Big Data Analytics

RESEARCH Open Access

Chinese text-line detection from web videos with fully convolutional networks

Chun Yang[1†], Wei-Yi Pei[1†], Long-Huang Wu[1] and Xu-Cheng Yin[1,2*]

*Correspondence:
xuchengyin@ustb.edu.cn
†Equal contributors
[1]Department of Computer Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
[2]Beijing Key Laboratory of Materials Science Knowledge Engineering, University of Science and Technology Beijing, Beijing 100083, China

Abstract

Background: In recent years, video becomes the dominant resource of information on the Web, where the text within video usually carries significant semantic information. Video text extraction and recognition plays an essential role in web multimedia understanding and retrieval for big visual data analytics and applications. To deal with challenging backgrounds and embedding noises, most conventional approaches usually tend to design sophisticated pre-processing and post-progressing steps before and after text detection. In this paper, we present a simple yet powerful pipeline that directly and uniformly detects Chinese text lines for embedded captions from web videos.

Results: In this Chinese text-line detection system, a fully convolutional network with local context is adopted to localize via an end-to-end learning way. The produced caption predictions are with the word level that could be directly fed into the character classifier. Text-line construction is then performed by heuristic strategies. A variety of experiments are conducted on several real-world web video datasets and demonstrated the effectiveness and efficiency of our proposed method.

Conclusion: The proposed system can directly detect the English word and Chinese characters in the caption text-lines without word or character segmentation with the high performance on real-world web video datasets.

Keywords: Video text detection, Text segmentation, Fully convolutional networks, Embedded captions, Web videos

Background

Nowadays, with the rapid development of digital multimedia technology, we are exposed to a huge number of images and videos in daily life and on the Web. Captions in movies, TV programs and short videos can supply rich semantic information which is very useful in video content analysis, indexing, retrieval and multi-language translation [1, 2]. A variety of research efforts have been made toward extracting video captions in various big visual data analytics and applications.

Generally speaking, the major related techniques for video caption extraction involve three aspects: caption detection, caption segmentation and optical character recognition (OCR). Most of existing video OCR systems are based on the combination of sophisticated pre-processing procedures for text extraction and traditional OCR engines [3]. Techniques from standard OCR, which focus on high resolution scans of printed (text) documents, are also applicable for video images. For video OCR, video frames have to be first identified which obtain visible textual information, then the text is localized and interfering background has to be removed, and geometrical transformations have to be applied before standard OCR engines can process the text successfully. Because of several grand challenges for text extraction from web videos, there are still some limitations for conventional Video OCR technologies [1]. In this paper, we focus on text detection (caption detection) from web videos.

Recently, deep Convolutional Neural Networks (CNN) have promoted generic object detection substantially. Proposal-based methods like Faster R-CNN [4], Single Shot MultiBox Detector (SSD) [5] and YOLO [6, 7] lead the state-of-the-art performance on generic object detection. For mainly two reasons, however, it is difficult to apply these general object detection systems directly to caption detection, which generally requires a higher localization accuracy. Firstly, in generic object detection, each object has a well-defined closed boundary, while such a well-defined boundary may not exist in caption text, since a text line or word is composed of a number of separate characters or strokes [8]. Secondly, text has a larger aspect ratio range than generic objects that text could be short or long in different directions.

Existing methods, either conventional or deep neural networks based ones, mostly consist of several stages and components, which can easily cause accumulated errors and probably waste time. Furthermore, different kinds of text patterns (especially for Chinese) and highly cluttered backgrounds pose main challenges of accurate text localization and text segmentation. As a result, the accuracy and efficiency of such methods are still far from satisfactory.

In this paper, we propose an accurate Chinese text-line caption detection pipeline that merges both the detection task and the text-line segmentation task. The pipeline utilizes a fully convolutional network (FCN) model that directly produces word or text-line level predictions, excluding text-segmentation algorithm and slow intermediate steps. The produced caption predictions are with the word level that could be directly fed into the character classifier. By adopting a pixel-wise classification approach we fix the drawbacks of proposal-based methods and remove intermediate steps.

