Scene Text Localization Using A Two-Class Detector Constructed By Filtered Channel Features

Takuro Oki[1] and Ryusuke Miyamoto[2]
[1]Department of Computer Science, School of Science and Technology, Meiji University
[2]Department of Fundamental Science and Technology, Graduate School of Science and Technology, Meiji University,
1-1-1 Higashimita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan
E-mail : [1]o_tkr@cs.meiji.ac.jp, [2]miya@cs.meiji.ac.jp

Abstract: Scene Text Recognition is one of the challenging tasks in the filed of image processing. To solve this problem, the authors proposed a novel framework to improve both detection accracy and processing speed. In the framework, a two-class detector with exhaustive search is adopted to extract text regions before applying a multi-class classifier. In this paper, we apply Filtered Channel Features that achieves much better accuracy than deep learning for pedestrian detection to constuct a two-class detector. In the framework, experimental results using ICDAR dataset show that the proposed scheme can reduce the area to be evaluated by a multi-class classifier to 30.9% while false negative rate is 15.7%.

1. Introduction

Scene Text Recognition(STR) is one of the challenging tasks in the filed of image processing and the ICDAR competition has been held several times to improve the accuracy of scene text reading. In this research field, two kinds of schemes are currently popular: derivatives from Optical Character Recognition(OCR) and schemes based on generic visual object recognition. The former schemes adopt segmentation based on ”color information”, ”binalization”, ”noise reduction”, etc., for localizing of text regions[1]. After this preprocessing, character recognition is performed in the same way as the traditional OCR using a multi-class classifier. In contrast, the latter schemes aggressively exploit scientific knowledge cultivated in the field of visual object detection and recognition that has remarkably advanced in recent years. These schemes use multi-class detector constructed by a machine learning algorithm with training samples. However, there remains some problems to adopts visual object detection and recognition techniques in STR: a large number of classes to be detected and recognized. For example, considering only alphabet, a number of classes to be detected is fifty-two classes at least. Typically, it is difficult for the multi-class classification to detect many targets accurately with quickly.

The authors have focused on the latter schemes and proposed a novel framework for STR using a two-class classifier[2]. In the proposed framework, character recognition is processed by the following steps: Character detection by character or non-character two-class detector with sliding window exhaustive search, Text region localization by merging outputs of detector, and Character recognition by multi-class classifier. By the framework, computational amount can be drastically reduced because computationally exhaustive multi-class classification is applied only to detected regions that has high possibility to include characters. In addition, once text regions are localized properly, several conventional character recognition schemes can be applied to classify characters included in these regions that have achieved high recognition accuracy. Therefore, the most significant task of the proposed framework is to have a good coverage of the correct text regions in the image to keep recognition accuracy at the third step, since missed regions will never be recovered.

To confirm effectiveness of the proposed framework, the previous work[2] have tried to use integral channel features[3] for construction of detector that shows good performance in the field of human detection, nevertheless, sufficient accuracy did not achieved. To improve the detection accuracy of the framework, this paper tries to apply Filtered Channel Features[4]. Filtered Channel Features achieves state-of-the-art and much better performance than deep learning in the field of pedestrian detection. The method is evaluated with Precision & Recall on a dataset that has been used in ICDAR 2013 Competition, and compared with existing methods. Moreover, we evaluate how much regions that are unlikely to contain characters are reduced properly before multi-class classification.

2. Related work

This section introduces a framework for scene text recognition proposed by the authors and Filtered Channel Features that shows much greater accuracy than deep learning for human detection.

2.1. A framework for scene text recognition using two-class detector for scene text localization

Fig.1 shows a framework that the authors try to construct to improve the performance of scene text recognition[2]. In this framework, exhaustive search by a two-class detector that classifies text regions and non-text regions in the first step. After this process, several regions that are expected to include some texts are extracted. The output of this process includes much rectangular regions corresponding to detection results by exhaustive search, so to obtain only significant regions a region merging scheme called cross-wise region merging is applied. After the detection of text regions, multi-class classifier is applied to determine what characters exist on these regions. By applying such two-step operation, the authors aim to obtained the following benefits:

- Speed up of total recognition process and
- improvement of recognition accuracy.

