International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Scene Text Recognition Using Part-Based Tree Structured Models and Linguistic Knowledge

Namrata R. Purohit[1], Dr. Vinayak G. Kottawar[2]

[1]Department of Computer Science and Engineering, Mahatma Gandhi Mission's College of Engineering, Nanded, Maharashtra, India

[2]Assistant Professor, Department of Computer Science and Engineering, Mahatma Gandhi Mission's College of Engineering, Nanded, Maharashtra, India

Abstract: ***Scene text recognition has nice interests from the computer vision society in recent years. In this paper, a unique scene text-recognition methodology is proposed, which is an integration of structure-guided character detection and linguistic knowledge. In this paper, a part-based tree structure is used to model each and every category of characters to detect and recognize characters at the same time. Since the character models make use of each the native appearance and global structure information, so the detection results are more reliable. For word recognition, combine the detection scores and language model into the posterior probability of character sequence from the Bayesian decision view. The final word-recognition result's obtained by increasing the character sequence posterior probability via Viterbi algorithm. Experimental results on a spread of difficult public datasets demonstrate that the projected methodology achieves progressive performance each for character detection and word recognition.***

Keywords: Character recognition, cropped word recognition, part-based tree-structured models (TSMs), posterior probability, scene text recognition

1. Introduction

In the past decade, the use of camera-based applications is increased, due to this growth readily available on smart phones and portable devices, understanding the photographs taken by these devices semantically has gained increasing attention from the computer vision community. Among all the information contained within the image, text that carries linguistics information, may provide valuable cues about the content of the image and therefore is very important for human as well as computer to know the scenes. In an experimental study, Judd et al. [1] found that given a picture containing text and other different objects, viewers tend to fixate on text. This additional demonstrates that text recognition is very important for humans to know the scenes. In fact, text recognition is indispensable for lots of applications like automatic sign reading, language translation, navigation, and so on. Generally speaking, to know the knowledge carried by text within the image, we'd like to recognize the text.

They adopt multistate window strategy to notice characters and directly extract options from the first image to acknowledge the characters. The performance of some recently a planned technique [5]–[7] is kind of promising. However, thanks to the at liberty lighting conditions, numerous fonts, deformations, occlusions, typically low resolution, and complicated background of text in natural scene pictures, the performance of scene text recognition continues to be unacceptable. It shows some samples of scene text pictures. As we are able to see, thanks to the high degree of intra-class variation of scene characters likewise because the quality of background, recognizing these text pictures is kind of difficult even for progressive OCR strategies.

In fact, characters area unit designed by humans and every class of characters has distinctive structure representing itself. Therefore, regardless of however the background changes or the character degrades, as long because the structure remains carved in stone, we tend to might acknowledge them by detective work the distinctive structure from untidy background.

Figure 1: Some scene text images from ICDAR 2003 [2] and SVT [3]. The characters in these images have different fonts, shadows, distortions, deformations, low resolutions, and occlusions

In alternative words, humans naturally create use of character-specific structure info once recognizing characters from scene pictures. Thus, a decent scene character-recognition technique ought to create use of each the native look and world structure info. In this paper, we tend to propose a unique scene text-recognition technique combining structure-guided character detection and linguistic data. As character detection and recognition is that the premise for word recognition, it plays a crucial role for the performance.

Thus, we tend to propose a unique and effective character detection approach. Completely different from standard multistate window character detection strategy, we tend to

use part-based tree-structure to model every class of characters thus on collectively notice and acknowledge characters. The characters might be recognized by detective work character-specific structures, seamlessly combining detection, and recognition along. As each the worldwide structure and therefore the native look information's contribute to the part-based tree-structured models (TSMs) for characters, the detection results area unit additional reliable. To acknowledge the scene text, we tend to mix the detection scores and language model into the posterior chance of character sequence from the Bayesian call read. The ultimate word-recognition result's obtained by increasing the character sequence posterior chance via Viterbi algorithmic rule. Since the aim of word-recognition is to grasp the words and therefore the case insensitive results might satisfy this would like, we tend to solely report case insensitive word-recognition ends up in this paper. We tend to measure our technique on a variety of difficult knowledge sets. Experimental results show that our technique achieves progressive performance each for character detection and word recognition.

