

# TEXT DETECTION IN MANGA BY COMBINING CONNECTED-COMPONENT-BASED AND REGION-BASED CLASSIFICATIONS

Yuji Aramaki, Yusuke Matsui, Toshihiko Yamasaki, Kiyoharu Aizawa

The University of Tokyo

## ABSTRACT

As manga (Japanese comics) have become common content in many countries, it is necessary to search manga by text query or translate them automatically. For these applications, we must first extract texts from manga. In this paper, we develop a method to detect text regions in manga. Taking motivation from methods used in scene text detection, we propose an approach using classifiers for both connected components and regions. We have also developed a text region dataset of manga, which enables learning and detailed evaluations of methods used to detect text regions. Experiments using the dataset showed that our text detection method performs more effectively than existing methods.

**Index Terms**— Manga, text regions, detection, connected components, deep features

## 1. INTRODUCTION

Manga are Japanese comics, which are read by many people all over the world. The globalization of manga makes it necessary to translate manga automatically. Moreover, it has become more usual to publish manga in electronic form. However, methods for searching manga are limited to their meta-data such as titles or the authors. It is more convenient and interesting to explore manga by text including the speeches of manga characters. This kind of information also leads to applications to understand manga context or systematize manga automatically, such as by layout generation [1], text ordering [2], and content-based retrieval [3]. All of these applications require identification of text or text regions in manga.

There are optical character recognition (OCR) systems for text recognition for general documents. They can analyze layouts and recognize the text in documents. However, a page of manga cannot be recognized directly by OCR. This is because the text and illustrations are combined in manga pages in complicated ways and drawn in abstracted forms. It is, therefore, very important to detect text regions in manga. Since OCR can recognize the identified text regions, we focus on detecting text regions.

We thank the authors of manga for the use of their images. This work was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE).

Several methods for detecting text regions in manga or comics have been proposed. Speech balloon detection (SBD) [4, 5] is based on connected components and filters the components using their inner structures. Text line detection (TLD) [6] also employs connected component labeling, and groups the components by their positions and alignment. However, these methods are heavily dependent on some heuristic assumptions, and cannot find various kinds of texts in manga such as onomatopoeias. In addition, the quantitative evaluations in those studies were not sufficient. For example, only 15 pages from just one title were used in the evaluation of SBD [5], and TLD [6] was evaluated with eBDtheque [7], a database of comics that includes only six images of manga.

Scene text detection is currently an active field of research. Many methods to detect text in scene images have been proposed [8, 9, 10]. They handle the diversity of text forms using approaches based on machine learning, and have achieved high performance of detection. However, those methods are not applicable to manga images, because most of these methods use color distributions or gradients, which do not exist in manga images.

Our contributions are two-folds. First, we have developed a method for detecting texts in manga images. Our detection method combines classifiers for connected components and for regions. We have also constructed a dataset of manga text regions, that includes various manga titles and includes not only the positions but also the types of each text region. With this dataset, we can provide training samples without noise for learning and adequately evaluate our detection method and previous approaches.

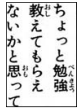



## 2. TEXT REGION DATASET

For large-scale learning and evaluation, we created a dataset of text regions in manga<sup>1</sup>. The dataset includes manga images with the positions, sizes, and categories of the text regions. The categories are essential for the training without noise such as background objects, and for detailed evaluation of how and which kinds of regions are correctly detected.

Our dataset is based on a subset of Manga109<sup>2</sup> [11], a

<sup>1</sup>The dataset is available at <https://www.hal.t.u-tokyo.ac.jp/~aramaki/>.

