Computational Interpretation of Comic Scenes

Miki Ueno

Abstract Understanding intellectual products such as comics and picture books is one of the important topics in the field of artificial intelligence. Hence, stepwise analysis of a comic story, i.e., features of a part of the image, information features, features relating to continuous scene etc., by human and by a combination of several classifiers was pursued. As the first step in this direction, several classifiers for comics are constructed in this study by utilizing a convolutional neural network, and the results of classification by a human annotator and by a computational method are compared.

Keywords Computational model of comics · Comic engineering · Deep learning

1 Introduction

Image recognition by deep learning has seen rapid development in recent years [1]. Previously, I had focused on modeling the pictures in and analysis features of comics [2]. However, numerous challenging tasks must be addressed in order to understand comics and animation, which are composed of pictures and natural languages. The aforesaid tasks are broadly classified into the following two types.

1. **Recognize deformed pictures.** Classification of deformed pictures is difficult [3], while a convolutional neural network (CNN) [4] [5] remarks the potential to achieve the same.
2. **Recognize each scene and the whole story.** Recognize the complex meaning of the story in the following three steps and construct a suitable layered classifier.

M. Ueno(✉)
Information and Media Center, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi 441-8580, Japan
e-mail: ueno@imc.tut.ac.jp

S. Omatu et al. (eds.), *DCAI, 13th International Conference*,
Advances in Intelligent Systems and Computing 474,
DOI: 10.1007/978-3-319-40162-1_42

A part of each scene. The name of the object and facial expressions.
Scene. Social relationships between the characters and the emotions related to each character
Inbetweening scenes. Interpretation of the story by generating intermediate frames between two images for sequential transition of the first frame to the second.

The previously used image recognition method involves machine learning based on appropriate manual vectors utilizing SIFT, HOG and Haar-like features. Then, a part of each scene is recognized, for example face detection, object recognition, etc. On the other hand, CNN, which is one of the methods of deep learning, especially for images, is applied to the image and feature vectors are automatically constructed, so that scene recognition is possible. There are a few studies on the feature vectors for comics [6] because of the rapid advances in deep learning. Thus, it is difficult to design the problems to be solved and datasets to be prepared; adequate discussion of the feature vectors is hence needed.

In this study, as the first step toward understanding comics by using computers, I constructed several classifiers for comics and compared the result classification by hand and by a computational method. The rest of the paper is structured as follows. Section 2 describes the features of comics. Section 3 shows the preliminary experiment carried out to classify images by hand, while Section 4 shows the experiment for the computational method. Section 5 describes the detailed comparison of the two experiments. Finally, section 6 concludes this research and gives brief insights into a future study.

2 Four-Scene Comics

Numerous genres and structures of comics exist across the globe. In this study, four-scene comics, which are structured with four continuous scenes, are considered. The length of four-scene comics is limited so as to ensure clear interpretation of the contents. Figure 1 shows the general structure of each page of the four-scene comics.

Story four-scene comics is one of the styles used in popular Japanese comics. The notable feature of this type of comics is that the characters are common among various short stories, and continuous small stories result in a whole story in the book. Therefore, it is easy to classify two stories by considering the whole series of sequential images. Although the fourth scene of a small story plays an important role to interpret stories, it may be difficult to classify two stories focused on the fourth scene that is selected randomly, because some of characters may be identical between stories; features of characters are similar, and new characters may have appeared in the middle of a story, in the case of works by the same artist.

