21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Scene Character Detection and Recognition Based on Multiple Hypotheses Framework

Rong Huang[1], Shinpei Oba[1], Shivakumara Palaiahnakote [2] and Seiichi Uchida[1]
[1]*Kyushu University, Fukuoka, Japan*
[2]*National University of Singapore, Singapore*

Abstract

To handle the diversity of scene characters, we propose a multiple hypotheses framework which consists of an image operator set module, an optical character recognition (OCR) module, and an integration module. Image operators detect multiple suspicious character areas. The OCR engine is then applied to each detected area and returns multiple candidates with weight values for future integration. Without the aid of heuristic constraints on area, aspect ratio or color etc., the integration module prunes the redundant detection and pads the missing detection based on the outputs of OCR. The experimental results demonstrate that the whole multiple hypotheses outperforms each operator's hypotheses and be comparable with existing methods in terms of recall, precision, F-measure and recognition rate.

1. Introduction

Scene character detection and recognition are well known difficult tasks since the character may suffer from background interferences, nonuniform lighting condition and shadow. It has been proved by experiment [1] that, in most situations, directly using the off-the-shelf optical character recognition (OCR) engines lead to abortive or incorrect results. Therefore, many researchers argued that the critical and imperative issue was to develop efficient preprocessing approaches geared towards OCR.

Along this line of consideration, a variety of impressive methods have been proposed as efforts to narrow the gap between capability of OCR and complexity of scene character. The approach proposed by Chen *et al.* [2] located the horizontal text regions and then segmented each textline into multiple binary images. The confidence value was definitively computed relying on language modeling and OCR statistics. Goto [3] employed Fisher's discriminant ratio to assess the DCT-based features. The experimental results revealed that there is actually a dependency between the frequency threshold and the width of character strokes. Wang *et al.* [4] extended the generic object recognition methods to the problem of spotting words in the wild. The word configuration was confirmed by optimizing a cost objective function. Recently, Yi *et al.* [5] designed the adjacent character and textline grouping algorithm to detect text strings with arbitrary orientations.

In this paper, inspired by the fact that hybrid is always endowed with robustness, we deviate from the routine of previous works and develop a novel multiple hypotheses framework which constitutes of an image operator set module, an OCR engine module, and an integration module. Multiple character detection results are produced by multiple image operators through changing their parameter. Different from many existing works which aim at improving the detection precision, our method performs more practical by introducing the OCR engine which recognizes each detected area and then returns multiple candidates with weight values. Consequently, both scene character detection and recognition can be completed. The integration module combines multiple detection results and draws upon the strength of OCR to achieve the refinement in terms of pruning the redundant detection and padding the missing detection. Benefiting from this cooperation and complementation mechanism, the proposed framework is competent to detect and recognize the scene character. Moreover, this framework works free from heuristic constraints on area, aspect ratio or color and so on.

We evaluate the performance of the proposed framework on the ICDAR 2011 competition dataset [8] and make comparison studies. Experimental results demonstrate that our method is comparable with existing methods and yields remarkable performance, what a single image operator can not achieve.

2. Multiple Hypotheses Framework

2.1. Outline

In this section, we introduce the proposed multiple hypotheses framework and show its structure in Fig.1.

978-4-9906441-1-6

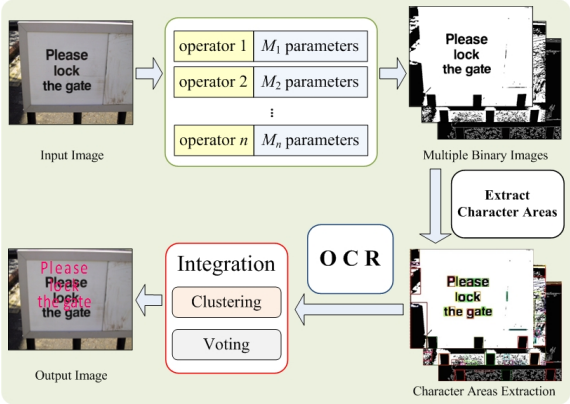

Figure 1. The proposed multiple hypotheses framework.

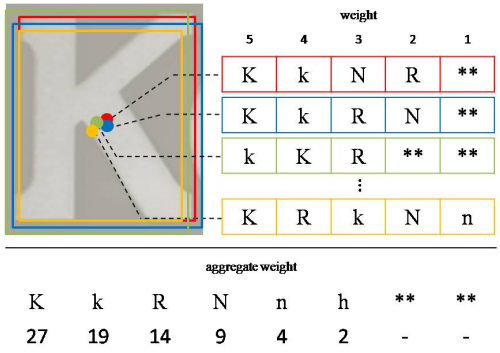

Figure 2. OCR recognition results, weights and aggregate weights.

Each hypothesis is generated from an image operator activated by a parameter. Benefiting from setting multiple parameters, we are liberated from the toil of parameter tuning. The proposed multiple hypotheses framework is so flexible that any image operators can be assigned according to specific application. However, we argue that it is better to select the operators having different characteristics since homogeneous character detection results may attenuate the integration module.

