PAPER

Scene Character Detection and Recognition with Cooperative Multiple-Hypothesis Framework

Rong HUANG[†a)], **Palaiahnakote SHIVAKUMARA**[††*], **Yaokai FENG**[†], ***Nonmembers***, ***and*** **Seiichi UCHIDA**[†], ***Senior Member***

SUMMARY To handle the variety of scene characters, we propose a cooperative multiple-hypothesis framework which consists of an image operator set module, an Optical Character Recognition (OCR) module and an integration module. Multiple image operators activated by multiple parameters probe suspected character regions. The OCR module is then applied to each suspected region and returns multiple candidates with weight values for future integration. Without the aid of the heuristic rules which impose constraints on segmentation area, aspect ratio, color consistency, text line orientations, etc., the integration module automatically prunes the redundant detection/recognition and pads the missing detection/recognition. The proposed framework bridges the gap between scene character detection and recognition, in the sense that a practical OCR engine is effectively leveraged for result refinement. In addition, the proposed method achieves the detection and recognition at the character level, which enables dealing with special scenarios such as single character, text along arbitrary orientations or text along curves. We perform experiments on the benchmark ICDAR 2011 Robust Reading Competition dataset which includes a text localization task and a word recognition task. The quantitative results demonstrate that multiple hypotheses outperform a single hypothesis, and be comparable with state-of-the-art methods in terms of recall, precision, F-measure, character recognition rate, total edit distance and word recognition rate. Moreover, two additional experiments are conducted to confirm the simplicity of parameter setting in this proposal.

key words: *cooperative multiple-hypothesis framework, scene character, OCR, integration, voting*

1. Introduction

Character detection and recognition in real-world scenes have been increasingly studied to cater to various applications such as content-based image retrieval task, automotive navigation system and aid for the visually impaired. As a type of descriptive feature, characters appearing in natural scenes usually indicate the pivotal semantic information.

Unfortunately, in contrast to the characters in scanner-based document images, scene characters commonly merge into the complex background with variable font, size, color and layout. In addition, scene characters may further suffer from background interferences, nonuniform lighting circumstance and shadow. It has been proved by Epshtein *et al.* [1] that, in most situations, directly using the off-the-shelf Optical Character Recognition (OCR) engines leads to a considerable amount of abortive or incorrect recognition results. Therefore, the imperative issue argued by many researchers is 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 the limited capability of OCR and the unconstrained appearances of scene character. Yi *et al.* [2] designed a bigram color uniformity model to cluster edge pixels into several boundary layers. Color assignment was exploited to filter out background interferences in each boundary layer, and then text strings were verified by a string fragment classifier. Yao *et al.* [3] proposed a bottom-up grouping and top-down pruning method to detect scene texts appearing in arbitrary orientations. To verify the suspected character, the components extracted from Stroke Width Transform (SWT) [1] were pruned by heuristic rules and a trained component level classifier. Character pair candidates were gathered by using a greedy hierarchical agglomerative clustering method. A chain level classifier worked for dispelling the ambiguity that one character might correspond to multiple chains. Pan *et al.* [4] estimated the text existing confidence and scale information for implementing local binarization. Non-text filtering relied on a new proposed conditional random field model. Finally, text lines were formed by a learning-based energy minimization method. However, the above three methods [2]–[4] were separated from the recognition stage and focused solely on the text detection issue.

At the opposite extreme, word recognition methods only aimed at recognizing the content in a cropped word image without designating an explicit detection stage. Wang *et al.* [5] proposed a pipeline word recognition, in which characters along the text line were located by a nearest neighbor classifier taking Histograms of Oriented Gradient [6] as features. To find out the most suitable word in the lexicon, configuration scores were evaluated by optimizing a cost objective function. Mishra *et al.* [7] presented a framework that explored both bottom-up and top-down cues to read texts in street images. The bottom-up cues referred to individual character detections and interactions between them, while lexicon-based prior knowledge, namely language statistics, were modeled to impose top-down cues.

Manuscript received October 18, 2012.
Manuscript revised April 22, 2013.

[†]The authors are with the Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka-shi, 819–0395 Japan.

[††]The author was with the Department of Computer Science, National University of Singapore, Singapore.

[*]Presently, with Systems and Information Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

a) E-mail: rong@human.ait.kyushu-u.ac.jp

DOI: 10.1587/transinf.E96.D.2235

Through the above survey, we found that in the scene character related research community, separate studies on detection and recognition occupied the mainstream. This status was also reflected in the well-known ICDAR Robust Reading Competition dataset. In its recent 2011 version, scene text detection and recognition were still treated in isolation [14].

Rarely, some scene character reading methods performed completely by considering both detection and recognition problems. The approach proposed by Chen *et al.* [9] probed the characters based on the fact that character strokes contain cross edges. Each normalized text line was segmented into multiple binary images. Confidence value was definitively computed based on language modeling and OCR statistics. Wang *et al.* [10] applied Random Ferns [11] to perform sliding window based multi-scale character detection. Then, lexicon-based Pictorial Structures [12] formulation searched an optimal configuration for forming a particular word. Neumann *et al.* [13] presented an end-to-end real-time scene text localization and recognition method. The probability of each extremal region being a character was estimated through a coarse-to-fine two-stage classification.