In our previous scheme[2], two-class detector is constructed by Integral Channel Featuresthat shows great performance for human detection. However, the classification ac-

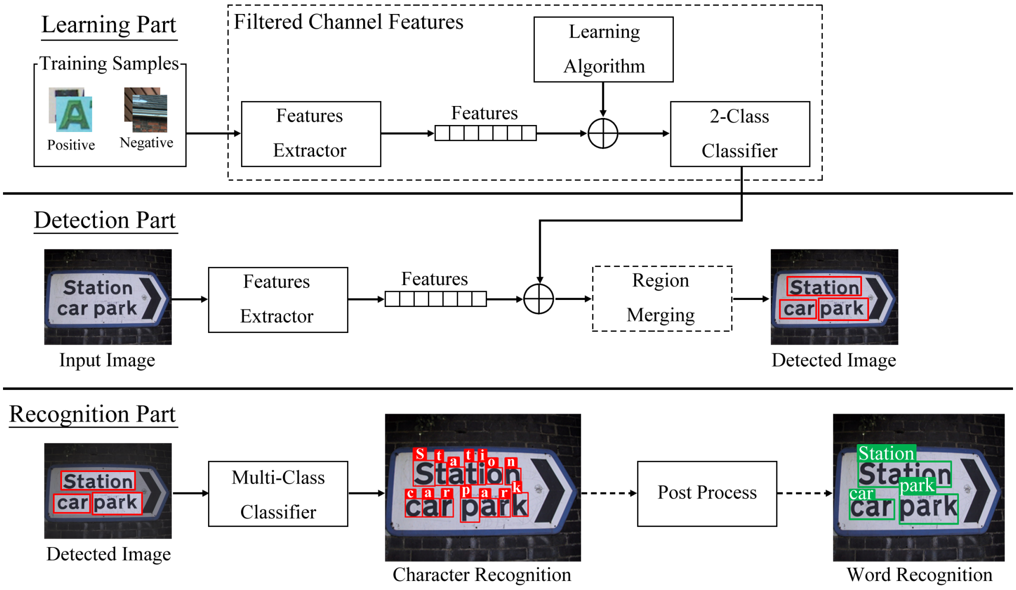

Figure 1. Scene text recognition using two-class detector for scene text localization

curacy is not sufficient. In this paper, we have tried to apply Filtered Channel Features that is one of its derivatives proposed in CVPR2015 shows better accuracy. The following subsection explains Filtered Channel Features.

2.2. Filtered Channel Features

Filtered Channel Features is a human detection scheme proposed by Zhang et. al and this scheme is designed considering Informed Haar-like features. Both schemes are extensions of Integral Channel Features[3] proposed by Dollár et. al. Surprisingly, it has shown that the accuracy for human detection by Filtered Channel Features is much better than deep learning.

In recent schemes that show good accuracy for object detection, it becomes important how to design a feature pool that contain several features required to construct weak classifiers. In also Filtered Channel Features, a feature pool is carefully designed and huge kinds of rectangular patterns called “filters” are prepared in it. Detection accuracy is improved by pooling because good weak classifiers can be selected in the training process by boosting. In the pooling process, many kinds of filter banks constructed by several filters are used. For filter generation, several schemes are proposed to improve classification accuracy such as CheckerBoards [4], Randomfilter [4], and LDCF [5],PcaForeground [6].

3. Proposal

This section describes how to apply the Filtered Channel Features for scene text localization and how to construct a two-class classifier.

3.1. How to construct a classifier

Filtered Channel Features requires filter banks that consists of several kinds of filters to construct a classifier. The subsection 3.1.1 explains a filter bank used in our implementation and the subsection 3.1.2 describes training samples used to construct a two-class classifier.

