Postprint of Uyghur Text Detection in Natural Scenes Based on an Improved Single Deep Neural Network

Authors: Peng Yong, Halidan Abudureyimu, Ding Weichao

Date: 2018-05-24T00:00:00+00:00

Abstract

To address the challenge of Uyghur text detection in natural scenes, an improved single deep neural network is proposed. The network architecture consists of a Uyghur text feature extraction component and a multi-layer feature fusion-based text detection component, trained in an end-to-end manner to predict the positions of Uyghur text bounding boxes and their confidence scores. The Uyghur text feature extraction component utilizes convolutional neural networks to extract multi-scale and multi-level Uyghur text features from natural scene images containing Uyghur text. The multi-layer feature fusion-based text detection component then employs the features extracted by the Uyghur text feature extraction component to predict the positions of text bounding boxes and the confidence scores of Uyghur text categories. Analysis reveals that, unlike Chinese and English text detection, Uyghur text exhibits more distinctive characteristics. To accommodate these characteristics, default boxes with multiple aspect ratios and sizes were designed, and the sizes of some convolutional kernels were adjusted. Experiments on a dataset of images containing Uyghur text in natural scenes demonstrate that the improved single deep neural network method considers the impact of multi-scale and multi-level features on detection accuracy, achieving accuracy and F-score of 0.7234 and 0.6115, respectively, thereby improving detection accuracy.

Full Text

Abstract

To address the challenges of detecting Uyghur text in natural scenes, this paper proposes an improved single deep neural network for Uyghur text detection. The network architecture comprises a Uyghur feature extraction component and a multi-layer feature fusion text detection component, trained end-to-end

to predict the location and confidence of Uyghur text bounding boxes. The feature extraction component utilizes convolutional neural networks to extract multi-scale and multi-level Uyghur features from natural scene images. The detection component then employs these extracted features to predict text box positions and category confidence scores. Analysis reveals that Uyghur text possesses more distinctive characteristics compared to Chinese and English text. To accommodate these features, we design default boxes with multiple aspect ratios and scales, and adjust the size of certain convolution kernels. Experiments on a dataset of natural scene images containing Uyghur text demonstrate that the improved single deep neural network effectively leverages multi-scale and multi-level image features to enhance detection accuracy, achieving precision and F-value of 0.7234 and 0.6115, respectively.

Keywords: Uyghur text detection; single deep neural network; multi-scale features

0 Introduction

Text serves as the most direct representation of high-level semantic information in images and plays an indispensable role in image understanding. Text detection in natural scenes has long been a challenging task in the field of optical character recognition. The difficulties include non-uniform illumination, low contrast, partial occlusion, multi-oriented text, and perspective distortion caused by camera angles, as illustrated in [Figure 1: see original paper].

Uyghur text detection faces additional challenges specific to the script. Uyghur is an agglutinative language with subject-object-verb (SOV) word order. Nouns inflect for number and case, with six grammatical cases: nominative, accusative, dative, genitive, ablative, and locative. The Uyghur alphabet consists of 32 letters (8 vowels and 24 consonants) with 128 character forms. Each letter appears in four different positional variants (final, medial, initial, and isolated forms) depending on its position within a word. Uyghur words are formed by connecting these characters into ligatures along a horizontal baseline. Furthermore, some Uyghur letters share the same basic shape and are distinguished only by dots placed above, below, or inside the character. These dots can be easily mistaken for noise [2]. The baseline feature represents a particularly distinctive characteristic of Uyghur script, as shown in [Figure 2: see original paper].

1 Related Work

Existing approaches to address these challenges can be categorized into three main groups: texture-based methods, connected component-based methods, and hybrid approaches.

Texture-based methods [3,4] exploit the observation that text regions exhibit distinct texture patterns compared to background areas. These approaches typically employ sliding windows to extract features such as Local Binary Patterns

(LBP) and Histograms of Oriented Gradients (HOG). However, they often lack robustness for multi-oriented text and significant scale variations, and incur high computational costs.

Connected component-based methods [5,6] extract candidate text regions using techniques like edge detection [7], Maximally Stable Extremal Regions (MSER [8–11]), and color-enhanced contrast regions. These candidates are then filtered using specially designed rules or trained classifiers (e.g., SVM) to eliminate non-text components.