<sup>2</sup><http://www.manga109.org/>

	Clean (Text only)	Dirty (With other objects)
Typical font	TC 	TD 
Atypical font	AC 	AD 

**Fig. 1:** Definition of region categories. The four categories are defined by the font of the texts and the presence of background scenes. ©Ken Akamatsu

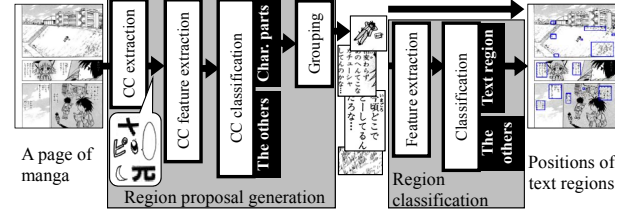
large dataset of manga that was released by our group. The 109 titles in this dataset were drawn by professional artists and there is great diversity of drawing styles. To construct the text region dataset, six titles in Manga109 were selected: DollGun (DG), Aosugiru Haru (AH), Lovehina (LH), Arisa 2 (A2), Bakuretsu KungFu Girl (BK), and Uchuka Katsuki Eva Lady (UK). In choosing these titles, we ensured a variety of drawing styles.

For each page of these manga titles, we annotated the locations and types of text regions. The locations are defined with the position of the upper left corner, and their width and height. For the types, we defined four categories of text regions. The categories are defined by a pair of criteria: the font of the texts and the presence of background scenes. For the categories, we assigned the names: TC, TD, AC, and AD. In Figure 1, we show an example for each category.

### 3. PROPOSED DETECTION METHOD

To detect texts in manga images, we developed a method that consists of two steps: *region proposal generation* and *region classification*. The region proposal generation, which is based on the classification of connected components, enables fast detection of various text regions. Our region classification makes use of deep learning, which is widely known to show high discrimination ability, and makes it possible to increase the precision. By connecting the two approaches, we can use features that cannot be captured by either one of them. The region classification is optional because deep learning has high costs for its computation.

A flowchart of our method is shown in Figure 2. In region proposal generation, we extract the features of connected components, classify them according to their features, and then group them into rectangular regions. After the region proposals have been generated, we compute the features of the regions using deep learning models, and classify them according to their deep features.



**Fig. 2:** Flowchart of our method. Our region proposal generation, based on connected components, is followed by region classification based on deep features of the regions. The regions produced in the first step can also be treated as the final output. ©Ken Yagami

#### 3.1. Region proposal generation

From a page of manga, we extract all the connected components. For the following classification, we compute the geometric features of each connected component, including area, perimeter, the side lengths of the outer rectangle, the area of the convex hull, the lengths of the axes of and the slope of the approximate ellipse, and the Euler characteristic.

Because each character or part of the character in manga is drawn as an independent connected component, we must classify each of the connected components as either part of a character or part of a noncharacter object. We use the Random Forest [12] with the geometric features of the connected components in our text region dataset. As the positive samples for training, we use only the connected components in the text regions in the TC and AC categories because the regions in the other categories contain connected components of noncharacter objects. For the negative samples, we choose only connected components outside all of the regions in any category. The selection of the samples is enabled by our region categories, and is crucial to avoid training using noise samples.

The connected components classified as part of a character are integrated into text regions. We propose two methods for grouping: *basic grouping* (BG) and *exhaustive grouping* (EG). In BG, we group the connected components having centers the distances of which are smaller than  $L_1$ , and obtain a region as the rectangle circumscribing the grouped connected components. We then integrate the regions having shortest distances that are smaller than  $L_2$ . On the other hand, EG is designed for higher comprehensiveness when region proposal generation is followed by region classification. EG includes as many regions that may be text regions as possible. In EG, we vary  $L_1$  for various scales, and in the second step keep the ungrouped regions as the proposals. In this paper,  $L_1$  was set to 40 in BG and was varied from 10 to 100 by 10 in EG, and  $L_2$  was set to 5.

We regard the grouped regions as the final results that can be forwarded to the region classification as region proposals.

### 3.2. Region classification

For the region proposals generated in the previous step, we apply classification per region rather than per connected component. Using this classification, we can remove falsely detected regions. As region features, we employ deep features, that are extracted using models constructed using deep learning. The Support Vector Machine (SVM) [13] approach is used to classify the deep features because SVM performed better than Random Forest in a preliminary experiment.

