by reducing feature generating.

C. Faster R-CNN

Although Fast R-CNN reduced CNN processing, the problem remains that it takes time for Selective Search processing. Ren et al. proposed Faster R-CNN [9] as a further improved method. Faster R-CNN uses CNN layers named Region Proposal Network (RPN) instead of Selective Search. The outline of Faster R-CNN is shown in **Fig. 2**. Faster R-CNN generates region proposals from feature maps generated by RPN. RPN scans the sliding window on the feature map and extracts object candidates. At this time, in order to detect long slender objects, each grid have some bounding boxes called "anchor boxes". Finally, like the Fast R-CNN, the region proposals are projected onto the feature map, and the object is detected by classifying region proposals. Since Faster R-CNN generates image features and region proposals using single CNN, there is an advantage that end-to-end training can be performed in addition to faster detection.

3.2. Single Shot MultiBox Detector

Faster R-CNN realized detection processing by single CNN. However, the network configuration is still complicated, because the process of generating image features and region proposals are separated. For this reason, processing time is insufficient for real time detection. To make the network configuration simpler, Wei et al. proposed Single Shot Multi Box Detector (SSD) [10]. The outline of SSD is shown in **Fig. 3**. SSD divides the input image into predefined small areas called "grids" and applies anchor boxes for each grid. Anchor boxes are trained to respond the objects close in size and aspect ratio. Finally, classification and rectangle estimation are performed for object proposals. In general object detection, SSD shows faster detection time than conventional methods and same detection accuracy as Faster R-CNN.

III. Experiment

In this section, we examine the change of detection rate for comic images by region proposals. In this experiment, we compare the detection results of detectors trained using same dataset for Fast R-CNN, faster R-CNN and SSD.

3.1. Dataset

We use images in Manga109 dataset [11] for training and evaluation of detectors. As a training dataset, we randomly select 19 titles from the 109 titles. From each title, we select 100 pages and manually define grand truth about 4 classes of objects (panel, speech balloon, character face, and text). The example image with annotation is shown in **Fig. 4**. In this paper, "character face" is defined as face areas of persons appearing in manga, and "text" is defined as characters included in speech balloons and onomatopoeias existing in panels. For evaluation, we select 5 titles whose authors are different from training manga images, and from each title we select 30 pages for grand truth.

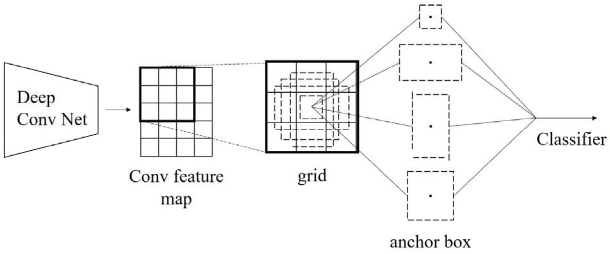
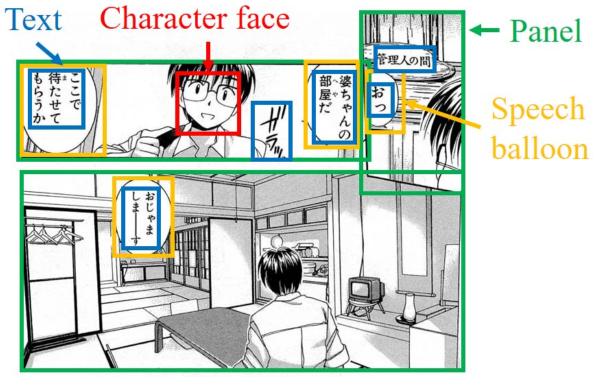


Fig .3. Outline of object detection by SSD



©Ken Akamatu

Fig .4. Annotation for manga objects used in this experiment

TABLE I. Detection results of Fast R-CNN, Faster R-CNN and SSD (AP)

	Fast R-CNN	*Faster R-CNN*	*SSD*
Panel layout	0.959	0.953	0.897
Speech balloon	0.969	0.961	0.907
Character face	0.810	0.816	0.765
Text	0.740	0.898	0.866
mAP	0.870	0.910	0.859

3.2. Training Parameter

In this section, we mention about the parameter of Fast R-CNN, Faster R-CNN, and SSD for training. As reference [5] stated, manga objects are located densely unlike general objects. Thus, the problem of class allocation for region proposals is occurred when detecting multiple object classes. To avoid allocation problem, we train 4 detectors corresponding to each class. We use VGG-16 model [12] for the CNN architecture, and use pre-trained ImageNet [13] models for the initial weights. Training iteration is set to 70000.

3.3. Evaluation

TABLE I shows the average precision (AP) and mean values of AP (mAP). In those results, threshold value of IoU is set to 0.5, which parameter commonly used in object detection. The examples of detection results are shown in **Fig. 5**. Fast R-CNN shows the maximum detection rate for panel and speech balloon. On the other hand, Faster R-CNN shows the maximum detection rate for character face and text.