For Uyghur text detection specifically, Fang et al. [2] utilized convolutional neural networks to detect Uyghur text in complex background images. Tursun et al. [12] employed Harris corner detection and mathematical morphology to generate candidate text regions, applying heuristic rules to remove typical non-text areas and leveraging the baseline feature of Uyghur text for candidate validation. Li et al. [13] used corner detection in homogeneous image space to rapidly obtain candidate text regions and proposed an improved LBP feature for classifying Uyghur text candidates, though this method' s effectiveness depends heavily on the completeness of the initial corner detection. Liu et al. [14] established a baseline structural feature based on detected texture and edge features, but this approach is vulnerable to complex backgrounds and unknown noise in natural scene images that can affect binarization and degrade baseline feature extraction, resulting in reduced detection accuracy.

2 Algorithm Design

The overall framework of our natural scene Uyghur text detection system is shown in [Figure 3: see original paper]. The process involves: (1) acquiring natural scene images containing Uyghur text and annotating the text regions; (2) feeding the samples into the improved single deep neural network' s feature extraction component to extract raw image features; (3) passing these features to the multi-layer feature fusion detection module for Uyghur text presence determination and localization.

The main contributions of this work are: a) An improved single deep neural network architecture that extracts multi-level and multi-scale features from natural scene Uyghur text images. b) Design of multi-scale, multi-aspect-ratio default boxes and adjustment of certain convolution kernel sizes to accommodate the characteristics of Uyghur text lines in natural scenes.

We adopt the single deep neural network structure from TextBoxes [15], which achieved good performance in English text detection, and enhance it by adding a text box layer, redesigning the aspect ratios of candidate boxes, and adjusting convolution kernel sizes to suit Uyghur text line detection. Our improved single deep neural network employs a feature pyramid detection approach that integrates feature maps from different convolutional layers for separate detection, then combines results through non-maximum suppression (NMS) to directly

predict object categories and bounding boxes. The overall network architecture is shown in [Figure 4: see original paper].

The network consists of 28 fully convolutional layers. The first 13 layers are from VGG-16, followed by 11 additional convolutional layers. Text box layers are connected to 6 convolutional layers. At each feature map location, a text box layer predicts a 30-dimensional vector comprising text presence scores and offsets for 5 default boxes. NMS is applied in the final stage to fuse all text box outputs.

The multi-layer feature fusion detection component is the key part of our improved network. Conditioned on input feature maps, it simultaneously predicts Uyghur text presence and text box locations. At each feature map location, it outputs classification scores and position offsets for associated default boxes through convolutional operations.

We utilize features extracted from conv4_3, conv7, conv8_2, conv9_2, conv10_2, and conv11_2 to generate 5 types of default boxes at each layer. These default boxes are mapped back to the original image. For a convolutional feature map of size W×H, the layer generates 5×W×H default boxes, as illustrated in [Figure 5: see original paper]. Two different 3×5 convolution kernels are applied to the output feature maps of these six convolutional layers. This rectangular receptive field design better fits Uyghur text' s larger aspect ratios. One branch outputs classification confidence (2 categories: Uyghur text and background), while the other outputs regression coordinates (4 values per default box).

The training objective function combines localization loss and confidence loss:

$$\text{loss}(x, c, l, g) = \frac{1}{N}(\text{loss}_{\text{conf}}(x, c) + \alpha \cdot \text{loss}_{\text{loc}}(x, l, g))$$

where the localization loss is:

$$\text{loss}_{\text{loc}}(x, l, g) = \sum_{i \in \text{pos}}^{N} \sum_{m \in \{c_x, c_y, w, h\}} x_{ij}^{k} \cdot \text{smooth}_{L1}(l_i^m - \hat{g}_j^m)$$

with: - $\hat{g}_j^{c_x} = (g_j^{c_x} - d_i^{c_x})/d_i^w$ - $\hat{g}_j^{c_y} = (g_j^{c_y} - d_i^{c_y})/d_i^h$ - $\hat{g}_j^w = \log(g_j^w/d_i^w)$ - $\hat{g}_j^h = \log(g_j^h/d_i^h)$

The confidence loss is:

$$\text{loss}_{\text{conf}}(x, c) = -\sum_{i \in \text{pos}}^{N} x_{ij}^{p} \log(\hat{c}_i^p) - \sum_{i \in \text{neg}} \log(\hat{c}_i^0)$$

where $\hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)}$.

