Text Detection and Recognition in Imagery: A Survey

Qixiang Ye, *Member, IEEE* and David Doermann, *Fellow, IEEE*

Abstract—This paper analyzes, compares, and contrasts technical challenges, methods, and the performance of text detection and recognition research in color imagery. It summarizes the fundamental problems and enumerates factors that should be considered when addressing these problems. Existing techniques are categorized as either stepwise or integrated and sub-problems are highlighted including text localization, verification, segmentation and recognition. Special issues associated with the enhancement of degraded text and the processing of video text, multi-oriented, perspectively distorted and multilingual text are also addressed. The categories and sub-categories of text are illustrated, benchmark datasets are enumerated, and the performance of the most representative approaches is compared. This review provides a fundamental comparison and analysis of the remaining problems in the field.

Index Terms—Text detection, text localization, text recognition, survey

1 Introduction

The problems of text detection and recognition in images and video have received increased attention in recent years, as indicated by the emergence of recent "robust reading" competitions in 2003, 2005, 2011, and 2013 [21], [43], [44], [144], along with bi-annual international workshops on camera-based document analysis and recognition (CBDAR) from 2005 to 2013. The emergence of applications on mobile devices [42], including the iPhone and Android platforms, which translate text into other languages in real time, has stimulated renewed interest in the problems.

Several primary reasons for this trend exists, including the demand of a growing number of applications. Text is one of the most expressive means of communications, and can be embedded into documents or into scenes as a means of communicating information. This is done in the way that makes it "noticeable" and/or readable by others. The collection of massive amounts of "street view" data is just one driving application. The second factor is the increasing availability of high performance mobile devices [26], [77] with both imaging and computational capability. This creates an opportunity for image acquisition and processing anytime, anywhere, making it convenient to recognize text in various environments. The third is the advance in computer vision and pattern recognition technologies, making it more feasible to address challenging problems.

While many researchers view optical character recognition (OCR) as a solved problem, text detection and recognition in imagery possess many of the same hurdles as computer vision and pattern recognition problems driven by lower quality or degraded data. Ample room for research exists, as suggested by the low detection rates (often less than 80 percent) [208] and recognition rates (often less than 60 percent) of state-of-the-art approaches [189], [204], [206]. By contrast, OCR typically achieves recognition rates higher than 99 percent on scanned documents [100]. Complex backgrounds, variations of text layout and fonts, and the existence of uneven illumination, low resolution and multilingual content present a much greater challenge than clean, well-formatted documents. Solving these problems requires the application of advanced computer vision and pattern recognition techniques.

Numerous methods have been proposed to detect and recognize text in scene imagery, yet we are unaware of comprehensive surveys of the subject during the past five years. Two surveys have been conducted on text information extraction [32] and camera-based document analysis [40], but most of the reviewed literature was published before 2003. Much of the work that has been published since then has made incremental advances to the state of the art, so establishing a baseline for future work remains important.

This paper attempts to establish this baseline by providing a comprehensive literature survey of text detection and recognition research. We summarize the problems and sub-problems, review the applications, and analyze the challenges. We then define various taxonomies to compare representative methods and approaches. We also highlight the state of the art by reporting performance of the representative approaches in publicly available datasets.

The paper is organized as follows. The remainder of Section 1 summarizes the problems and the progress made during the past decade. Related background is analyzed in Section 2. Methodologies, sub-problems and relevant issues are presented in Sections 3, 4 and 5, respectively. Datasets and evaluation are presented in Section 6 and the paper is

- *Q. Ye is with the Department of Electronics, Electrical and Communication Engineering, University of Chinese Academy and Sciences, Beijing, China. E-mail: qxye@ucas.ac.cn.*
- *D. Doermann is with the Institute of Advanced Computer Studies, University of Maryland, College Park, MD 20742. E-mail: doermann@umiacs.umd.edu.*

Manuscript received 12 Jan. 2014; revised 22 Sept. 2014; accepted 22 Oct. 2014. Date of publication 2 Nov. 2014; date of current version 5 June 2015. Recommended for acceptance by E. G. Learned-Miller.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2366765

summarized with discussions about remaining problems and future directions in Section 7.

1.1 Overview of the Problem

Although the recognition of text gives rise to many applications, the fundamental goal is to determine whether or not there is text in a given image, and if there is, to detect, localize, and recognize it. In the literature, various stages of these fundamental tasks are referred to by different names including text localization [14], which aims to determine the image positions of candidate text, text detection, which determines whether or not there is text using localization and verification procedures, and text information extraction [32], [85], which focuses on both localization and binarization. Tasks such as text enhancement are used to rectify distorted text or improve resolution prior to recognition. Other references include scene text recognition [100] and text recognition in the wild [173], which restrict analysis of images to text in natural scenes. Suffice it to say that the primary goals of text detection, localization and recognition are essential for an "end-to-end" system.

1.2 Summary of Progress in the Past Decade

Early text detection and recognition research was a natural extension of document analysis and recognition research, moving from scanned page images to camera captured imagery, focusing on basic preprocessing, detection and OCR technology [17]. Recently, the application of sophisticated computer vision and learning methods has resulted from the realization that the problems do not lend themselves to a sequential series of independent solutions. The trend is to integrate the detection and recognition tasks into an "end-to-end" text recognition system [118].

In the early years, researchers extensively investigated graphic overlay text in video as a way to index video content. Scene text, especially video scene text, has been regarded as presenting a more difficult challenge yet very little work had been done with it [32]. Recently, researchers have explored approaches that prove effective for text captured in various configurations, in particular, incidental text in complex backgrounds. Such approaches typically stem from advanced machine learning and optimization methods, including unsupervised feature learning [123], convolutional neural networks (CNN) [173], [176], deformable part-based models (DPMs) [195], belief propagation [100] and conditional random fields (CRF) [96], [133].

2 Background

To understand the overall value of text detection and recognition approaches, it is useful to provide background information about the underlying problems, applications and technical challenges.

2.1 Text in Imagery

Graphic text and scene text are considered two basic classes of text, where the former refers to machine print text overlaid graphically and the latter refers to text on objects, captured in its native environment. Graphic text is usually machine printed, found in captions, subtitles and annotations in video and born-digital images on the web and in email [129]. Scene text, however, includes text on signs, packages and clothing in natural scenes, and is more likely to include handwritten material [117].

(a) (b) (c)

Fig. 1. Text in imagery. (a) Video graphical text. (b) Point-and-shoot scene text. (c) Incidental scene text.

Most recent research has focused on scene text, and, to portray the challenges more accurately, it helps to further distinguish between images where the primary purpose of the image is to capture text, and images where the text is embedded in the scene. Although a continuum exists between the two, we refer to the former as *point-and-shoot* text and the latter as *incidental* text, as shown in Fig. 1.

2.2 Applications

Over the past two decades, there have been numerous text related applications for both images and video, which can be broadly categorized as multimedia retrieval, visual input and access, and industrial automation.

Multimedia retrieval. Text in web images is relevant to the content of the web pages. Video captions usually annotate information about where, when and who of the happening events [8], [49]. Recognizing text and extracting keywords in such multimedia resources enhances multimedia retrieval.

Visual input and access. The expansion of mobile devices containing digital cameras has made imaging devices widely available. With an embedded module, mobile devices automatically input name cards, whiteboards and slide presentations [10], [40], [41]. Without being forced to input by keyboard, users feel more comfortable and work more efficiently.

Signs in natural scenes carry significant information. Automatic sign recognition and translation systems enable users to overcome language barriers [26]. Carnegie Mellon University developed an early PDA-based sign recognizer [26], and recent platforms include iOS and Android, which can instantly recognize and translate text into another language [202].

According to the World Health Organization,[1] approximately 39 million legally blind and 285 million visually impaired people live in the world. Developing personal text-to-speech devices assists them in understanding grocery signs, product and pharmaceutical labels, and currency and ATM instructions [37], [77]. The University of Maryland [77] and City University of New York [174] have developed text recognition prototypes for people who are

1. http://www.who.int/blindness

TABLE 1
Challenges in Text Detection and Recognition

Category	Sub-category
Environment	Scene complexity Uneven lighting
Image acquisition	Blurring/degradation Perspective distortion
Text content	Variation of aspect ratio Multi-oriented/curved text Variation of fonts Multilingual environments

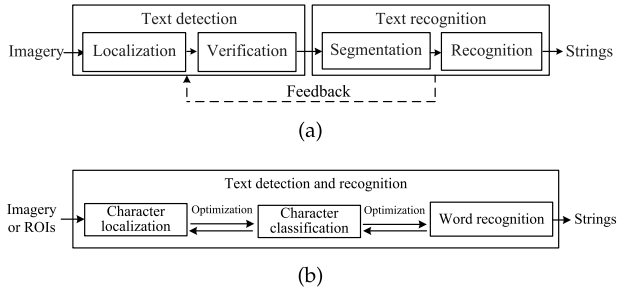


Fig. 2. Frameworks of two commonly used text detection and recognition methodologies. (a) Stepwise methodology. (b) Integrated methodology.

visually impaired. The Kurzweil National Federation of the Blind (KNFB) reader[2] runs on mobile platforms, enabling people who are visually impaired to "read" text from indoor scenes.

Industrial automation. Recognizing text on packages, containers, houses, and maps has broad applications related to industrial automation. For example, recognition of addresses on envelopes is applied in mail sorting systems. Automatic identification of container numbers improves logistics efficiency [39]. Recognition of house numbers and text in maps benefits automatic geocoding systems [168].

2.3 Challenges

The complexity of environments, flexible image acquisition styles and variation of text contents pose various challenges, which are categorized in Table 1 and analyzed as follows.

Scene complexity. In natural environments, numerous man-made objects, such as buildings, symbols and paintings appear, that have similar structures and appearances to text. Text itself is typically laid out to facilitate legibility. The challenge with scene complexity is that the surrounding scene makes it difficult to discriminate text from non-text.

Uneven lighting. When capturing images in the wild, uneven lighting is common due to the illumination and the uneven response of sensory devices. Uneven lighting introduces color distortion and deterioration of visual features, and consequently introduces false detection, segmentation and recognition results.

Blurring and degradation. With flexible working conditions and focus-free cameras, defocusing and blurring of text images occur [40]. Image/video compression and decompression procedures also degrade the quality of text, in particular, graphical video text. The typical influence of defocusing, blurring and degradation is that they reduce characters' sharpness and introduce touching characters, which makes basic tasks such as segmentation difficult [40].

Aspect ratios. Text such as traffic signs, may be brief, while other text, such as video captions, may be much longer. In other words, text has different aspect ratios. To detect text, a search procedure with respect to location, scale and length needs to be considered, which introduces high computational complexity.

Distortion. Perspective distortion occurs when the optical axis of the camera is not perpendicular to the text plane, as shown in Fig. 1b. Text boundaries lose rectangular shapes and characters distort, decreasing the performance of recognition models trained on undistorted samples.

Fonts. Characters of italic and script fonts might overlap each other, making it difficult to perform segmentation [132]. Characters of various fonts have large within-class variations and form many pattern sub-spaces, making it difficult to perform accurate recognition when the character class number is large.

Multilingual environments. Although most of the Latin languages have tens of characters, languages such as Chinese, Japanese and Korean (CJK), have thousands of character classes. Arabic has connected characters, which change shape according to context. Hindi combines alphabetic letters into thousands of shapes that represent syllables [99]. In multilingual environments, OCR in scanned documents remains a research problem [99], while text recognition in complex imagery is more difficult.

3 Methodologies

In this section, we analyze two commonly used methodologies in complete text detection and recognition systems: stepwise and integrated. As shown in Fig. 2a, stepwise methodologies have separated detection and recognition modules, and use a feed-forward pipeline to detect, segment and recognize text regions. Integrated methodologies, by contrast, have a goal of recognizing words where the detection and recognition procedures share information with character classification and/or use joint optimization strategies, as shown in Fig. 2b. Some stepwise approaches utilize a feedback procedure from text recognition to reduce false detections, and some integrated approaches use a preprocessing step to localize regions of interest. The key difference lies in the fact that the latter uses recognition as a key focus.