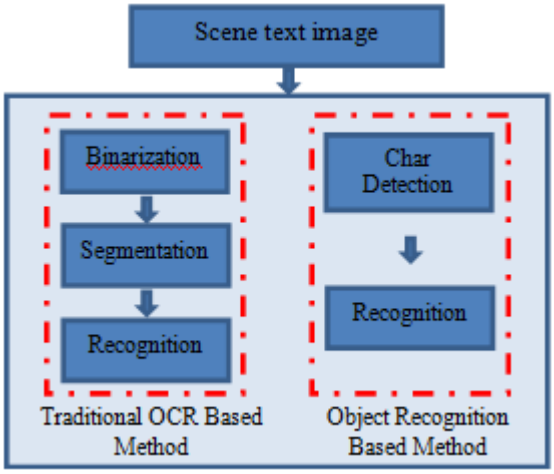

Figure 2: Illustration of the traditional OCR-based method and object recognition-based method

2. Related work

Most of the scene text detection algorithms in the literature can be classified into Region-based and Connected Component (CC)-based approaches. Region-based methods adopted a sliding window scheme, which is basically a brute force approach which requires a lot of local decisions. Therefore, the region-based methods have focused on an efficient binary classification (text versus non text) of a small image patch. Text Localization is of fundamental importance in image understanding and content based retrieval. For instance the localization must always be achieved prior to Optical Character Recognition (OCR). Stability of such method includes robustness to noise and blurness because they accomplish features assembled throughout the region of interest. The second approach used is localizing the individual characters using the local parameters of an image (intensity, stroke-width, color, gradient etc). Feature extraction also plays a vital role in image localization process. The main goal of feature extraction is to maximize the recognition rate with minimum number of elements used in it. After analyzing existing feature descriptor methods it is found experimentally Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for character detection and best suited for the proposed system. Many researchers have made research related to this but no technique is almost perfect and they found need to improve the work in more areas at different instants and techniques.

2.1 Traditional OCR-based Methods

For traditional OCR-based methods as shown they focus on the binarization process, which segments text from background, and then the binary image could be segmented into individual characters, which could be recognized by the OCR engine. Various approaches have been proposed to binarize images with low quality or complex background. Gllavata et al. [14] first used a color quantizer to determine the color of text and background, and then adopted the modified k-means algorithm to classify pixels into text and background. Song et al. [15] used color-based k-means clustering to segment text from background. The performance of color clustering methods is dependent on the color consistency and also sensitive to noise and text resolution. Chen et al. [16] used mixtures of Gaussians to model the gray level distributions in the image and assigned the pixels to one of the Gaussian layer based on the prior of the contextual information modeled by a Markov random field. Ye et al. [17] proposed to train the Gaussian mixture models (GMM) of intensity and hue components in HSI color space using sampled pixels and utilized the GMM together with the spatial connectivity information to segment text pixels from the background. Li et al. [18] proposed to integrate local visual information and contextual label information into a conditional random field to segment text from complex background. Shi et al. [19] proposed to binarize video text images using graph cut algorithm based on the automatic acquired hard constraint seeds. Recently, Shivakumara et al. [20] introduced a novel ring radius transform and the concept of medial pixels on characters with broken contours in the edge domain for reconstruction to improve the character-recognition rate in video images. Feild et al. [21] proposed to use bilateral regression to model smooth color changes across an image region without being corrupted by neighboring image regions and use feedback from a recognition system to choose the best foreground region. Compared with the conventional scanned document binarization methods such as Otsu and Niblack, some recent binarization or preprocessing methods could indeed improve the scene text-recognition rates. However, since text in natural images has unconstrained resolution, illumination condition, size and font style, and the binarization results are unsatisfactory. Moreover, the loss of information during the binarization process is almost irreversible, which means if the binarization result is poor; the chance of correctly recognizing the text is very small. As shown in Fig. 3, the binarization results are very disappointing, making it almost impossible for the following segmentation and recognition.

Fig. 3. The binarization results are quite disappointing, making it difficult for the following segmentation and recognition. (a) Scene text images. (b) Binarization result.

