Scene Character and Text Recognition: The State-of-the-Art

Chongmu Chen[1], Da-Han Wang[2], and Hanzi Wang[1](✉)

[1] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, Xiamen, Fujian, China
Chongmu.chen@gmail.com, hanzi.wang@xmu.edu.cn
[2] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, Fujian, China
wangdh@xmut.edu.cn

Abstract. Scene text recognition is gaining renewed interest owing to the increase of scene image based applications and new intelligent devices. Unlike recognition of printed text, scene text recognition is challenging due to the complexity of scene images. To provide an overview of the techniques and inspire future research, this paper reviews the advances in scene character and text recognition, with emphasize on character feature representation methods and word recognition models. The papers published in the most recent conferences ECCV 2014, ACCV 2014, ICIP 2014, and ICPR 2014 are also reviewed in this paper to provide the state-of-the-art of scene character and text recognition. The state-of-the-art performance is provided to show the achieved performance so far and demonstrate the potential of deep learning based methods.

Keywords: Scene character recognition · Scene text recognition · Character feature representation · Word recognition models

1 Introduction

Scene text recognition has attracted renewed interests in recent years due to the wide use of intelligent devices such as smart phones. In a typical application, for example, one needs to recognize text in an image captured by a mobile phones and translate text into other languages. As a result, numerous researchers devoted to the research of scene text recognition [1].

However, scene text recognition is a challenging problem due to the complexity of scene images such as background clutter, illumination changes, and the variation of text position, size, font, color and line orientation. Figure 1 shows some scene images that contain text. We can see that unlike the recognition of text in printed documents (this is also called Optical Character Recognition, OCR), which contain clean and well-formatted text, scene text detection and recognition is a more challenging problem.

Y.-J. Zhang (Ed.): ICIG 2015, Part III, LNCS 9219, pp. 310–320, 2015.
DOI: 10.1007/978-3-319-21969-1_27

Fig. 1. Some scene images that contain text.

In scene text recognition, there are four main problems: (1) text detection, (2) text recognition, (3) full image word recognition, and (4) isolated scene character recognition. Text detection is to locate text regions in an image; while text recognition, given text regions, is usually referred to as cropped word recognition. Full image word recognition usually includes both text detection and text recognition in an end-to-end scene text recognition system. Isolated character recognition is usually a basic component of a scene text recognition system. Hence the problem of isolated scene character recognition is also a fundamental problem and has attracted increasing attentions recently.

Numerous methods have been proposed for these problems, and have achieved great progresses in the past decade. Ten years ago two papers [2] and [3] addressed these problems and provided comprehensive surveys on text information extraction and camera-based document analysis. Recently, Zhang et al. [4] reviewed the text detection problem. More recently, Ye et al. [1] presented more general and extensive discussions on all the problems in scene text recognition.

However, we notice that, there are still some issues remained to be addressed. First, athough in [1] the achieved progresses of scene word recognition and end-to-end recognition are surveyed, the state-of-the-art of scene character recognition is ignored. Second, the papers published in some recent literature such as ECCV 2014, ACCV 2014, ICIP 2014, and ICPR 2014 are not included in [1]. In fact, some papers in these conferences have renewed the state-of-the-art. Third, the paper [1] describes the problems of scene word recognition and end-to-end recognition briefly while the methods proposed for these problems are not categorized in details.

In this paper, we focus on the problems of scene character recognition and scene text recognition (mainly in the context of cropped word recognition), and present the state-of-the-art of these problems. We do not review the text

detection problem and end-to-end recognition problem due to the space limitation. We review the most recently published papers on scene text recognition in ECCV 2014, ACCV 2014, ICIP 2014, and ICPR 2014. Specifically, we review the papers in two aspects: character feature representation and word recognition model, which are two important issues in scene text recognition.

The rest of the paper is organized as follows. In Sect. 2, we first review some public databases used for scene character and text recognition. In Sect. 3, we then introduce scene character recognition methods, focusing on the issue of character feature representation, and provide the state-of-the-art performance achieved so far. In Sect. 4, we review the problems of scene word recognition, focusing on the multiple information integration methods, and provide the state-of-the-art performance achieved so far. In Sect. 5, we conclude the paper with some discussion.