Summarily, the contributions of this paper are threefold. Firstly, we design an end-to-end trainable neural network model for video caption detection, which costs short training time and has high generalization ability. This model can easily spread from Chinese to many other languages. Secondly, our framework directly detects English words or Chinese characters in video captions without text-line segmentation and sophisticated pre-processing procedures. Thirdly, a variety of experiments on real web videos show that our model achieves highly competitive results while keeping its computational efficiency.

Related work

The caption is essentially a kind of text in video and added manually for special purpose. Generally speaking, available text detection approaches can be grouped into two categorizations, i.e., conventional methods and deep CNN based methods.

Conventional text detection approaches mainly rely on well-designed feature such as texture, edge, color, motion and some other text representative features to discriminate text from background. The corresponding methods can be roughly divided into four groups: edge-based, texture-based, corner-based and stroke-based methods. In recent years, connected components (CCs) based approaches have obtained great success, which is motivated by an observation that text strokes generally have homogeneous intensity / color. For example, Yin et al. [9, 10] proposed robust text detection approaches by extracting Maximally Stable Extremal Regions (MSERs) as character candidates to seek text areas.

Recently, the deep CNN based framework has become increasingly prevalent and opened a brand new dimension in the field of text detection. Through a combination of MSERs and deep CNN, Huang et al. [11] constructed a powerful classifier to better detect text regions. In [12], Jaderberg et al. proposed a text detection and recognition system based on region proposal mechanism and deep CNN. The Connectionist Text Proposal Network (CTPN) [8] innovatively combined Faster R-CNN with LSTM and constructed a bottom-to-up model to detect horizontal text lines. Inspired by Faster R-CNN, Ma et al. [13] introduced a novel rotated anchors based framework for arbitrary-oriented text detection. Different from previous mentioned methods, Zhang et al. [14] proposed to utilize segmentation networks, namely Fully Convolutional Network [15], to generate the text prediction maps and geometric approaches for inclined proposals. Approaches above achieved desirable results on a variety of standard benchmarks.

The architecture of the proposed network in our paper is inspired by a deconvolution network for Semantic Segmentation [16] and EAST [17]. Deconvolution network plays an important role in image semantic segmentation. By adopting the idea of up-sampling with deconvolution, we can perform pixel-wise prediction, which mitigates the limitations of the existing methods based on fully convolutional networks by integrating deep deconvolution network and proposal-wise prediction.

Methods

As shown in Fig. 1, our framework generally includes two major steps, i.e., character candidate classification, boundary expansion (text-line construction). The core steps (character candidate classification) can be further divided into three major parts, feature extraction, multi-level feature merging and pixel-wise prediction. The model is a fully convolutional neural network that is composed mainly of convolution, pooling and deconvolution. By combining down-sampling and up-sampling, the proposed model implements an end-to-end trainable network. The post-processing parts performs boundary expansion aiming to predict the geometric shapes and group text lines for final text recognition.

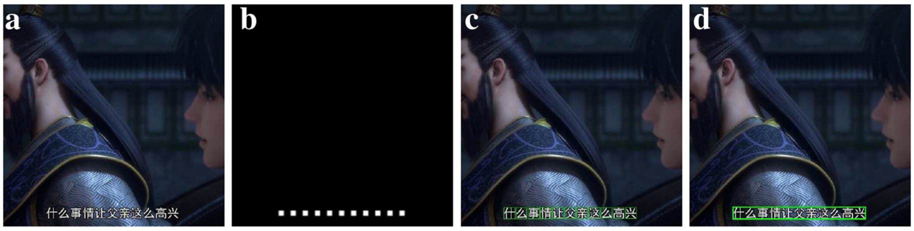


Fig. 1 The procedure of our proposed framework: (**a**) The input image, (**b**) the outputs of character candidate classification, (**c**) word/character extraction after boundary expansion, and (**d**) the finally constructed text-line

Network architecture

The proposed model is designed to become an end-to-end fully convolutional neural network. Instead of adopting a proposal-based framework, direct prediction of each pixel avoids replicate computation in overlapping area and breaks constrains brought by limited proposals' aspect ratio. Furthermore, the end-to-end pixel-wise prediction network also provides a favorable natural characteristic for the whole pipeline and the multi-orientation objects detection task.

In the feature extraction stage, we explore two well-known models pre-trained on ImageNet [18], namely ResNet-50 [19] and PVANET [20]. ResNet is a recent more powerful object detection network, which explicitly reformulates the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions [19]. PVANET is the deep but lightweight neural network for real-time object detection [20].