3.1 Stepwise Methodologies

Stepwise methodologies have four primary steps: localization, verification, segmentation, and recognition. The localization step coarsely classifies components and groups them into candidate text regions, which are further classified into text or non-text regions during verification. The underlying assumption is that various text regions might be regarded as a kind of uniform pattern, therefore, there must exist properties or features that are invariant over this pattern. The segmentation step separates the characters so that exclusive, accurate outlines of image blocks remain for the

2. http://www.knfbreader.com

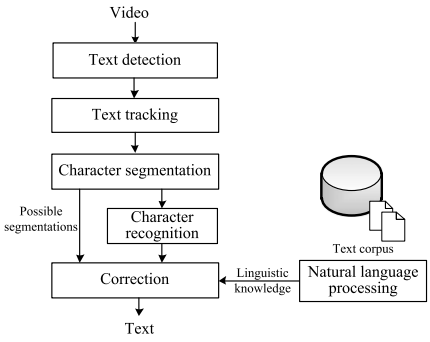


Fig. 3. Flowchart of the stepwise video text recognition approach with detection, tracking, segmentation, recognition and language processing [126].

recognition step. Finally, the recognition step converts image blocks into characters. In some approaches, the verification and/or segmentation step could be ignored, or additional steps might be included to perform text enhancement and/or rectification.

In [126], a stepwise approach including detection, tracking, segmentation, recognition, and correction was proposed, as shown in Fig. 3. Text detection is performed with a convolutional neural network [67] trained on raw pixel values, and the detected components of local maximal responses are grouped as text. A tracking process is integrated to determine the start and end frame of localized text. A segmentation step based on the Shortest Path method is proposed to calculate separations that enable accurate CNN based character recognition. A language model is then used to remove recognition ambiguities and segmentation errors.

Yao et al. [175], [197] developed an orientation robust, multilingual approach. Stroke pixels are grouped into connected components (CCs), which are filtered with a decision forest trained on component features of shape, occupation ratio, axial ratio, width variation, and component density. Filtered connected components are then aggregated into multi-oriented chains with a hierarchical clustering algorithm, and verified by a decision forest classifier trained on region features including color, density, stroke, and structure. The chains that pass verification are enhanced by a low rank structure recovery algorithm, and are then fed to an OCR module to produce recognition results.

3.2 Integrated Methodologies

With an integrated methodology, character classification responses are considered the primary cues, and shared with detection and recognition modules [51].

Using character classification responses as primary features requires the discrimination of characters from the background as well as from each other, which is a complex multi-class problem. Solutions require not only robust character recognition models but also appropriate integration strategies, such as holistic matching, i.e., "word spotting" [148], joint optimization [173] and/or decision delay [102], [188].

Word spotting looks to match specific words in a given lexicon with image patches by character and word models. As shown in Fig. 4, Wang and Belongie proposed a word spotting approach by training character models with histogram of oriented gradient (HOG) features and a nearest

Fig. 4. Illustration of a word spotting approach [148]. Characters are recognized with HOG features and Random Ferns classifiers, and words are modeled with a pictorial structure model (Courtesy of K. Wang).

neighbor classifier [118] (random ferns classifiers in [148]). They use the multi-scale sliding window classification to obtain character responses, and the non-maximum suppression to localize character candidates. They employ the pictorial model that takes the scores and locations of characters as input to determine an optimal configuration of a particular word from a small lexicon.

Wang et al. [173] proposed combining a multi-layer CNN with unsupervised feature learning to train character models, which are used in both text detection and recognition procedures. As shown in Fig. 5, they run CNN based sliding window character classification and use the responses to localize candidate text lines. They then integrate the character responses with character spacings and a defined lexicon using a beam search algorithm [15] to recognize words.

Neumann and Matas [188] proposed a decision delay approach by keeping multiple segmentations of each character until the last stage when the context of each character is known. They detect character segmentations using extremal regions. Based on the segmentations, a directed graph is constructed with character classification scores, character intervals and language priors. A dynamic programming algorithm is used to select the path on the graph with the highest score. The sequence of regions and their labels induced by the optimal path are the outputs, i.e., a word, a sequence of words or a non-text region.

3.3 Comparison of the Methodologies

Stepwise methodologies typically employ a coarse-to-fine strategy, which first localizes text candidates, and then verifies, segments, and recognizes them. One attractive feature is that most of the background is filtered in the coarse localization step, which greatly reduces the computational cost, and consequently guarantees computational efficiency. The other attractive feature is that it processes oriented text as the text orientations are estimated in the localization step. Given language independent features or multilingual OCR modules [12], [45], [80], it processes multilingual text. The disadvantages are twofold. The first is the increase in complexity when integrating different techniques from all steps. The second is the difficulty in optimizing parameters for all steps, which might introduce error accumulation.

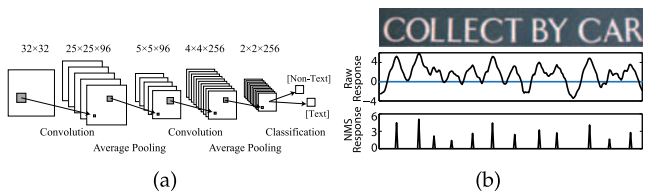


Fig. 5. The CNN based integrated detection and recognition approach [173]. (a) CNN for character detection. (b) CNN responses for recognition. (Courtesy of T. Wang).

By contrast, the goal of integrated methodologies is to identify specific words in imagery with character and language models. Integrated methodologies can avoid the challenging segmentation step or optimize it with character and word recognition, which makes it less sensitive to complex backgrounds and low resolution text. The disadvantage lies in that the multi-class character classification procedure is computationally expensive when considering a large character class number and a large amount of candidate windows. In addition, the increase of word class number could significantly decrease the detection and recognition performance, so the generality is often limited to a small lexicon of words.

4 Fundamental Sub-Problems

In this section, sub-problems including text localization, verification, segmentation, and recognition are described. Each approach is reviewed with respect to its primary contribution. The approaches that make multiple contributions are analyzed with respect to each contribution.

4.1 Text Localization

The objective of text localization is to localize text components precisely as well as to group them into candidate text regions with as little background as possible. For text localization, connected component analysis (CCA) and sliding window classification are two widely used methods, and color, edges, strokes, and texture are typically used as features.

4.1.1 Methods

Connected component analysis. CCA could be regarded as a graph algorithm, where subsets of connected components are uniquely labeled based on heuristics about feature consensus, i.e., color similarity and spatial layout. In implementations of CCA, syntactic pattern recognition methods are often used to analyze the spatial and feature consensus, and to define text regions. Considering the complexity of fine-turning the syntactic rules, a new trend is to perform CCA with statistical models [109], [138], [182], e.g., using an AdaBoost classifier on pairwise spatial features to learn the CCA models [182]. The use of statistical models in CCA significantly improves its adaptivity.

Sliding window classification. In the sliding window classification method, multi-scale image windows that are classified into positives are further grouped into text regions with morphological operations [130], CRF [148] or graph methods [123], [173]. The advantage of this method lies in the simple and adaptive training-detection architecture. Nevertheless, it is often computationally expensive when complex classification methods are used and a large number of windows need to be classified.

4.1.2 Features

For text localization, color [174], edge [28] and texture features [19] were conventionally used, and stroke [47], [107], [163], point [152], region [137], [138], [150], [164], [182] and character appearance features [94], [196], [198], [199] have recently been explored.

Color features. Text is often produced in a consistent and distinguishable color so that it contrasts with the background [40]. Under this assumption, color features could be used to localize text [2], [22], [54], [63], [82], [92], [96], [109], [150]. As a 20-year old method, color-based text localization operates often simply and efficiently, although it is sensitive to multi-color characters and uneven lighting, which can seriously degrade color features.

An early color-based text localization approach is from Jain and Yu [2]. They used color reduction to generate color layers, a clustering algorithm to obtain CCs, and connected CCs into text candidates with color similarity and component layout analysis. In other work [95], it was shown that the use of a mean-shift algorithm to generate color layers could improve the robustness to complex backgrounds.

To be adaptive to color variation, color features are extracted in converted or combined color spaces or described with mixture models [27], [74], [76], [174]. In [7], Garcia and Apostolidis performed text extraction with a k-means clustering algorithm in the hue-saturation-value (HSV) color space. Karatzas and Antonacopoulos [33] extracted text components with a split-and-merge strategy in the hue-lightness-saturation (HLS) color space. Chen et al. [26] proposed using Gaussian mixture models in R, G, B, hue and intensity channels to localize text.

Edge/Gradient features. The family of edge/gradient-based approaches assumes that text exhibits a strong and symmetric gradient against its background. Thus, those pixels with large and symmetric gradient values could be regarded as text components. In [4], [12], [23], [27], [80], [114], [177], [181] edge features are used to detect text components, and in [12], [24], [71], [98] gradient features are used.

Wu et al. [4] proposed using Gaussian derivatives to extract horizontally aligned vertical edges, which are aggregated to produce chips corresponding to text strings if "short paths" exist between edge pairs. In recent work [167], Phan et al. proposed grouping horizontally aligned components of "gradient vector flow" into text candidates based on spatial constraints of sizes, positions and color distances.

Compared with color features, gradient/edge features are less sensitive to uneven lighting and multi-color characters [9]. They are combined with such classifiers as artificial neural networks [14], [16] or Adaboost [28], [68] to perform sliding window based text localization. However, they often have difficulty when discriminating text components with complex backgrounds having a strong gradient.

Texture features. When characters are dense, text could be considered as a texture [29]. Texture features including Fourier Transform [116], Discrete Cosine Transform (DCT) [8], Wavelet [5], [49], LBP, and HOG [113] have been used to localize text. Such features are usually combined with a multi-scale sliding window classification method to perform text localization. Texture features are effective for detecting dense characters, although they might not detect sparse characters, i.e., signs in scene images which lack significant texture properties.

Li et al. pioneered the text localization method with Wavelet texture features [5]. They proposed using mean, second and third order central moments of wavelet coefficients and a neural network to classify image windows, of

which negative and isolated positive windows are filtered and connected positive windows are retained as text. Zhong et al. [8] pioneered text localization in the JPEG/MPEG compressed domain, using DCT features. They detected image patches of high horizontal spatial intensity variation as text components, aggregating such components into regions with morphological operations and verifying the regions by thresholding spectrum energy. Goto and Tanaka [93] proposed using DCT features and Fisher discriminant analysis (FDA) to localize text in scene images. Kim et al. [92] employed the LBP to describe the texture property around background-text transition pixels. Kim et al. [19] proposed using SVMs and texture templates to perform text localization. Pixels that are classified into positives are connected into text regions by a mean shift algorithm.

Stroke width transform (SWT). SWT is a local image operator that computes the width of the most likely stroke containing the pixel [107]. SWT outputs a map, where each element corresponds to the stroke width value of a pixel. Stroke-based features have been shown to be competitive for localizing high resolution scene text [60], [62], [107], [156], in particular, when they are combined with appropriate learning methods [153], [156] or enhanced with other cues such as edge orientation variance (EOV) and opposite edge pairs (OEPs) [155] or combined with spatial-temporal analysis [160]. More recently, Mosleh et al. [163] improved SWT by introducing a Bandlet-based edge detector which enhances text edges as well as dismisses noisy and foliage edges, and consequently applies to low resolution text.

Point and region features. Concerning the observation that dense presences of corner points exist in text regions, Harris corners were employed to perform video text localization [108], [152]. In [152], corner points are aggregated into candidates, which are further discriminated with a Decision Tree classifier utilizing geometry and optical flow features.

MSERs-based text localization has been widely explored [78], [112], [122], [137], [164], [182], [195]. The main advantage of this representation is rooted in the effectiveness of using MSERs as character/component candidates. It has been observed that text components usually have significant color contrast with backgrounds and tend to form homogenous color regions. The MSER algorithm that adaptively detects stable color regions provides a viable solution for localizing text [198], [208]. The approach [208] that uses a pruning algorithm to select appropriate MSERs as character candidates and hybrid features to validate the candidates achieved state-of-the-art performance in the ICDAR'13 competition.

Hybrid features. Text from various categories have different characteristics. Text objects, such as video captions, have dense characters and strong gradients, while others may have sparse characters but color distinguishes them from their surroundings. To improve the robustness on various text categories, hybrid features have been applied in text localization [1], [24], [73], [90], [115], [122], [138], [171], [174]. As an early work, Jain et al. [1] proposed using a combination of color features and gray value variation. In a recent work, Lee et al. [130] proposed using a hybrid of features from gradients, Gabor filter energy, variance of Wavelet coefficients and edge intervals. The features over 16 spatial scales are integrated and fed to AdaBoost for classification.

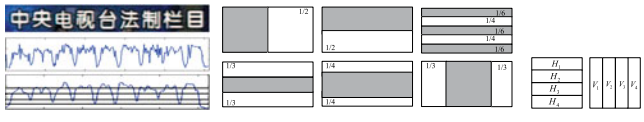


Fig. 6. (a) Global text feature extraction [49]. (b) and (c) Sub-regions for local feature extraction [174], [182].

4.2 Text Verification

The text localization often introduces false positives because a small piece of components/patches may not contain sufficient information for classification. After text localization, holistic features of text regions are available for precise classification and verification.

