SReN: Shape Regression Network for Comic Storyboard Extraction

Zheqi He, Yafeng Zhou, Yongtao Wang, and **Zhi Tang**
Institute of Computer Science & Technology, Peking University, Beijing, China
{hezheqi, zhouyafeng, wangyongtao, tangzhi}@pku.edu.cn

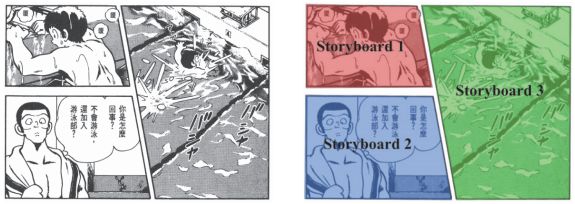
Abstract

The goal of storyboard extraction is to decompose the comic image into storyboards, which is the fundamental step of comic image understanding and producing digital comic documents suitable for mobile reading. Most of existing approaches are based on hand crafted low-level visual patters like edge segments and line segments, which do not capture high-level vision information. To overcome this drawback of the existing approaches, we propose a novel architecture based on deep convolutional neural network, named Shape Regression Network (SReN), to detect storyboards within comic images. Firstly, we use Fast R-CNN to generate rectangle bounding boxes as storyboard proposals. Then we train a deep neural network to predict quadrangles for these proposals. Unlike existing object detection methods which only output rectangle bounding boxes, SReN can produce more precise quadrangle bounding boxes. Experimental results on 7382 comic pages, demonstrate that SReN outperforms the state-of-the-art methods by more than 10% in terms of F1-score and page correction rate.

Introduction

Comic is a kind of entertainment publication popular among people of different ages around the world. As shown in Fig. 1, the storyboard is the basic semantic unit of a comic. Hence, decomposing the comic image into storyboards is the fundamental step to understand content of comic. In addition, decomposing the comic image into storyboards is the key technique to produce digital comic documents suitable for reading on mobile devices with small screen.

Most of previous storyboard extraction methods (Wang, Zhou, and Tang 2015; Li et al. 2015) use only hand crafted low-level visual patterns, such as edge segments, line segments or connected component. These methods can work effectively under certain assumptions. However, they may fail to handle the comic image with complex layout. For example, when the storyboards with missing borderlines or when more than two storyboards are overlapped with each other, these methods are tend to fail. The most important reason is that low-level visual patterns can not represent image content well.



(a) (b)

Figure 1: (a) Original comic page; (b) Comic storyboards (Source:Mitsuru Adachi, Rough,vol.1, p.52).

Recently, deep learning methods (Girshick 2015; Liu et al. 2015) have been applied to object detection and gain the state-of-the-art performances in many challenges. The effective feature learning capability of deep neural network make great contribution to high-level vision tasks. However, these methods can only obtain rectangle bounding box of objects, which are not precise enough for many application tasks. For example, the tasks of comic storyboard detection or traffic sign detection. It is better to use parameterized shape like triangle, quadrangle or ellipse to represent the detected results.

In this paper, we propose a novel architecture based on deep convolutional neural network named SReN [1] to detect storyboards within comic images. SReN can regress parameters of quadrangles, which improve localization accuracy for object detection. Experiments conducted on a dataset of 7382 comic pages, and the results demonstrate that our method outperforms the state-of-the-art methods.

SReN: A Storyboard Extraction Architecture

In this section, we introduce our new storyboard extraction architecture SReN, which consists of two main steps: generate storyboard proposals and train shape regression network. The architecture is illustrated in Fig. 2.

Generating storyboard proposals

We use comic images to train a Fast R-CNN model which detect rectangle bounding boxes r for storyboards. Since

[1] Supplemental materials of this research are available in http://philokey.github.io/sren.html

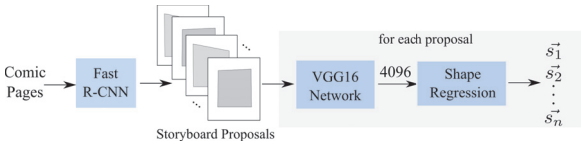


Figure 2: The architecture of SReN

Fast R-CNN can only generate rectangle bounding boxes, we use corresponding exterior rectangles as ground truth for each storyboard. But the bounding box often miss some parts of a storyboard, in order to obtain the complete storyboard, we enlarge r by a factor of 1.1 to generate storyboard proposal p as the input of our SReN. Another problem is that storyboards are often various in sizes, to reduce the interference of this, we normalize the vertexes of storyboard proposals into $[-1, 1]$, that is

$$x' = \frac{x - c_x}{w}, \quad y' = \frac{y - c_y}{h}, \tag{1}$$

where (x', y') is the vertex of the regression target, (x, y) is the vertex of the original storyboard, (c_x, c_y) is the center of the storyboard proposal, w and h is the width and the height of the storyboard proposal.