Although the above three methods [9], [10], [13] introduced the practical recognition stage, they still followed the conventional routine that correctly detecting characters first, and then tailoring them to a successor recognition stage. An apparent drawback of this separative structure is that errors at the predecessor stage will be propagated to the successor stage. Therefore, most existing methods put effort into the predecessor stage, namely the character detection stage, to pursue the high precision for easing the impact of the error propagation. However, as mentioned earlier, precise scene character localization or segmentation without false positive is an extremely hard task, so that without exception, most existing methods refined the detection result aided by a series of heuristic rules such as imposing constraints on segmentation area, aspect ratio, color consistency and text line orientations. More severely, each heuristic rule will definitively affect the final result. We argue that assuming the modality of character appearance in advance will limit the generality of a proposal. This is because different from the superimposed captions or the texts on journal cover [15]–[17] which often appear in canonical horizontal or vertical orientations with high contrast and strong rectangular boundaries, characters found in unconstrained natural scene take on a great variety of appearances.

In this paper, we realize both character detection and recognition through a cooperative multiple-hypothesis framework, which copes with the unconstrained appearances of scene character by taking multiple image operators as a countermeasure. This idea is inspired by the fact that cooperation is always endowed with robustness. The proposed framework consists of an image operator set module, an OCR engine module and an integration module. The image operator set, like bag-of-operators, may include multiple general character detection approaches, such as binarization, edge detector, morphological operation, SWT, extremal region detector and Scale Invariant Feature Transform (SIFT). Each image operator activated by multiple parameters produces multiple suspected character regions. After the Connected Component (CC) analysis, each CC is fed into the OCR engine which implements character recognition and returns multiple candidates with weight values. The integration module collects the multiple recognition results and achieves the refinement in terms of pruning the redundant detection/recognition and padding the missing detection/recognition through a voting stage.

The proposed framework deviates from the conventional separative structure and displays cooperative behavior. The cooperation mechanism is twofold. First, extremely precise character detection is no longer dominated pursuit. Alternatively, recognition stage is effectively leveraged to achieve the refinement. This characteristic bridges between detection and recognition stage and enables us to alleviate the dependence on heuristic rules. Second, a single detection/recognition result will be no longer decisive. Instead the majority of multiple results are taken into account. This characteristic renders us robust outputs without the aid of heuristic rules. Therefore, the proposed framework not only possesses the cooperation mechanism between detection and recognition stage but also works free from heuristic rules. It is worthwhile to highlight that the processing target is a single CC throughout the three modules so that our method can handle special scenarios such as single character, text along arbitrary orientations, text along curves or more complex layout.

We evaluate the performance of the proposed framework on the benchmark ICDAR 2011 Robust Reading Competition dataset [14] and make comparative studies. Experimental results demonstrate that our method is comparable with existing methods and yields the remarkable performance which a single hypothesis can not achieve.

The rest of the paper is organized as follows. Section 2 gives the overall description of the proposed cooperative multiple-hypothesis framework and a corresponding concrete implementation. In Sect. 3, we exhibit the experimental results. Section 4 concludes this paper and outlines the future work.

2. Cooperative Multiple-Hypothesis Framework

2.1 Overview

In this section, we introduce the structure of the proposed cooperative multiple-hypothesis framework as illustrated in Fig. 1. Each hypothesis is generated from an image operator activated by an explicit parameter. The terminology "hypothesis" implies the ideal assumption that under the current parameter setting, the image operator can produce optimal suspected character segmentations. The core idea of the proposed framework is to approximate the optimal segmentation by combining multiple quasi-optimal hypothe-

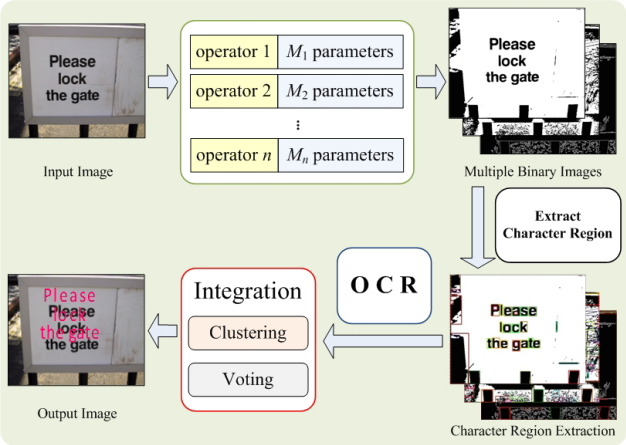


Fig. 1 Diagram of the proposed cooperative multiple-hypothesis framework.

ses. In addition, benefiting from setting multiple parameters, we are liberated from the toil of parameter tuning. Moreover, the proposed multiple-hypothesis framework is so flexible that any image operators can be designated according to specific applications. We argue that it is better to select operators having different characteristics since homogeneous character detection results may attenuate the integration module. However, setting an exact criterion for image operator selection is a quite hard issue due to the unconstrained appearances of scene characters. The proposed cooperative multiple-hypothesis framework can clear this hurdle by leveraging the complementary property among image operators to alleviate the dependence on such unpredictable criteria. This complementary property will be demonstrated by subsequent experiments.

It is worthwhile to point out the distinction between our proposal and the multiple-hypothesis utilized by literatures [8], [9], [18]–[20]. The method proposed by Clark *et al*. extracted rectangular or quadrilateral regions by first generating many hypotheses for possible quadrilaterals and then rejecting irrelevant ones. However, this method assumed that texts represented a page, which would not always be satisfied in natural scenes. Methods [9], [18], [19] clustered pixels into multiple layers and then produced multiple binary images by assuming that one layer corresponded to text and all other layers corresponded to background. Obviously, only one hypothesis correctly marked out characters. Therefore, their integration stage actually selects a best case rather than exerts the effectiveness of the cooperation mechanism. Neumann's method [20] applied multiple hypotheses to purify content-based text line, but no cooperation mechanism exists between their detection and recognition stage.