Knowledge based methods. Prior knowledge about color, size and space consensus, and projection profile[3] have been used to perform text verification. In [2], thresholds on horizontal and vertical projection profiles are used to verify text candidates. In [71], thresholds are used on edge-area/text-area, text-block-width and text-block-height. In [80], [115], [141], thresholding is used on projection profiles, character distances, straightness and edge density. In [92], aspect ratio is used to verify localized text regions. In [68], syntactic rules are used on edge count, horizontal profile, connected component height and width. In [22], syntactic rules about region contrast, structure, alignment, and the recognition results of characters are used. In [73], the proportion of width and length of minimum bounding rectangles (MBRs), the proportion of text and background pixels in MBRs are used.

Knowledge based verification is simple and intuitive. However, it is difficult to translate prior knowledge of text into well-defined syntactic rules. If the rules are strict, they may fail to keep text that doesn't comply with all the rules. If the rules are loose, they may introduce numerous false detections.

Feature discrimination methods. Various features including structure [159], intensity and shape features [74], Wavelet [5], [49], LBP [104] and HOG texture descriptors [171], [186], Gabor strokes [151], [174], and hybrids [166] were used to perform text discrimination. For text discrimination a prerequisite of which is that features extracted from image regions of different aspect ratios are normalized to the same dimensionality. One way to obtain normalized features is to extract global features that are independent of a region aspect ratio, as illustrated in Fig. 6a. The other is to divide image regions into an equal number of sub-regions of different sizes, as shown in Figs. 6b and 6c, and extract local features having the same dimensionality from the sub-regions.

Ye et al. [49] proposed extracting global Wavelet and cross line features to represent text. A forward search algorithm is applied to select features and an SVM classifier is trained to identify true text from the candidates. In [159], global features concerning the height and width ratio, solidity value, stroke width and gradient variation are trained with a kernel SVM to discriminate text. In [27], edge, gradient and texture features are combined and trained with a multilayer perceptron (MLP) for text verification. In [163], features of intensity,

3. A project profile is defined as the vector of the sums of the pixel intensity or gradient or edge count etc.

mean and variance of stroke width and bounding box aspect-ratio are extracted to represent CCs. Such features from the CCs are fed to a *k*-means algorithm for classification.

Yi and Tian [174] proposed dividing text regions into sub-regions, as shown in Fig. 6b, for text discrimination. Pixel of interests are located with maximums of anti-compatible Gabor filters, and sub-region-based statistical features of oriented histogram, gradient and stroke width are then extracted and classified with an SVM. Koo and Kim [182] proposed splitting each component into eight square sub-regions, from which the following features are extracted and classified with a multilayer perceptron classifier: 1) the number of foreground pixels, 2) the number of vertical white-black transitions, and 3) the number of horizontal black-white transitions. They use the average of the classification responses from all sub-regions for text/non-text classification.

4.3 Text Segmentation

Before detected text regions are recognized by an OCR module, certain approaches use binarization, text line segmentation and character segmentation algorithms to obtain the precisely bounded characters. Segmentation has been identified as one of the most challenging problems [66], and recent approaches often integrate the segmentation step with the recognition step, or use word matching to avoid the segmentation problem.

4.3.1 Text Binarization

Text binarization operates to extract text pixels and remove the background pixels. Algorithms related to adaptive thresholding [14], [121], probability models [35], [149] and clustering [63], [73], [147] have been used in this problem.

Adaptive thresholding approaches segment text according to their respective local features, and thus are adaptive to backgrounds [45], [89], [92]. Nevertheless, it is difficult to select a reliable threshold value for degraded text where the pixels at the text boundary often blend with the background. In this case, Gaussian mixture models could be applied given the context that a significant amount of foreground pixels are sampled to build the models [35], [148], [206].

Inspired by the success of CRF models for solving image segmentation problems, Mishra [135] and Kim and Lee [185] formulated the text binarization problem in optimal frameworks and used an energy minimization to label text pixels. In [185], Lee and Kim proposed using a two-stage CRF model to label coherent groups of text regions based on the hierarchical spatial structures of segmented characters.

When extracting degraded text in video, clustering methods are preferred [54], [61], [63], [147]. In [54], [63], Mancas-Thillou and Gosselin leveraged multiple color metrics and clustering to extract text pixels. They also complemented the color metrics with the spatial information obtained from Log-Gabor filters. In [147], Wakahara and Kita leveraged a "clustering and classification" strategy to extract degraded text from the background. They generated binarized images by *k*-means clustering, divided each binarized image into a sequence of "single-character-like" images, calculated the SVM response, and finally selected a single binarized image with the maximum SVM response as binarized text.

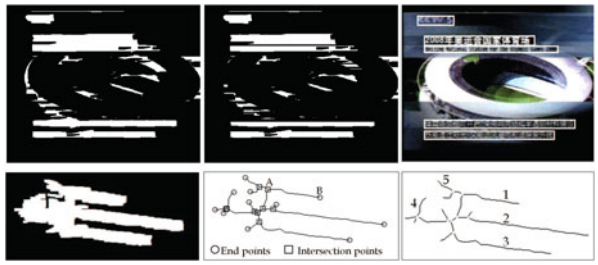


Fig. 7. Text line segmentation with projection profile (first row) [23], and skeleton analysis (second row) (Courtesy of Phan and Tan [141]).

When evaluating various binarization approaches using OCR software and the ICDAR text recognition benchmarks, the approach that uses local binarization to generate seed pixels and a graph-cut algorithm to perform final segmentation achieves state-of-the-art performance [187].

4.3.2 Text Line Segmentation

The function of text line segmentation is to convert a region of multiple text lines into multiple sub-regions of single text lines. For horizontal text, the projection profile analysis of text components, as shown in Fig. 7 (first row), presents a simple but effective method [23], [82], [115]. For skewed or perspective distorted text, however, the projection profile analysis method is useless before estimating the text orientation.

A recent advance of text line segmentation comes with the emergence of the skeleton analysis method [141]. Text skeletons are extracted from connected components, as shown in Fig. 7 (second row), and a text line is defined as a continuous path on the skeleton from an intersection point to either an end point or an intersection point. The "path" corresponding to a text line does not include any other points in the middle. Given these definitions, a text region Is segmented to text lines using a skeleton-cut algorithm.

4.3.3 Character Segmentation

Character segmentation separates a text region into multiple regions of single characters. Vertical projection profile analysis was an early method for character segmentation. However, it is often difficult to determine an optimal projection threshold when degradation or touching characters exist. With a high threshold, true segmentations might be missed, as shown in Fig. 8 (top left), while with a low threshold, many false segmentations might be detected, as shown in Fig. 8 (top right).

Adaptive methods, including the adaptive morphological operation [46], clustering [142] and optimization methods [97], [139], have been steadily developed. Phan et al. [139] investigated the gradient vector flow features and a minimum cost path optimization method for character segmentation, as shown in Fig. 8 (second row). A two-pass path search algorithm is applied where the forward search localizes potential cuts and the backward direction removes the false cuts, i.e., those that pass through the characters.

4.4 Text Recognition

Text recognition converts image regions into strings. In recent research, word recognition has been central to text

Fig. 8. Character segmentation with the projection profile analysis (first row) and the path optimization method (second row) [139] (Courtesy of Phan).

recognition because words are well-formulated with statistical models in terms of low-level features and high-level language priors. This is consistent with psycholinguistic research, where words have been the elementary units when studying human visual cognition [55]. It would seem that recognizing text at higher levels, such as clauses or sentences, has seldom been investigated because they are less tractable than words. It was demonstrated that recognizing degraded text at the character level is difficult due to the lack of language priors.

4.4.1 Character Recognition

To recognize characters of a single font, general features, such as Gabor features, and simple classifiers, such as linear discriminant analysis (LDA), are often used [26]. When multiple fonts or distorted characters present, however, the within-class diversity makes it difficult to model characters of the same class [30], [58], [88], [132], [170]. One solution is to have a specified classifier for each of them [170], [183]. Other solutions include aligning characters using unsupervised [123] or representative learning [207], discriminative feature pooling [205], image rectification algorithms [132], or deformable models [195].

Sheshadri and Divvala [170] applied an exemplar SVM to recognize distorted characters in scene images, which makes individual decisions for each classifier and relies on decision calibration to reach a systemic consensus. Two decision calibrations for SVM scores and affine transformation estimation are used to process different distortions. Their approach achieves state-of-the-art performance in the Chars74k dataset.

Part based implicit models have explored for distorted character recognition [192], [195]. Shi et al. [195] proposed using deformable part based models and sliding window classification to localize and recognize characters in scene images. Characters are divided into parts, each of which moves in a local domain with penalty parameters. Trained on the Chars74k dataset, the DPMs effectively recognize characters with distortion and with a variety of fonts.

In [207], a learned representation named Strokelets was proposed for character recognition. Strokelets captures the structural characteristics of characters at multiple scales, ranging from local primitives, like bar, arc and corner to whole characters. A histogram feature named Bag-of-Strokelets is formed by binning the Strokelets and is trained with Random Forest for recognition. This approach has robustness to distortion and generality to variant languages.

4.4.2 Word Recognition

Concerning degraded and/or distorted text, as shown in Fig. 9, it is not unusual for a recognition model to assign

Fig. 9. Word examples from dataset ICDAR'11 (first row) and IIIT5k (second row) (Table 2).

different labels to identical characters. This is particularly common given distortions or lack of training data for particular fonts [100]. In this case, the character segmentation and character recognition can be integrated with language priors using optimization methods including Bayesian inference [25], [57], [64], [100], Integer programming [145], Markov [36], [83], [119], [206], CRF [161], [195], and graph models [56], [70], [123], [141], [143], [158], [189].

Weinman et al. proposed a probabilistic inference method [100] that integrates similarity, language priors and lexical decision to recognize scene text. The inference procedure is accelerated with sparse belief propagation, which is an optimization method for shortening messages by reducing the dependency between weakly supported hypotheses. Their approach has been shown to be effective for eliminating unrecoverable recognition errors and improving accuracy.

Mishra et al. [161] presented a framework that utilizes both bottom-up (character) and top-down (language) cues for text recognition. They use sliding window classification to obtain local maximum character detections, and a CRF model to jointly model the strength of the detections and the interactions among them. Shi et al. [195] proposed using DPMs to detect and recognize characters, then building a CRF model on the potential character locations to incorporate the classification scores, spatial constraints, and language priors for word recognition (Fig. 10).

Higher order language models (*n*-grams) have been explored to enforce recognition accuracy. In [165], word recognition is performed by estimating the maximum a posteriori (MAP) under the joint distribution of character appearance and the language model. The MAP inference is performed with weighted finite-state transducers (WFSTs). Large dictionaries have also been adopted to enforce a high-order language model [162]. The use of a large dictionary not only facilitates weak character detections but also enforces recognizing non-dictionary words, such as business names and street names [178].

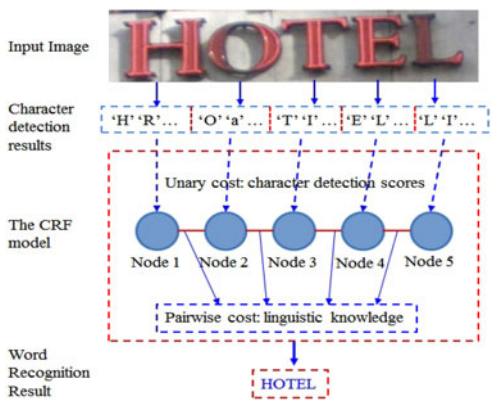


Fig. 10. Word recognition with a CRF model [195]. (Courtesy of C. Shi).

4.4.3 "End-to-End" Recognition

Given imagery with complex backgrounds as input, an end-to-end recognition system embodies the localization, detection and recognition functions to convert all text regions in the imagery into strings. Considering a small lexicon, word spotting offers an effective strategy for realizing end-to-end recognition. The motivation of word spotting is that "the whole is greater than the sum of parts", and the task looks to match specific words in a given lexicon with image patches using character and word models [118], [179]. Considering an open lexicon, however, word spotting strategies are impracticable because of the large search space. In this case, systems require strong character representation [173], [202], large scale language models [202], [204], and sophisticated optimization strategies [189], [206].

In [118], Wang and Belongie proposed a word spotting approach based on an optimal configuration of character response, character layout and lexicon. In [179], Goel et al. proposed a word spotting approach by transforming the lexicon into a collection of synthetic word images, then converting the text recognition task into a problem of retrieving the best match from the lexicon image set with a weighted dynamic time warping (wDTW) approach.

Neumann and Matas [189] introduced an end-to-end approach that integrates character detection and recognition based on oriented stroke features. Strokes are detected by convolving the image gradient field with a set of oriented bar filters. Characters are detected and recognized as image regions, which contain strokes of specific orientations in a relative position. Dynamic Programming is adopted to optimize recognition responses, character spacing and 3-gram language priors, i.e., character triplets.

