Manga FaceNet: Face Detection in Manga based on Deep Neural Network

Wei-Ta Chu
National Chung Cheng University
Chiayi, Taiwan
wtchu@ccu.edu.tw

Wei-Wei Li
National Chung Cheng University
Chiayi, Taiwan
welcometoway@gmail.com

ABSTRACT

Among various elements of manga, character's face plays one of the most important role in access and retrieval. We propose a DNN-based method to do manga face detection, which is a challenging but relatively unexplored topic. Given a manga page, we first find candidate regions based on the selective search scheme. A deep neural network is then proposed to detect manga faces of various appearance. We evaluate the proposed method based on a large-scale benchmark, and show performance comparison and convincing evaluation results that have rarely done before.

CCS CONCEPTS

•**Computing methodologies** → **Object detection;**

KEYWORDS

Manga, face detection, convolutional neural network

ACM Reference format:
Wei-Ta Chu and Wei-Wei Li. 2017. Manga FaceNet: Face Detection in Manga based
on Deep Neural Network. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 5 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079031

1 INTRODUCTION

Face detection is a fundamental step to many computer vision and multimedia applications. This topic has been widely studied for natural images. However, much fewer studies have been proposed for manga (Japanese comics). Manga is one of the biggest book sales in the world. Although the book market slumped, in Japan the market of compiled manga books keep creating record-high sales and reach around 2.4 billion US dollars in year 2014 [1]. As more and more manga books are digitized, efficient access and retrieval of manga is urgently demanded [6].

There are at least three differences between faces in natural images and in manga. First, we focus on the largest comics market, i.e., Japanese comics (manga). In most manga, only black-and-white and sometimes gray information is available, which is different from color information in natural images. Second, there are extreme variations in faces in different manga. Figure 1(a) and Figure 1(b) show two normal frontal faces, while Figure 1(c) and Figure 1(d) show drastically different visual appearance especially on eyes. Third, manga faces do not entirely possess properties of human faces. The spatial layout, visual appearance, and expression of manga faces may not physically reasonable (Figure 1(e) and Figure 1(f)). The methods proposed for human face detection are thus not able to be directly employed in manga face detection.

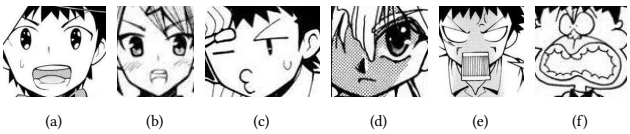


Figure 1: Sample manga faces.

Some works were proposed to detect manga faces. Sun and Kise [8] extracted Haar-like features, and concatenated a sequence of weak classifiers to construct a face detector. For a limited dataset, frontal manga faces can be detected. Focusing on colorful comics images, Takayama et al. [9] detected skin color and the jaw contour to find character's face. The symmetric property of face was then adopted to filter out noises. This method may not be generic to faces in different poses. In [12], the deformable part model (DPM) was used to consider pose and spatial variations. This approach achieved performance better than the conventional HOG-based (histogram of oriented gradient) approach. Chu and Chao [3] attempted to avoid visual variations by first detecting character's eyes, and then expanded the eye regions to find the face.

The aforementioned methods are mostly ad-hoc approaches and are hard to be generalized. In addition, most of them were not evaluated on a large-scale dataset, making the conclusion not convincing enough. Thanks to the recently proposed Manga109 dataset [6], now we have more resource to train the face detector, and can evaluate the proposed method at a larger scale. We propose a deep-based manga face detection framework. The selective search [10] scheme is adopted to first detect regions with objects. Each region is then examined by a deep neural network that extracts features and classifies the input region as a face or not.

Contributions of this paper are twofold. First, we construct a deep manga face detection framework to resist visual variations. Second, we evaluate performance of the proposed framework based on large-scale benchmark to facilitate fair comparison in the future.

The rest of this paper is organized as follows. Section 2 provides details of the deep framework. Section 3 shows performance of the

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079031

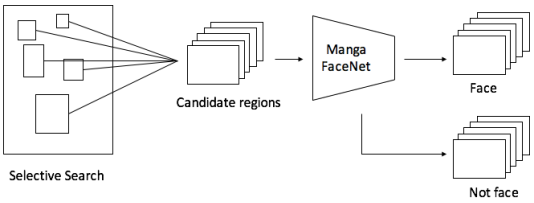


Figure 2: Overview of the manga face detection framework.

Figure 3: Left: ground truth; right: the regions found by selective search with large overlap with the ground truth. All regions are resized to 40×40.

proposed method as well as comparison with existing methods, and conclusion is given in Section 4.