2.2 Details

We design a concrete character detection and recognition system based on the proposed cooperative multiple-hypothesis framework. Three image operators, fixed-threshold binarization, Niblack binarization and SWT are employed. Each selected operator possesses its own characteristics. Fixed-threshold is a global binarization method while Niblack is a well-known local counterpart. Niblack determines thresholds for each pixel by combining the local average and standard deviation. We select Niblack from the operator set since it can still achieve stable binarization even if the original image suffers from low contrast, high complexity or noise. Note that integral image is used to yield a fast Niblack independent of the local window size [21]. We pass up the Sauvola local binarization technique since it is more sensitive to the change of background than Niblack [22]. The third operator SWT can capture the skeleton of character by calculating stroke width between pairs of parallel edges and then grouping pixels with similar stroke width.

The number of parameters for each image operator, M_1, M_2 and M_3, is set to ten empirically. Setting ten parameters for each operator represents the compromise between computational cost and detection/recognition performance, which is shown in a subsequent experiment.

The detailed procedures are described as follows:

[**Step 1**]: Preprocessing: color to grayscale conversion. Since there exist characters merging into the surrounding background due to the low contrast, we take Grundland's decolorization approach [23] as a countermeasure. This approach can well preserves equiluminant chromatic contrasts in grayscale.

[**Step 2**]: Each image operator activated by multiple parameters transforms the grayscale image into binary/stroke width image.

For the fixed-threshold binarization operator, its ten parameters, namely threshold values, are uniformly distributed over the grayscale interval $0 \sim 255$.

In the Niblack binarization, the local threshold value can be expressed as $m \pm k \cdot s$, where m, s are the local mean and standard deviation, respectively. The parameter k is orderly assigned by ten numbers taking -0.9 as initial value and 0.9 as interval.

In SWT, edge detecting must be first performed for the subsequent stroke width calculation. In general, Canny operator is employed. To ensure that noisy edges are not broken up into multiple edge fragments, a process of Canny known as non-maximal suppression exhibits hysteresis controlled by two thresholds. We fix Canny's lower threshold value at 30 empirically, and change its upper one in turn from 50 to 195 taking 25 as interval. Different from a binary image, each pixel of SWT image stores the stroke width value which is a critical cue to detect character. Thus, one reasonable rule is preserved to immediately filter out the pixel where the stroke width ratio of this pixel and its neighbors exceeds 3.0 [1]. A family of heuristic rules including limitations on aspect ratio, area and text line used in [1] are abandoned.

The approach of parameter setting is flexible, in the sense that the performance will not sharply deteriorate as long as multiple parameters cover the primary zone of the

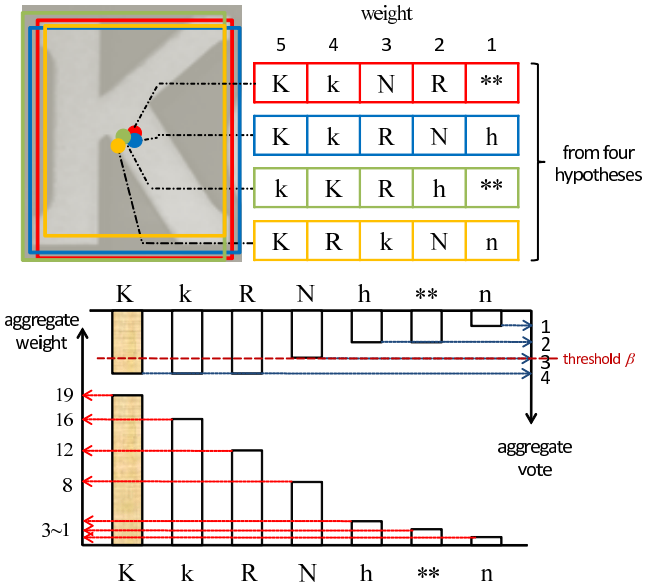


Fig. 2 An example: the recognition results of a CC and its weight/vote aggregate.

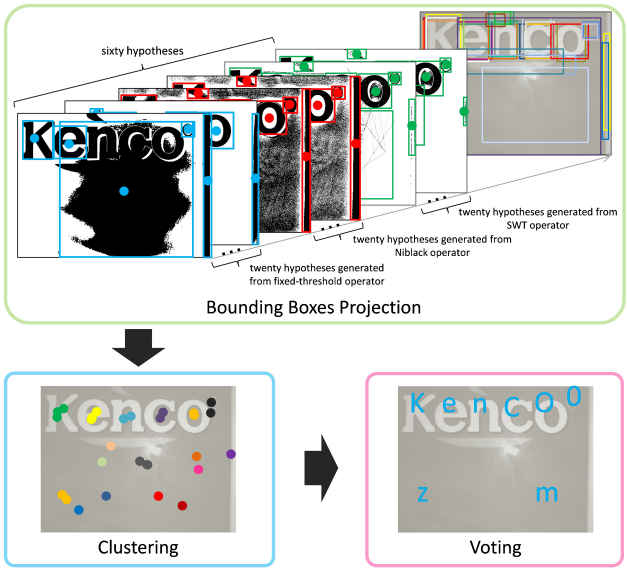


Fig. 3 Diagram of the integration module.

value range. This appealing merit will be demonstrated by a subsequent experiment.

Not that we should apply each image operator twice to accommodate both bright text on dark background and vice-versa. Therefore, from a scene grayscale image, we obtain sixty binary/stroke width images which correspond to sixty hypotheses under the setting $M_1 = M_2 = M_3 = 10$.

[**Step 3**]: Extracting CCs and their bounding boxes. The connected component analysis is used to segment a transformed image into CCs.

Since characters "i" and "j" annoyingly produce two separated CCs, their corresponding bounding boxes will merge together if all following three criteria are satisfied:

(1) The vertical interval between two CCs is inferior to one third of the height of either CC.
(2) Two CCs are alike in width.
(3) The height of lower CC is t times larger than that of upper one, where t is a constant between 2 and 3.

