Signal Processing: Image Communication 60 (2018) 131–143

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

End-to-end subtitle detection and recognition for videos in East Asian languages via CNN ensemble

Yan Xu [a,b], Siyuan Shan [a], Ziming Qiu [c], Zhipeng Jia [b,d], Zhengyang Shen [e], Yipei Wang [a], Mengfei Shi [f], Eric I-Chao Chang [b,*]

[a] State Key Laboratory of Software Development Environment and Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education and Research Institute of Beihang University in Shenzhen, Beijing Advanced Innovation Centre for Biomedical Engineering, Beihang University, Beijing 100191, China
[b] Microsoft Research Asia, Beijing 100080, China
[c] Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, USA
[d] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[e] Department of Computer Science, University of North Carolina at Chapel Hill, USA
[f] Beijing No.8 High School, Beijing 100032, China

ARTICLE INFO

Keywords:
Subtitle text detection
Subtitle text recognition
Synthetic training data
Convolutional neural networks
Video sequence information
East Asian language

ABSTRACT

In this paper, we propose an innovative end-to-end subtitle detection and recognition system for videos in East Asian languages. Our end-to-end system consists of multiple stages. Subtitles are firstly detected by a novel image operator based on the sequence information of consecutive video frames. Then, an ensemble of Convolutional Neural Networks (CNNs) trained on synthetic data is adopted for detecting and recognizing East Asian characters. Finally, a dynamic programming approach leveraging language models is applied to constitute results of the entire body of text lines. The proposed system achieves average end-to-end accuracies of 98.2% and 98.3% on 40 videos in Simplified Chinese and 40 videos in Traditional Chinese respectively, which is a significant outperformance of other existing methods. The near-perfect accuracy of our system dramatically narrows the gap between human cognitive ability and state-of-the-art algorithms used for such a task.

1. Introduction

Detecting and recognizing video subtitle texts in East Asian languages (e.g. Simplified Chinese, Traditional Chinese, Japanese and Korean) is a challenging task with many promising applications like automatic video retrieval and summarization. Different from traditional printed document OCR, recognizing subtitle texts embedded in videos is complicated by cluttered backgrounds, diversified fonts, loss of resolution and low contrast between texts and backgrounds [1].

Given that video subtitles are almost always horizontal, subtitle detection can be partitioned into two steps: subtitle top/bottom boundary (STBB) detection and subtitle left/right boundary (SLRB) detection. These four detected boundaries enclose a bounding box that is likely to contain subtitle texts. Then the texts inside the bounding box are ready to be recognized.

Despite the similarity between video subtitle detection and scene text detection, the instinctive sequence information of videos makes it necessary to address these two tasks respectively [2]. As illustrated in Fig. 1, for most videos with single-line subtitles in East Asian languages, texts at the subtitle region exhibit homogeneous properties throughout the video, including consistent STBB position, color and single character width (SCW). Meanwhile, the non-subtitle region varies unpredictably from frame to frame. With the assistance of this valuable sequence information, we put forward a suitable image operator that can facilitate the detection of STBB and SCW. We call this image operator the *Character Width Transform* (CWT), as it exploits one of the most distinctive features of East Asian characters—consistent SCW.

Considering the complexity of backgrounds and the diversity of subtitle texts, adopting a high-capacity classifier for both text detection and recognition is imperative. CNNs have most recently proven their mettle handling image text detection and recognition [3,4]. By virtue of their special bio-inspired structures (i.e. local receptive fields, weight sharing

[*] Corresponding author.
E-mail addresses: xuyan@buaa.edu.cn (Y. Xu), shansiliu@outlook.com (S. Shan), zq415@nyu.edu (Z. Qiu), v-zhijia@microsoft.com (Z. Jia), zyshen@unc.cs.edu (Z. Shen), yipeiwang@buaa.edu.cn (Y. Wang), shimengfei2012@outlook.com (M. Shi), echang@microsoft.com (E.I.-C. Chang).
Abbreviations: STBB, subtitle top/bottom boundary; SLRB, subtitle left/right boundary; CWT, Character Width Transform; SCW, single character width

https://doi.org/10.1016/j.image.2017.09.013
Received 8 April 2017; Received in revised form 30 September 2017; Accepted 30 September 2017
Available online 16 October 2017
0923-5965/© 2017 Elsevier B.V. All rights reserved.