2.2 Object Recognition-based Methods

On the other hand, object recognition-based methods assume that scene character recognition is quite similar to object recognition with a high degree of intra-class variation. For scene character recognition, these methods directly extract features from original image and use various classifiers to recognize the character. Chen et al. [22] proposed a local intensity normalization method to handle lighting variations, then used a Gabor transform to obtain local features and finally adopted a linear discriminant analysis for feature selection. Weinman et al. [26] proposed to incorporate character appearance, bigram frequencies, similarity and lexicons into the recognition process. Smith et al. also proposed to incorporate character similarity information to improve recognition performance. Based on bag-of-visual-words framework, De Campos et al. [23] benchmarked the performance of various features to assess the feasibility of posing the problem as an object-recognition task and showed that geometric Blur [7] and shape context in conjunction with nearest neighbor (NN) classifier, performed better than other methods. Wang et al. [24] proposed to use histograms of oriented gradients (HOG) in conjunction with an NN classifier and reported better performance. Newell and Griffin proposed two extensions of HOG descriptor to include features at multiple scales and their method achieved promising performance on two data sets, chars74k and ICDAR03-CH [2], using the same evaluation framework as De Campos. Coates et al. [25] took an unsupervised approach to learn features from unlabeled data and the character-recognition results on the ICDAR03-CH [2] are quite promising. While for object recognition-based scene text recognition, since there are no binarization and segmentation stages, as shown, most of the existing methods adopt multiscale sliding window strategy to get the candidate character detection results. Then, word-recognition strategies, such as pictorial structures or CRF are used to get the final word-recognition results from the candidate character detection results. Elagouni et al. adopted the convolutional neural networks (CNN) to get the candidate character detections and a graph model is used to determine the best sequence of characters. Novikova et al. proposed to use maximally stable extremal regions as character candidates and formulate the problem of word recognition as the maximum a posteriori inference in a unified probabilistic framework. Recently, Wang et al. [18] used CNN to train the text detection and character-recognition modules and recognize the words with non-maximal suppression (NMS) and beam search with the help of the lexicon. Weinman et al. [26] proposed to use probabilistic methods to coarsely binarized a given text region and jointly perform word and character segmentation during the recognition process, which achieved state-of-the-art performance.

2.3 Structure-based Model for Object Detection

Structure-based model, which captures the local appearance properties and the deformable configuration of an object, has inspired great interest for object detection, since Felzenszwalb and Huttenlocher proposed the pictorial structures framework for object recognition. In this framework, objects are represented by a collection of parts arranged in a deformable configuration, which has been proved to be effective for many applications. To deal with object by significant variations, Felzenszwalb et al. proposed to use mixtures of star structured model defined by a root filter plus a set of part filters and deformation models. Impressively, their method won the first place on PASCAL visual object detection challenge 2008 and 2009. Later, Yang and Ramanan proposed to detect articulated pose of human using flexible mixtures-of-parts. Tree structure is used to model co-occurrence and spatial relations, and the model could be efficiently optimized with dynamic programming. Recently, Zhu and Ramanan proposed to jointly address the tasks of face detection, pose estimation, and landmark estimation using mixtures of trees with a shared pool of parts. Although their model is only trained with hundreds of faces, it compares favorably with the commercial systems trained with billions of examples. Characters are designed by humans and each class of characters has a unique structure representing itself. Thus, we should try to use the global structure as well as local appearance information for character detection and recognition. The proposed approach is closely related to those methods in. We use modified part-based TSM to detect candidate characters.

3. Design and Implementation

The proposed research started with the objective of processing and refining image dataset that we are using in the proposed framework and algorithm. In this process, following steps and processes are evolved which lead to development of the research work. To find the proposed objectives the proposed work mainly works upon two algorithms.

Text Segmentation and Localization: In this phase, we have to segment the characters in the image by bounding box method or by bounding the characters by edges and after that finally we are in position to normalize the image to find out correct characters present in the input images and localize the text in the input image.

Recognition: On the basis of above all procedure of the proposed algorithm finally the recognition and text localization of input images will be done in this phase.

4. System Overview

As mentioned above, a good scene character-recognition method should make use of both the local appearance and global structure information. Motivated by the recent progress in object detection using part-based TSM, we find that we could adopt these models to use both the global structure and local appearance information of characters.

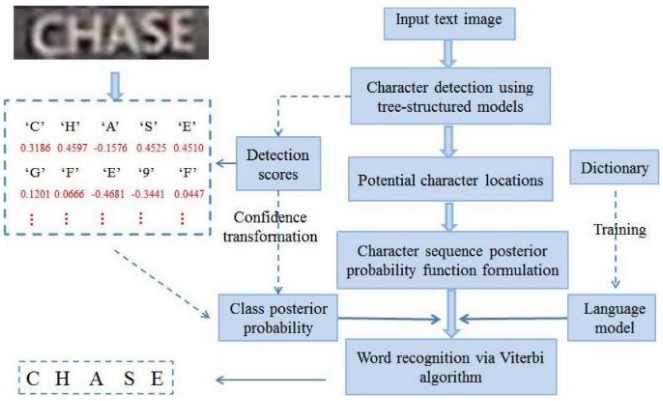

Figure 4: Flowchart of the proposed system

The flowchart of the proposed method is shown. Given an input text image, first, we use part-based TSM for all the categories of characters to detect the character-specific structures, based on which we get the potential character locations. Then, we convert the candidate detection scores to posterior probabilities via confidence transformation. For word recognition, we combine the detection scores and language model into the posterior probability of the character sequence from the Bayesian decision view. Bigram, trigram, and even higher order language model could be incorporated. The final word-recognition result is obtained by finding the most probable character sequence via Viterbi algorithm.