[**Step 4**]: OCR. CCs are fed into OCR one by one. OCR returns N candidates with weight values for each CC as shown in Fig. 2. In this example, we color the bounding boxes for highlighting that they stem from four different hypotheses. OCR is forced to produce five candidates with decreasing weight values (a larger weight value means a higher confidence level). The symbol "**" represents that the OCR recognition result is non-character.

[**Step 5**]: Integration. As illustrated in Fig. 3, all bounding boxes are projected onto the original scene image. Then, these bounding boxes are clustered into several subsets in terms of the adjacency principle. The detailed clustering procedure is elaborated as follows:

(1) Extract the centroid coordinate of each bounding box.
(2) Compute the Euclidean distance between pairs of CCs.
(3) Roughly overlapped bounding boxes are considered as the sibling derived from the same suspected character region thus merging into one subset. Concretely, if the distance between two centroid coordinates is below a threshold value α, the corresponding bounding boxes are clustered into the same subset.
(4) Record the number of subsets.
(5) Increase the threshold value α appropriately and repeat the manipulations (2)~(4).
(6) Stop the iteration when the number of subsets converged.

Here, the threshold value α just plays a "catalytic role" independent of the final detection and recognition performance. The number of subsets always converges within finite iteration loops (around one to two orders of magnitude).

After the above clustering procedure, we obtain a fused image on which the location of each subset is labeled by the centroid coordinates. Note that each subset containing one or more bounding boxes corresponds to a suspected character CC. As displayed in Fig. 2, four color bounding boxes around an actual character "K" are grouped together as one subset.

For each subset, majority voting combination scheme sums the weight values of identical candidates, and then elects a winner under the following two conditions:

(1) maximum aggregate weight value.
(2) aggregate vote exceeding a threshold β.

Then, the corresponding winning candidate is output as final result. An example can be found in the bottom half of Fig. 2, where the threshold β is set to 3, and the selected column pair is filled with yellow texture. If no aggregate votes exceed the threshold, the subset is eliminated. Note that this elimination is useful to remove non-character subsets (redundant detection/recognition).

This integration module exploits the strength of OCR engine in terms of its highly confident recognition outputs. For an actual character region, OCR returns identical candidates over multiple hypotheses (see an example in Fig. 2). However, when recognizing a non-character region, OCR will produce conflicting and cluttered candidates so that no aggregate votes can exceed the threshold β. Therefore, our proposal not only behaves practically but also removes the erroneous false positive bounding boxes automatically.

3. Experiment

3.1 Experiment Overview

We evaluated the performance of the cooperative multiple-hypothesis framework on ICDAR 2011 Robust Reading Competition dataset [14] including two scenarios, namely a text localization task and a word recognition task. In the text localization task, the ground-truth CCs with the area smaller than 10×10 were wiped off one by one in advance. We deem this manipulation impartial since humans may be hard to perceive the content in such a small area as well. In the word recognition task, a correct output means that both the missing detection/recognition and the redundant detection/recognition are successfully suppressed. We performed two comparative studies: (1) between the top-ranked methods listed in [14] and our proposal; (2) between a single hypothesis and multiple hypotheses. In addition, experiments on parameter setting were conducted to confirm its prospective simplicity in our proposal. It is worthwhile to mention that a commercial OCR engine (OCR library version 7.0 developed by Media Drive Corporation) without the aid of lexicon is employed to realize recognition at the character level.

To evaluate the detection and recognition performance quantitatively, we employed six canonical metrics, namely recall r, precision p, F-measure F, Character Recognition Rate CRR, Total Edit Distance TED [14] and Word Recognition Rate WRR. Here, the first four metrics are known as the metrics to evaluate the character detection performance, while the last two are used for estimating the precision of word recognition. Recall is the ratio of the number of successfully extracted regions to the number of ground-truth character regions. Precision is the ratio of the number of correctly extracted character regions to the number of whole extracted regions. F-measure is the harmonic mean between r and p: $F = 2pr/(p + r)$. Character recognition rate which directly reflects the performance of OCR is estimated by the ratio of the number of correctly recognized characters to the number of correctly extracted character regions. Total edit distance accumulates the normalized edit distance for each of the ground-truth word and the corresponding recognition result. The last one, word recognition rate, is the ratio of the number of correctly recognized words to the total number of cropped word images.

Table 1 Quantitative detection and recognition results.

Method	Character Location Task (%)				Word Recognition Task	
	R	*P*	*F*	*CRR*	*TED*	*WRR* (%)
Kim's Method	62.47	82.98	71.28			
Yi's Method	58.09	67.22	62.32			
TH-TextLoc System	57.68	66.97	61.98		176.23	41.20
Neumann's Method	52.54	68.93	59.63		429.75	33.11
TDM_IACS	53.52	63.52	58.09			
LIP6-Retin	50.07	62.97	55.78			
KAIST AIPR System	44.57	59.67	51.03		318.46	35.60
ECNU-CCG Method	38.32	35.01	36.59			
Text Hunter	25.96	50.05	34.19			
All Hypotheses	**59.32**	**59.55**	**59.44**	**80.28**	**448.88**	**34.44**
Multiple Hypotheses on Fixed-threshold	57.96	66.65	62.00	83.36	565.90	21.91
Single Hypothesis on Fixed-threshold	50.08	46.36	48.15	79.14	655.59	13.14
Multiple Hypotheses on Niblack	62.73	58.69	60.64	82.22	503.64	22.72
Single Hypothesis on Niblack	66.14	19.27	29.84	75.34	542.36	14.06
Multiple Hypotheses on SWT	58.66	32.86	42.12	78.02	748.17	6.93
Single Hypothesis on SWT	72.41	13.99	23.45	71.84	769.11	1.94