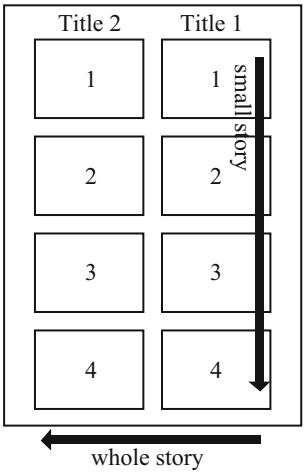

Fig. 1 The structure of four scene comics

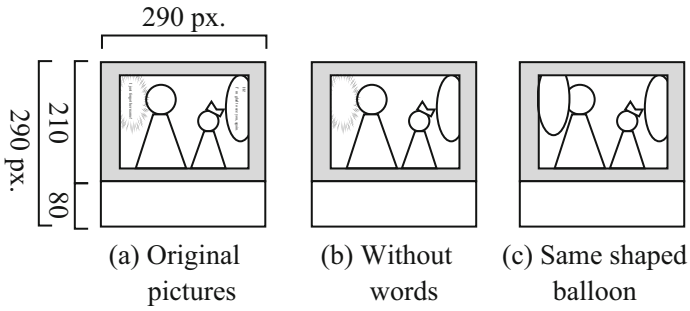

Fig. 2 Each example of three types dataset

3 Preliminary Experiment

In this preliminary experiment, two *story four-scene comics* by same artist are used in order to consider interpretation of comics by human. In order to prevent interpret stories by the title of a small story and inbetweening meanings, only the fourth scene of each small story is given to the annotator.

Dataset:

Two classes, "*Konpeito 1*" [7] and "*Ringo no ki no shitakko de*" [8] by the same artist are used, which have the number of 182 and 178 images respectively.

Table 1 The mean accuracy rate by the human annotator

Size of image	Mean accuracy rate
32 × 32 px.	0.68
64 × 64 px.	0.83
128 × 128 px.	0.93

Six small stories of from the beginning of "*Konpeito 1* were not included in this dataset because the numbers of images in the dataset should be limited to 360 which can be divided by ten. The fourth scene of each small story with 290 × 210 pixels are resized to 290 × 290 pixels added 80 blank pixels to the bottom of the image. Subsequently, they are reduced to 32 × 32 pixels. All the images are in grey scale format.

Comics contain complex information based on both pictures and natural languages. I assumed that picture information was more important for understanding comics. To confirm this, I created three types of datasets by reducing information about natural languages as follows.

– Original pictures
– Without words inside balloons
– Same shaped balloons without words; shapes of all balloons are replaced into same ones manually.

Figure 2 shows each example of these three types of dataset. A preliminary experiment with "same shaped balloons" dataset type is carried out as follows.

1. Give 90% of the dataset with the name of the story to the annotator. The size of the image is 32 × 32 pixels.
2. Annotator labels 10% of the dataset as test data without the name of the story. The size of the image is 32 × 32 pixels.

The accuracy rate is calculated as : the number of the correct label of the book title is divided by the number of test data. As a result, the mean accuracy rate obtained after repeating the experiment three times. It is difficult for human to recognize what objects appear in image with 32 × 32 px. Thus, the accuracy rate of three types of the size in the same dataset are compared. Table 1 shows the mean accuracy rate for three sizes of images in the same datasets. As larger the size, the higher the accuracy rate. I found that characters appearing is the most important information for the human annotator.

4 Experiment

The experiment involves classifying two works by the same artist among three types of datasets by Chainer [9], which is a flexible framework of neural network, written

Table 2 Parameters of neural network

The number of output units	2
Batch size	20
The number of epoch	100
Dropout ratio	0.5
Activation function	ReLU function
Loss function	Softmax cross entropy
Optimizer	Adam. default parameter

Table 3 CNN layer parameters

	CNN1	CNN2
Filter size	5×5	5×5
Padding size	0	0
Stride	1	1

Table 4 Pooling layer parameters

	POOLING1	POOLING2
Filter size	2×2	3×3
Padding size	0	0
Stride	2	3

Table 5 The mean accuracy rate by deep learning

Types of datasets	Mean accuracy rate
Original pictures	0.72
Without words	0.83
Same shaped balloon	0.84

in Python. Table 2 shows the parameter of the neural network. Table 3, Table 4 shows the CNN layer parameters, the pooling layer parameters respectively.

The network architecture and the data are described below.

Architecture:

Input - CNN1 - ReLU - MAX POOLING1 - CNN2 - ReLU - MAX POOLING2 - LINEAR1 - ReLU- Dropout [10] - LINEAR2 - Output

Data:

The same dataset is used in Section 3. Each of the three types of datasets is randomly divided into two groups as follows.