2 The Databases for Scene Character and Text Recognition

In this subsection, we summarize some publicly available datasets that are commonly used for scene character and text recognition. The most widely used datasets for scene character and/or text recognition include the ICDAR2003 dataset [5], the ICDAR2011 dataset [6], the Chars74K dataset [7], the Street View Text (SVT) dataset [8], and the III5K-Word dataset [9]. Among them, the ICDAR2003 and ICDAR2011 datasets are used for the "Robust OCR", "Robust Reading and Text Locating", and "Robust Word Recognition" competitions organized jointly with the ICDAR main conference. Hence the ICDAR2003 and ICDAR2011 datasets contain natural scene images, from which words and character samples can be cropped. The original SVT dataset contains natural scene images and cropped words only. Later, Mishra et al. provided the character level annotations of the test set of the SVT dataset [9]. The III5K-Word dataset is composed of word images, and character level annotations are provided by the authors. The Chars74K dataset is composed of isolated scene character samples.

Besides the datasets mentioned above, there are also some other datesets for research of scene text detection and recognition (see [1] for a brief review). Since this paper focuses on the state-of-the-art of scene character and text recognition, we only introduce the commonly used datasets for evaluation of scene character and text recognition.

The ICDAR2003 dataset contains 507 natural scene images including 258 training images and 249 test images in total. There are totally 1,156 words (including 6,185 character samples) cropped from the training set of the ICDAR2003 dataset, and 1,107 words (including 5,379 character samples) cropped from the test set of the ICDAR2003 dataset. The ICDAR2011 dataset contains 229 images for training (including 846 words) and 255 images for test (including 1,189 words). The SVT dataset is composed of 100 training images (including 258 words) and 249 test images (including 647 words that contains 3,796 character samples). For the ICDAR2011 dataset, only the words in the

images can be cropped because the images are annotated at word level only. The IIIT5K-Word dataset is the largest and most challenging dataset for word recognition so far. This dataset includes 5000 word images, where 2000 images are used for training and 3000 images for test. The Chars74K dataset contains nearly 74 thousand scene character samples.

In summary, for evaluating scene character recognition methods, the ICDAR2003 and Chars74K datasets are commonly used. For evaluating scene text recognition methods, the ICDAR2003, ICDAR2011, SVT, and III5K-Word datasets are usually used.

3 The State-of-the-Art Scene Character Recognition Methods

For scene character recognition, two important issues may affect the performance of scene character recognition: character feature representation methods and character classification methods. Whereas, much more attentions are paid to feature representation. For the character classification methods, the support vector machine (SVM) classifier (with a linear kernel or RBF or chi-square kernel) is one of the most popular one. Some other classifiers such as the random ferns classifier (FERNS) [8], the nearest neighbor (NN) classifier [8], random forest [10] and the convolutional neural network classifier (CNN) [11] have been adopted.

Since much more attentions are paid to character feature representation methods, in the following we mainly review the papers related to feature representation. We categorize the existing methods in three main kinds: HOG and its variants, mid-level character structural features, and deep learning based methods. Table 1 shows the state-of-the-art performance achieved so far. From the results, we can see that, the deep learning based feature learning methods achieve the highest performance.

3.1 HOG and Its Variants

The Histograms of Oriented Gradients (HOG) features have been shown to be effective and have been used in object detection [24], and for scene character feature representation [8,25]. Although HOG is very simple and is effective in describing local features (such as edges), HOG ignores the spatial and structural information. Hence some methods are proposed to improve HOG. For example, Yi et al. [22] improve the HOG features by global sampling (called GHOG) or local sampling (called LHOG) to better model character structures. Tian et al. [20] propose the Co-occurrence of Histogram of Oriented Gradients (called CoHOG) features, which capture the spatial distribution of neighboring orientation pairs instead of only a single gradient orientation, for scene text recognition. The CoHOG method improves HOG significantly in scene character recognition. Later, the authors of [20] propose the pyramid of HOG (called

Table 1. The state-of-the-art methods and their results for scene character recognition(%).