Here, N represents the number of matched default boxes, g denotes ground-truth box parameters, α is a balance factor set to 1, l represents predicted boxes, d represents default boxes, (c_x, c_y) are default box center coordinates, and w, h are default box width and height. $x_{ij}^p = 1$ indicates that the i-th default box matches the j-th ground-truth box of class p; otherwise $x_{ij}^p = 0$.

2.1 Dataset Construction

Due to the lack of publicly available annotated Uyghur text datasets in natural scenes, we constructed our own dataset collected by the Pattern Recognition and Intelligent Control Laboratory of Xinjiang University' s Electrical Engineering College from street views in Urumqi, with a small portion downloaded from the internet. The images include billboards, signs, road markers, and bulletin boards. Most text is Uyghur, mixed with small amounts of Chinese and minimal English/Arabic numerals. Text varies in size, color, and location. Since the network requires uniform input sizes, all images were resized to 350×500 pixels.

The original 660 images were randomly split into 550 training samples and 110 test samples. Data augmentation was applied to the training set, yielding 3,930 augmented images. All images were annotated using the Label-Image tool, with Uyghur text coordinates manually labeled for evaluation. The final dataset contains 4,590 JPG images with 4,590 XML label files. Training and test set details are shown in .

Data Augmentation: Original images were transformed through color variations, contrast changes, and addition of Gaussian noise at four different scales.

Training Process: Training images were uniformly resized to 350×500 pixels with random horizontal flipping. Stochastic Gradient Descent (SGD) was used with weight decay 0.0005, momentum 0.9, initial learning rate 0.01 (reduced by 0.1 every 40,000 iterations). Sample training and test images are shown in [Figure 6: see original paper] and [Figure 7: see original paper].

Matching Strategy: Ground-truth boxes are matched with default boxes based on Intersection over Union (IoU) overlap. Pairs with IoU > 0.6 are considered positive samples (Uyghur text), while unannotated Chinese/English regions are treated as negative samples (background).

Hard Negative Mining: After matching, we control the negative-to-positive sample ratio (3:1) by sorting negative predictions by confidence and selecting the top ones, ensuring stable and efficient optimization.

3 Experiments and Analysis

We formulate Uyghur text detection as a binary classification problem (text vs. background) with bounding box regression, adopting a fine-tuning strategy on VOC2007-format data.

Experimental Platform: Ubuntu 16.04, Caffe framework, 8GB RAM, Intel i7 CPU, NVIDIA GeForce GTX 1080 GPU with 8GB memory.

3.1 Dataset and Evaluation Metrics

Common evaluation metrics for object detection include:

$$\text{Precision} = \frac{\text{Number of correctly detected text boxes}}{\text{Total number of detected text boxes}}$$

$$\text{Recall} = \frac{\text{Number of correctly detected text boxes}}{\text{Total number of text boxes in dataset}}$$

$$F\text{-value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.2 Feature Extraction

CNNs exhibit translation invariance, scale invariance, and robustness to various distortions in object detection. Since feature detection layers learn from training data, manual feature design and extraction are unnecessary.

[Figure 8: see original paper] shows a test image, while [Figure 9: see original paper] visualizes feature maps from various convolutional layers during detection. Different layers capture features at different scales: conv1_1 through conv2_2 learn basic features like color and edges (low-level global features), while deeper layers extract high-level local features.

[Figure 10: see original paper] displays learned convolution kernels from our trained network, demonstrating that the kernels are automatically learned from training data rather than manually designed.

3.3 Results Analysis

Training results are summarized in , showing average detection precision versus iteration count. Overall precision increases with training iterations, reaching approximately 0.57 at peak.

Impact of Network Structure: Performance comparison using different feature map layers for text box prediction is shown in (IoU = 0.5). Adding conv11_2 features improves prediction performance by incorporating more multi-level and multi-scale features, along with appropriately designed default boxes and adjusted kernel sizes.