In the multi-level feature merging stage, we design a merging strategy for adaptively fusing convolutional features in the network. Unlike generic objects with a well-defined closed boundary, such a well-defined boundary does not exist in caption text. Whether English or Chinese caption text-line could be divided into several words or Chinese characters which further is composed of a number of separate characters or strokes. All these components are surrounded by the complicated video background in most cases. To deal with the problem, our proposed network adopts the idea of fusing convolutional features from four different levels (multi-level feature merging) whose sizes are 1/32, 1/16, 1/8, 1/4 of the input image respectively. The large receptive field in the high-level convolutional stage ensures the network has ability to see the long text and the low-level convolutional features guarantee the gap between words can be distinguished. Coarse-to-fine structures of captions are reconstructed progressively through a sequence of deconvolution operations. The detailed design of network is shown in Fig. 2.

For learning of the very deep CNN model, the deeper the network is, the harder it is to be trained. Differences in the data distribution, the ground truth distribution, and the loss function will all lead to difficulties in network convergence, even with the initialization of a pre-trained model. In remedy of this, we add a batch normalization layer after each convolutional layer to keep data in the network following the same distribution. It is a key process to make the network steadily converge.

In order to save computation, our network tends to keep the number of channels of convolutions in the up-sampling branch small and outputs prediction in the 1/4 of input image size. Therefore, the classification task outputs a blob of size $S/4 \times S/4$ with 1 channel, where S is the input image size. Note that it is not necessary to add a regression task to locate the word in our model. On the contrary, we determine the geometry of English words and Chinese characters by accurate classification as described in the next.

Network processing

In the following, the two main parts of our proposed network, i.e., the classification loss and lab map generation, are described in details.

Classification loss

For a typical video image, the distribution of caption/non-caption pixels is heavily unbalanced. In most cases, almost over 95% of the ground truth is non-caption. There are

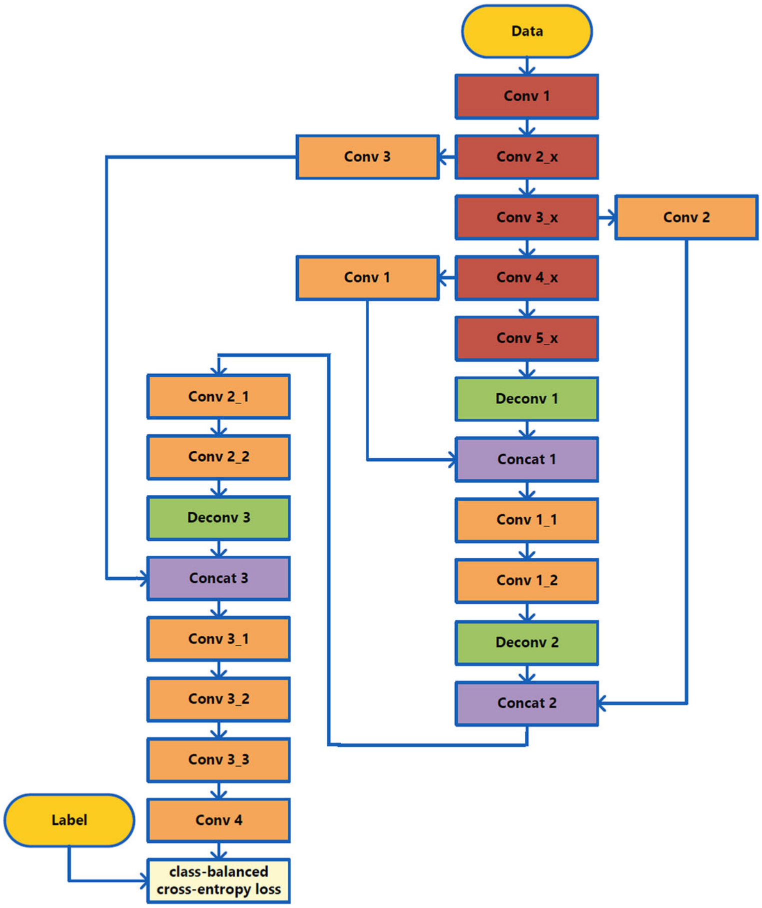


Fig. 2 Details of the fully convolutional network, where the first five convolutional units represent the 5 convolutional stages of PVANET or ResNet-50, and the batch normalization layers and the pooling layers are omitted in the figure for simplicity