2 MANGA FACE DETECTION

2.1 System Overview

Figure 2 shows overview of the proposed method. Given a manga page, we first employ the selective search scheme [10] to detect regions probably containing objects. Each region is then examined by the proposed deep neural network, named Manga FaceNet, to see whether this region is a manga face or not.

Recently the power of deep learning has been demonstrated in many domains. Not only for image classification [5] or object detection [7] for natural images, now the effectiveness of deep learning on sketch or line drawings have also been demonstrated [11] [13]. We therefore propose to construct a deep neural network called Manga FaceNet to do this task.

Before training, we randomly select 66 manga titles from the Manga109 dataset, and from each title we select 50 manga pages as the evaluation dataset. From each page, we manually define the bounding box of each manga face.

To train the proposed network, we extract positive samples and negative samples in the following way. Given each manga page, we employ the selective search scheme [10] to find regions of objects. The regions that have more than 70% overlap with any truth bounding box and contain frontal faces are considered as positive samples. On the other hand, the regions that have less than 30% overlap with any truth bounding box are considered as negative samples. Figure 3 illustrates results of sampling by showing just a few positive samples because of space limitation. We can see that the selective search may detect several regions corresponding to the same ground truth.

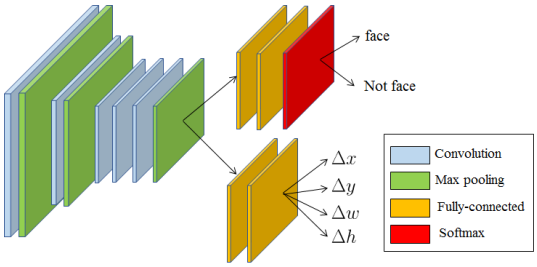


Figure 4: Structure of the Manga FaceNet.

2.2 Manga FaceNet

Based on the training data, we construct a face detector based on convolutional neural networks (CNN). Figure 4 shows structure of the proposed Manga FaceNet. Given a candidate region found by selective search, we resize it into 40×40 pixels and input it to a CNN consisting of five convolutional layers to do feature extraction. The output of the fifth convolutional layer is flattened to be a vector, which is then input to two branches. The top branch is a classification network consisting of two fully-connected layers. The activation function of the last fully-connected layer is softmax, and thus it outputs the probability of a given region being a manga face or not. To train the top branch of Manga FaceNet, the loss function is set as:

$$L_1 = \frac{-1}{N} \sum (y_g \log y_p + (1 - y_g) \log(1 - y_p)), \quad (1)$$

where y_p is the predicted probability of the given data being a manga face, and y_g is the ground truth where $y_g = 1$ if the given region is really a face, and $y_g = 0$ otherwise.

For the bottom branch of Manga FaceNet, we attempt to further consider the spatial displacement of a given training region to its corresponding ground truth, in order to more finely evaluate the goodness of a region being a manga face. Associated with each region, the spatial displacement and aspect ratio to the corresponding ground truth are further considered. More particularly, taking the original manga page as a coordinate system, assume that the left-top corner of a given region is at (x', y'), and its width and height are w' and h', respectively. For its corresponding ground truth, assume its left-top corner is at (x, y), and its width and height are w and h, respectively. Motivated by the settings in [7], the horizontal and vertical spatial displacements are calculated and normalized as $\Delta x = \frac{x'-x}{w}$, and $\Delta y = \frac{y'-y}{h}$, respectively. The width difference Δw and the height difference Δh are calculated as $\Delta w = \log \frac{w'}{w}$, and $\Delta h = \log \frac{h'}{h}$, respectively. Similar to the top branch of Manga FaceNet, we design two fully-connected layers to estimate the spatial displacement and aspect ratio change. Given a candidate region, if the predicted value of spatial displacement and aspect ratio difference are $\Delta\hat{x}$, $\Delta\hat{y}$, $\Delta\hat{w}$, and $\Delta\hat{h}$, respectively. The loss function to train the bottom branch of Manga FaceNet is the mean square error:

$$L_2 = \frac{1}{N} \sum (\Delta x - \Delta\hat{x})^2 + (\Delta y - \Delta\hat{y})^2 + (\Delta w - \Delta\hat{w})^2 + (\Delta h - \Delta\hat{h})^2, \quad (2)$$

where N is the number of training regions.

Table 1: Detailed Manga FaceNet configuration.

input (40×40 gray images)				
conv3-32 maxpooling dropout(0.25)	conv3-64 maxpooling dropout(0.25)	conv3-128	conv3-128	conv3-128 maxpooling dropout(0.25)
fully-connected (256 nodes) (both branches) dropout(0.5)(both branches)				
fully-connected (4 nodes) (bottom branch) fully-connected (2 nodes) – softmax (top branch)				

Overall, the top branch and the bottom branch of the Manga FaceNet are jointly trained by considering the integrated loss

$$L = \lambda_1 L_1 + \lambda_2 L_2, \tag{3}$$

where both weighting parameters λ_1 and λ_2 are currently set as 1.