Method	Chars74K-15	ICDAR03-CH	SVT-CHAR	III5K
Deep CNN [12]	–	91.0	80.3	–
Maxout+Hybrid HMMs [13]	–	89.8	–	–
Feature learning (CNN) [14]	–	81.7	–	–
ConvCoHOG+Linear SVM [15]	–	81	75	–
CoStrokes [16]	67.5	82.7	–	–
PHOG+Chi-Square SVM [17]	–	79.0	74.7	75.8
Stroke Bank [18]	65.9	79.8		
Feature Pooling+L2 SVM [19]	64	79	–	–
CoHOG + Linear SVM [20]	–	79.4	75.4	–
HOG+AT+Linear SVM [21]	**68**	73	–	–
GHOG+Chi-Square SVM [22]	62	76	–	–
LHOG+Chi-Square SVM [22]	58	75	–	–
MSER [23]	-	67	–	–
HOG+NN [8]	58	52	–	–
MKL [7]	55	-	–	–
HOG+FERNS [8]	54	64	–	–
GB+RBF SVM [7]	53	-	–	–
ABBYY [7]	31	21	–	–

PHOG) [17] to encode the relative spatial layout of the character parts, and propose the convolutional CoHOG (called ConvCoHOG) to extract richer character features. These methods effectively improve the performance of scene character recognition.

3.2 Mid-level Character Structural Features

Character structure information is important to character representation and has been exploited in [10,19,26–28,30]. In [26–28], the authors propose to use part-based tree-structured features, which are originally designed for face detection [29] for representing character features. The part-based tree-structured features are designed directly according to the shape and structure of each character class. Yao et al. [10] propose to use a set of mid-level detectable primitives (called strokelets), which capture substructures of characters, for character representation. The strokelets are used in conjunction with the HOG features for character description, as supplementary features to the HOG features. In [19], a discriminative feature pooling method that automatically learns the most informative sub-regions of each scene character is proposed for character feature representation. Zhang et al. [30] propose to use sparse coding based features for capturing

character structures. The basis idea of [30] is to learn common structures with sparse coding and to capture character structures using histograms of sparse codes.

Recently, Gao et al. [18] propose a stroke bank based character representation method. The basic idea is to design a stroke detector for scene character recognition. In [16], Gao et al. propose to learn co-occurrence of local strokes by using a spatiality embedded dictionary, which is used to introduce more precise spatial information for character recognition. The results demonstrate the effectiveness of the two methods.

It is interesting to find that, some character feature representation methods mentioned above explore the mid-level features to describe character structures. Such as strokelets extracted in Yao et al. [10], the sub-regions learned by [19], and the stoke bank designed in [18], and the sub-structures learned by [30], they are all mid-level features. These learned mid-level features have shown their effectiveness in scene characeter/text recognition.

3.3 Deep Learning Based Methods

The deep learning methods have also been adopted for feature learning of scene characters. Coates et al. [14] propose a unsupervised feature learning method using convolutional neural networks (CNN) for scene character recognition. Recently, in ECCV 2014, Jaderberg et al. [12] develop a CNN classifier that can be used for both text detection and recognition. The CNN classifier has a novel architecture that enables efficient feature sharing using a number of layers in common for character recognition. The performance achieved by Jaderberg et al. [12] on both scene character recognition and text recognition is pretty high and is the best among the existing methods so far (see Sect. 4 for the performance of text recognition achieved by [12]).

4 The State-of-the-Art Scene Text Recognition Methods

Since the scene text recognition methods in end-to-end recognition systems are similar to those in cropped word recognition. In this paper, we mainly focus on the state-of-the-art of cropped word recognition. In the following, we review the methods in cropped word recognition methods.

4.1 A Typical Word Recognition Procedure

In a typical word recognition system, there are mainly two steps. The first step is character detection, which aims to simultaneously detect and recognize characters. In this step, a 63-class (10 digits, 52 English letters, and the outlier/backgound class) classifier is used to obtain character candidates and classify them. For generating character candidates, two strategies have been used: one is the sliding window strategy (such the work in [8,11,30], etc.), and one is to detect character candidates using the character detector/classifier directly (such as the

work in [10,26], etc.). In this step, the character feature representation methods play an important role in scene text recognition, and the performance of character classification highly affects the performance of scene text recognition.

The second step is the word formation step, which aims to combine character candidates to yield the word recognition result. In this step, multiple information can be integrated to help improve the performance of scene text recognition. An information integration model or a word recognition model can be used to integrate multiple information, which raises another important issue in scene text recognition. In the next subsection, we will briefly review the word recognition model (or score function or object function) in the literature. Figure 2 shows the scene text recognition procedure presented in [26], showing the results of the two steps.