several strategies which can be referred to solve this problem, such as Online Hard Negative Mining [21] and Weighted Softmax Loss. In this work, we use the class-balanced cross-entropy loss function which is effective for several systems [17, 22, 23]. The function is defined with

$$L_{cls} = -\beta \sum_{j=1}^{|Y|} Y_j \log y_j - (1-\beta) \sum_{j=1}^{|Y|} Y_j \log(1-y_j) \tag{1}$$

where y_j is the prediction of the j_{th} pixel, and Y_j is the ground truth of the j_{th} pixel. β is a class-balance parameter between positive samples and negative samples, and is defined as

$$\beta = \frac{|Y_-|}{|Y|} = 1 - \frac{|Y_+|}{|Y|} \tag{2}$$

where $|Y_+|$ and $|Y_-|$ represent the caption and non-caption ground truth pixel sets, respectively.

Label map generation

By considering the short distance between words, especially Chinese characters, we shrink the margin by $0.3l$ (l the shortest side length of the word/character, see Eq. 3) of each English word or Chinese character in the ground truth label to get a clear gap between words. Experiment results also empirically show that this strategy does make the network learn boundary information more easily and output a word-level prediction accurately. According to the proposed architecture, the size of the label mask is 1/4 of the original image. The value of the pixel in the positive area is set to be 1 and the rest are set to be 0. An example of the pixel-level label map is shown in Fig. 3. l is defined as the shortest side length of the English word or the Chinese character. For a rectangle $R(X_{lt}, Y_{lt}, X_{rb}, Y_{rb})$, l is simply computed by

$$l = min(X_{rb} - X_{lt}, Y_{rb} - Y_{lt}). \tag{3}$$

Boundary expansion

The task of boundary expansion generally includes two steps, i.e., word/character extraction, and text-line construction.

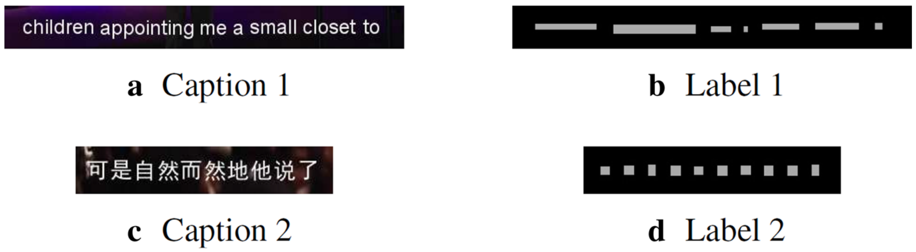
Word/Character extraction

After getting the word-level prediction output by the class-balanced cross-entropy loss, a post-processing step, namely boundary expansion, is used to determine the bounding box of each word. We first use a threshold λ to remove some noise and get the positive region. Then, we get the preliminary box by using a rectangle to enclose the positive region tightly. According to the operation of label mask generation described above, we directly reverse the process to enlarge the box by $\varepsilon \times l$. In our work, we set λ to 0.7 and ε to 0.75. When we get the bounding box, $r\left(x^i_{lt}, y^i_{lt}, x^i_{rb}, y^i_{rb}\right)$, in the output image, we compute the final bounding box, $R\left(\hat{x}^i_{lt}, \hat{y}^i_{lt}, \hat{x}^i_{rb}, \hat{y}^i_{rb}\right)$, in size of input image as

$$\hat{x}^i_{lt} = 4\left(x^i_{lt} - kl\right), \qquad \hat{y}^i_{lt} = 4\left(y^i_{lt} - kl\right) \tag{4}$$

$$\hat{x}^i_{rb} = 4\left(x^i_{rb} + kl\right), \qquad \hat{y}^i_{rb} = 4\left(y^i_{rb} + kl\right) \tag{5}$$

where $\left(x^i_{lt}, y^i_{lt}\right)$ and $\left(x^i_{rb}, y^i_{rb}\right)$ are the left top coordinate and the bottom right coordinate of i_{th} box in the output image, and $\left(\hat{x}^i_{lt}, \hat{y}^i_{lt}\right)$ and $\left(\hat{x}^i_{rb}, \hat{y}^i_{rb}\right)$ are the left top coordinate and the bottom right coordinate of corresponding box in the input image, respectively. l is a parameter defined as the shortest side of box in the output image (see Eq. 3). The processing of both the English word and the Chinese character shares this same equation.