3.1.1 Preparation of a filter bank

In this paper, a filter bank consists of 16 kinds of filters that are designed manually. Each filter is composed of rectangular regions whose size is 6×6 pixels. The total size of a filter is designed that its size becomes from 1×2 to 4×3 when a 6×6 rectangular region is represented by $1\times$. Fig. 2 shows some examples of filters used in this paper. In the feature calculation of Filtered Channel Features, ternary operations represented by $+1$, -1, and 0 are applied though the original Viola-Jones for face detection uses binary operations corresponding to $+1$ and -1 to evaluate the difference of the luminance at the feature computation of a Haar-like feature. In the Fig. 2, white, black, and red regions correspond to $+1$, -1, and 0, respectively.

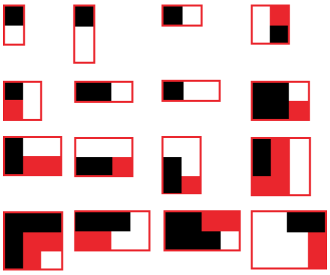

Figure 2. 16 kinds of filters used in this paper.

3.1.2 Training samples to construct a two-class classifier

Training samples are collected from the ICDAR 2013 data set: 4786 positive samples and 12000 negative samples are extracted from 229 training images. Figs. 3 and 4 show some positive and negative samples, respectively. In this process, the number of negative samples are reduced to 12000 by random sampling because the number of negative samples are too larger than the number of positive samples if any process is applied to adjust the number of both training samples. In addition, hard examples that it is difficult for the constructed classifier to discriminate are obtained by applying the constructed classifier to residual negative samples. After the generation of hard examples, training process is executed again using newly created negative samples that consists of randomly selected negative samples and hard examples obtained by the above operation. In this paper, this training iteration is repeated 3 times and 27000 negative samples are used to improve classification accuracy of the constructed classifier.

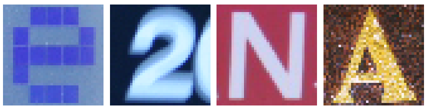

Figure 3. Positive samples.

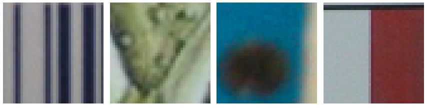

Figure 4. Negative samples.

4. Evaluation

This section evaluates our proposal by using the following criteria: the same criterion as ICDAR 2013[7] and false negative rate vs. area reduction rate. The former is used to compare our proposal to other existing schemes, and the latter is used to evaluate the suitability of the detection based on Filtered Channel Featuresas a two-class classifier for the proposed framework.

4.1. Accuracy comparison to other schemes in the ICDAR competition

This subsection evaluates the proposed scheme based on Precision, Recall, and H-mean, where many-to-one matches[7] are used to compute these scores.

By the Table 1, the proposed scheme shows better accuracy than other schemes in ICDAR 2003. On the other hand, comparing it to the state-of-the art schemes in ICDAR 2013, the accuracy by our proposal is as same as $Text\ Detection$ and H-mean is quite worse than other top-level schemes. The reason is that our proposal shows very lower Precision though it marks high recall. However, our proposal is not so bad because the proposed scheme obtains text regions with only exhaustive search by two-class detector and without any other pre- or post-processing using dictionaries to compensate detected regions. If pre- or post-processing widely used in the state-of-the-art schemes is applied to our scheme, it is expected that Precision and H-mean are improved drastically.

Table 1. Accuracy comparison to other schemes.

Method Name	Recall	Precision	H.mean
OurProposal	**60.30**	**61.77**	**61.03**
OurOld[2]	49.65	48.52	49.08
ICDAR2013			
USTB_TexStar	66.5	88.5	75.9
Text Spotter	64.8	87.5	74.5
TH-TextLoc	65.19	69.96	67.49
Text Detection	53.42	74.15	62.10
$Baseline$	34.7	60.8	44.2
Inkam	35.27	31.20	33.11
ICDAR2003			
Ashida	41.7	55,3	47.5
H.W.David	46.6	39.6	42.8
Wolf et al.	44.9	19.4	27.1
Todoran	17.9	14.3	15.9

4.2. False negative vs. area reduction rate

The primary aim of our framework is to reduce area of regions to be evaluated by a multi-class classifier by preprocessing with a two-class detector that extracts text regions from non-text regions. Therefore, we think that false negative rate vs. area reduction rate is a suitable criterion for adequate evaluation of our scheme.