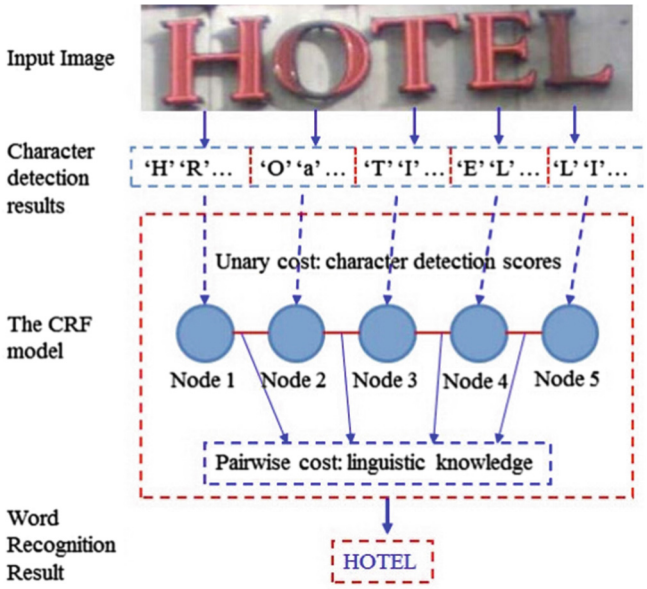

Fig. 2. A typical scene text recognition procedure. The images are referred to paper [26]. In this paper, a CRF model is used as the information integration model.

4.2 Information Integration Model/Word Recognition Model

Regarding the word recognition model for yielding word recognition results, Wang et al. [8] apply a lexicon-driven pictorial structures model to combine character detection scores and geometric constraints. Mishra et al. [25] build a conditional random field (CRF) model to integrate bottom-up cues (character detection scores) with top-down cues (lexicon prior). Similarly, Shi et al. [26] use a CRF model to get final word recognition results. In [11,12,31–33], heuristic integration models (summation of character detection scores) are used to integrate character detection result. In [28], a probabilistic model is proposed to

combine the character detection scores and a language model from the Bayesian decision view. In those works, the parameters of the word recognition model are set empirically.

In [30], Zhang et al. apply the lexicon-driven pictorial structures model similar to that in [8] for word recognition. However, they improve it by taking into account the influence of the word length (i.e., the number of characters in the word) to word recognition results. Moreover, they propose to learn parameters using the Minimum Classification Error (MCE) training method [34] to optimize scene text recognition. For searching the optimal word as the recognition result, the dynamic programming algorithm is commonly used, such as the work in [8,12,30], etc.

4.3 Word Spotting Based Methods Versus Open Vocabulary Based Methods

For cropped word recognition, the existing scene text recognition methods can be categorized into two kinds: word spotting based methods [8,11,12,30,32,35] and open vocabulary based methods [9,10,19,25,26,31,36–39]. For word spotting based methods, a lexicon is provided for each cropped word image, and the optimal word is the one yielding the maximum matching score. This is similar to a word spotting procedure. For open vocabulary based methods, language prior or language model is obtained using a general larger corpus, from which the language prior or language model can be estimated.

Since the work of Wang et al. [8], most papers on scene text recognition report results using lexicons consisting of a list of words (which can be 50 words containing the ground truth word or the words created from all the words in the test set, called Full lexicons). That is, for open vocabulary based methods, one needs to retrieve the word with the smallest edit distance in the lexicon as the recognition result, such as [25,26], etc.

4.4 The State-of-the-art Performance of Scene Text Recognition

We show the state-of-the-art performance of scene text recognition in Table 2. In the table, SVT, I03, I11, and III5K denotes the SVT, ICDAR2003, ICDAR2011, and III5K-Word dataset, respectively. In the end of each name of the dataset, the number "50" means using the lexicon consisting of 50 words; the word "Full" means using the lexicons created from words of the test set; and the word "Med" means using the Medium lexicon provided by the authors of [9].