Method Comparison: compares different methods on our dataset. While our recall is lower than Faster-RCNN, our F-value and precision are higher than both Faster-RCNN and TextBoxes at IoU thresholds of 0.5 and 0.6, demonstrating superior overall performance. Sample successful detections are shown in [Figure 11: see original paper].

4 Conclusion

This paper presents an improved single deep neural network for Uyghur text detection in natural scenes. Experiments demonstrate that the improved network effectively leverages multi-scale and multi-level image features to enhance detection accuracy. Future work will focus on: (1) developing polygonal bounding box regression to handle multi-oriented text and reduce accuracy loss from rectangular box regression; (2) fusing traditional features with deep neural network features for Uyghur text detection.

References

[1] Eli Jume, Halidan A, Huang Hao. Recognition of extracting Uyghur texts from videos images [J]. Computer Engineering and Application, 2011, 47(36): 190-192.

[2] Fang Shancheng, Xie Hongtao, Chen Zhineng, et al. Detecting Uyghur text in complex background images with convolutional neural network [J]. Multimedia Tools & Applications, 2017, 76(13): 1-21.

[3] Li Huiping, Doermann D, Kia O. Automatic text detection and tracking in digital video [J]. IEEE Trans on Image Processing, 2000, 9(1): 147-56.

[4] Lee J J, Lee P H, Lee S W, et al. AdaBoost for text detection in natural scene [C]// Proc of IEEE International Conference on Document Analysis and Recognition. 2011: 429-434.

[5] Yao Cong. Detecting texts of arbitrary orientations in natural images [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Computer Society, 2012: 1083-1090.

[6] Huang Weilin, Lin Zhe, Yang Jianchao, et al. Text localization in natural images using stroke feature transform and text covariance descriptors [C]// Proc of IEEE International Conference on Computer Vision. 2014: 1241-1248.

[7] Lu Shijian, Chen Tao, Tian Shangxuan, et al. Scene text extraction based on edges and support vector regression [J]. International Journal on Document Analysis & Recognition, 2015, 18(2): 125-135.

[8] Kethineni V, Velaga S M. Text detection on scene images using MSER [J]. International Journal of Research in Computer and Communication Technology, 2015, 4(7): 2320-5156.

[9] Liu Shun, Xie Hongtao, Yin Jian, et al. Uyghur language text detection in images [C]// Proc of the 8th International Conference on Digital Image Processing. International Society for Optics and Photonics. 2016: 1003345.

[10] Liu Shun, Xie Hongtao, Zhou Chuan, et al. Uyghur language text detection in complex background images using enhanced MSERs [C]// Proc of International Conference on Multimedia Modeling. Cham: Springer, 2017: 385-395.

[11] Ha Ennan, Ji Lixin, Gao Chao. Scene text detection based on object proposals [J/OL]. Application Research of Computers, 2018(01): 1-7 [2018-02-28]. http://sq.lib.xju.edu.cn:80/rwt/CNKI/http/NNYHGLUDN3WXTLUPMW4A/kcms/detail/51.1196.TP.20170

[12] Tursun Dilmurat, Li Kai, Gulxax Halelbek, et al. A joint approach of harris corners detection and baseline searching for localization of Uyghur text lines in image sequences [J]. Journal of Information Hiding & Multimedia Signal Processing, 2016, 7(2): 352-361.

[13] Li Mingqing, Halidan · Abudureyimu, Yan Ke. An Uyghur text algorithm based on improved Local Binary Pattern Features [J]. Journal of Henan University of Science and Technology: Natural Science, 2015, 36(3): 43-47.

[14] Liu Chang, Song Yifan, Zhao Zhicheng, et al. Fast Uyghur text detection in videos based on learning of baseline feature [C]// Proc of Visual Communications and Image Processing. 2016: 1-4.

[15] Liao Minghui, Shi Baoguang, Bai Xiang, et al. TextBoxes: a fast text detector with a single deep neural network [EB/OL]. (2016-11-21). arXiv: 1611.06779.

[16] Ren Shaoqing, He Kaiming, R Girshick, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

Note: Figure translations are in progress. See original paper for figures.