It is worthwhile to point out the difference between our proposal and the multiple hypotheses utilized by literatures [2, 6]. The former classified pixels into K classes and then produces the multiple binary images by assuming that one class corresponded to text and all other classes corresponded to background. Obviously, only one hypothesis marks out characters correctly. Therefore, their integration stage actually selects a best result rather than exerts the effectiveness of the cooperation and complementation mechanism. The latter combined multiple character hypotheses prior to the OCR module. This means that the strong power of OCR is not adequately exploited.

2.2. Details

We design a concrete character detection and recognition method based on the proposed multiple hypotheses framework. Three image operators, namely, fixed-threshold binarization, Niblack binarization and Canny edge detector are employed. Each operator possesses its own characteristics. Fixed-threshold is a global binarization method while Niblack is known as a local counterpart. Niblack determines a pixel threshold relying on the local average and standard deviation. We select Niblack as a member of the operator set since it can still achieve stable binarization even if the original image suffers from low contrast, high complexity or noise. Moreover, Canny edge detector samples the crucial gradient information which can sketch the contours of character. The number of parameter for each image operator, M_1,M_2 and M_3, are set to ten empirically.

The detail procedures are described as follows:

[**Step 1**]: Preprocessing: color to grayscale conversion. Since there exist characters immersing in the surrounding background due to the low contrast, we take Grundland's decolorization approach [7] as counter-measure.

[**Step 2**]: Performing each image operator in turn on the grayscale image with the change of parameter. Under the setting $M_1 = M_2 = M_3 = 10$, we obtain thirty binary images from a grayscale image.

[**Step 3**]: Extracting connected component (CC) and bounding box. The connected component analysis is used to segment a binary image into CCs. Note that for the closed loop characters like "O" or "B", we eliminate their annoying interior bounding boxes.

[**Step 4**]: Using OCR. The CCs are fed into OCR one by one. OCR returns N candidates with weights as shown in Fig.2. In this example, we color the bounding boxes for indicating that they stem from different hypotheses. OCR is forced to produce $N = 5$ candidates with different weights (a larger weight means a higher confidence). The symbol "**" represents non-character.

[**Step 5**]: Integration. Clustering among bounding boxes is performed. A block diagram shown in Fig.3 illustrates this idea clearly. Detailed description is given in the following: At first, we collect bounding boxes from all binary images and number them. After extracting each center coordinate, the Euclidean distance is pairwise computed. Neighbouring bounding boxes are considered to be generated from the same single character area thus being merged into a class. Increase the threshold value α appropriately and repeat the above manipulations until the number of class is no longer changed. After doing so, we get a fused image labeling

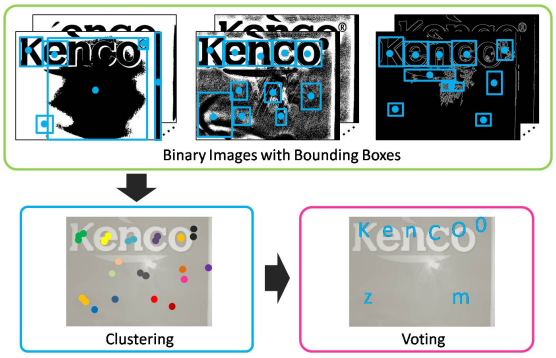

Figure 3. The block diagram of the integration stage.

with the class location. Note that each class contains one or more bounding boxes.

The voting algorithm which sums identical candidate's weight is applied on all classes in turn. If the maximum aggregate weight is greater than a threshold β, its corresponding character is output as final result. Otherwise, this class is eliminated. Note that this elimination is useful to remove non-character class.

This integration module exploits the strength of OCR engine in terms of its highly confident recognition outputs. For an actual character area, OCR returns multiple similar results as shown in Fig.2. However, when recognizing a non-character area, OCR will produce multiple conflicting results so that no aggregate weights can exceed the threshold β. Therefore, our proposal is not only practical but also enables removing the erroneous bounding boxes automatically.

3. Experiment

We evaluated the effectiveness of the proposed multiple hypotheses framework on ICDAR 2011 competition test dataset [8]. In advance, we wiped off the CC whose area was smaller than 10×10. This manipulation is impartial since humans may be hard to perceive the content in such a small area as well. We made performance comparisons between the whole thirty multiple hypotheses and each ten image operator's hypotheses. To this end, when a single operator was activated, the threshold β should be adjusted in inverse proportion to the number of operators. In our case, the threshold became $\frac{1}{3}\beta$.

We measured the performance through four metrics, namely recall r, precision p, F-measure F and recognition rate R. F-measure is the harmonic mean between r and p: $F = 2pr/(p+r)$. Recognition rate is the ratio of the number of correctly recognized character to the number of correctly extracted character region.