5. Character detection using part-based TSM

We propose to recognize characters by detecting part based tree-structures, which seamlessly combines detection and recognition together. Briefly shows how to train the TSM for character it shows how to recognize the characters using the trained character models. Next, we will give details about the model, the inference of the character specific structures and the learning of the parameters of the models.

5.1 Model

To make use of both global structure and local appearance information of characters, we build a part-based TSM for each category of characters.

1) Model for Characters: We represent each category of characters by a tree $T_k = (V_k, E_k)$, where k is the index of the model for different structures, V_k represents the nodes, and E_k specifies the topological relations of nodes [7]. Each node represents a part of the character. Let I represents the input image and $l_i = (x_i, y_i)$ denotes the location of part i . Then, the score of the configuration of all the parts-

$L = \{l_i, i \in V_k\}$ could be defined as

$$S(L, I, k) = S_{App}(L, I, k) + S_{Str}(L, k) + \alpha_k \quad \ldots(1)$$

Where

$$S_{App}(L, I, k) = \sum_{i \in V_k} w_i^k . \phi(I, l_i) \quad \ldots(2)$$

$$S_{Str}(L, k) = \sum_{ij \in E_k} w_{ij}^k . \psi(l_i - l_j) \quad \ldots(3)$$

As we can observe, the total score of a configuration *L* for model *k* consists of the local appearance score in (2), the structure or shape score in (3), and the bias α_k . Here, α_k is a scalar bias or prior associated with character k. Next, we will give details about the appearance model and the shape model.

2) Local Appearance Model: Equation (2) is the local appearance model, which reflects the suitability of putting the part-based models on the corresponding positions. w_i^k represents the filter or the model for part *i* , structure *k*, and $\phi(I, l_i)$ denotes the feature vector extracted from location l_i. Thus, the score of placing part w_i^k on position li is actually the filter response of template w_i^k . We choose HOG [9] as the local appearance descriptor due to its good performance on many computer vision tasks. For color image, we choose the color channel with the largest gradient magnitude to calculate the HOG features.

3) Global Structure Model: Equation (3) is the structure or shape model, which scores the character-specific global structure arrangement of configuration *L*. Here, we set $\psi(l_i - l_j) = [dx\ dx^2\ dy\ dy^2]$, where $dx = x_i - x_j$ and $dy = y_i - y_j$ are the relative distance from part *i* to part *j*. Each term in the sum acts as a spring that constrains the relative spatial positions between a pair of parts.

5.2 Inferring Character-specific Structures

Inferring the character-specific structure corresponds to maximizing *S(L, I, k)* in (1) over all the possible configurations of all the parts, L and all the classes of characters k. Since the TSMs for all the characters are independent from each other, we could maximizing *S(L, I)* for all the structures in parallel. Thus, for each structure, we need to maximize *S(I)* over *L* S∗(I) = max L S(I, L).

6. Word-Recognition Model

The text images are normalized to the same height while preserving the aspect ratio. Since some lowercase character might only have half the size of the uppercase ones, image pyramid is used to deal with characters with different sizes. Although the character detection step provides us with a set of windows containing characters with high confidence,

inevitably it also produces some false positives and ambiguities between the similar characters.

If we only use the detection result to recognize the word, the results would be incorrect. Thus, we need to make use of other information, such as language model to eliminate these ambiguities. To this end, we combine the detection scores and language model from the Bayesian decision view.

7. Results and Discussion

We collected over 45 images from various dataset. We used our proposed approach of scene text localization and recognition and found that the efficiency of finding and localization of characters, text and image is quiet higher for every input image at various instants and rotation.

We give an image as an input to the system, Shown in fig. 5. By using trained, extraction of characters i.e. text localization is done. For this, we have employed canny edge detection, region filtering and finally stroke width technique to extract text regions from MSER, whose results are shown in fig. 6.