Weinman et al. [206] proposed an end-to-end approach that uses combined approaches for text detection, uses probabilistic methods for text binarization, and jointly optimizes character segmentation and word recognition with a semi-Markov model.

A recent large lexicon end-to-end text recognition system, Google PhotoOCR [202], takes advantage of substantial progress in deep learning, large scale language modeling and careful engineering. Two combined approaches are used to localize text, and a beam search algorithm is adopted to optimize segmentation, localization and language priors. A deep neural network character classifier is trained on as many as two million examples, and a language model is learned utilizing a corpus of more than a trillion tokens. The system was tested with 29 languages with Latin script, Greek, Hebrew and four Cyrillic languages, and it demonstrated state-of-the-art performance.

5 Special Issues

Although the reviewed approaches achieved promising results compared to high-resolution point-and-shoot text, incidental text in uncontrolled environments remains very challenging. The following sections analyze and review issues relevant to this kind of text: text enhancement, multi-orientation, perspective distortion, and multilingual content. Special issues relevant to video text detection and recognition are also analyzed.

5.1 Text Enhancement

Text enhancement uses image processing, learning [11], [105] or reconstruction methods [134], [193], [197] to improve text resolution or recover degraded text.

Deconvolution is conventionally used for general image deblurring. Nevertheless, research [120], [131], [157] has shown that deconvolution has difficulties when processing text images, as they do not respect the special properties of text regions. In [157], Cho et al. improved the deconvolution method by introducing an auxiliary image to enforce domain-specific properties, i.e., strokes of similar widths and uniform colors, using an iterative optimization algorithm.

Learning-based methods have demonstrated positive results concerning text enhancement [11]. Nevertheless, if image degradation, e.g., a compression artifact, is not included in the training data, it is regarded as image data and enhanced. Caner and Haritaoglu [105] proposed an approach that is insensitive to training data. They modeled a predictive relationship between the degraded images and high-resolution training counterparts using encoding of a multi-resolution histogram, and they employed this relationship in a probabilistic scheme to generate high resolution text.

In [193], Shivakumara et al. proposed to use ring radius, i.e., a kind of stroke property, to reconstruct broken characters in video. They used normalized gradient features together with the Canny edge map to extract character contours. A ring radius transform (RRT) was then applied to identify medial pixels, with which they utilized the symmetry property between the inner and outer contours of a broken character to reconstruct the character.

Sparse reconstruction based text enhancement has also been investigated [134], [197]. The underlying assumption of such methods is that over-complete basis-based sparse coefficients that encode a given image are equivalent to modified basis-based coefficients that encode a blurred image. Following the "analysis-by-synthesis" strategy, an explicit model computes sparse reconstruction coefficients of the blurred/degraded image. The coefficients are then used to combine elements of the basis to yield an enhanced image.

5.2 Text in Video

Considering text in video, the multi-frame integration strategy is commonly used for improving text resolution, depressing video backgrounds or enforcing text recognition results [6], [14], [20], [48], [60], [69], [93], [103], [110], [190]. Multi-frame up-sampling and frame selection (finding a frame in which text is most clearly displayed) were also used to obtain a single higher resolution text frame from multiple frames [48], [60], [103].

Despite the effectiveness of multi-frame integration for static video captions [14], [69], [103], there are challenges when addressing moving video captions and video scene text. For moving text, solutions requires well designed tracking algorithms to guarantee precise text region registration at pixel level or sub-pixel level [3], [6]. False tracking could blend text with its background and further degrade text.

Spatial-temporal analysis has also been developed for video text detection and/or recognition. In [152], Zhao et al. proposed to detect moving video captions based upon well designed spatial-temporal features. The motion features,

extracted by optical flow, combines with spatial domain text features, i.e., Harris corners and region properties, to detect the moving video captions. A decision tree is adopted to learn the classification criteria in multiple video frames. In [160], stroke based spatial localization is used to provides a rough estimate of caption regions, as block-aligned. After spatial localization, spatial-temporal features are integrated to provide a more accurate estimation of video captions. Temporal information is also used in text segmentation and post-processing.

5.3 Multi-Orientation

Text orientation needs to be estimated in the detection procedure, so that skewed text is corrected to a horizontal orientation before the recognition procedure. The existing approaches that process multi-oriented text usually use bottom-up or top-down methods. Bottom-up methods include agglomerative clustering, dominant orientation analysis [18], [33], [140], region growing [169], boundary growing [172], and Hough Transform [81], [150]. Top-down methods include skeleton segmentation [141] and spanning tree partition [138].

In [175], an agglomerative clustering method was applied to aggregate text components into component chains. By aggregating components in arbitrary orientations, the approach can detect multi-oriented text. In [150], Yi and Tian proposed using a Hough Transform to fit text lines to the centroids of components. Text orientations are reflected with Hough Transform parameters. In [203], Kang and Doermann proposed using projection profile and correlation clustering of MSER components to localize multi-oriented text. In [141], Shivakumara et al. proposed a skeleton analysis method for segmenting a complex CC into constituent parts, by which oriented text lines are segmented and localized. Pan et al. [138], proposed using a minimum spanning tree (MST) algorithm to construct graphs corresponding to multi-oriented text. They then used edge cutting to partition the MST trees into sub-trees, each of which corresponds to a text line.

Curved text represents a special case of multi-oriented text that is difficult to process. Shivakumara et al. [194] proposed using Quad Tree and region growing to detect curved text in video. Chiang and Knoblock [106] proposed using raster maps to localize curved text and estimate character orientation. In later work [124], they proposed a curvature estimation algorithm to group characters from curved text into strings.

5.4 Perspective Distortion

Perspective distortion occurs when the optical axis of the camera is not perpendicular to the text plane, as shown in Fig. 11a. Characters in perspectively distorted text lose their common shapes, and therefore introduce challenges to recognition models. The affine transform [26], homography [34], [65], [125], borderline analysis [38], [75] and curve surface projection [146], have been used to correct distorted text. However, these approaches often require such assumptions as the existence of a rectangle boundary of text or the availability of camera parameters [13], [26].

Fig. 11. Perspective correction of text with a homography operation [65]. (a) Text of perspective distortion. (b) Corrected text.

Chen et al. [26] proposed using an affine transform to rectify distorted text regions. Given the calibrated camera parameters, they detect vanishing points of upper and lower bounders of a text line, and calculate affine parameters of the text line. Ye et al. [65] proposed using the correspondence of feature points and a plane-to-plane homography operation to rectify perspectively distorted text, as shown in Fig. 11. Cambra and Murillo [125] improved the approach of Ye et al. by integrating a feature point detection algorithm and a constrained searching procedure to optimize the homograph parameters.

To process text of complex distortions, such as warp [146], Liang et al. presented a rectification framework that extracts the 3D document shape from a single 2D image [72]. However, two basic assumptions are required: 1) the image contains sufficient characters, and 2) the image is either flat or smoothly curved. For scene text of few characters, however, such assumptions usually do not hold [52], [53], [79], [86], [101].

Zhang and Sun [201] proposed a multi-distortion de-warping (MDD) model based on transform invariant low-rank textures. The idea behind their approach holds that if an image has repetitive patterns, then mathematically it is a low-rank matrix in its front view. One advantage of using the matrix rank as a constraint lies in that it is a reliable indicator for both short strings and multiple text lines as long as the arrangement of text strings has regular patterns [201]. The other advantage is that it requires few assumptions, processing images with both multiple text lines and short text strings.

In addition to various rectification strategies, feature invariants are also employed to recognize distorted text. In [191], Phan et al. adopted densely extracted SIFT descriptors to recognize distorted text. In [86], Zhou et al. proposed the cross ratio spectrum and dynamic time warping to recognize distorted characters. In [101], a clustering based approach was proposed to index the cross ratio spectrum. The cross ratio spectrum of all character templates is clustered, and a query image is compared with the cluster centroids for recognition.

5.5 Multilingual Content

Various language specific approaches have been proposed to detect and recognize text, including English [170], Farsi/Arabic [111], Chinese [26], Japanese [127], Kanji [84], Korean [114], Urdu [128], and Devanagari and Bangla [87]. Language independent approaches [45], [76], [154], [180] have been also considered.

With respect to text detection, Lyu [45] et al. showed that features from contrast, color consistency and orientation are

language independent, while stroke density, aspect ratio and stroke statistics are language dependent. Liu et al. [76] developed language independent features including distances between centroids of characters, areas of characters and the ratio of foreground to background pixels. Zhou et al. [154] demonstrated that general texture features, i.e., HOG, LBP and means of gradients, remain independent of language given training samples from each language.

Existing approaches exclusively validate multilingual detection capability, however, few of them involve multilingual recognition capability. The pattern differences in detection arise from the font sizes and stroke distributions and are not as critical as the pattern diversity in recognition [45], which is related to character structures, shapes and class number. Considering English and Chinese, English has 62 alphanumeric characters of single components while Chinese has more than 2,500 characters [59], most of which have multiple components and complex structures. Google researchers showed that the classifier used by Tesseract OCR module could be adapted to languages, including simplified Chinese, a mixture of European languages, and Russian [99]. In recent end-to-end PhotoOCR system [202], they extended the multilingual capability to 29 languages by integrating multiple detection approaches and employing deep learning based recognition.

6 Evaluation

With so many approaches and datasets, reproducing all of them and comparing them with each dataset are problematic. Therefore, we survey published results to approximate the performance of representative approaches. Such an evaluation cannot precisely characterize how well these approaches will compare in the field. A few factors complicate the evaluation. First, protocols are frequently inconsistent for different error tolerances. Second, the experimental results are based on different training sets. Using a large and well-formatted training set might enhance the performance. Last, domain knowledge related to the text orientation, colors, sizes, and language priori is used in some approaches, which improves the performance. Nevertheless, it also reduces the generalization capability of these approaches.

6.1 Datasets

In Table 2, we collected commonly used datasets and summarized their features including the text categories, sources, tasks, orientations, languages, and information of training/test samples. Selected sample images are shown in Fig. 12.

The MSRA-I, Tan, ICDAR'11 and ICDAR'13 datasets include graphic text in video, web images and email.

The ICDAR'03/05[4] and ICDAR'11/13[5] datasets are prepared for scene text, covering tasks of text localization, character segmentation and word recognition.

The Chars74k dataset works for character recognition in natural scene images.

4. The ICDAR'03 and ICDAR'05 "robust reading" competitions use the same dataset.

5. ICDAR'13 dataset is a subset of ICDAR'11 dataset, with duplicated images removed, as well as ground truth errors corrected.

The VIDI data set was created by Weinman et al. from images of text on signs from around a city [100]. It consists of 95 grayscale sign images with ground truth labels and ground truth character bounding boxes. There are a total of 215 words in the test set, and a training set of synthetic character images from different fonts. The IIIT5K Word dataset provides cropped words (localized text) and is used independently to evaluate character segmentation and/or recognition approaches.

The OSTD, Tan and MSRA-II datasets that include multi-oriented text, and the NEOCR and KIST datasets that include incidental text, combine challenges from cluttered backgrounds and perspective distortions. The MSRA-II, Tan and Pan datasets that contain English and Chinese, the KAIST dataset that contains English and Korean, and the NEOCR dataset that contains eight European languages provide multilingual evaluation environments.

ICDAR'13 dataset includes 28 video sequences prepared to evaluate video scene text detection, tracking and recognition. The video sequences were collected in different countries including Spain, France and the United Kingdom, and were captured using a variety of hardware including mobile phones, hand-held cameras and head-mounted cameras. Recognizing text in such video sequences corresponds to certain tasks like searching for a shop in the street or finding their way inside a building.

6.2 Evaluation Protocols

In this section, we summarize protocols for text detection and recognition evaluation. For text detection the ICDAR protocols are most commonly adopted, and for text recognition the word recognition accuracy is typically used.

Overlap ratio detection protocol. Text detection precision is defined as the ratio between the area of intersection regions and that of detected text regions. Recall is defined as the ratio between the area of intersection regions and that of ground truth regions. Hua et al. improved the protocol by assigning a difficulty level to each ground truth element [31]. Yao et al. [175] further improved the protocol for oriented text. If the included angle between an estimated rectangle and a ground truth rectangle is less than $\pi/8$ and their overlap ratio exceeds 0.5, the estimated rectangle is a correct detection, as shown in Fig. 13a.

The overlap ratio protocol is related to the areas of text. However, it cannot give intuitive evaluation, i.e., the number of correct detections and the number of false alarms [50]. A more flexible protocol was proposed in the ICDAR'03/05 competitions [21], [43].