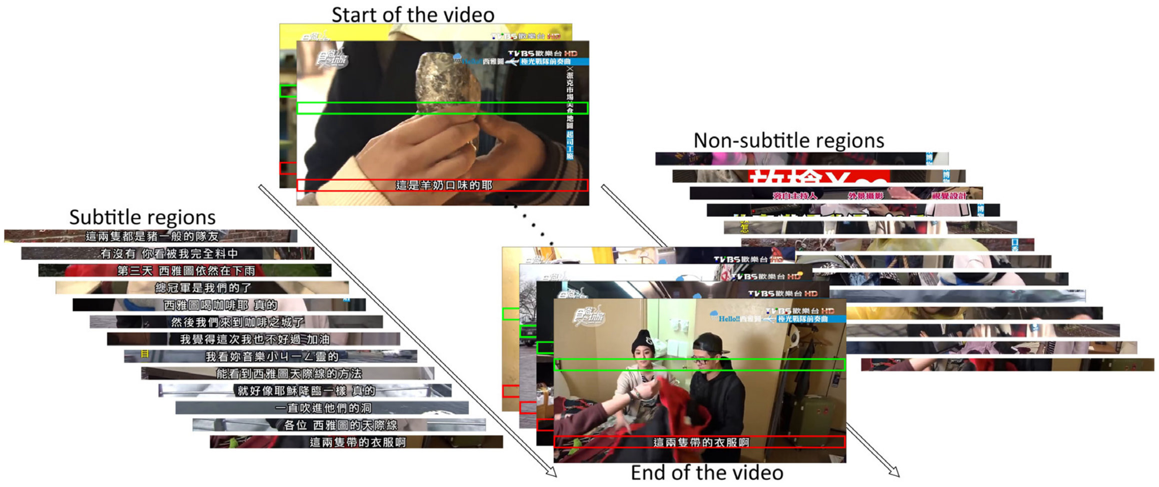

Fig. 1. Illustration of the consistent STBB position throughout the video. The red box denotes the subtitle region, while the green box denotes the non-subtitle region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and sub-sampling), CNNs are extremely robust to noise, deformation and geometric transformations [5] and thus are capable of recognizing characters with diverse fonts and distinguishing texts from cluttered backgrounds. Besides, the architecture of CNNs enables efficient feature sharing across different tasks: features extracted from hidden layers of a CNN character classifier can also be used for text detection [4]. Additionally, the fixed input size of typical CNNs makes them especially suitable for recognizing East Asian characters whose SCW is consistent.

In view of the straightforward generation pipeline of video subtitles, it is technically feasible to obtain training data by simulating and recovering this generation pipeline. To be more specific, when equipped with a comprehensive dictionary, several fonts and numerous random backgrounds, machines can produce huge volumes of synthetic data covering thousands of characters in diverse fonts without strenuous manual labeling. As a cornucopia of synthetic training data meet the "data-hungry" nature of CNNs, models trained merely on synthetic data can achieve competitive performance on real-world datasets.

Another observation is that the recognition performance degrades with the burgeoning number of character categories (as in the case of East Asian languages). In a similar circumstance, Jaderberg et al. [6] attempt to alleviate this problem with a sophisticated incremental learning method. Here we propose a more straightforward solution: instead of using a single CNN, we independently train multiple (ten in this paper) CNN models that consolidate a CNN ensemble. These models are complementary to each other, as the training data is shuffled respectively for training different models.

In this paper, by seamlessly integrating the above-mentioned cornerstones, we propose an end-to-end subtitle text detection and recognition system specifically customized to videos with a large concentration of subtitles in East Asian languages. Firstly, STBB and SCW are detected based on a novel image operator with the sequence information of videos. SCW being determined at an early stage can provide instructive information to improve the performance of the remaining modules in the system. Afterwards, SLRB is detected by a SVM text/non-text classifier (it takes CNN features as input) and a horizontal sliding window (its width is set to SCW). According to the detected top, bottom, left and right boundaries, the video subtitle is successfully detected. Finally, single characters are recognized by the CNN ensemble and the text line recognition result is determined by a dynamic programming algorithm leveraging a 3-gram language model. We show that the CNN ensemble produces a recognition accuracy of 99.4% on a large real-world dataset including around 177,000 characters in 20,000 frames. This dataset with ground truth annotations has been made publicly available.[1]

Our contribution can be summarized as follows:

- We propose an end-to-end subtitle detection and recognition system for East Asian languages. By achieving 98.2% and 98.3% end-to-end recognition accuracies for Simplified Chinese and Traditional Chinese respectively, this system remarkably narrows the gap to human-level reading performance.[2]
- We define a novel image operator whose outputs enable the effective detection of STBB and SCW. The sequence information is integrated throughout the video to increase the reliability of the proposed image operator. This module achieves a competitive result on a dataset including 1097 videos.
- We leverage a CNN ensemble to perform the classification of East Asian characters across huge dictionaries. The ensemble reduces the recognition error rate by approximately 75% in comparison with a single CNN. CNNs in our system serve both as text detectors and character recognizers.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 describes the synthetic data generation scheme, the CNN ensemble and the end-to-end system. In Section 4, the proposed system and each module in it are evaluated on a large dataset, and the experimental results are presented. In Section 5, observations from our experiments are discussed. A conclusion and discussion of future work are given in Section 6.

2. Related work

In this section, we focus on reviewing relevant literature on image text detection and recognition. As for other text detection and recognition methods, several review papers [1,7–10] can be referred to.

2.1. Image text detection

Generally, text detection methods are based on either connected components or sliding windows [4]. Connected component based methods, like Maximally Stable Extremal Regions (MSER) [11–13], enjoy their computational efficiency and high recall rates, but suffer from a large number of false detections. Methods based on sliding windows [3,4,14–17] adopt a multi-scale window to scan through all locations of an image, then apply a trained classifier with either hand-engineered features or learned features to distinguish texts from non-texts. Though this kind of method produces significantly less false

[1] https://drive.google.com/file/d/0B0x5IW_m4AC5M0RuY1JiUWJIcUU/view?usp=sharing.

[2] Human-level reading performance is 99.6% according to the experiment in Section 4.1.