Figure 5: Input image

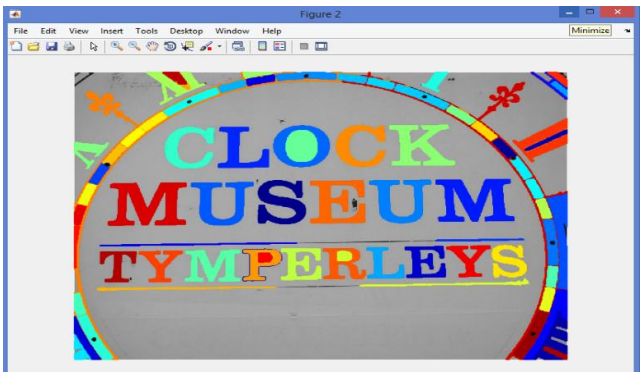

Figure 6: Image using MSER features

Stroke width is useful discriminator for text in images, is the variation in stroke width within each text candidate. We consider patterns in the intermediate image as strokes (as shown in fig 7). And try to divide it as individual character by cropping the selected part (as shown in figure 8). Final cutouts for the characters are shown in fig 9.

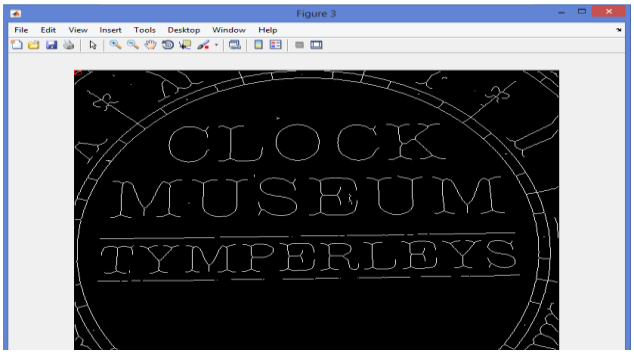

Figure 7: Image after finding strokes (Intermediate image)

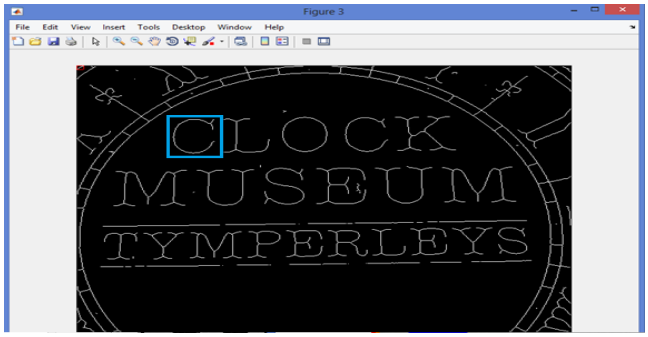

Figure 8: Image after finding strokes: selection of each character

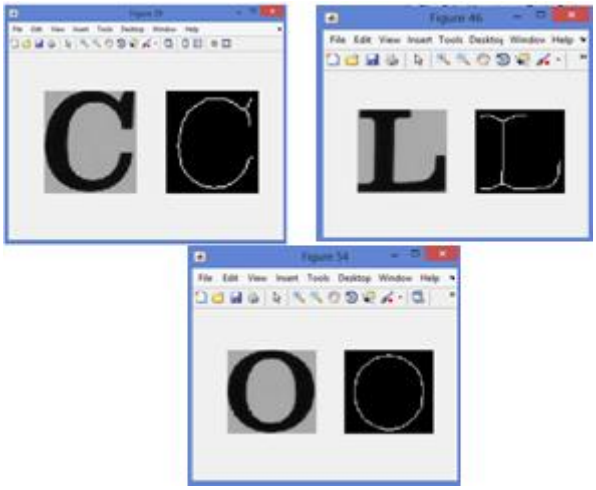

Figure 9: Characters extracted from the text

With the help of these cutouts, we recognize the actual characters, final outcome is shown below in fig 10.

Figure 10: Recognizing the text from the scene text image

The efficiency can be defined as the total number of accuracy based on the features, segmentation and recognition of image and some other properties. Higher will be the efficiency

higher will be the results accurate and effective. Efficiency = Σ total number of favorable condition on the basis of features/ Total number of conditions.

The efficiency of the proposed system for all input images taken in the dataset and it is found that the performance of proposed work is quite higher for every images and overall its total average efficiency is around 66% which is quite efficient and optimist for any system.