We adopt two models for deep feature extraction. The first is the model made for ImageNet classification [14] (ImN). This model has five convolutional and three fully connected (FC) layers, and was trained on large-scale natural images. The deep features are extracted from its sixth layer as a 4096-dimensional vector. The second model was made using Illustration2vec [15] (I2v). I2v is based on VGG models [16] but replaces their FC layers with convolutional layers to be adapted to more detailed parts. Though I2v was originally designed for tag prediction, we use a separate model prepared for feature extraction.

SVM is used for the classification. In the training step, the text regions in any category were used as the positive samples, and the regions sampled outside all of the text regions were used as the negative samples. In the test phase, the trained model classifies the region proposals. The regions classified as text regions are the final detection results.

## 4. EXPERIMENTS

We conducted several experiments to compare the proposed detection methods with one of the scene text detection methods (STD) [8], SBD [5] and TLD [6]. TLD was modified for vertical writing, that is typical in Japanese. The proposed methods are the pairs of either the basic grouping (BG) or exhaustive grouping (EG), and either the proposed region classification with ImN, or with I2v, or no classification. For training, we used 100 pages randomly extracted from our dataset, which contain 644, 176, 66 and 304 text regions in the TC, TD, AC and AD category, respectively. The parameters of the learning algorithms were optimized for each condition.

Our evaluation method [17] is the same as that used in the ICDAR 2013 robust reading competition [18]. For each detected region, a ground-truth region corresponds only if both of the two parameters,  $t_p$  and  $t_r$ , are larger than the ratio of the overlapped area to the area of the detected region and to the area of the ground-truth region, respectively. It is also possible to evaluate without any parameters. The method covers the cases in which more than one detected region corresponds to one or more ground-truth regions, and vice versa.

### 4.1. Experiment (I)

We first evaluate the whole dataset. For test, we randomly selected 100 pages from the pages that were not used for train-

**Table 1:** Parameter-free evaluation results. The proposed methods achieved the higher performance than STD [8], SBD [5] and TLD [6].

Method	Precision	Recall	F-measure
STD [8]	0.165	0.051	0.078
SBD [5]	0.180	0.102	0.130
TLD [6]	0.095	0.095	0.095
BG	0.169	0.496	0.252
BG + ImN	0.451	0.481	<b>0.466</b>
BG + I2v	<b>0.715</b>	0.191	0.301
EG	0.068	<b>0.851</b>	0.126
EG + ImN	0.177	0.806	0.291
EG + I2v	0.557	0.192	0.285

ing. In the test set, there were 670, 160, 61 and 303 text regions in the TC, TD, AC and AD category, respectively.

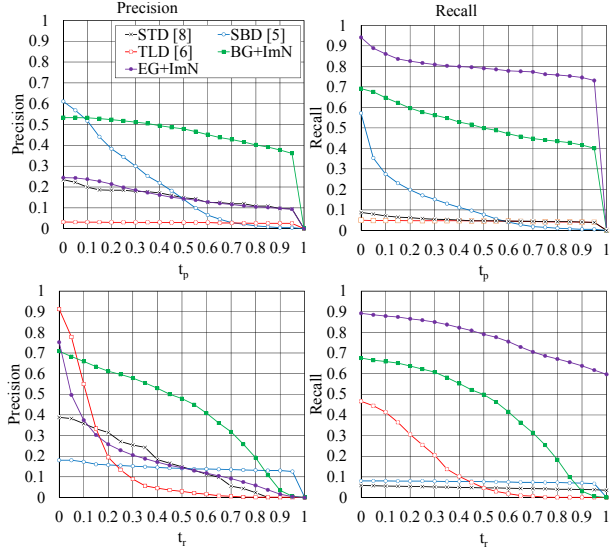
Table 1 shows the results of the parameter-free evaluation. F-measures of the variation except for EG of our methods exceeded those of STD [8], SBD [5] and TLD [6]. Adding region classification with ImN drastically improved the precision while the recall was retained. BG followed by ImN achieved the highest F-measure of all of the methods. Although this type of region classification decreased the recall, the precision was extremely improved by I2v.