Table 1 shows detailed configurations of Manga FaceNet. The input region is processed through five convolutional layers, as shown in the second row of Table 1, from left to right. The convolutional parameters are denoted as "conv⟨receptive field size⟩ - ⟨number of channels⟩". The ReLU activation function is used in all convolutional layers. The first, the second, and the fifth convolutional layers are followed by max pooling and dropout with ratio 0.25. Results of convolution are input to two fully-connected layers, where the first one consists of 256 nodes, and the second one consists of 4 nodes and 2 nodes for the bottom branch and the top branch, respectively. The activation function of the final fully-connected layer of the top branch is softmax.

We adopt mini-batch of size 100, and the learning process updates network parameters for 60 epochs. The learning algorithm is RMSprop, with the learning rate 0.001. We employ a strategy similar to the "image-centric" sampling strategy [4] to the train the network. A mini-batch contains positive samples and negative samples both sampled from the same manga title, i.e., more like "manga-title-centric" sampling.

3 EVALUATION

3.1 Dataset and Evaluation Settings

As we mentioned before, we evaluate the proposed system based on a large-scale manga benchmark, i.e., the Manga109 dataset [6]. We randomly select 66 titles from the 109 titles, and from each title we select 50 manga pages and manually define ground truths of manga faces. There are 14,405 faces in total.

According to the framework shown in Figure 2, we first detect object regions by the selective search scheme, and then estimate the probability of each region being a face. If the probability is larger than a threshold τ, we say the test region is a manga face.

One may wonder that if the selective search scheme is able to detect regions covering manga faces. To verify this, we calculate the ratio of manga faces that can be included in the regions found by selective search. According to our experiment, this ratio is around 92%. In the following experiments, we will use precision and recall values to measure performance of manga face detection. Because of the designed procedure shown in Figure 2, the value 0.92 is therefore the upper bound of the recall value we can obtain.

Table 2: Performance comparison between methods.

Methods	Precision	Recall	F-measure
OpenCV (pre-trained)	0.48	0.04	0.07
OpenCV (trained with manga) [8]	0.15	0.95	0.26
Eye-based method [3]	0.42	0.55	0.48
Manga FaceNet (top branch only)	0.57	0.70	**0.63**
Manga FaceNet (both branches)	0.79	0.56	**0.66**
Manga FaceNet (both) + VGG	0.97	0.46	**0.62**

3.2 Performance of Manga Face Detection

Table 2 shows performance comparison between different manga face detection methods. As we expect, the Adaboost method originally designed for detecting real human faces and is embedded in the OpenCV library does not work well (the first row). To decrease the mismatch problem between real human faces and manga faces, we retrain the OpenCV model based on our manga data. This is more like the method proposed in [8], though the training data are not the same. From the second row of Table 2, we see that the recall rate can be largely improved, while the precision rate largely decreases. Overall, only 0.26 F-measure can be obtained by the Adaboost method. The eye-based method [3] works better, but is still not promising. Intuitively the shapes of eyes is similar to round or elliptical. However, eyes are the most important features to show different artists' drawing styles or to show different characters. Accurately detecting eyes is thus not a trivial problem.

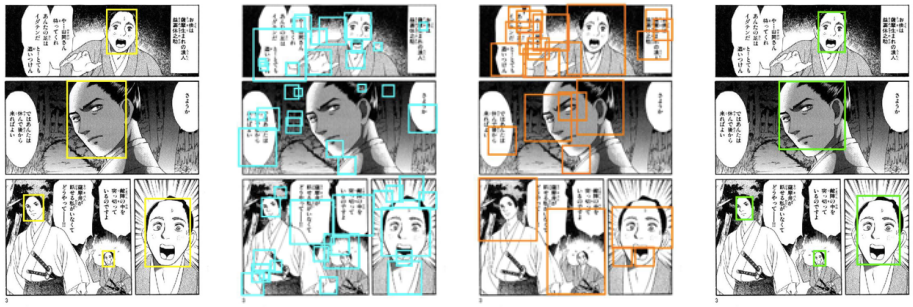
The fourth row shows performance of the proposed Manga FaceNet (top branch only), with threshold τ mentioned above set as 0.8. We see significant improvement can be obtained by the proposed deep-based method. When both branches are considered in training, the best performance as F-measure equal to 0.66 can be obtained (the fifth row). When a large number of manga pages (and thus a large number of manga faces) are available, precision of face detection is more important than recall in some applications, e.g., style analysis proposed in [3]. Therefore, for the regions determined as faces by Manga FaceNet, we further extract CNN features based on the VGG-f framework [2], and then construct a support vector machine to further verify them as face or non-face. As shown in the last row of Table 2, this process largely improves precision, with the cost of recall decrease.