From the table, we can see that the PhotoOCR method presented in [33] report the highest performance on SVT-50, achieving accuracy of 90.3 % on SVT-50. On I03-50 and I03-Full, the method proposed in [12] performs the best, achieving accuracy of 96.2 % and 91.5 % on I03-50 and I03-Full, respectively. It is worth noting that both [33] and [12] adopt deep learning based methods. This demonstrates the potential advantages of the deep learning based methods. Only a few works report performance on I11-50, I11-Full, III5K-50 and III5K-Med. On I11-50 and I11-Full, Shi et al. [28] report promising performance, achieving

Table 2. The state-of-the-Art performance of scene text recognition. (%)

Method	SVT-50	I03-50	I03-Full	I11-50	I11-Full	III5K-50	III5K-Med
K. Wang et al. [8]	57	76	62	–	–	–	–
Mishra et al. [25]	73.26	81.78	–	–	–	68.25	55.50
Mishra et al. [9]	73.57	80.28	–	–	–	66	57.5
Novikova et al. [39]	72.9	82.8	–	–	–	–	–
T. Wang et al. [11]	70	90	84	–	–	–	–
Yildirim et al. [37]	–	85.70	–	–	–	–	–
Shi et al. [26]	73.51	87.44	79.30	87.04	82.87	–	–
Goel et al. [35]	77.28	89.69	–	–	–	–	–
Weinmann et al. [40]	78.05	–	–	–	–	–	–
Shi et al. [27]	74.65	84.52	79.98	–	–	–	–
Shi et al. [28]	73.67	87.83	79.58	87.22	83.21	–	–
Zhang et al. [30]	74.34	88.24	80.56	–	–	–	–
Yao et al. [10]	75.89	88.48	80.33	–	–	80.2	69.3
Lee et al. [19]	80	88	76	88	77	–	–
Bissacco et al. [33]	90.3	–	–	–	–	–	–
Su et al. [32]	83	92	82	91	83	–	–
Alsharif et al. [13]	74.3	93.1	88.6	–	–	–	–
Jaderberg et al. [12]	86.1	96.2	91.5	–	–	–	–

accuracy of 87.22 % and 83.21 % on I11-50 and I11-Full, respectively. On III5K-50 and III5K-Med, Yao et al. [10] report promising results, achieving accuracy of 80.2 % and 69.3 % on III5K-50 and III5K-Med, respectively.

5 Conclusions

This paper reviews the state-of-the-art of scene character and text recognition, with emphasize on character feature representation and word recognition models. The performance of scene character recognition and text recognition obtained by the recently proposed methods on both scene character recognition and text recognition are reviewed, including the most recent papers in ECCV 2014, ACCV 2014, ICIP 2014, and ICPR 2014. From the reported results, we can see that the deep learning based methods achieve the highest performance, indicating that this type of methods open a new direction for scene character and text recognition. Character feature representation, as a basic component of scene character and text recognition systems, will also be an important research direction in the future.

Acknowledgment. This work was supported by the National Natural Science Foundation of China under Grants 61305004 and 61472334, by the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 20110121110033, and by the Fundamental Research Funds for the Central Universities under Grant 20720130720.

References

1. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey, IEEE. Trans. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2015)
2. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. Pattern Recognit. **37**(5), 977–997 (2004)
3. Liang, J., Doermann, D., Li, H.: Camera-based analysis of text and documents: a survey. Int. J. Doc. Anal. Recognit. **7**(2–3), 84–104 (2005)
4. Zhang, H., Zhao, K., Song, Y.-Z., Guo, J.: Text extraction from natural scene image: a survey. Neurocomputing **122**, 310–323 (2013)
5. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R.: ICDAR 2003 robust reading competitions. In: Proceedings of International Conference on Document Analysis and Recognition (2003)
6. Shahab, A., Shafait, F., Dengel, A.: ICDAR 2011 robust reading competition challenge 2: reading text in scene images. In: Proceedings ICDAR, pp. 1491–1496 (2011)
7. de Campos, T.E., Babu, B.R., Varma, M.: Character recognition in natural images. In: Proceedings of International Conference on Computer Vision Theory and Applications, Lisbon (2009)
8. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: Proceedings ICCV, pp. 1457–1464 (2011)
9. Mishra, A., Alahari, K., Jawahar, C.V.: Scene text recognition using higher order language priors. In: Proceedings BMVC, pp. 1–11 (2012)
10. Yao, C., Bai, X., Shi, B., Liu, W.: Strokelets: a learned multi-scale representation for scene text recognition. In: Proceedings CVPR (2014)
11. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: Proceedings ICPR, pp. 3304–3308 (2012)
12. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 512–528. Springer, Heidelberg (2014)
13. Alsharif, O., Pineau, J.: End-to-end text recognition with hybrid HMM maxout models. In: International Conference on Learning Representations (2014)
14. Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D.J., Ng, A.Y.: Text detection and character recognition in scene images with unsupervised feature learning. In: Proceedings ICDAR, pp. 440–445 (2011)
15. Su, B., Lu, S., Tian, S. , Lim, J.-H., Tan, C.L.: Character recognition in natural scenes using convolutional co-occurrence HOG. In: Proceedings ICPR, pp. 2926–2931 (2014)
16. Gao, S., Wang, C., Xiao, B., Shi, C., Zhou, W., Zhang, Z.: Learning co-occurrence strokes for scene character recognition based on spatiality embedded dictionary. In: Proceedings ICIP, pp. 5956–5960 (2014)
17. Tan, Z.R., Tian, S., Tan, C.L.: Using pyramid of histogram of oriented gradients on natural scene text recognition. In: Proceedings ICIP, pp. 2629–2633 (2014)
18. Gao, S., Wang, C., Xiao, B., Shi, C., Zhang, Z.: Stroke bank: a high-level representation for scene character recognition. In: Proceedings ICPR, pp. 2909–2913 (2014)
19. Lee, C.-Y., Bhardwaj, A., Di, W., Jagadeesh, V., Piramuthu, R.: Region-based discriminative feature pooling for scene text recognition. In: Proceedings CVPR, pp. 4050–4057 (2014)
20. Tian, S., Lu, S., Su, B., Tan, C.L.: Scene text recognition using co-occurrence of histogram of oriented gradients. In: Proceedings ICDAR, pp. 912–916 (2013)