We compared two of our methods (BG + ImN and EG + ImN) and the other methods [8, 5, 6] in detail. In the detailed evaluation, either  $t_p$  or  $t_r$  is fixed to 0.5 and the other is varied in the range from 0.0 to 1.0 by 0.05. Figure 3 shows the results of the detailed evaluation. Because SBD [5] tends to produce larger rectangles than the defined text regions, the performance at low  $t_p$  were relatively high. On the other hand, the output rectangles of TLD [6] are often smaller than the ground-truth regions, which makes the performance relatively high at low  $t_r$ . However, even at low  $t_p$  and  $t_r$ , our method (BG + ImN) showed much higher recall than and precision comparable with those of the previous methods.

As shown in Figure 4, our method produced more reasonable regions than the previous methods [5, 6]. See the supplemental materials for more results.

### 4.2. Experiment (II)

We separated the test dataset according to the region categories, and analyzed the differences in the parameter-free evaluation for the subsets in each category. The results for the subsets per category are shown in Figure 5. Our method (BG+ImN) outperformed the other methods for all categories. We found that the recalls by our method for regions in all categories ranked in the descending order of TC, TD, AC, and AD. Note that SBD [5] and TLD [6] could not correctly detect even regions in the TC category, which are the main targets of the two methods.



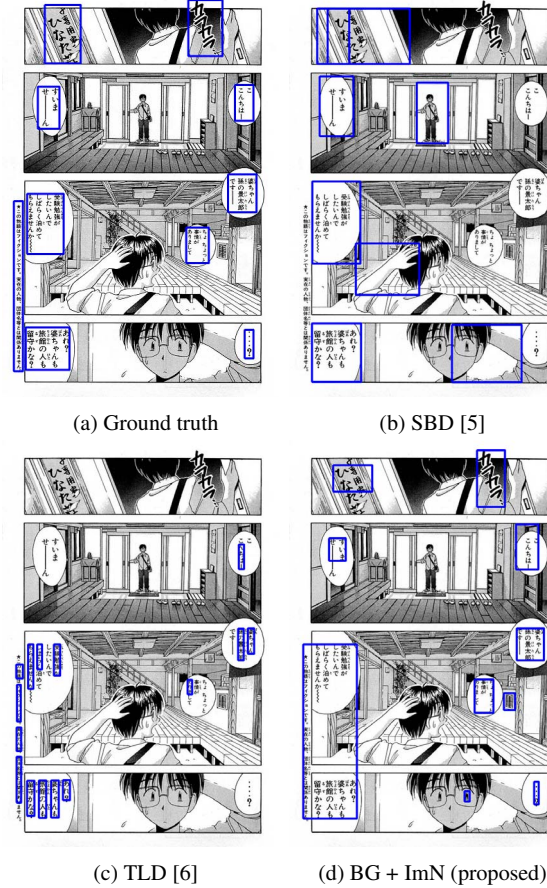
**Fig. 3:** Precision and recall when either  $t_p$  or  $t_r$  is fixed to 0.5 and the other is varied in the range from 0.0 to 1.0 by 0.05. Our method achieved higher performance than the other methods.

The differences in detection performance between titles were also evaluated. We randomly extracted 50 pages from each title, and applied the detection methods to them. The pages were totally different from those for training, and contain 491, 1107, 793, 567, 745 and 717 text regions for DG, AH, LH, A2, BK and UK, respectively. Figure 6 shows the parameter-free evaluation results for the subsets per title. Although there were some differences in performance depending on the titles, the proposed method (BG+ImN) achieved the highest precision and recall for all titles.