a Caption 1

b Label 1

c Caption 2

d Label 2

Fig. 3 Captions and their corresponding label maps, where (**a**) and (**c**) are the English and Chinese caption area extracted from one frame, and (**b**) and (**d**) are their corresponding labels

Text-line construction

Since the proposed framework is a text-line-level caption detector, text-lines can be constructed by directly grouping continuous detected words or characters in each line via heuristic strategies, respectively. Firstly, we sort all the word/character boxes by their y coordinates and calculate each gap distance between two continuous boxes. Secondly, according to the gap values we use a threshold to make these boxes split into several vertical groups which means there are as many as text lines as groups. Next, in each vertical groups, we iteratively perform the same procedure like before. Thirdly, we directly group the word box groups got in Step 2 into text lines by a bounding box. Finally, we will get all the text lines for the input image.

Results

A variety of experiments are conducted on two different tasks, caption text-line detection, and word extraction (segmentation) English and Chinese characters. These experimental results show that the proposed framework is effective for detecting and segmenting text from complex web videos.

Datasets

In order to perform qualitative and quantitative experiments, the proposed framework is evaluated on three representative and comprehensive datasets (Dataset I, II, and III). These three datasets have 15 video sequences respectively. Dataset I is composed of video sequences whose captions are English. Dataset II is composed of video sequences whose captions are Chinese. Dataset III is a bilingual caption video sequences that contain 42% English and 58% Chinese captions at the same time. Note that all video sequences in these three datasets are subsampled every 10 frames. To demonstrate the effectiveness, the three datasets covers Hollywood movies, animations, news and entertainment shows, where captions lie on different locations in a frame, and a variety of font styles and sizes are involved. These 30 video sequences in three datasets total about 20 min, contain 57282 frame images and 5367 characters, and have a resolution that varies between 496×352 and 1920×1280. Typical frame samples are shown in Fig. 4.

Implementation details

In our experiments, the proposed deep CNN model is optimized by stochastic gradient descent(SGD) with a mini-batch size of 8. The input of the network is 512×512 cropped images from the training data. The step learning rate policy with gamma is set as 0.5 and the step size is 20,000 in Caffe [24]. The weight decay and momentum are set to 5×10^{-4} and 0.9, respectively. The network runs 30000 iterations for training. Here, we do not set the learning rate of the deconvolutional layer to 0 like [14], and actually set its weights initialized by the bilinear method.

All the experiments are conducted on Caffe and run on a workstation with 3.5 GHz 12-core CPU, 128 G RAM, GTX1080 GPU and Manjaro Linux 64-bit OS. The whole training time is about 6 h.

Results for text-line localization

In our experiments, we use the metric for text detection with the text bounding box, where the percentage of overlapping of the text bounding box area in the ground truth and