= not given

3.2 Quantitative Evaluation Results and Comparative Studies

In this subsection, we exhibited the quantitative evaluation results in Table 1 and then conducted two comparative studies. In the text localization task, our method achieves the recall of 59.32%, the precision of 59.55% and the F-measure of 59.44%, which are comparable with the top-ranked methods. Particularly, the character recognition rate of 80.28% indicates that if a character region is correctly detected, its recognition accuracy becomes dramatically high. Unfortunately, the recognition rate of character level is not a part of ICDAR competition [14]. In the word recognition task, the proposed method achieves the total edit distance of 448.88 and word recognition rate of 34.44%. Surprisingly, our method gets better word recognition rate than Neumann's, but fails in terms of total edit distance. The reason of this apparent contradiction is that more redundantly recognized characters may still remain without the aid of text line orientations or lexicon, and thus improperly contribute a higher edit distance. Especially, there is no assumption in the location of characters in cropped word images, which also affected the total edit distance.

The results listed in the bottom half of Table 1 show the performance comparison between a single hypothesis and multiple hypotheses. For the single hypothesis, the preci-

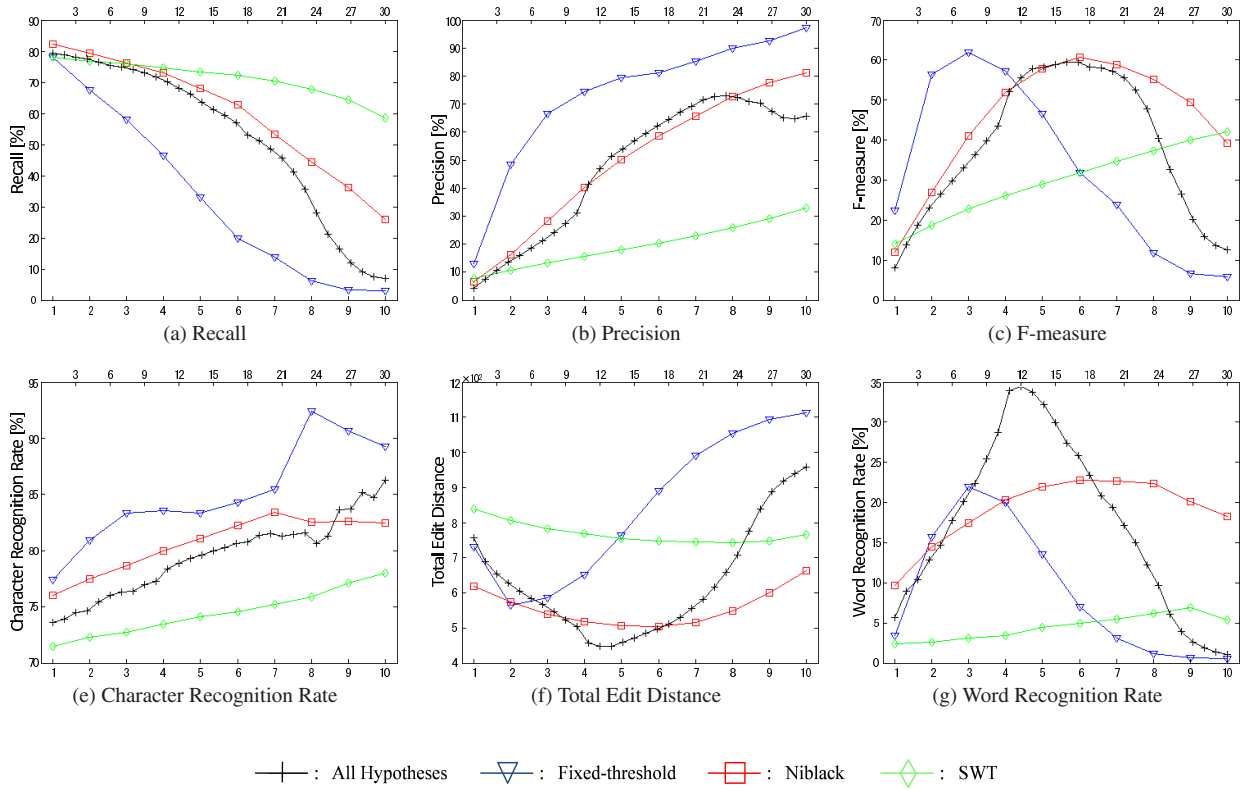


Fig. 4 Performance evolution versus the increase of threshold β.

sion deteriorates sharply. This is because most redundant CCs can not be filtered out due to the absence of integration. On the other hand, it is an undeniable fact that the integration module may drop the recall somewhat. Multiple hypotheses performs much better in the word recognition task since the cropped word images are generally full of conspicuous text strings without severe interferences, which becomes so different from the born natural scene images. More interestingly, single hypothesis on fixed-threshold and Niblack achieves word recognition rate of 13.14% and 14.06%, respectively, which is almost consistent with the fact that 13.41% of the cropped word images appear in a standard font and size on a uniform single colored background [14]. The higher word recognition rate achieved by multiple hypotheses demonstrates that it can handle more complex appearance.