ICDAR'03 detection protocol. The match m_a between two text bounding rectangles r and r' is defined as twice the area of intersection divided by the sum of the areas, as

$$m_a(r, r') = \frac{2a(r \cap r')}{a(r) + a(r')}, \tag{1}$$

where $a(r)$ is the area of rectangle r. The best match for a rectangle r in a set of rectangles R is defined as:

$$m(r, R) = max\{m_a(r, r') | r' \in R\}. \tag{2}$$

TABLE 2
Datasets

Dataset (Year)	Categories	Source	Task	Image number (training/test)	Text number (training/test)	Shooting manner	Orientation/ distortion	Language	Location
ICDAR'03 (2003) [21]	Scene text	Camera	D/R	509 (258/251)	2,276 (1,110/1,156)	P	Horizontal	English	http://algoval.essex.ac.uk/icdar/Datasets.html
MSRA-I (2004) [31]	Graphic & scene text	Video frames	D	45	158	–	Horizontal	English, Spanish, Chinese	http://www.cs.cityu.edu.hk/~liuwy/PE_VTDetect/
Char74k (2009) [88]	Character	Camera	R	74,107	74,107	–	Horizontal	English, Kanada	http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/
VIDI (2009) [100]	Cropped scene text	Camera	R	95	215 (synthetic training images)	–	Horizontal	English, Kanada	http://www.cs.grinnell.edu/~weinman
KIST (2010) [109]	Scene text	Camera, mobile phone	D	3,000	>5,000	P, I	Distortion	English, Korean	http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database
SVT (2010) [118]	Scene text	Video frames	D/R	350 (100/250)	904 (257/647)	I	Horizontal	English	http://vision.ucsd.edu/~kai/grocr/
Tan (2011) [172]	Graphic & scene text	Video frames	D	520	220	–	Multi-oriented	English, Chinese	–
Pan (2010) [113]	Scene text	Camera	D/R	487 (248/239)	–	P	Multi-oriented	English, Chinese	http://liama.ia.ac.cn/wiki/projects:pal:home
NEOCR (2011) [136]	Scene text	Camera	D/R	659	5,238	I	Multi-oriented, distortion	Eight languages	http://www.iapr-tc11.org/mediawiki/index.php
OSTD (2011) [150]	Scene text	Camera	D	89	218	P	Multi-oriented	English	http://media-lab.engr.ccny.cuny.edu/cyi/project_scenetextdetection.html
ICDAR'11 (2011) [129]	Scene text	Camera	D	484 (229/255)	2,037 (848/1,189)	P	Horizontal	English	http://robustreading.opendfki.de/
	Graphic text	Camera	S/R	522 (420/102)	4,501 (3,583/918)	P	Distortion	English	
IIIT5K Word (2012) [161]	Graphic & scene text	Web & camera	R	5,000 cropped images	5,000 (2,000/3,000)	–	Distortion	English	http://cvit.iiit.ac.in/projects/SceneTextUnderstanding
MSRA-II (2012) [175]	Scene text	Camera	D	500 (300/200)	1,719 (1,068/651)	P	Multi-oriented	English, Chinese	http://pages.ucsd.edu/~ztu/Download_front.htm
ICDAR'13 (2013) [184]	Scene text	Camera	D/S/R	462 (229/233)	848/1,095	P	Horizontal	English	http://dag.cvc.uab.es/icdar2013competition
	Graphic text	Web	D/S/R	551 (410/141)	4,501 (3,564/1,439)	-	Multi-oriented	English	
	Video scene text	Camera	D/S/R	28 videos (13/15)	–	I	Multi-oriented	Spanish, French, English	

(In the Task column, 'D', 'S' and 'R' respectively denote 'Detection', 'Segmentation' and 'Recognition'. In the Shooting manner column, 'P' and 'I' respectively denote 'point-and-shoot' and 'incidental' text).

Precision and recall are defined as:

$$Precision = \frac{\sum_{r_e \in E} m(r_e, T)}{|E|}, \tag{3}$$

$$Recall = \frac{\sum_{r_t \in T} m(r_t, E)}{|T|}, \tag{4}$$

where T and E are the sets of ground-truth and estimated rectangles, respectively. r_e and r_t denote a detected rectangle and a ground-truth rectangle, respectively. The harmonic measure f is adopted to combine the precision and recall figures:

$$f = \frac{1.0}{\frac{\alpha}{Precision} + \frac{1.0-\alpha}{Recall}}, \tag{5}$$

where the parameter α is usually set as 0.5 to give equal importance to precision and recall.

ICDAR'11 (DetEval) detection protocol. In [50], Wolf and Jolion proposed the DetEval protocol that comprises the area overlap and the object level evaluation. As shown in Fig. 13b, it supports one-to-one and one-to-many matches among the ground truth and detections, and considers over-split or over-merge of detections. The protocol was adopted in ICDAR'11 and ICDAR'13 "Robust Reading" competitions.

ICDAR'13 video text protocol. The ICDAR'13 "Robust Reading" competition [184] includes protocols for video text detection and tracking: the multiple object tracking

Fig. 12. Sample images of text from the ICDAR'11, MSRA-I, MSRA-II, SVT, and NECOR datasets.

precision (MOTP) and the multiple object tracking accuracy (MOTA). The average tracking accuracy (ATA) metric provides a spatial-temporal measure that penalizes fragmentation while accounting for the number of correctly detected and tracked words, false negatives, and false positives.

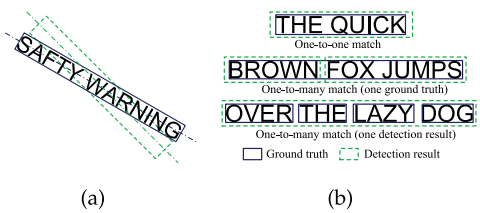


Fig. 13. Two protocols for text detection. (a) Overlap area protocol [175]. (b) DetEval protocol [50].

Word recognition accuracy. Given cropped word images, word recognition accuracy is defined as:

$$WRA = \frac{|C|}{|T|}, \tag{6}$$

where C and T are the correctly recognized word number and the ground truth number, respectively.

"End-to-end" recognition protocol. A strict notion of end-to-end recognition was defined in the ICDAR'03 Robust Reading competition: the detected rectangles must have an area match score m_a greater than 0.5, and the recognized word must match exactly [21]. With this definition, the standard

TABLE 3
Text Detection Performance

Literature	Year	Brief Description	Dataset/ Protocol	Precision	Recall	f	Highlight
Ashida [21]	2003	L:Clustering and rectangle grouping, CCA V:SVM on hybrid features	ICDAR'03	0.550	0.460	0.500	ICDAR'03 competition winner
Becker [43]	2005	L:Binarization analysis, CCA V:Geometry constrains	ICDAR'03	0.620	0.670	0.620	ICDAR'05 competition winner
Epshtein [107]	2010	L:Stroke Width Transform, Hierarchical clustering V:Geometry constrains	ICDAR'03	0.730	0.600	0.660	Introducing Stroke Width Transform
Pan [138]	2011	L:HOG classification with WaldBoost, CCA and Spanning Tree clustering V:Classification of hybrid features	ICDAR'03	0.674	0.697	0.685	Detecting multi-oriented text
Tan [141]	2011	L:Fourier-Laplacian filtering, skeleton analysis V:Geometry constrains	MSRA-I video text	0.810	0.930	0.870	Detecting multi-oriented text
			ICDAR'03/ overlap ratio	0.760	0.860	0.810	
			Tan (video text)	0.740	0.810	0.770	
Koo [182]	2011	L:MSERs, Agglomerative clustering controlled by an Adaboost classifier V:Perceptron learning on mesh features	ICDAR'11 scene text	0.830	0.625	0.713	ICDAR'11 competition winner
Mosleh [163]	2012	L:Bandlet based Stroke Width Transform, CCA V:Hybrid features' clustering	ICDAR'03	0.760	0.660	0.710	Bandlet Stroke Width Transform
Yao [175]	2012	L:Stroke Width Transform, CCA V:Hybrid features and Random Forest	ICDAR'03/ different f measure	0.688	0.660	0.660	Detecting multilingual, multi-oriented text
			MSRA-II	0.630	0.630	0.600	
Ye [198]	2013	L:MSERs, component hypothesis extension V:Integrated discrimination of structural and shape features	ICDAR'11 scene text	0.892	0.623	0.733	Integrated localization and verification
Yin [208]	2013	L:MSERs pruning, single-link clustering algorithm with learned distance parameters V:Bayes classifier, hybrid character features	ICDAR'13 scene text	0.885	0.665	0.759	ICDAR'13 competition winner
			ICDAR'13 graphic text	0.938	0.824	0.877	

('L' and 'V' respectively denote 'Text Localization' and 'Text Verification').

measures of recognition precision, recognition recall and f (Eq. (5)) are used to evaluate end-to-end recognition.

6.3 Performance

Table 3 reports text detection performance of ten approaches including the ICDAR competition winners.

The stroke width transform is regarded as an important language independent approach, and the 'stroke' is regarded as one of the most effective features to detect text, in particular to detect high resolution text. Epshtein et al. [107] developed the classic SWT approach, and the Bandlet based SWT [163] approach reports better performance on the ICDAR'03 scene text detection benchmark.

With the ICDAR'11 scene text dataset, the MSER based detection approach with learned CCA models [182], [198] achieves a solid performance. With the ICDAR'13 dataset, a state-of-the-art performance is achieved by introducing a MSER pruning strategy and hybrid feature based verification [198], [208].

The approach that uses Fourier-Laplacian filtering, component skeleton analysis and geometry feature based verification [141] reports the best performance with the Tan dataset, which contains multilingual and multi-orientated video text. The approach based on SWT, hierarchical clustering and Random Forest classification of hybrid features [175] shows state-of-the-art performance with the MSRA-II dataset, which includes multilingual and multi-oriented scene text.

In the ICDAR'13 video scene text detection and tracking tasks, the TextSpotter by Neumann et al. [164], [188], [189] outperforms the baseline algorithm that uses OCR in the

TABLE 4
Cropped Word (Localized Text) Recognition Performance

Literature	Year	Brief Description	Dataset/ Protocol	Lexicon	WRA	Highlight
TH-OCR [129]	-	Recognition using commercial OCR software	ICDAR'11	–	0.412	ICDAR'11 competition winner
Wang [148]	2011	HOG and Random Ferns based character model, pictorial model optimization with a small lexicon	ICDAR'03 SVT	50/1,156 50	0.760/0.620 0.570	Word spotting
Wang [173]	2012	CNN based character modeling, Beam search based optimization with a lexicon	ICDAR'03 SVT	50/1,156 50	0.900/0.840 0.700	Using CNN
Mishra [161]	2012	Recognition by integrating language prior and appearance features using CRF	ICDAR'03 SVT	50 50	0.818 0.732	Integrating bottom-up and top-down features
Novikova [165]	2012	Large-lexicon driven recognition, weighted finite- state transducers based inference	ICDAR'03 ICDAR'11 SVT	1156/90k 90k 50	0.828/0.785 0.667 0.729	Large lexcion
Goel [179]	2013	Holistic recognition by gradient based features and dynamic matching	ICDAR'03 SVT	50 50	0.897 0.773	Word spotting
Shi [195]	2013	Deformable character models, pairwise language priors, CRF based optimization	ICDAR'03 ICDAR'11 SVT	50/1,156 50/1,189 50	0.874/0.793 0.870/0.829 0.735	Deformable character models
PhotoOCR [202]	2013	Deep Neural Network character models, Beam search based optimization with a large lexicon	ICDAR'13 SVT	100k 50	0.823 0.904	ICDAR'13 competition winner

MOTA and ATA scores, and achieves a slightly better MOTP. The means for MOTP, MOTA and ATA metrics are 0.67, 0.27 and 0.12, respectively [184].

With respect to cropped word recognition, the performance is improved by integrating appearance models with language priors, as shown in Table 4. Recent approaches further improved the performance by using top-down and bottom-up cues [161], and high order language priors [165]. The deformable model based approach [195] shows high performance given a small lexicon, and the PhotoOCR approach using deep learning and a very large lexicon [202] reports the state-of-the-art performance.

With respect to end-to-end recognition, as shown in Table 5, the performance of all stepwise and integrated approaches is very low. With a small lexicon, the CNN based approach [173] reports 67 percent recognition accuracy with the ICDAR'03 dataset. With open vocabulary (or a very large lexicon), however, the highest end-to-end recognition accuracy remains lower than 50 percent, indicating that it remains an open problem.

7 Summary

This paper has described problems related to automatic text detection and recognition in imagery. As the first comprehensive survey in the past five years, it has analyzed recent approaches, classified them according to as many as criteria, and illustrated performance for the most representative approaches. In the past decade, research in this field has progressed as improved methods emerge. However, the low end-to-end recognition performance shows that ample room remains for future research, which raises the following discussions.