Fig. 7 shows curves corresponding to false negative rate and area reduction rate. By this figure, the proposed scheme can reduce area of regions to be evaluated by a multi-class classifier to 15.7% of input images when false negative rate is 30.9%. This result means that the proposed scheme can reduce false negative rate by 15% at the same area reduction rate.

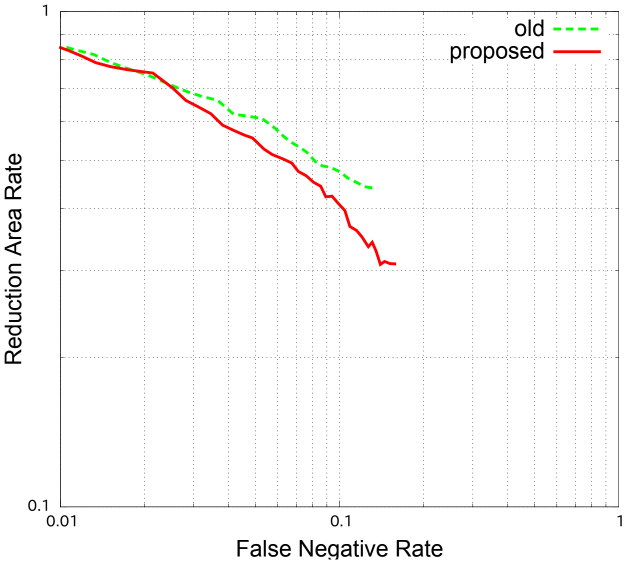

Figure 7. Reduced Rate vs False Negative Rate

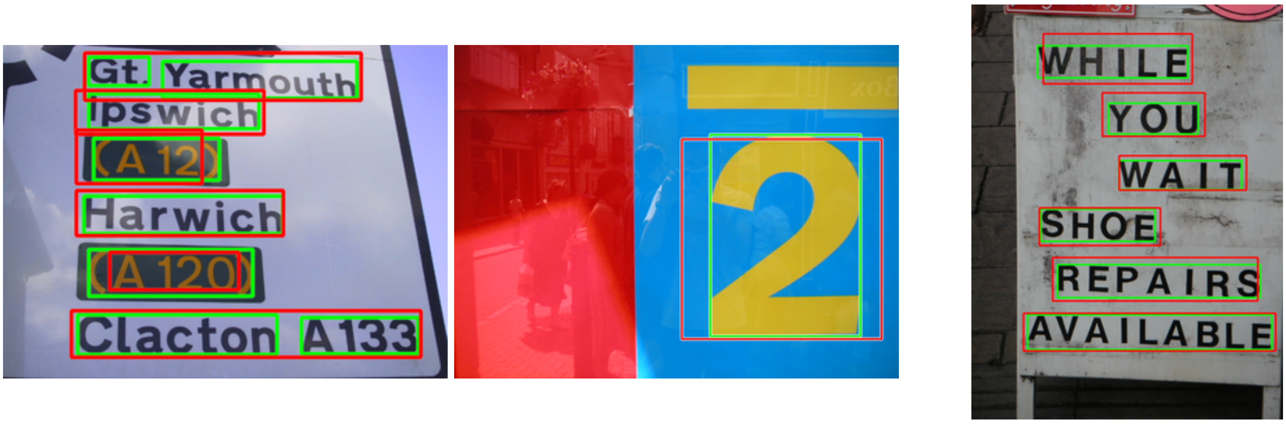

Figure 5. Examples of correct detection.