Figure 6 further shows performance variations of Manga FaceNet (MF) and Manga FaceNet with VGG (MF+VGG), in terms of PR curves obtained based on different thresholds τ's. The MF+VGG approach hardly achieves high recall, but its precision is significantly higher than the MF approach. In addition, we see the MF+VGG approach relatively achieves more stable performance (precision ranges from 0.94 to 0.98, while recall ranges from 0.41 to 0.54).

Figure 5 shows sample manga face detection results for a manga page. It can be easily seen that the proposed method can much more accurately detect manga faces than the method originally designed for human face detection.

3.3 Discussion

Although the proposed Manga FaceNet largely improves detection performance, it is still far from perfect. The top row of Figure 7



(a) (b) (c) (d)

Figure 5: Sample detection results. (a) Ground truth; (b) OpenCV trained with manga; (c) the method in [3]; (d) our results.

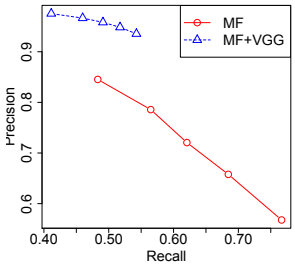


Figure 6: PR curves of two variants of the proposed method.

Figure 7: Sample face detection results. Top row: detected faces; middle row: false alarms; bottom row: miss.

shows correctly detected faces. The middle row shows false alarms. We observe that speech balloons or regions with large white area and symmetry may be falsely detected as faces. The bottom row shows miss cases. Apparently side faces are often missed in the current work. This is also a challenging issue in real human face detection. In the future, we will jointly consider side faces in the deep framework and try to improve recall rates.

4 CONCLUSION

We have presented a deep-based face detection method specially for Manga. Given a manga page, we first find candidate regions by the selective search scheme, and then determine each region as a manga face or not by the proposed Manga FaceNet. In addition to classification label, we also jointly consider spatial displacement and aspect ratio in the proposed network. Being able to model high variations on visual appearance and expression, the proposed method significantly outperforms the methods designed for real human faces and the conventional manga face detection methods. In the future, we would like to design a more elegant framework to enhance the ability of detecting side faces, in order to improve recall values.

ACKNOWLEDGMENTS

The work was partially supported by the Ministry of Science and Technology of Taiwan under the grant MOST105-2628-E-194-001-MY2, MOST104-2221-E-194-014, and MOST103-2221-E-194-027-MY3.

REFERENCES

[1] Anime News Network 2015. *Japanese Manga Book Market Rises to Record 282 Billion Yen.* Anime News Network. http://www.animenewsnetwork.com/news/2015-01-23/japanese-manga-book-market-rises-to-record-282-billion-yen/.83614.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *Proceedings of British Machine Vision Conference.*

[3] W.-T. Chu and Y.-C. Chao. 2014. Line-based Drawing Style Description for Manga Classification. In *Proceedings of ACM International Conference on Multimedia.* 781–784.

[4] R. Girshick. 2015. Fast R-CNN. In *Proceedings of International Conference on Computer Vision.* 1440–1448.

[5] A. Krizhevsky, I. Sutskever, and G.E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of Neural Information Processing Systems.*

[6] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa. 2016. Sketch-based Manga Retrieval Using Manga109 Dataset. *Multimedia Tools and Applications* (2016).

[7] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of Neural Information Processing Systems.*

[8] W. Sun and K. Kise. 2010. Similar Partial Copy Detection of Line Drawings Using a Cascade Classifier and Feature Matching. In *Proceedings of International Workshop on Computational Forensics.* 121–132.

[9] K. Takayama, H. Johan, and T. Nishita. 2012. Face Detection and Face Recognition of Cartoon Characters Using Feature Extraction. In *Proceedings of IIEEJ Image Electronics and Visual Computing Workshop.*

[10] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. 2013. Selective Search for Object Recognition. *International Journal of Computer Vision* 104, 2 (2013), 154–171.

[11] F. Wang, L. Kang, and Y. Li. 2015. Sketch-based 3D Shape Retrieval Using Convolutional Neural Networks. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition.* 1875–1883.

[12] H. Yanagisawa, D. Ishii, and H. Watanabe. 2014. Face Detection for Comic Images with Deformable Part Model. In *Proceedings of IIEEJ Image Electronics and Visual Computing Workshop.*

[13] F. Zhu, J. Xie, and Y. Fang. 2016. Learning Cross-Domain Neural Networks for Sketch-Based 3D Shape Retrieval. In *Proceedings of AAAI Conference on Artificial Intelligence.* 3683–3689.