7.1 What Is Text?

For the problem "What is text?", the answer could be "structured edges", "a series of uniform color regions", "a group of strokes" or "a kind of texture". However, there are many objects in natural scenes, such as leaves, fences or windows, that have similar edges, strokes or texture properties with text, making it difficult to design effective feature representation to discriminate text. A better assumption could be that "Text is a hybrid of edges, CCs, strokes and texture". Based on this assumption, several hybrid approaches were proposed for text detection.

"Text is a character composite" seems to be a more precise answer. Characters are well-defined patterns, and many effective methods have been developed to recognize characters [59]. Thus, integrated approaches that share character classification results with both the detection and recognition problems have been investigated [118], [173], [188].

7.2 How Does Text Differ?

Text appearance varies significantly for different character combinations. Fig. 14 compares the average text, average face and average pedestrian. The average face and the average pedestrian keep the basic shapes, while the average text looks like noise. This is because both the component number and the component appearance of text are variable, which implies that text detection is not a simple two-class problem. So, a mechanical transplanting of popular object detection methods to the text detection problem is unlikely to provide favorable results.

The text objects admit a great deal of font, color, aspect ratio, clutter background, and distortion. Text recognition confronts challenges beyond those in general object recognition. Humans, however, have little difficulty reading text.

TABLE 5
"End-to-End" Text Recognition Performance

Literature	Year	Methodology	Brief Description	Dataset/ Protocol	Lexicon	WRA/f	Highlight
Tesseract [204]	-	Stepwise	MSER based detection,and Tesseract OCR software based recognition	ICDAR'11	–	0.314	–
Wang [148]	2011	Integrated	HOG and Random Ferns based character model, pictorial model optimization with a small lexicon	ICDAR'03 SVT	1,156 50	0.510 0.380	Word spotting
Wang [173]	2012	Integrated	CNN based character modeling, Beam Search based optimization with a lexicon	ICDAR'03 SVT	1,156 50	0.670 0.460	Word spotting using CNN
Neumann [164]	2012	Stepwise	MSER based character localization, structural feature based character filtering, exhaustive search for grouping, and OCR based recognition	ICDAR'11	-	0.372	Open vocabulary recognition
Neumann [189]	2013	Integrated	Oriented stroke based detection and recognition, dynamic optimization of character modeling, spacing, and the 3-gram language model	ICDAR'11	30K	0.452	Integrated approach using stroke features
Weinman [206]	2014	Stepwise	CCA based detection, Gaussian mixture model based segmentation, Expectation-Maximization fitting for correction, integrated character and word recognition with a Semi-Markov model	ICDAR'11	$244K$	0.386	Open vocabulary recognition
Feild [204]	2014	Stepwise	MSER based detection, probabilistic syllable character model, web-based lexicon, probabilistic context-free grammar for word recognition	ICDAR'11	$13.5M$	0.475	Open vocabulary recognition

The gap between human and machine techniques could be that the former can seamlessly integrate multi-level information from strokes, characters, words, sentences, and language context. Evidence from the cognitive sciences has shown that a hierarchical recursive architecture is used by the human brain when recognizing text [55]. By contrast, most current automated approaches use information from one or two levels and few of them use a recursive procedure.

7.3 Remaining Problems

The gap between the technical status and the required performance indicates that text detection and recognition remain unsolved problems. While great progress has been made there are still numerous research opportunities. Here we summarize some of the more prevalent problems and provide possible research directions.

End-to-end recognition. Compared with the performance of OCR on clean documents, end-to-end text recognition performance is still far behind. Improvement will come not only from stronger character recognition models, but also from well-designed information sharing, feedback and optimization strategies. Recently developed large scale deep learning has substantially improved character classification performance by learning hierarchical multi-scale representations [205], [207]. The integration of deep learning with optimized segmentation, recognition and high order language models could further boost the performance.

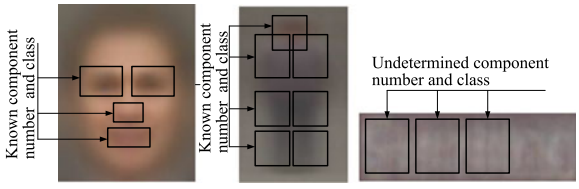


Fig. 14. Comparison of average face, average pedestrian and average text. Each average image is the mean of randomly selected 2,000 aligned samples.

Open vocabulary recognition. Existing word spotting approaches with small lexicons report encouraging performance with the ICDAR benchmarks. However, the utility of these approaches is limited because general lexicons are unlikely to contain the proper nouns and other words that appear in scene or graphical text imagery. To overcome the limitation, it is useful to incorporate an open vocabulary, i.e., large scale web-based language information [204], [206]. It is also useful to develop approximation methods to efficiently use large scale language information for recognition.

Processing incidental text. Incidental text suffers from image degradation, distortions, font variations and cluttered backgrounds. Many approaches can tackle single issues, yet few approaches process a combination of them. To address the general problem of incidental text detection and recognition, improved invariant features must be designed or learned, state-of-the-art enhancement and rectification methods must be integrated, and new sensors must be applied [200].

Processing multilingual text. Text from various languages shows different characteristics. Recognition of text from East Asia countries, e.g., China, Japan and Korea (CJK), was considered an extremely difficult problem due to the large number of character classes, complicated character structures, the similarity among characters, and the variability of fonts [59]. Using a single method with fixed parameters to recognize text from all the languages remains difficult to achieve. One possible solution is to use a common trainable method to specify a model for each kind of language [99] and a configurable method [91] to manage the models.

Real-time detection and recognition. Video from mobile devices has presented an important source for text detection and recognition applications. Often it is desirable to port approaches to the mobile devices and process video data in real time. Nevertheless, many approaches have only been applied to captured images with offline processing modes, and the real-time efficiency requirements from mobile device applications are often ignored. Applying text detection on a frame-by-frame basis makes little sense for video sequences as it ignores any temporal cues [184]. Combining text detection and text recognition with text tracking algorithms will not only improve detection and recognition accuracy but also enhance real-time performance.

ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to Prof. Rama Chellepa, Dr. Jie Chen, the associate editor and the reviewers for their comments and suggestions. They would also like to thank Tao Wang of Stanford, Kai Wang of UCSD, Chongzhao Shi of Chinese Academy of Sciences, and Chew Lim Tan of the National University of Singapore for providing images. The partial support of this research by the US Government through NSF Awards IIS-0812111 and IIS -1262122, the National Natural Science Foundation of China (NSFC) Award 61271433 is gratefully acknowledged.

REFERENCES

[1] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," *Pattern Recognit.*, vol. 28, pp. 1523–1535, 1995.
[2] A. K. Jain and B. Yu, "Automatic text location in images and video frames," *Pattern Recognit.*, vol. 31, no. 12, pp. 2055–2076, 1998.
[3] H. Li and D. Doermann, "Text enhancement in digital video using multiple frame integration," in *Proc. ACM Multimedia Conf.*, 1999, pp. 19–22.
[4] V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: An automatic system to detect and recognize text in images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1224–1229, Nov. 1999.
[5] H. Li and D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 147–156, Jan. 2000.
[6] H. Li and D. Doermann, "Super resolution-based enhancement of text in digital video," in *Proc. IEEE Int. Conf. Pattern Recognit.*, pp. 847–850, vol. 1, 2000.
[7] C. Garcia and X. Apostolidis, "Text detection and segmentation in complex color images," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2000, pp. 2326–2330.
[8] Y. Zhong, H. J. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 385–392, Apr. 2000.
[9] J. Gao and J. Yang, "An adaptive algorithm for text detection from natural scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 84–89.
[10] I. Haritaoglu, "Scene text extraction and translation for handheld devices," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. 408–413.
[11] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.
[12] M. Cai, J. Song and M. R. Lyu, "A new approach for video text detection," in *Proc. IEEE Int. Conf. Image Process.*, 2002, pp. 117–120.
[13] P. Clark and M. Mirmehdi, "Recognizing text in real scenes," *Int. J. Doc. Anal. Recognit.*, vol. 4, pp. 243–257, 2002.
[14] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 256–268, Apr. 2002.
[15] C. L. Liu, M. Koga, and H. Fujisawa, "Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1425–1437, Nov. 2002.
[16] X. Tang, X. Gao, J. Liu, H. Zhang, "A spatial-temporal approach for video caption detection and recognition," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 961–971, Jul. 2002.
[17] A. Vinciarelli, "A survey on off-line word recognition," *Pattern Recognit.*, vol. 35, no. 7, pp. 1433–1446, 2002.
[18] D. Crandall, S. Antani, and R. Kasturi, "Extraction of special effects caption text events from digital video," *Int. J. Document Anal. Recognit.* vol. 5, pp. 138–157, 2003.
[19] K. I. Kim, K. Jung, and H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1631–1639, Dec. 2003.
[20] C. W. Lee, K. Jung and H. J. Kim, "Automatic text detection and removal in video sequences," *Pattern Recognit. Lett.*, vol. 24, pp. 2607–2623, 2003.
[21] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2003, pp. 682–687.
[22] K. Wang and J. A. Kangas, "Character location in scene images from digital camera," *Pattern Recognit.*, vol. 36, no. 10, pp. 2287–2299, 2003.
[23] Q. Ye, W. Wang, W. Gao, and W. Zeng, "A robust text detection algorithm in images and video frames," in *Proc. Joint Conf. Inf., Commun. Signal Process. Pac. Rim Conf. Multimedia*, 2003, pp. 802–806.
[24] E. K. Wong and M. Chen, "A new robust algorithm for video text extraction," *Pattern Recognit.*, vol. 36, no. 6, pp. 1397–1406, 2003.
[25] D. Zhang and S. F. Chang, "A Bayesian framework for fusing multiple word knowledge models in videotext recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. 528–533.
[26] 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.
[27] D. Chen, J. M. Odobez, and H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognit.*, vol. 37, no. 3, pp. 596–608, 2004.
[28] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 366–373.
[29] J. Gllavata, R. Ewerth, and B. Freisleben, "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2004, pp. 425–428.
[30] H. Hase, T. Shinokawa, S. Tokai, and C. Y. Suen, "A robust method of recognizing multi-font rotated characters," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2004, pp. 363–366.
[31] X. Hua, W. Liu, and H. Zhang, "An automatic performance evaluation protocol for video text detection algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 4, pp. 498–507, Apr. 2004.
[32] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: A survey," *Pattern Recognit.*, vol. 37, pp. 977–997, 2004.
[33] D. Karatzas and A. Antonacopoulos, "Text extraction from web images based on a split-and-merge segmentation method using colour perception," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2004, pp. 634–637.
[34] G. K. Myers, R. C. Bolles, Q. T. Luong, J. A. Herson, and H. B. Aradhye, "Rectification and recognition of text in 3-D scenes," *Int. J. Doc. Anal. Recognit.*, vol. 7, pp. 147–158, 2004.