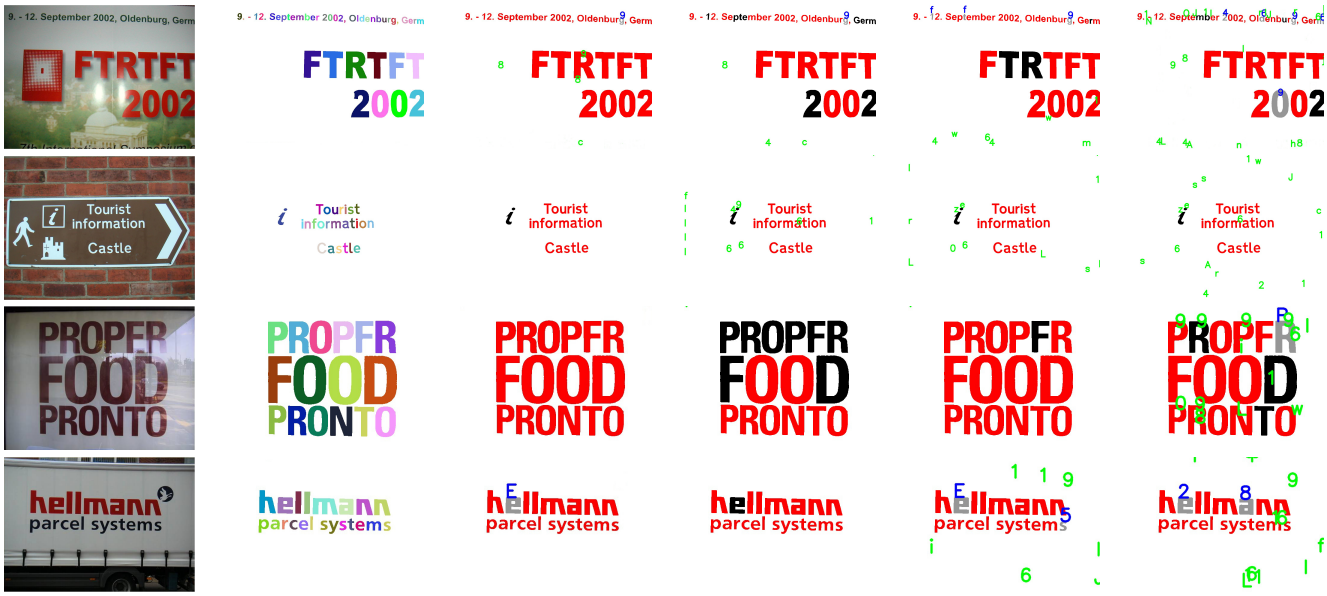
Six curve charts in Fig. 4 display the performance evolution versus the increase of threshold β. The top horizontal axis represents the global vote threshold β (for thirty parameters), while the bottom horizontal axis corresponds to the sectional vote threshold (for ten parameters). The vote threshold is used to filter out false positive CC subsets automatically. Each evaluation metrics is marked on the vertical axis.

We can see that Niblack and SWT perform well in recall, while fixed-threshold achieves higher precision. On the other hand, the results of all hypotheses are dragged down due to the mediocre performance of SWT. This tendency is also reflected in the bottom half of Table 1. The reason why twenty hypotheses by SWT do not achieve their desired effect is that too many fragment-like CCs are incorrectly identified as characters. This situation does not mean that integration module becomes completely invalid since the precision of SWT is on the rise with the increase of threshold (see Fig. 4 (b)). Note that if we directly discard SWT's results, F-measure and recognition rate of all hypotheses can climb up to around 61% and 82%, respectively. The SWT's weak results was finally reserved in order to demonstrate that even if an image operator was defected, the other ones could complement its failure to some extent. We argue that the complementary property among image operators is meaningful since there is still lack of axiomatic criteria to evaluate which type of image operator will be appropriate for a given natural scene image.

The evolution of F-measure and word recognition rate share the similar curve shape. The performance is poor on smaller threshold values since a considerable amount of redundant outputs are not filtered out. With the increase of threshold, the performance gets better, and finally deteriorates on larger threshold values. As we can find, setting threshold in the middle, namely 12∼15, usually returns remarkable performance.

3.3 Detection and Recognition Examples

We showed detection and recognition examples in this subsection. Four examples of the text localization task were displayed in Fig. 5. To clearly exhibit results, we colored



(i) Original image (ii) Ground-truth image (iii) All hypotheses (iv) Fixed-threshold (v) Niblack (vi) SWT

Fig. 5 Examples of the scene character detection and recognition. Red = correct detection and recognition; black = missing detection and recognition; gray = correct detection but false recognition; blue = the false recognition result; green = redundant detection and recognition.

the correctly detected and recognized characters red on the ground-truth image, while the false or missing ones were colored by gray and black, respectively. A blue character associated with the gray ground-truth character represents the false recognition. In addition, a green character stands for the redundant detection/recognition. As expected, the proposed cooperative multiple-hypothesis framework has the cooperation mechanism. The first row shows that the missing detection/recognition (the black ground-truth character) can be padded through the integration module. On the other hand, the second row indicates that the redundant detection/recognition (the green character) can be suppressed automatically. In the third row, we gave a nice example of overcoming specularities. A character may segment into fragmentary pieces due to the non-uniform lighting circumstance. The local threshold method, say Niblack, can well handle this type of cases. The fourth row shows that our method can detect and recognize the characters with uncommon font.

Moreover, we provided examples related to the word recognition task in Fig. 6. The recognized results were separately printed below the corresponding cropped word images. Since the selected three image operators work in the form of segmentation then CC analysis pipeline, the extracted CC is vulnerable to obstacle (see example “PIZZA”), connected characters (see example “Shop” and “STAR”), separative characters (see example “OF”). We would suffer another reverse when encountering the common difficulties such as severe blur (see example “ESC”), highly decorated characters (see example “venve”).

Fig. 6 Examples of the cropped word recognition. Red = correct detection and recognition; black = missing detection and recognition; gray = correct detection but false recognition; blue = the false recognition result; green = redundant detection and recognition.