From this result, Selective Search in Fast R-CNN is effective for extracting the objects with clear boundaries such as panel and speech balloon, since it extracts regions by image segmentation. In addition, for character face and text, whose boundaries are ambiguous, RPN in Faster R-CNN is effective.

As shown in image **Fig. 5 (f)**, SSD cannot detect some objects. Thus, its mAP is lower than other methods. The reason is that SSD detects objects by dividing the input image into grid. Therefore, the objects with a small ratio to the whole image cannot be detected. As the way of solving this problem, we propose that dividing an image into panels and detecting objects from each panel. **Fig. 6** shows the detection results of SSD from images cut out for each panel. From this result, it is expected that detection accuracy of SSD is improved by dividing image into panels.

IV. Conclusion

In this paper, we examined the effectiveness of object proposals for manga object detection. By the experimental results, it is shown that Selective Search is effective for objects with clear boundaries, and RPN is effective for objects whose boundaries are ambiguous. In addition, SSD is difficult to detect manga objects in whole image, so process of dividing image into small regions is necessary.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 17K00511.

References

[1] K. Arai, T. Herman, “Method for Real Time Text Extraction from Digital Manga Comic,” International Journal of Image Processing Vol 4, No. 6, pp. 669-676, 2011.

[2] A. K. N. Ho, J.C. Burie, J.M. Ogier, “Panel and speech balloon extraction from comic books,” Document Analysis Systems (DAS), 2012 10th IAPR International Workshop, 2012.

[3] W.T. Chu, W.W. Li, “Manga Face Net: Face Detection in Manga based on Deep Neural Network,” Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 412-415, June 2017.

[4] M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. A. Daumé III, L. Davis, “The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives,” in Confernce on Computer Vision and Pattern Recognition, IEEE, 2017.

©Atushi Sasaki

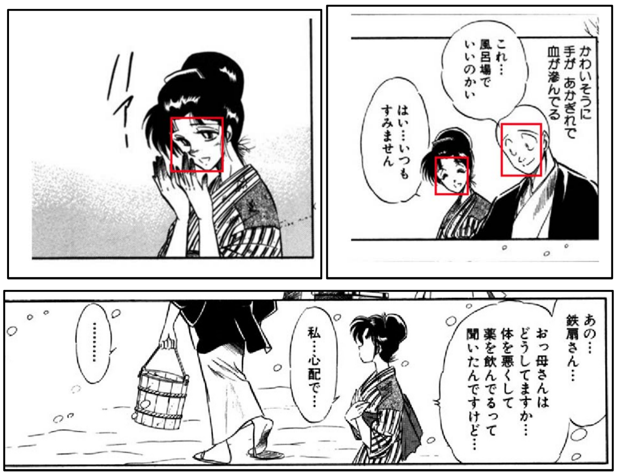

かわいそうに 手が あかぎれて 血が滲んでる
これ… 風呂場で いいのかい
はい…いつも すみません
あの… 鉄扇さん… おっ母さんは どうしてますか… 体を悪くして 薬を飲んでるって 聞いたんですけど…
私…心配で…
……

Fig .6. Detection results of SSD from images cut out for each panel

[5] T. Ogawa, T. Yamasaki, K. Aizawa, “Parallel Detectors for Manga Objects,” in Conference on Forum on Information Technology 2017, CH-007, Sep 2017.

[6] R. Girshick, J. Donahue, T. Darrel, J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Conference on Computer Vision and Pattern Recognition, pp. 580-587, IEEE, 2014.

[7] J. R.R. Uijlings, K. E.A. V. D. Sande, T. Gevers, A. W.M. Smeulders, “Selective search for object recognition,” International Journal of Computer Vision, Vol. 104, No. 2, pp. 154-171, 2013.

[8] R. Girshick, “Fast r-cnn,” in Inernational Conference on Computer Vision, pp. 1440-1448, IEEE, 2015.