[35] Q. Ye, W. Gao, and Q. Huang, "Automatic text segmentation from complex background," in *Proc. IEEE Int. Conf. Image Process.*, 2004, pp. 2905–2908.
[36] D. Chen and J. Odobez, "Video text recognition using sequential Monte Carlo and error voting methods," *Pattern Recognit. Lett.*, vol. 26, no. 9, pp. 1386–1403, 2005.
[37] N. Ezaki, K. Kiyota, B. T. Minh, M. Bulacu, and L. Schomaker, "Improved text-detection methods for a camera-based text reading system for blind persons," in *Proc. IEEE Int. Conf. Document Anal. Recognit.*, 2005, pp. 257–261.
[38] S. Ferreira, V. Garin, and B. Gosselini, "A text detection technique applied in the framework of a mobile camera-based application," in *Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2005, pp. 133–139.
[39] Z. He, J. Liu, H. Ma, and P. Li, "A new automatic extraction method of container identity codes," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 1, pp. 72–78, Mar. 2005.
[40] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: A survey," *Int. J. Doc. Anal. Recognit.*, vol. 7, pp. 84–104, 2005.
[41] L. Lin and C. L. Tan, "Text extraction from name cards using neural network," in *Proc. Int. Joint Conf. Neural Netw.*, 2005, pp. 1818–1823.
[42] C. Liu, C. Wang, and R. Dai, "Text detection in images based on unsupervised classification of edge-based features," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2005, pp. 610–614.
[43] S. Lucas, "ICDAR 2005 text locating competition results," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2005, pp. 80–84.
[44] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J. Jolion, L. Todoran, M. Worring, and X. Lin, "ICDAR 2003 robust reading competitions: Entries, results, and future directions," *Int. J. Doc. Anal. Recognit.*, vol. 7, pp. 105–122, 2005
[45] M. R. 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.
[46] S. Nomura, K. Yamanak, O. Katai, H. Kawakami, T. Shiose, "A novel adaptive morphological approach for degraded character image segmentation," *Pattern Recognit.*, vol. 38, no. 11, pp. 1961–1975, 2005.
[47] K. Subramanian, P. Natarajan, M. Decerbo, and D. Castanon, "Character-stroke detection for text-localization and extraction," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, pp. 33–37, 2005.
[48] L. Tang and J. R. Kender, "A unified text extraction method for instructional videos," in *Proc. IEEE Int. Conf. Image Process.*, 2005, pp. 1216–1219.
[49] Q. Ye, Q. Huang, W. Gao and D. Zhao, "Fast and robust text detection in images and video frames," *Image Vis. Comput.*, vol. 23, pp. 565–576, 2005.
[50] C. Wolf and J. M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *Int. J. Doc. Anal. Recognit.*, vol. 28, no. 4, pp. 280–296, 2005.
[51] W. Wu, D. Chen, and J. Yang, "Integrating co-training and recognition for text detection," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1166–1169.
[52] S. Lu and C. L. Tan, "Camera text recognition based on perspective invariants," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2006, pp. 1042–1045.
[53] S. Omachi, M. Iwamura, S. Uchida, and K. Kise, "Affine invariant information embedment for accurate camera-based character recognition," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2006, pp. 1098–1101.
[54] C. Mancas-Thillou and B. Gosselin, "Spatial and color spaces combination for natural scene text extraction," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 985–988.
[55] M. J. Traxler and M. A. Gernsbacher, *Handbook of Psycholinguistics*. Amsterdam, The Netherlands: Elsevier, 2006.
[56] S. Wachenfeld, H. Klein, and X. Jiang, "Recognition of screen-rendered text," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2006, pp. 1086–1089.
[57] J. J. Weinman and E. Learned-Miller, "Improving recognition of novel input with similarity," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 1, pp. 308–315.
[58] S. Uchida, M. Iwamura, S. Omachi, and K. Kise, "OCR fonts revisited for camera-based character recognition," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2006, pp. 1134–1137.
[59] R. Dai, C. Liu, and B. Xiao, "Chinese character recognition: History, status and prospects," in *Proc. Front. Comput. Sci. China*, 2007, pp. 126–136.
[60] V. C. Dinh, S. S. Chun, S. Cha, H. Ryu, and S. Sull, "An efficient method for text detection in video based on stroke width similarity," in *Proc. Asia Conf. Comput. Vis.*, 2007, pp. 200–209.
[61] J. Lim, J. Park, and G. Medioni, "Text segmentation in color images using tensor voting," *Image Vis. Comput.*, vol. 25, no. 5, pp. 671–685, 2007.
[62] K. Subramanian, P. Natarajan, M. Decerbo, and D. Castanon, "Character-stroke detection for text localization and extraction," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2007, pp. 33–37.
[63] C. Mancas-Thillou and B. Gosselin, "Color text extraction with selective metric-based clustering," *Comput. Vis. Image Understanding*, vol. 107, no. 1/2, pp. 97–107, 2007.
[64] J. J. Weinman, E. Learned-Miller, and A. Hanson, "Fast lexicon-based scene text recognition with sparse belief propagation," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2007, pp. 979–983.
[65] Q. Ye, J. Jiao, J. Huang, and H. Yu, "Text detection and restoration in natural scene images," *Int. J. Vis. Commun. Image Representation*, vol. 18, no. 6, pp. 504–513, 2007.
[66] L. Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "reCAPTCHA: Human-Based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.
[67] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2008, vol. 2, pp. 290–294.
[68] S. M. Hanif, L. Prevost, and P. A. Negri, "A cascade detector for text detection in natural scene images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[69] D. Kim and K. Sohn, "Static text region detection in video sequences using color and orientation consistencies," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[70] S. Lee and J. Kim, "Complementary combination of holistic and component analysis for recognition of low-resolution video character images," *Pattern Recognit. Lett.*, vol. 29, no. 4, pp. 383–391, 2008.
[71] M. Li and C. Wang, "An adaptive text detection approach in images and video frames," in *Proc. Int. Joint Conf. Neural Netw.*, 2008, pp. 72–77.
[72] J. Liang, D. DeMenthon, and D. Doermann, "Geometric rectification of camera-captured document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 591–605, Apr. 2008.
[73] F. Liu, X. Peng, T. Wang, and S. Lu, "A density-based approach for text extraction in images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[74] Z. Liu and S. Sarkar, "Robust outdoor text detection using text intensity and shape features," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[75] H. Liu, Q. Wu, H. Zha, and X. Liu, "Skew detection for complex document images using robust borderlines in both text and non-text regions," *Pattern Recognit. Lett.*, vol. 29, no. 13, pp. 1893–1900, 2008.
[76] X. Liu, H. Fu, and Y. Jia, "Gaussian Mixture modeling and learning of neighboring characters for multilingual text extraction in images," *Pattern Recognit.*, vol. 41, no. 2, pp. 484–493, 2008.
[77] X. Liu and D. Doermann, "A camera phone based currency reader for the visually impaired," in *Proc. ACM SIGACCESS Conf. Comput. Accessibility*, 2008, pp. 305–306.
[78] D. Nister and H. Stewenius, "Linear time maximally stable extremal regions," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 183–196.
[79] P. Roy, U. Pal, J. Llados, and M. Delalandre, "Convex hull based approach for multi-oriented character recognition from graphical documents," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[80] P. Shivakumara, W. Huang, and C. L. Tan, "Efficient video text detection using edge features," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008 pp. 1–4.
[81] C. Singh, N. Bhatia, and A. Kaur, "Hough transform based fast skew detection and accurate skew correction methods," *Pattern Recognit.*, vol. 41, no. 12, pp. 3528–3546, 2008.
[82] 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. Int. Conf. Comput. Inf. Sci.*, 2008, pp. 185–190.
[83] J. J. Weinman, E. Learned-Miller, and A. Hanson, "A discriminative semi-Markov model for robust scene text recognition," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[84] L. Xu, H. Nagayoshi, and H. Sako, "Kanji character detection from complex real scene images based on character properties," in *Proc. IAPR Int. Workshop Doc. Anal. Syst.*, 2008, pp. 278–285.

[85] J. Zang and R. Kasturi, "Extraction of text objects in video documents: Recent progress," in *Proc. IAPR Int. Workshop Doc. Anal. Syst.*, 2008, pp. 5–17.

[86] P. Zhou, L. Li, and C. L. Tan, "Character recognition under severe perspective distortion," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1-4.

[87] U. Bhattachatya, S. K. Parui, and S. Mondal, "Devenagari and Bangla text extraction from natural scene images," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 57–61.

[88] T. E. de Campos, B. R. Babu, and M. Varma, "Character recognition in natural images," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2009.

[89] Y. Chen and B. Wu, "A multi-plane approach for text segmentation of complex document images," *Pattern Recognit.*, vol. 42, no. 7, pp. 1419–1444, 2009.

[90] S. M. Hanif and L. Prevost, "Text detection and localization in complex scenes using constrained adaboost algorithm," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 1–5.

[91] J. Jiao, Q. Ye, and Q. Huang, "A configurable method for multi-style license plate recognition," *Pattern Recognit.*, vol. 42, no. 3, pp. 504–513, 2009.

[92] 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.

[93] H. Goto and M. Tanaka, "Text-tracking wearable camera system for the blind," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 141–145.

[94] N. Mavaddat, T. K. Kim, and R. Cipolla, "Design and evaluation of features that best define text in complex scene images," in *Proc. IAPR Conf. Mach. Vis. Appl.*, 2009, pp. 94–97.

[95] N. Nikolaou and N. Papamarkos, "Color reduction for complex document images," *Int. J. Imag. Syst. Technol.*, vol. 19, pp. 14–26, 2009.

[96] Y. Pan, X. Hou, and C. L. Liu, "Text localization in natural scene images based on conditional random field," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 6–10.

[97] P. Roy, U. Pal, J. Llados, and M. Delalandre, "Multi-oriented and multi-sized touching character segmentation using dynamic programming," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 11–15.

[98] P. Shivakumara, T. Phan, and C. L. Tan, "A gradient difference based technique for video text detection," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 150–160.

[99] R. Smith, D. Antonova, and D. Lee, "Adapting the Tesseract open source OCR engine for multilingual OCR," in *Proc. Joint Workshop Multilingual OCR Anal. Noisy Unstruct. Text Data*, 2011.

[100] J. 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.

[101] P. Zhou, L. Li, and C. L. Tan, "Character recognition under severe perspective distortion," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 1–4.

[102] T. Yamazoe, M. Etoh, T. Yoshimura, and K. Tsujino, "Hypothesis preservation approach to scene text recognition with weighted finite-state transducer," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 359–363.

[103] J. Yi, Y. Peng, and J. Xiao, "Using multiple frame integration for the text recognition of video," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2009, pp. 71–75.

[104] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A two-stage scheme for text detection in video images," *Image Vis. Comput.*, vol. 28, no. 9, pp. 1413–1426, 2010.

[105] G. Caner and I. Haritaoglu, "Shape-DNA: Effective character restoration and enhancement for Arabic text documents," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 2053–2056.

[106] Y. Chiang and C. A. Knoblock, "An approach for recognizing text labels in raster maps," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 3199–3202.

[107] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2963–2970.

[108] X. Huang and H. Ma, "Automatic detection and localization of natural scene text in video," *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 3216–3219.

[109] S. Lee, M. Cho, K. Jung, and J. Kim, "Scene text extraction with edge constraint and text collinearity," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3983–3986.

[110] X. Liu, W. Wang, and T. Zhu, "Extracting captions in complex background from videos," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 3232–3235.

[111] M. Moradi, S. Mozaffari, and A. A. Orouji, "Farsi/Arabic text extraction from video images by corner detection," in *Proc. Iranian Mach. Vis. Image Process.*, 2010, pp. 1–6.

[112] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asia Conf. Comput. Vis.*, 2010, pp. 770–783.

[113] Y. Pan, C. L. Liu, and X. Hou, "Fast scene text localization by learning-based filtering and verification," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 2269–2272.

[114] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, and S. Hwang, "Automatic detection and recognition of Korean text in outdoor signboard images," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1728–1739, 2010.

[115] P. Shivakumara, W. Huang, T. Phan, and C. Tan, "Accurate video text detection through classification of low and high contrast images," *Image Vis. Comput.*, vol. 43, no. 6, pp. 2165–2185, 2010.

[116] P. Shivakumara, T. Phan, and C. Tan, "New Fourier-statistical features in RGB space for video text detection," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 20, no. 11, pp. 1520–1532, Nov. 2010.

[117] P. Shivakumara, A. Dutta, U. Pal, and C. L. Tan, "A new method for handwritten scene text detection in video," in *Proc. IEEE Int. Conf. Front. Handwritten Recognit.*, 2010, pp. 387–392.

[118] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 591–604.

[119] J. J. Weinman, "Typographical features for scene text recognition," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 3987–3990.

[120] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 157–170.

[121] Z. Zhou, L. Li, and C. L. Tan, "Edge based binarization for video text images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 133–136.

[122] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Proc. IEEE Int. Conf. Image Process.*, 2011, pp. 2609–2612.

[123] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu, and A. Y. Ng, "Text detection and character recognition in scene images with unsupervised feature learning," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 440–445.

[124] Y. Chiang and C. A. Knoblock, "Recognition of multi-oriented, multi-sized, and curved text," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1399–1403.

[125] A. B. Cambra and A. C. Murillo, "Towards robust and efficient text sign reading from a mobile phone," in *Proc. Workshop IEEE Int. Conf. Comput. Vis.*, 2011, pp. 64–71.

[126] K. Elagouni, C. Garcia, and P. Sbillot, "A comprehensive neural-based approach for text recognition in videos using natural language processing," in *Proc. ACM Conf. Multimedia Retrieval*, 2011.

[127] M. Iwamura, T. Kobayashi, and K. Kise, "Recognition of multiple characters in a scene image using arrangement of local features," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1409–1413.

[128] A. Jamil, I. Siddiqi, F. Arif, and A. Raza, "Edge-based features for localization of artificial Urdu text in video images," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1120–1124.

[129] D. Karatzas, S. Robles Mestre, J. Mas, and F. Nourbakhsh, "ICDAR 2011 robust reading competition: Challenge 1: Reading text in born-digital images," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1491–1496.

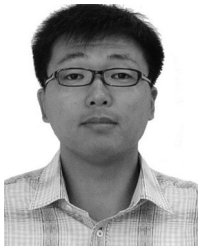
[130] J. Lee, P. Lee, S. Lee, A. Yuille, and C. Koch, "AdaBoost for text detection in natural scene," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 429–434.