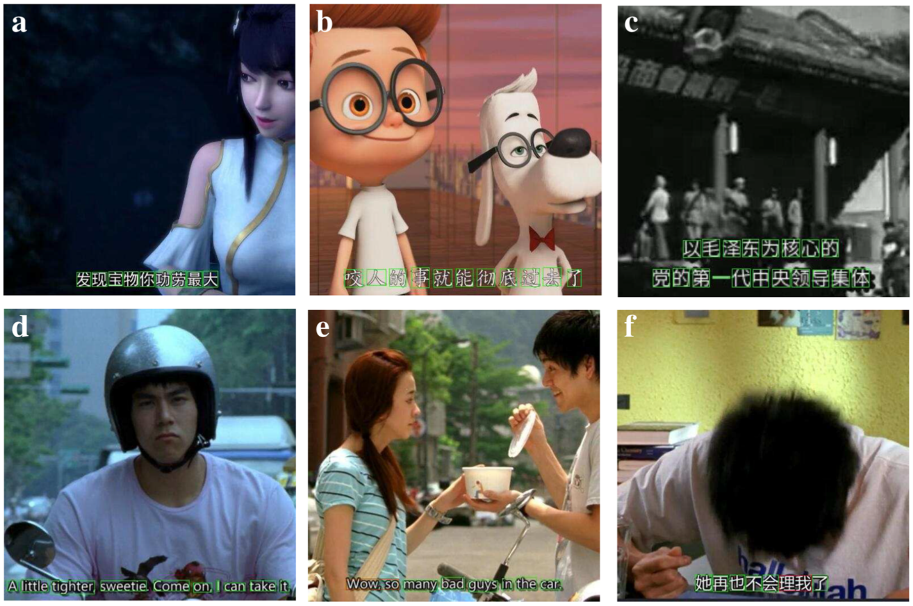


Fig. 4 Typical examples for word/character extraction, where (Chinese character or English word) detection results are indicated by bounding boxes: (**a**)-(**c**) Chinese text-line examples, (**d**)-(**e**) English text-line examples, and (**f**) a wrong case for character extraction (with a red bounding box)

the experimental results determines recall and precision. For our proposed method, the text-line region is constructed by detected words and characters. We compare the proposed method with MSER-based method proposed by Yin et al. [9]. As shown in Table 1, our method with PVANET or ResNet-50 as the feature extracted network obtains significant improvements on recall, and obtains the increase of 4.21%, 4.69% and 3.78% in $F1$-measure on three datasets, respectively.

Results for word/character extraction

We also evaluate the word/character extraction (segmentation) performance with the ground truth manually obtained at the pixel level. In the same way, two metrics (Precision and Recall) are used in the evaluation, i.e., $Precision = N_{rp}/N_{gp}$, and $Recall = N_{rp}/N_{tp}$, where N_{rp} is the number of text pixels segmented correctly, N_{gp} is the number of text pixels in ground truth, and N_{tp} is the total number of text pixels segmented by the proposed method. Both English and Chinese captions follow the same steps. Here, we compare our results with the segmentation results generated by Yin et al. method as well [9]. Like most text detection methods, the approach of Yin et al. segments the detected text line and localizes the pixels for the detected text. Table 2 presents the corresponding results of word/character segmentation results.

Table 1 Text-line detection (text detection) performance comparison with Recall (%), Precision (%), and F1-measure (%) on three datasets of web videos

Method	Dataset I			Dataset II			Dataset III		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Yin et al.	88.24	82.28	85.15	86.71	83.73	85.19	86.42	82.67	84.50
Ours+ResNet-50	87.79	88.32	87.46	86.63	89.51	88.05	86.11	89.03	87.54
Ours+PVANET	86.13	92.85	89.36	87.31	92.62	89.88	85.11	91.70	88.28

Table 2 Word/character segmentation comparisons on three datasets of web videos (%)

Method	Dataset I			Dataset II			Dataset III		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Yin et al.	79.82	74.28	76.95	80.15	75.38	77.69	80.02	74.37	77.09
Ours+ResNet-50	82.33	84.85	83.57	82.39	85.62	83.97	82.11	81.88	81.99
Ours+PVANET	83.02	84.97	83.98	80.63	86.20	83.32	81.98	85.65	83.74

In the comparison experiments on the same datasets, the proposed approach has the better performance than MSER-based methods. our method with PVANET (as the feature extracted network) obtains an increase of 7.03%, 5.63% and 6.65% in F1-measure on three datasets, respectively. The results has also demonstrated the strength of pixel-level prediction. Moreover, typical detection examples under various challenging caseson the three datasets are also shown in Fig. 4.

Discussions

For many conventional text segmentation approaches, there will be error accumulation during the denoise, binarization and other sophisticated post-processing steps. As for text-line detection and segmentation methods, e.g., our proposed method, this problem can be easily solved. Furthermore, we can get much more precise position of text-lines, which was demonstrated in the text-line Localization task. Through observation of experimental results, we also found that the proposed approach is much more robust in cases of low resolutions and complex backgrounds for web videos.

Benefiting from GPU acceleration, for a 512 × 512 image input, the proposed method achieves 21 *fps* in average on GTX1080, which is nearly real-time on GPU. We also empirically check the results for text detection and segmentation from web videos, and find that failed cases are always occurred when words in different lines are grouped. In general, words are easily grouped together if these words are surrounded by other text regions in web videos. This is also one topic for the future research.