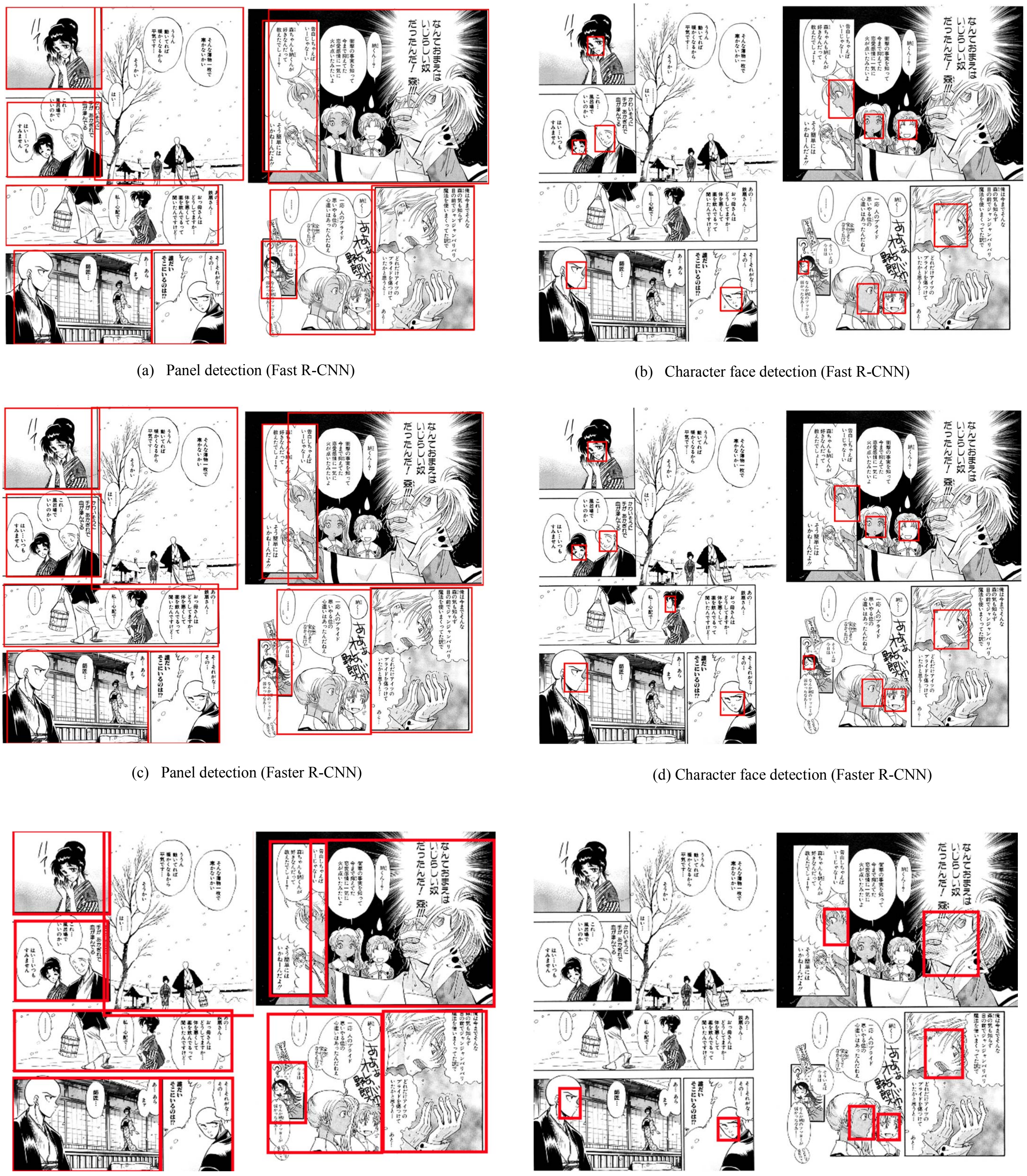
[9] S. Ren, K. He, R. Girshick, J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks” in Advances in Neural Information Processing Systems, pp. 91-99, 2015.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A. C. Berg, “Ssd: Single shot multibox detector,” in European Conference on Computer Vision, pp. 21-37, Springer, 2016.

[11] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yaasaki, K. Aizawa, “Sketch-based manga retrieval using manga109 dataset,” Multimedia Tools and Applications, pp. 1-28, 2016.

[12] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” International Conference on Learning Representations, 2015.

[13] J. Deng, W. Dong, R. Socher, L.J Li, K. Li, L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vison and Pattern Recognition 2009 (CVPR2009) IEEE Conference on, pp. 248-255, IEEE, 2009.


そんな薄物一枚で寒かないかい
ううん 動いてれば暖かくなるから平気です…
そうかい
はい……
かわいそうに手が あかぎれで血が滲んでる
これ…風呂場でいいのかい
はい…いつもすみません
あの…鉄眉さん…
おっ母さんはどうしてますか…体を悪くして薬を飲んでるって聞いたんですけど…
私…心配で…
……
そ…それがな…その…
誰だい そこにいるのは!?
あ…あら まァ
師匠…
なんでおまえはいじらしい奴だったんだ！ 森!!!
納くん…?
衝撃の事実を知って今まで抑えてた恋愛感情に一気に火が点いたみたいよ
告白しちゃえばいーじゃなーい
森ちゃんも納くんが好きなんだって教えたでしょー?
そう簡単にはいかねーんだよ!!
俺は今までそんな森の気も知らず目の前でジャンジャンバリバリ魔法を使いまくってた訳で
どれだけアイツのプライドを傷つけていたかと思うと…
あと…
ああ オレのバカ野郎
納く…
一応 人のプライド思いやる位の心遣いはあったんだねえ
全然できてなかったけど
……
その頃の森
そういえば今日は
なんか納のツッコミが弱かったなあー…

(a) Panel detection (Fast R-CNN)

(b) Character face detection (Fast R-CNN)

(c) Panel detection (Faster R-CNN)

(d) Character face detection (Faster R-CNN)

(e) Panel detection (SSD)

(f) Character face detection (SSD)

Fig .5. Examples of object detection from manga images.