[131] A. Levin, Y. Weiss, B. Freeman, and F. Durand, "Efficient marginal likelihood optimization in blind deconvolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2657–2664.

[132] J. Liu, H. Li, S. Zhang, and W. Liang, "A novel italic detection and rectification method for Chinese advertising images," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 698–702.

[133] C. Liu, C. Yang, X. Q. Ding, and K. Wang, "An improved scene text extraction method using conditional random field and optical character recognition," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 708–712.
[134] Y. Lou, A. L. Bertozzi, and S. Soattom, "Direct sparse deblurring," *J. Math. Imag. Vis.*, vol. 39, no. 1, pp. 1–12, 2011.
[135] A. Mishra, K. Alahari, and C. V. Jawahar, "An MRF model for binarization of natural scene text," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 11–16.
[136] R. Nagy, A. Dicker, and K. Meyer-Wegener, " NEOCR: A configurable dataset for natural image text recognition," in *Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2011, pp. 150–163.
[137] L. Neumann and J. Matas, "Text localization in real-world images using efficiently pruned exhaustive search," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 687–691.
[138] Y. F. Pan, X. Hou, and C. L. 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.
[139] T. Phan, P. Shivakumara, B. Su, and C. L. Tan, "A gradient vector flow-based method for video character segmentation," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1024–1028.
[140] D. Rajendran, P. Shivakumara, B. Su, S. Lu, and C. L. Tan, "A new Fourier-moments based video word and character extraction method for recognition," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1165–1169.
[141] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian approach to multi-oriented text detection in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 412–419, Feb. 2011.
[142] P. Shivakumara, S. Bhowmick, B. Su, C. L. Tan, and U. Pal, " A new gradient based character segmentation method for video text recognition," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 126–130.
[143] P. Shivakumara, T. Phan, S. Lu, and C.L. Tan, "Video character recognition through hierarchical classification," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 131–135.
[144] A. Shahab, F. Shafait, and A. Dengel, "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 1491–1496.
[145] D. Smith, J. Field, and E. Learned-Miller, "Enforcing similarity constraints with integer programming for better scene text recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 73–80.
[146] N. Stamatopoulos, B. Gatos, I. Pratikakis, and S. J. Perantonis, "Goal-oriented rectification of camera-based document images," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 910–920, Apr. 2011.
[147] T. Wakahara and K. Kita, "Binarization of color character strings in scene images using K-means clustering and support vector machines," in *Proc. IEEE Int, Conf. Doc. Anal. Recognit.*, 2011, pp. 3183–3186.
[148] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1457–1464.
[149] X. Wang, L. Huang, and C. Liu, "A novel method for embedded text segmentation based on stroke and color," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 151–155.
[150] 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, Sep. 2011.
[151] C. Yi and Y. Tian, "Text detection in natural scene images by stroke Gabor words," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 177–181.
[152] X. Zhao, K. H. Lin, Y. Fu, Y. Hu, Y. Liu, and T. S. Huang, "Text from corners: A novel approach to detect text and caption in videos," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 790-799, Mar. 2011.
[153] Y. Zhao, T. Lu, and W. Liao, "A robust color-independent text detection method from complex videos," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 374–378.
[154] G. Zhou, Y. Liu, Q. Meng, and Y. Zhang, "Detecting multilingual text in natural scene," in *Proc. IEEE 1st Int. Symp. Access Spaces*, 2011, pp. 116–120.
[155] B. Bai, F. Yin, and C. L. Liu, "A fast stroke-based method for text detection in video," in *Proc. IAPR Int. Workshop Doc. Anal. Syst.*, 2012, pp. 69–73.
[156] A. R. Chowdhury, U. Bhattacharya, and S. K. Parui, "Scene text detection using sparse stroke information and MLP," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 294–297.
[157] H. Cho, J. Wang, and S. Lee, "Text image deblurring using text-specific properties," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 524–537.
[158] K. Elagouni, C. Garcia, F. Mamalet, and P. Sebillot, "Combining multi-scale character recognition and linguistic knowledge for natural scene text OCR," in *Proc. IAPR Int. Workshop Doc. Anal. Syst.*, 2012, pp. 120–124.
[159] M. Felhi, N. Bonnier, and S. Tabbone, "A skeleton based descriptor for detecting text in real scene images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 282–285.
[160] X. Liu and W. Wang, "Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 482–489, Apr. 2012.
[161] A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2687–2694.
[162] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–11.
[163] A. Mosleh, N. Bouguila, and A. Ben Hamza, "Image text detection using a bandlet-based edge detector and stroke width transform," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–2.
[164] L. Neumann and J. Matas, "Real-time scene text location and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3538–3545.
[165] T. Novikova, O. Barinova1, P. Kohli, and V. Lempitsky, "Large-lexicon attribute-consistent text recognition in natural images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 752–765.
[166] Y. Pan, C. L. Liu, and X. Hou, "Fast scene text localization by learning-based filtering and verification," in *Proc. IEEE Int. Conf. Image Process.*, 2012, pp. 2269–2272.
[167] T. Q. Phan, P. Shivakumara, and C. L. Tan, "Text detection in natural scenes using gradient vector flow-guided symmetry," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 3296–3299.
[168] P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, vol. 4, pp. 3288–3291.
[169] N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein, and C. L. Tan, "A new method for arbitrarily-oriented text detection in video," in *Proc. IAPR Int. Workshop Doc. Anal. Syst.*, 2012, pp. 74–78.
[170] K. Sheshadri and S. K. Divvala, "Exemplar driven character recognition in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–10.
[171] C. Shi, B. Xiao, C. Wang, and Y. Zhang, "Graph-based background suppression for scene text detection," in *Proc. IAPR Int. Workshop Doc. Anal. Syst.*, 2012, pp. 210–214.
[172] P. Shivakumara, R. Sreedhar, T. Phan, S. Lu, and C. L. Tan, "Multioriented video scene text detection through Bayesian classification and boundary growing," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 22, no. 8, pp. 1227–1235, Aug. 2012.
[173] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolution neural networks," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 3304–3308.
[174] 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, Sep. 2012.
[175] C. Yao, X. Zhang, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1083–1090.
[176] Y. Zhu, J. Sun, and S. Naoi, "Recognizing natural scene characters by convolutional neural network and bimodal image enhancement," *in Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2012, pp. 69–82.
[177] S. Banerjee, K. Mullick, and U. Bhattacharya, "A robust approach to extraction of texts from camera captured images," in *Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2013, pp. 30–46.
[178] J. L. Feild and E. Learned-Miller, "Improving open-vocabulary scene text recognition," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 604–608.
[179] V. Goel, A. Mishra1, K. Alahari, and C. V. Jawaharl, "Whole is greater than sum of parts: Recognizing scene text words," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 398–402.
[180] L. Gomez and D. Karatzas, "Multi-script text extraction from natural scenes," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 467–471.

[181] R. Huang, P. Shivakumara, and S. Uchida, "Scene character detection by an edge-ray filter," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 462–466.
[182] H. Koo and D. H. Kim, "Scene text detection via connected component clustering and non-text filtering," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2296–2305, Jun. 2013.
[183] K. Kuramoto, W. Ohyama, T. Wakabayashi, and F. Kimura, "Accuracy improvement of viewpoint-free scene character recognition by rotation angle estimation," in *Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2013, pp. 60–70.
[184] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez, S. Robles, J. Mas, D. Fernandez, J. Almazan, and L.P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 1484–1493.
[185] S. H. Lee and J. H. Kim, "Integrating multiple character proposals for robust scene text extraction," *Image Vis. Comput.*, vol. 31, pp. 823–840, 2013.
[186] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "T-HOG: An effective gradient-based descriptor for single line text regions," *Pattern Recognit.*, vol. 46, no. 3, pp. 1078–1090, 2013.
[187] S. Milyaev, O. Barinova, T. Novikova, P. Kohli, and V. S. Lempitsky, "Image binarization for end-to-end text understanding in natural images," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 128–132.
[188] L. Neumann and J. Matas, "On combining multiple segmentations in scene text recognition," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 523–527.
[189] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 97–104.
[190] T. Q. Phan, P. Shivakumara, T. Lu, and C. L. Tan, "Recognition of video text through temporal integration," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 589–593.
[191] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 569–576.
[192] J. H. Seok and J. H. Kim, "Scene text recognition with a Hough forest implicit shape model," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 599–603.
[193] P. Shivakumara, T. Q. Phan, S. Bhowmick, C. L. Tan, and U. Pal, "A novel ring radius transform for video character reconstruction," *Pattern Recognit.*, vol. 46, no. 1, pp. 131–140, 2013.
[194] P. Shivakumara, H. T. Basavaraju, D. S. Guru, and C. L. Tan, "Detection of curved text in video: Quad tree based method," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 594–598.
[195] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detections," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2961–2968.
[196] Y. Terada, R. Huang, Y. Feng, and S. Uchida, "On the possibility of structure learning-based scene character detector," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 472–476.
[197] C. Yao, X. Zhang, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Rotation-invariant features for multi-oriented text detection in natural images," *PLoS ONE*, vol. 8, no. 8, p. e70173, 2013.
[198] Q. Ye and D. Doermann, "Scene text detection via integrated discriminition of component appearance and consensus," in *Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2013, pp. 47–59.
[199] C. Yi and Y. Tian, "Text extraction from scene images by character apperance and structure modeling," *Comput. Vis. Image Understanding*, vol. 117, no. 2, pp. 182–194, 2013.
[200] H. Wang, Y. Landa, M. F. Fallon, and S. J. Teller, "Spatially prioritized and persistent text detection and decoding," in *Proc. Int. Workshop Camera-Based Doc. Anal. Recognit.*, 2013, pp. 3–17.
[201] X. Zhang and F. Sun, "Multiple geometry transform estimation from single camera-captured text image," in *Proc. IEEE Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 538–542.
[202] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: Reading text in uncontrolled conditions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 785–992.
[203] L. Kang and D. Doermann, "Orientation robust text line detection in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4034–4041.
[204] J. L. Feild, "Improving text recognition in images of natural scenes," Ph.D. dissertation, Computer Science, Univ. Massachusetts Amherst, Amherst, MA, USA, 2014.
[205] C. Leey, A. Bhardwaj, W. Di, V. Jagadeesh, and R. Piramuthu, "Region-based discriminative feature pooling for scene text recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4050–4057.
[206] J. J. Weinman, Z. Butler, D. Knoll, and J. Feild, "Toward integrated scene text reading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 3, no. 2, pp. 375–387, Feb. 2014.
[207] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4042–4049.
[208] X. C. Yin, X. Yin, K. Huang, and H. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.



Qixiang Ye (M'10) received the BS and MS degrees in mechanical and electrical engineering from the Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2006. He has been an associate professor with the University of Chinese Academy of Sciences since 2009, and was a visiting assistant professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park until 2013. His research interests include image processing, visual object detection and machine learning. He pioneered the Kernel SVM based pyrolysis output prediction software which was put into practical application by SINOPEC in 2012. He developed two kinds of piecewise linear SVM methods which were successfully applied into visual object detection. He has published more than 50 papers in refereed conferences and journals, and received the Sony Outstanding Paper Award. He is a member of the IEEE.

David Doermann (F'13) received the BSc degree in computer science and mathematics from Bloomsburg University, Bloomsburg, PA, in 1987, the MS degree from the University of Maryland, College Park, in 1989, the PhD degree from the University of Maryland, in 1993, and the honorary doctorate of technology sciences from the University of Oulu, Oulu, Finland, in 2002, for his contributions to digital media processing and document analysis research. He is a senior research scientist with the University of Maryland Institute for Advanced Computer Studies (UMIACS). Since 1993, he has been the director with the Laboratory for Language and Media Processing, UMIACS, and an adjunct member of the graduate faculty. His team focuses on topics related to document image analysis and multimedia information processing. Recent intelligent document image analysis projects include page decomposition, structural analysis and classification, page segmentation, logo recognition, document image compression, duplicate document image detection, image-based retrieval, scene text detection, generation of synthetic OCR data, and signature verification. In video processing, projects have centered on the segmentation of compressed domain video sequences, structural representation and classification of video, detection of reformatted video sequences, and the performance evaluation of automated video analysis algorithms. He has more than 30 journal publications and 125 refereed conference papers. He is a founding co-editor of the *International Journal on Document Analysis and Recognition*, was the general chair or co-chair of more than a half dozen international conferences and workshops, and was the general chair of the International Conference on Document Analysis and Recognition in 2013. He is a fellow of the IEEE and IAPR.