Conclusions

This paper proposes a novel Chinese text-line caption extraction (detection and segmentation) pipeline with fully convolutional networks. This method can directly detect the English word and Chinese characters in the caption text-lines without word or character segmentation. The outputs of the proposed framework can directly and easily be fed to OCR for text recognition. Moreover, the proposed framework validates its high performance on several real-world web video datasets. In the future, we will focus on extending the proposed framework for multi-oriented text detection, and combine the networks of detection and recognition into one unified end-to-end framework.

Acknowledgements
This research has been supported by University of Science and Technology Beijing and funded by National Natural Science Foundation of China.

Funding
The research is partly supported by National Natural Science Foundation of China (61473036).

Availability of data and materials

Please contact author for data requests.

Authors' contributions

Conceived the research: X-CY. Designed the experiments: CY and W-YP. Performed the experiments: W-YP and L-HW. Analyzed the data: CY, W-YP and L-HW. Contributed materials/analysis tools: W-YP and L-HW. Wrote the paper: CY, L-HW, W-YP and X-CY. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 December 2017 Accepted: 27 December 2017
Published online: 05 January 2018

References

1. Yin XC, Zuo ZY, Tian S, Liu CL. Text detection, tracking and recognition in video: A comprehensive survey. IEEE Trans Image Process. 2016;25(6):2752–73.
2. Tian S, Yin XC, Su Y, Hao HW. A unified framework for tracking based text detection and recognition from web videos. IEEE Trans Pattern Anal Mach Intell. 2017. Accepted and published online (https://doi.org/10.1109/TPAMI.2017.2692763).
3. Lienhart R, Wernicke A. Localizing and segmenting text in images and videos. IEEE Trans Circ Syst Video Technol. 2002;12(4):256–68.
4. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. Montreal: Neural Information Processing Systems Foundation, Inc. 2015. p. 91–9.
5. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. Ssd: Single shot multibox detector. In: European Conference on Computer Vision. Amsterdam: Springer. 2016. p. 21–37.
6. Redmon J, Divvala SK, Girshick RB, Farhadi A. You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR. Las Vegas: IEEE Computer Society. 2016. p. 779–88.
7. Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR. Honolulu: IEEE Computer Society. 2017.
8. Tian Z, Huang W, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: European Conference on Computer Vision. Amsterdam: Springer. 2016. p. 56–72.
9. Yin XC, Yin X, Huang K, Hao HW. Robust text detection in natural scene images. IEEE Trans. Pattern Analysis and Machine Intelligence. 2014;36(5):970–83.
10. Yin XC, Pei WY, Zhang J, Hao HW. Multi-orientation scene text detection with adpative clustering. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1930–7.
11. Huang W, Qiao Y, Tang X. Robust scene text detection with convolution neural network induced mser trees. In: European Conference on Computer Vision. Zurich: Springer. 2014. p. 497–511.
12. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Reading text in the wild with convolutional neural networks. Int J Comput Vis. 2016;116(1):1–20.
13. Ma J, Shao W, Ye H, Wang L, Wang H, Zheng Y, Xue X. Arbitrary-oriented scene text detection via rotation proposals. 2017. arXiv preprint arXiv:1703.01086.
14. Zhang Z, Zhang C, Shen W, Yao C, Liu W, Bai X. Multi-oriented text detection with fully convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society. 2016. p. 4159–167.
15. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):640–51.
16. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE Computer Society. 2015. p. 1520–1528.
17. Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J. East: An efficient and accurate scene text detector. 2017. arXiv preprint arXiv:1704.03155.
18. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–52.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society. 2016. p. 770–8.
20. Kim KH, Hong S, Roh B, Cheon Y, Park M. Pvanet: Deep but lightweight neural networks for real-time object detection. 2016. arXiv preprint arXiv:1608.08021.
21. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society. 2016. p. 761–9.

22. Xie S, Tu Z. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE Computer Society. 2015. p. 1395–1403.
23. Yao C, Bai X, Sang N, Zhou X, Zhou S, Cao Z. Scene text detection via holistic, multi-channel prediction. 2016. arXiv preprint arXiv:1606.09002.
24. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. Orlando: ACM. 2014. p. 675–8.