3.4 Experimental Results on Scene Characters with special Scenarios

The experimental results on scene character with special scenarios were given in this subsection. We remark that the top-ranked methods listed in Table 1 achieved the outstanding performance with the aid of either straight text line or heuristic rules. It is worthwhile to emphasize that our

(i) Original image (ii) Ground-truth image (iii) Resulting image

Fig. 7 Examples of the special scene character scenarios. Red = correct detection and recognition; black = missing detection and recognition; gray = correct detection but false recognition; blue = the false recognition result; green = redundant detection and recognition.

proposal abandons such heuristic rules instead of filtering out the redundant detection/recognition and padding missing ones automatically. Figure 7 depicted three examples with special scenarios. All three natural scene images were harvested from the Internet, and we manually prepared the corresponding ground-truth images. In the top one, a single number "5" occupies a significant part of the scene. Yi's method [24] listed in Table 1 (the second rank) could not deal with this case since their method becomes low effective if text string contains less than three characters. Characters with different size appear along the curve layout in the middle natural scene image. Benefitting from detecting and recognizing targets at the character level, our method could handle this curve layout as verified in Fig. 7. In the bottom natural scene image, three characters are distributed without specific text lines. Although its background is so pure, text line-based methods will miss the number "1" since it deviates from the text line.

3.5 Experiments for Parameter Setting

Following two experiments were conducted to demonstrate the simplicity of parameter setting in our proposal. The so-called simplicity is reflected in the sense that uniformly setting ten parameters over the value range renders us reasonable and stable performance.

The first experiment tested the performance evolution versus the number of parameters. As shown in Fig. 8, the performance boosts dramatically when two parameters are exploited. The performance improvement gradually tends towards convergence after eight or nine parameters has been employed. Setting more than ten parameters will only feedback negligible improvement at the cost of proportional increase of the total number of hypotheses as shown in the bottom half of Fig. 8. Therefore, using ten parameters for each image operator achieves a trade off between computational cost and detection/recognition performance.

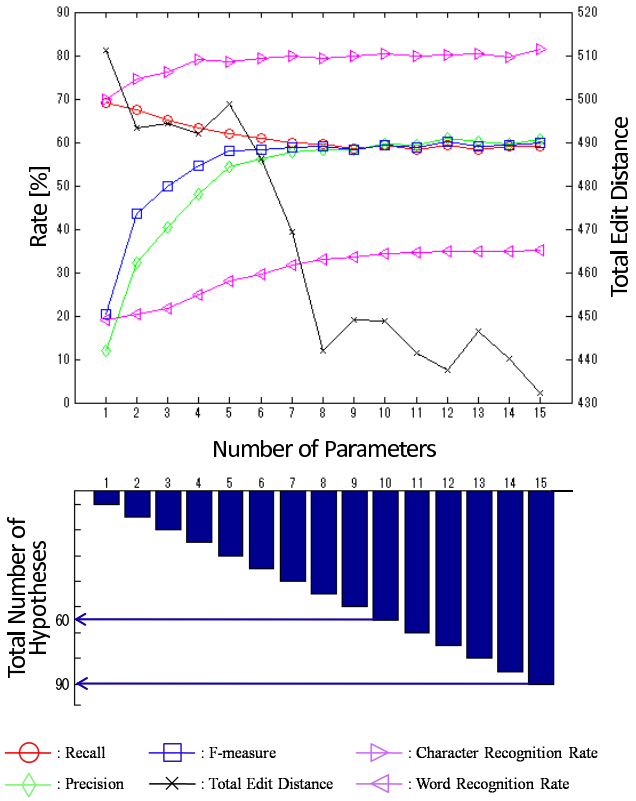


Fig. 8 Performance evolution versus the number of hypotheses.

Table 2 Parameter setting trials.

Trial Number	Operator	Parameters Setting	Results			
			F (%)	*CCR* (%)	*TED*	*WCR* (%)
Original	F	25 50 …… 250	59.44	80.28	448.88	34.44
	N	-0.9 -0.7 …… 0.9				
	S	50 75 …… 195				
Trial-1	F	15 30 …… 150	58.13	79.24	459.92	33.76
	N	-0.7 -0.6 …… 0.2				
	S	40 50 …… 130				
Trial-2	F	100 115 …… 235	57.94	79.01	462.48	33.64
	N	-0.2 -0.1 …… 0.7				
	S	110 120 …… 200				
Trial-3	F	100 110 …… 190	59.92	81.03	452.19	34.27
	N	-0.4 -0.3 …… 0.5				
	S	115 120 …… 160				

F = Fixed-threshold; N = Niblack; S = SWT

In the second experiment, we deliberately changed the parameter setting in terms of parameter range and interval as given in Table 2. In our original setting, multiple parameters were uniformly distributed over the range. Parameters in the trial-1 gather at the small value side, while they moved to the large value side in the trial-2. In the last trial, we made them

centralize more densely around the middle value range. The data listed in the last four columns demonstrate that changes in parameter range and interval will not greatly affect the final results. Considering the generality, we still suggest uniform distributed parameters as standard setting, even F, CR and TED of trial-3 exceed those of original setting slightly.

4. Conclusion

In this paper, we proposed a novel cooperative multiple-hypothesis framework to detect and recognize characters appearing in the natural scene. The framework consists of an image operator set module, an OCR module and an integration module. To evaluate the effectiveness of the proposed framework realistically, we built a concrete system which selected fixed-threshold, Niblack and SWT as image operators. By effectively leveraging the strength of OCR engine, the cooperation mechanism in the integration module filters out the redundant detection/recognition and pads the missing detection/recognition automatically. The performance of our proposal is still comparable with the existing top-ranked works even if we do not apply heuristic rules such as imposing constraints on segmentation area, aspect ratio, color consistency and text line orientations. Experiments on parameter setting demonstrate that the detection/recognition performance is not strongly dependent on the parameter tuning. Moreover, our method works at the character level so that it enables handling special scenarios such as single character, text along arbitrary orientations, text along curves or more complex layout.