Training data. 90 % of the number of the dataset
Test data. 10 % of the number of the dataset

Table 5 shows the result of the mean accuracy rate of 10 times of 100 epochs of learning.

5 Discussion

Comparison of the results presented in Section 3 and 4 indicates that the accuracy in the computational method is higher than that with the human annotator under the

condition of the 32×32 pixels. Thus, it can be said that the CNN accurately obtains features of scenes of comics.

The human could learn the features of the two types of books focused on a certain character. In the books prepared for the experiment, numerous characters appeared in one scene, while there were a few scenery images and abstract images. Therefore, the human annotator found that the female protagonist of each story appeared frequently in the scenes. However, the size of image is so small that it is difficult to recognize the characters that appear. In addition, the annotator sometimes cannot identify the characters appearing in each book given that the characters are similar due to the same artist. However, the accuracy rate even for 32×32 pixels is quite good. It indicates that human classified two books even if they cannot obtain what objects appear. Thus, other information might contributes to classify.

On the other hand, the result might indicate that the computational method learned based on the arrangement of objects such as characters and balloons. The size of the scene in four scene comics is the same, and it is smaller than that in other types of comics. The balloon object is generally located at the end of the frame. Thus, the location of the other objects is limited.

Considering the effect by reducing natural words, the mean accuracy rate for same shaped balloon dataset is the highest, while that of original pictures dataset is the lowest. From this result, it can be said that pictorial information is sufficient to classify these two works by deep learning, because the size of the image is too small to read the character words. Namely, information about the character word is regarded as noise. In the future research, information about pictures and natural languages will be considered in detail.

6 Conclusion

In this study, as the first step toward understanding comics by using computers, three types of datasets are prepared and classifiers are constructed by using a CNN. Comparing the result of classification of two books by the human annotator and by computational method, I found that the CNN is efficient to be applied to grey scaled unphotographic complicated images such as scenes of comics. Furthermore, I indicated the differences of efficient features to interpret comics between humans and computers, and I discussed the effect of reducing information of scenes of comics; i.e., size and words. Features of human annotator recognition and that of computational method may not be the same. However, in order to interpret story of comics, I believe that only simple image recognition is not sufficient but it requires combination of several classifiers for various information of comics referring to human recognition. Thus, I continue analyzing the differences between information focused by human annotator and filters learned by deeplearning.

Future studies would involve the following:

– Consideration of suitable parameters and a layered architecture.
– Detailed analysis for detecting features between human annotators and the computational method.
– Change in the size of the images and multimodal information composed of images and natural languages.

Acknowledgment I wish to thank Mr. Suenaga for preparing the dataset.

References

1. Le, Q.V.: Building high-level features using large scale unsupervised learning. In: Acoustics, Speech and Signal Processing (ICASSP), pp. 8595–8598 (2013)
2. Ueno, M., Mori, N., Matsumoto, K.: 2-Scene comic creating system based on the distribution of picture state transition. In: Advances in Intelligent Systems and Computing, vol. 290, pp. 459–467 (2014)
3. Eitz, M., Hays, J., Alexa, M.: How Do Humans Sketch Objects? ACM Trans. Graph. (Proc. SIGGRAPH) **31**(4), 44:1–44:10 (2012)
4. Fukushima, K., Miyake, S.: Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition **15**(6), 455–469 (1982)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
6. Tanaka, T., Toyama, F., Miyamichi, J., Shoji, K.: Detection and Classification of Speech Balloons in Comic Images. The journal of the Institute of Image Information and Television Engineers **64**(12), 1933–1939 (2010)
7. Fujino, H.: Konpeito ! 1 (Confetti ! 1). Houbunsha (2007)
8. Fujino, H.: Ringo no ki no shitakko de (Under the apple tree). Houbunsha (2005)
9. Chainer. http://chainer.org/
10. Hinton, G.E., et al.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)