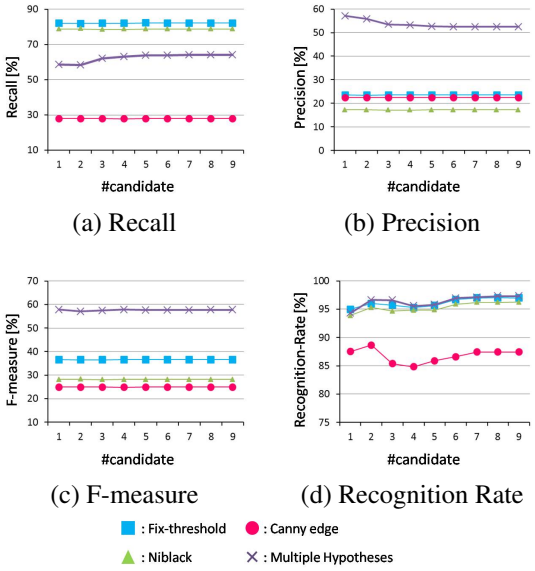

Figure 5. Quantitative evaluation results.

We exhibited the quantitative evaluation results in Fig.5. The horizontal axis represents the candidate number N while the vertical axis stands for the percentage of each metric. Note that each curve in Fig.5 labels the integrated results. For example, each Niblack green triangle is the result integrated from ten hypotheses whereas a multiple hypotheses cross represents the results generated by thirty hypotheses.

As shown in Fig.5(a), fixed-threshold and Niblack worked well in terms of recall, while multiple hypotheses was dragged down due to the bad recall of Canny. On the other hand, Fig.5(b) shew the precision of multiple hypotheses which was nearly two times higher than that of each operator. F-measure value and recognition rate were given in Fig.5(c) and Fig.5(d), respectively. The reason why Canny's ten hypotheses did not perform well may be that the OCR is hard to directly recognize the edge contour character. The immediate improvement way is to cascade a mathematical morphology operator or to use a stronger OCR engine. The results without Canny will be given later.

Our F-measure value is around 58% that is comparable with the top-ranked methods listed in ICDAR 2011 competition report [8]. Table 1 in [8] indicated that only one method achieved the exceptionally high F-measure value (around 70%). These listed methods achieved the remarkable performance with the aid of either the normalized text line or heuristic constraints. However, our proposal works free from such heuristics. More importantly, we achieved the character recognition rate around 95%. It means that if a character area is

Figure 4. Recognition results for two images.

correctly detected, its recognition result is dramatically high. Unfortunately, there is no character-level recognition rate provided by [8]. In addition, it is promising that F-measure value can be further boosted after replacing the naive image operator with strong counterpart like the Stroke Width Transform (SWT) [1].

In Fig.4, we displayed the result images on which the recognized characters were marked. As anticipation, the result images shown in Fig.4(a) indicate that the proposed multiple hypotheses framework has the cooperation and complementation mechanism which not only filters out the redundant detection but also pads missing detection (pointed out by blue arrows). However, in rare cases, multiple operators are not compatible with each other. It leads to a regrettable result as shown in Fig.4(b). The reason is that an operator (in this sample, Canny) fails in detecting the character region so that the all candidates' aggregate weights are inferior to the threshold β.

Note that if we directly discard Canny's results, F-measure value and recognition rate can climb up to around 62% and 98%, respectively. We finally reserve Canny's extremely weak results in order to demonstrate that even if an image operator is defected completely, the other operators can complement its failure to some extent.

4. Conclusion

In this paper, we propose a novel multiple hypotheses framework which filters out the redundant detection and pads the missing detection through the integration module. For evaluating the strength of the proposed framework, we build a concrete character detection and recognition system by filling three image operators, namely fixed-threshold, Niblack and Canny. The experimental results demonstrate that there exist the cooperation and complementation mechanism in the multiple hypotheses framework. In addition, the performance is still comparable with existing works even if we do not use the heuristic constraints. Our future work is to formulate the rules of operator selection and develop a more strong combination scheme instead of the simple voting.

References

[1] B. Epshtein, E. Ofek and Y. Wexler. "Detecting Text in Natural Scenes with Stroke Width Transform". *CVPR*, 2010.

[2] D. T. Chen, J. M. Odobez and H. Bourlard. "Text Detection and Recognition in Images and Video Frames". *Pattern Recognition*, 37(3):595-608, 2004.

[3] H. Goto. "Redefining the DCT-based feature for scene text detection". *IJDAR*, 11(1):1-8, 2008.

[4] K. Wang and S. Belongie. "Word Spotting in the Wild". *ECCV*, 2010.

[5] C. C. Yi and Y. L. Tian. "Text String Detection From Natural Scenes by Structure-based Partition and Grouping". *IEEE Trans. on Image Processing*, 20(9):2594-2605, 2011.

[6] T. Hiroaki and F. Katsuhito. "Word Extraction Method by Generating Multiple Character Hypotheses". *DAS*, 2008.

[7] M. Grundland and N. A. Dodgson. "Decolorize: fast, contrast, enhancing, color to grayscale conversion". *Pattern Recognition*, 40(11):2891-2896, 2007.

[8] S. Asif, S. Faisal and D. Andreas. "ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images". *ICDAR*, 2011.