The proposed cooperative multiple-hypothesis framework can be regarded as a bridge between the scene character detection and recognition. This conception is very flexible and has a wide room for further improvement. Our future work is to formulate the rules of operator selection and develop a stronger combination scheme instead of simple voting.

References

[1] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," CVPR, June 2010.
[2] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," IEEE Trans. Image Process., vol.21, no.9, pp.4256–4268, May 2012.
[3] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," CVPR, June 2012.
[4] Y. Pan, X. Hou, and C. Liu, "A hybrid approach to detect and localize texts in natural scene images," IEEE Trans. Image Process., vol.20, no.3, pp.800–813, March 2011.
[5] K. Wang and S. Belongie, "Word Spotting in the Wild," ECCV, Sept. 2010.
[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," CVPR, June 2005.
[7] A. Mishra, K. Alahari, and C. Jawahar, "Top-down and bottom-up cues for scene text recognition," CVPR, June 2012.
[8] P. Clark and M. Mirmehdi, "Recognising text in real scenes," IJDAR, vol.4, pp.243–257, 2002.
[9] D. Chen, J. Odobez, and H. Bourlard, "Text detection and recognition in images and viedo frames," Pattern Recognt., vol.37, no.3, pp.595–608, March 2004.
[10] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," ICCV, Nov. 2011.
[11] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," ICCV, Oct. 2007.
[12] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," International Journal of Computer Vision, vol.61, no.1, pp.55–79, Jan. 2005.
[13] L. Neumann and J. Matas, "Real-time scene text localization and recognition," CVPR, June 2012.
[14] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene image," ICDAR, Sept. 2011.
[15] K. Jung, K. Kim, and A. Jain, "Text information extraction in images and video: A survey," Pattern Recognit., vol.37, no.5, pp.977–997, May 2004.
[16] W. Kim and C. Kim, "A new approach for overlay text detection and extraction from complex video scene," IEEE Trans. Image Process., vol.18, no.2, pp.401–411, Feb. 2009.
[17] X. Zhao, K. Lin, Y. Fu, et al., "Text from corners: A novel approach to detect text and caption in videos," IEEE Trans. Image Process., vol.20, no.3, pp.790–799, March 2011.
[18] C. Li, X. Ding, and Y. Wu, "Automatic text location in natural scene images," ICDAR, Sept. 2001.
[19] E. Kim, S. Lee, and J. Kim, "Scene text extraction using focus of mobile camera," ICDAR, Sept. 2009.
[20] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," ACCV, Nov. 2010.
[21] F. Shafait, D. Keysers, and T. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images," Document Recognition and Retrieval Conference, Jan. 2008.
[22] J. He, Q. Do, A. Downton, and J. Kim, "A comparison of binarization methods for historical archive documents," ICDAR, Aug. 2005.
[23] M. Grundland and N.A. Dodgson, "Decolorize: fast, contrast, enhancing, color to grayscale conversion," Pattern Recognit., vol.40, no.11, pp.2891–2896, Nov. 2007.
[24] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," IEEE Trans. Image Process., vol.20, no.9, pp.2594–2605, Sept. 2011.

Rong Huang received his B.E. and completed the master course from East China University of Science and Technology, Shanghai, China, in 2008 and 2010, respectively. He is supported by China Scholarship Council (CSC) and currently pursuing the Ph.D. degree at Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan. His research interests include image processing, pattern recognition and multimedia security.

Palaiahnakote Shivakumara is visiting Senior Lecturer in the Department of Computer Systems and Information Technology, Faculty of Computer Science and Information Technology, University of Malaya. He received B.Sc., M.Sc., M.Sc Technology by research and Ph.D. degrees in computer science respectively in 1995, 1999, 2001 and 2005 from University of Mysore, Mysore, Karnataka, India. From 1999 to 2005, he was Project Associate in the Department of Studies in Computer Science, University of Mysore, where he conducted research on document image analysis, including document image mosaicing, character recognition, skew detection, face detection and face recognition. He worked as a Research Fellow in the field of image processing and multimedia in the Department of Computer Science, School of Computing, National University of Singapore, from 2005-2007. He also worked as a Research Consultant in Nanyang Technological University, Singapore for a period of 8 months on image classification in 2007. He worked as a Research Fellow (RF) in National University of Singapore (NUS) from 2008 to 2013 on video text extraction and recognition. He has published around 120 research papers in national, international conferences and journals. He has been reviewer for several conferences and journals. His research interests are in the area of image processing, pattern recognition, including video text extraction and recognition, document image processing.

Yaokai Feng received his B.E. and M.E. degrees in Computer Science from Tianjin University, China, in 1986 and 1992, respectively. He received his Ph.D. degree in Information Science from Kyushu University, Japan, in 2004. Now, he is an assistant professor at Kyushu University, Japan. His current research interests include database, pattern recognition, information retrieval and network security. He received MIRU2011 Excellent Paper Award. He is a member of IPSJ and IEEE.

Seiichi Uchida received B.E., M.E., and Dr. Eng. degrees from Kyushu University in 1990, 1992 and 1999, respectively. From 1992 to 1996, he joined SECOM Co., Ltd., Japan. Currently, he is a professor at Kyushu University. His research interests include pattern recognition and image processing. He received 2002 IEICE PRMU Research Encouraging Award, 2008 IEICE Best Paper Award, MIRU2006 Nagao Award (best paper award), MIRU2011 Excellent Paper Award, 2007 IAPR/ICDAR Best Paper Award, and 2010 ICFHR Best Paper Award. Dr. Uchida is a member of IEEE and IPSJ.