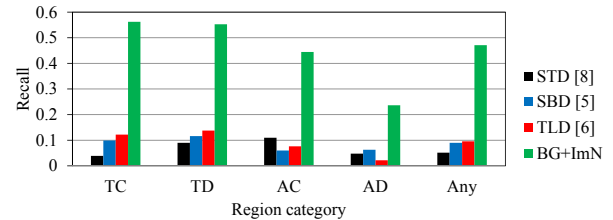
## 5. CONCLUSIONS AND LIMITATIONS

We have developed a method to detect text regions in manga. The classification based on connected components produced region proposals efficiently, and the following classification of the regions using deep features removed false detections effectively. The dataset of manga text regions that we have prepared made possible not only the learning without noise but also the adequate evaluation of the detection methods. The experiments with our dataset showed that our methods achieved recall up to 0.851 and F-measure up to 0.466, which are much higher than those of the previous methods.

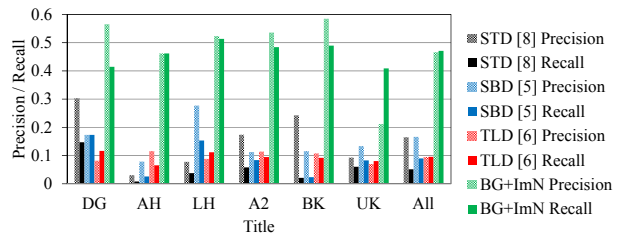
The proposed method has a few limitations. Although the region classifier is effective in removing misdetections, the region classification is much more computing intensive than the region proposal generation. OCR of text regions is beyond this research and requires more investigation.



**Fig. 4:** Detection examples for LH. ©Ken Akamatsu



**Fig. 5:** Detection recall for the subsets per category. Our method outperformed the other methods in all categories.



**Fig. 6:** Results for subsets by title. The proposed method produced relatively stable performance for all titles.

## 6. REFERENCES

- [1] Y. Cao, R. W. Lau, and A. B. Chan, "Look over here: Attention-directing composition of manga elements," *Transactions on Graphics*, vol. 33, no. 4, pp. 94, 2014.
- [2] S. Kovanen and K. Aizawa, "A layered method for determining manga text bubble reading order," in *International Conference on Image Processing*. IEEE, 2015, pp. 4283–4287.
- [3] T.-N. Le, M. M. Luqman, J.-C. Burie, and J.-M. Ogier, "Content-based comic retrieval using multilayer graph representation and frequent graph mining," in *International Conference on Document Analysis and Recognition*. IEEE, 2015, pp. 761–765.
- [4] K. Arai and H. Tolle, "Method for real time text extraction of digital manga comic," *International Journal of Image Processing*, vol. 4, no. 6, pp. 669–676, 2011.
- [5] H. Tolle and K. Arai, "Manga content extraction method for automatic mobile comic content creation," in *International Conference on Advanced Computer Science and Information Systems*. IEEE, 2013, pp. 321–328.
- [6] C. Rigaud, D. Karatzas, J. Van De Weijer, J.-C. Burie, and J.-M. Ogier, "Automatic text localisation in scanned comic books," in *International Conference on Computer Vision Theory and Applications*, 2013.
- [7] C. Guerin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel, "ebdtheque: a representative database of comics," in *International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1145–1149.
- [8] L. Gómez and D. Karatzas, "Multi-script text extraction from natural scenes," in *International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 467–471.
- [9] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-orientation scene text detection with adaptive clustering," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1930–1937, 2015.
- [10] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al., "Icdar 2015 competition on robust reading," in *International Conference on Document Analysis and Recognition*. IEEE, 2015, pp. 1156–1160.
- [11] Y. Matsui, K. Ito, Y. Aramaki, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *arXiv preprint arXiv:1510.04389*, 2015.
- [12] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.
- [13] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [15] M. Saito and Y. Matsui, "Illustration2vec: a semantic vector representation of illustrations," in *SIGGRAPH Asia 2015 Technical Briefs*. ACM, 2015, p. 5.
- [16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations*, 2015.
- [17] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal on Document Analysis and Recognition*, vol. 8, no. 4, pp. 280–296, 2006.
- [18] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez i Bigorda, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazan Almazan, and L.-P. de las Heras, "Icdar 2013 robust reading competition," in *International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1484–1493.