21. Mishra, A., Alahari, K., Jawahar, C.V.: Image retrieval using textual cues. In: Proceedings ICCV, pp. 3040–3047 (2013)
22. Yi, C., Yang, X., Tian, Y.: Feature representations for scene text character recognition: a comparative study. In: Proceedings ICDAR, pp. 907–911 (2013)
23. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 770–783. Springer, Heidelberg (2011)
24. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings CVPR, pp. 886–893 (2005)
25. Mishra, A., Alahari, K., Jawahar, C.V.: Top-down and bottom-up cues for scene text recognition. In: Proceedings CVPR, pp. 2687–2694 (2012)
26. Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S., Zhang, Z.: Scene text recognition using part-based tree-structured character detection. In: Proceedings CVPR, pp. 2961–2968 (2013)
27. Shi, C., Wang, C., Xiao, B., Gao, S., Hu, J.: End-to-end scene text recognition using tree-structured models. Pattern Recognit. **47**(9), 2853–2866 (2014)
28. Shi, C., Wang, C., Xiao, B., Gao, S., Hu, J.: Scene text recognition using structure-guided character detection and linguistic knowledge. IEEE Trans. Circuits Syst. Video Technol. **24**(7), 1235–1250 (2014)
29. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proceedings CVPR, pp. 2879–2886 (2012)
30. Zhang, D., Wang, D.-H., Wang, H.: Scene text recognition using sparse coding based features. In: Proceedings ICIP, pp. 1066–1070 (2014)
31. Neumann, L., Matas, J.: Scene text localization and recognition with oriented stroke detection. In: Proceedings ICCV, pp. 97–104 (2013)
32. Su, B., Lu, S.: Accurate scene text recognition based on recurrent neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9003, pp. 35–48. Springer, Heidelberg (2015)
33. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: PhotoOCR: reading text in uncontrolled conditions. In: Proceedings ICCV, pp. 785–792 (2013)
34. Juang, B.-H., Chou, W., Lee, C.-H.: Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. **5**(3), 257–265 (1997)
35. Goel, V., Mishra, A., Alahari, K., Jawahar, C.V.: Whole is greater than sum of parts: recognizing scene text words. In: Proceedings ICDAR, pp. 398–402 (2013)
36. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: Proceedings CVPR, pp. 3538–3545 (2012)
37. Yildirim, G., Achanta, R., Susstrunk, S.: Text recognition in natural images using multiclass hough forests. In: Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2013)
38. Feild, J.L., Learned-Miller, E.G.: Improving open-vocabulary scene text recognition. In: Proceedings ICDAR, pp. 604–608 (2013)
39. Novikova, T., Barinova, O., Kohli, P., Lempitsky, V.: Large-lexicon attribute-consistent text recognition in natural images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 752–765. Springer, Heidelberg (2012)
40. Weinman, J.J., Butler, Z., Knoll, D., Feild, J.L.: Toward integrated scene text reading. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 375–387 (2014)