8. Conclusion

The main objective of this paper is to localize and recognize real time scene text images. It has been found that each technique has its own benefits and limitations, no technique are best for every case. The proposed algorithm was made over each images stored in the set of data set of Real Time Scene Text images at various instants and rotation. The proposed algorithm is the combination of TSM transformation and linguistic algorithm. Moreover, various other methods like preprocessing, binarization, noise removal, segmentation, localization and recognition would be done effectively. The results clearly depicts that the value of efficiency for all the images stored in the set of dataset is quiet high and approximately an average of 66 % efficiency for the entire process. Furthermore, there is need to improve the results more by introducing more feature extraction phase. There is also need to use of some more technique and maybe comparison of different technique. As a future scope, one may also choose some more parameters to compare the results.

References

[1] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in Proc. IEEE 12th ICCV, Oct. 2009, pp. 2106–2113.
[2] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in Proc. 7th ICDAR, vol. 2. 2003, pp. 682–687.
[3] J. Gao and J. Yang, "An adaptive algorithm for text detection from natural scenes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 2. Dec. 2001, pp. 1–6.
[4] X. Chen and A. Yuille, "Detecting and reading text in natural scenes," in Proc. IEEE CVPR, vol. 2. Jul. 2004, pp. 366–373.
[5] M. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 2, pp. 243–255, Feb. 2005.
[6] Q. Ye, Q. Huang, W. Gao, and D. Zhao, "Fast and robust text detection in images and video frames," Image Vis. Comput., vol. 23, no. 6, pp. 565–576, Jan. 2005.
[7] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in Proc. IEEE CVPR, Jun. 2010, pp. 2963–2970.
[8] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in Proc. IEEE CVPR, Dec. 2012, pp. 3538–3545.
[9] 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, Mar. 2011.
[10] Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang, and S. Tang, "A novel image text extraction method based on K-means clustering," in *Proc. 7th IEEE/ACIS ICIS*, May 2008, pp. 185–190.
[11] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhong, "Scene text recognition using part-based tree-structured character detection," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2961–2968.
[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th ICCV*, Oct. 2009, pp. 2106–2113.
[13] 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, Mar. 2011.
[14] J. Gllavata, R. Ewerth, T. Stefi, and B. Freisleben, "Unsupervised text segmentation using color and wavelet features," in Image and Video Retrieval. New York, NY, USA: Springer-Verlag, 2004.
[15] Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang, and S. Tang, "A novel image text extraction method based on K-means clustering," in Proc. 7th IEEE/ACIS ICIS, May 2008, pp. 185–190.
[16] D. Chen, J. Olobez, and H. Bourlard, "Text segmentation and recognition in complex background based on Markov random field," in Proc. 16th Int. Conf. Pattern Recognit., vol. 4. 2002, pp. 227–230.
[17] Q. Ye, W. Gao, and Q. Huang, "Automatic text segmentation from complex background," in Proc. ICIP, vol. 5. Oct. 2004, pp. 2905–2908
[18] M. Li, M. Bai, C. Wang, and B. Xiao, "Conditional random field for text segmentation from images with complex background," Pattern Recognit. Lett., vol. 31, no. 14, pp. 2295–2308, Oct. 2010.
[19] C. Shi, B. Xiao, C. Wang, and Y. Zhang, "Adaptive graph cut based binarization of video text images," in Proc. 10th IAPR Int. Workshop DAS, May 2012, pp. 58–62.
[20] P. Shivakumara, T. Phan, S. Bhowmick, C. Tan, and U. Pal, "A novel ring radius transform for video character reconstruction," Pattern Recognit., vol. 46, no. 1, pp. 131–140, 2012.
[21] J. Feild and E. G. Learned-Miller, "Scene text recognition with bilateral regression," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep. UM-CS-2012-021, 2012.
[22] X. Chen, J. Yang, J. Zhang, and A. Waibel, "Automatic detection and recognition of signs from natural scenes," IEEE Trans. Image Process., vol. 13, no. 1, pp. 87–99, Jan. 2004.
[23] T. de Campos, B. Babu, and M. Varma, "Character recognition in natural images," in Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl., Feb. 2009, pp. 1–4.
[24] A. Newell, and L. Griffin, "Multiscale histogram of oriented gradient descriptors for robust character recognition," in Proc. IEEE ICDAR, Sep. 2011, pp. 1085–1089.

[25] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, et al., "Text detection and character recognition in scene images with unsupervised feature learning," in Proc. IEEE ICDAR, Sep. 2011, pp. 440–445.
[26] J. Weinman, E. Learned-Miller, and A. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 10, pp. 1733–1746, Oct. 2009.