Training shape regression network

Firstly, we sort the vertexes of regression target $\{\vec{t_1}, ..., \vec{t_4}\}$, where $\vec{t_i} = (t_{ix}, t_{iy})$, according to their polar angles. Then we use p and its regression target as the input of VGG16 network to generate a feature f with 4096 dimensions. Finally we add a fully connected layer to regress the vertexes of the storyboard $\{\vec{s_1}, ..., \vec{s_4}\}$, where $\vec{s_i} = (s_{ix}, s_{iy})$. Like Fast R-CNN, we use the loss function:

$$L(\vec{t_i}, \vec{s_i}) = \sum_{i=1}^{4} smooth_{L_1}(t_{ix} - s_{ix}) + smooth_{L_1}(t_{iy} - s_{iy}), \tag{2}$$

where

$$smooth_{L_1}(a) = \begin{cases} 0.5a^2, & \text{if } |a| < 1 \\ |a| - 0.5, & \text{otherwise.} \end{cases} \tag{3}$$

Experiments

We construct a dataset of 29845 labeled comic pages (contains 169421 storyboards) from 103 different comic books, which come from different Japanese, Hong Kong and European comics. We randomly select 15087 of the labeled comic pages to train SReN, we use another 7375 comic pages to validate the training result and conduct experiments on the remaining 7382 comic pages.

We evaluate the results on two levels: storyboard level and page level. On the storyboard level, we use precision, recall and F1 score as evaluation metrics. On the page level, we use page correction rate (PCR) as evaluation criterion, i.e., the ratio of comic pages in which all storyboards are correctly detected. To be more specific, each detected storyboard is represented by a quadrangle, if intersection-over-union (IoU) between the detected storyboard and the corresponding ground truth is more than 90%, we regard it as a correct detection. IoU for each ground truth and detected storyboards in a comic page is defined as following,

$$IoU = \max_i \frac{p \cap D_i}{p \cup D_i}, \tag{4}$$

where D_i is a set of detected storyboards, p is the manually label for each storyboard within the page. The intersection and the union operation are calculated in terms of area.

We compare our method with Fast R-CNN without shape regression and two low-level visual patterns based methods: TCRF (Li et al. 2015) and ESA (Wang, Zhou, and Tang 2015). Experimental results are listed in Table 1, which indicate that: 1) the deep learning based methods can achieve much better results than the hand crafted low-level visual patterns based methods; 2) among the deep learning based methods, Fast R-CNN with SReN is better than the original Fast R-CNN by the effective shape regression.

Table 1: Results on 7382 comic pages

Method	Precision	Recall	F1 score	PCR
ESA	0.835	0.700	0.762	0.418
TCRF	0.699	0.64	0.668	0.399
Fast R-CNN	0.807	0.799	0.803	0.518
SReN	**0.888**	**0.879**	**0.883**	**0.640**

Conclusions and Future Work

We propose a novel deep architecture to detect storyboards within comic images, named SReN. The main contribution is to use a shape regression network to obtain the vertexes of quadrilateral storyboards in comic pages. Experimental results demonstrate that SReN performs better than two state-of-the-art storyboards extraction methods. In the future, we will test our idea on other kinds of shapes, like ellipse and triangle, which can be used in traffic sign detection and other applications. We also want to investigate how to design an end-to-end model to automatically detect and regression the target shape.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grants 61300061 and 61673029.

References

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Li, L.; Wang, Y.; Suen, C. Y.; Tang, Z.; and Liu, D. 2015. A tree conditional random field model for panel detection in comic images. *Pattern Recognition* 48(7):2129–2140.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; and Reed, S. 2015. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*.

Wang, Y.; Zhou, Y.; and Tang, Z. 2015. Comic frame extraction via line segments combination. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 856–860. IEEE.