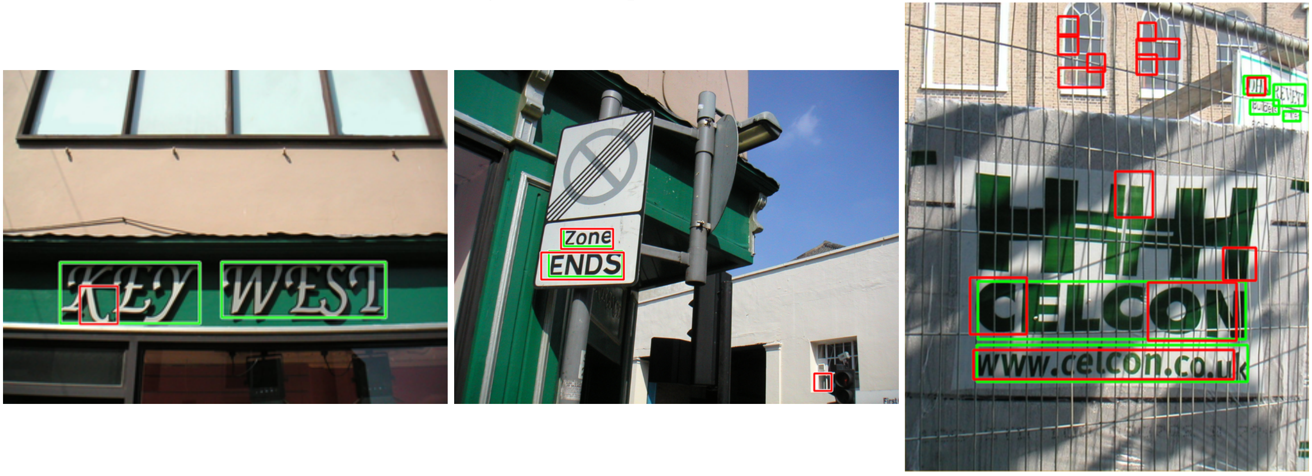

Figure 6. Examples of false detection.

4.3. Detection examples

Finally, detection examples by the proposed scheme are shown in Figs. 5 and 6. In these schemes, green and red bounding boxes show ground truth and detected regions, respectively. As you can see, Fig. 5 shows correctly detected results and Fig. 6 includes falsely detected regions. In these figures, red and green rectangles show detection results and ground truth, respectively.

By the results, the proposed scheme can detect text regions if background is flat and deformation of characters is not strong. However, the proposed scheme can not treat heavily deformed characters such as ornamental characters. In addition, the detection accuracy becomes worse if partial occlusion occurs or the color of background is similar to the foreground characters.

5. Conclusion

In this paper, we have tried to apply Filtered Channel Features that shows excellent accuracy for pedestrian detection to the framework proposed in [2]. Experimental results using ICDAR dataset, is evaluated by the following two kinds of methods. The former aims to compare our proposed schemes with other existing schemes. Experimental results show that the proposed scheme marks lower accuracy than the state-of-the-art schemes though it achieves high recall. The scheme seems not so good, but if pre- or post-processing is applied to our scheme in the same way as the state-of-the-art schemes, accuracy by our scheme will be able to improved drastically. The latter aims to confirm the effectiveness of reducing the area of regions to be evaluated by multi-class classifier in our proposed framework. As the results, the proposed scheme can reduce the area of regions to 30.9% while false negative rate is 15.7%. Considering these experimental results, it is shown that the proposed framework is feasible for scene text recognition.

References

[1] Jerod J. Weinman, Zachary Butler, Dugan Knoll, and Jacqueline Feild, “Toward integrated scene text reading,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 375–387, Feb. 2014.

[2] R. Miyamoto and T. Oki, “Feasibility study of crosswise region merging for scene text loacalization with two-class detector,” in *World Multi-Conference on Systemics, Cybernetics and Informatics*, 2015.

[3] P. Dollar, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.

[4] S. Zhang, R. Benenson, and B. Schiele, “Filtered channel features for pedestrian detection,” *Computing Research Repository*, 2015.

[5] W. Nam, P. Dollár, and J.H. Han, “Local decorrelation for improved pedestrian detection,” in *Adv. Neural Inf. Process. Syst.*, 2014.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014.

[7] C. Wolf and J.M. Jolion, “Object count/area graphs for the evaluation of object detection and segmentation algorithms,” *Proc. Int. J. Doc. Anal. Recognit.*, vol. 8, no. 4, pp. 280–296, 2006.