

Charcterness: An Indicator of Text in the Wild

Yao Li, Wenjing Jia, Chunhua Shen, Anton van den Hengel

Abstract—Text in an image provides vital information for interpreting its contents, and text in a scene can aide with a variety of tasks from navigation, to obstacle avoidance, and odometry. Despite its value, however, identifying general text in images remains a challenging research problem. Motivated by the need to consider the widely varying forms of natural text, we propose a bottom-up approach to the problem which reflects the ‘charcterness’ of an image region. In this sense our approach mirrors the move from saliency detection methods to measures of ‘objectness’. In order to measure the charcterness we develop three novel cues that are tailored for character detection, and a Bayesian method for their integration. Because text is made up of sets of characters, we then design a Markov random field (MRF) model so as to exploit the inherent dependencies between characters.

We experimentally demonstrate the effectiveness of our charcterness cues as well as the advantage of Bayesian multi-cue integration. The proposed text detector outperforms state-of-the-art methods on a few benchmark scene text detection datasets. We also show that our measurement of ‘charcterness’ is superior than state-of-the-art saliency detection models when applied to the same task.

Index Terms—Charcterness, scene text detection, saliency detection

I. INTRODUCTION

Human beings find the identification of text in an image almost effortless, and largely involuntarily. As a result, much important information is conveyed in this form, including navigation instructions (exit signs, and route information, for example), and warnings (danger signs *etc.*), amongst a host of others. Simulating such an ability for machine vision system has been an active topic in the vision and document analysis community. Scene text detection serves as an important pre-processing step for end-to-end scene text recognition which has manifested itself in various forms, including navigation, obstacle avoidance, and odometry to name a few. Although some breakthrough have been made, the accuracy of the state-of-the-art scene text detection algorithms still lag behind human performance on the same task.

Visual attention, or visual saliency, is fundamental to the human visual system, and alleviates the need to process the otherwise vast amounts of incoming visual data. As such it has been a well studied problem within multiple disciplines, including cognitive psychology, neurobiology, and computer vision. In contrast with some of the pioneering saliency detection models [1]–[5] which has achieved reasonable accuracy in predicting human eye fixation, recent work has focused on *salient object detection* [6]–[19]. The aim of salient object detection is to highlight the whole attention-grabbing object

Y. Li, C. Shen and A. van den Hengel are with the University of Adelaide, Australia. W. Jia is with the University of Technology, Sydney, Australia.

W. Jia’s contribution was made when visiting the University of Adelaide. All correspondence should be addressed to C. Shen (chhshen@gmail.com).

with well-defined boundary. Previous saliency detection approaches can be broadly classified into either local [12], [16], [19] or global [7], [9], [10], [14], [15], [17] methods. Saliency detection has manifested itself in various forms, including image retargeting [20], [21], image classification [22], image segmentation [23].

Our basic motivation is the fact that *text attracts human attention*, even when amongst a cluttered background. This has been shown by a range of authors including Judd *et al.* [5] and Cerf *et al.* [24] who verified that humans tend to focus on text in natural scenes.

Previous work [25]–[28] has also demonstrated that saliency detection models can be used in early stages of scene text detection. In [25], for example, a saliency map obtained from Itti *et al.* [1] was used to find regions of interest. Uchida *et al.* [28] showed that using both SURF and saliency features achieved superior character recognition performance over using SURF features alone. More recently, Shahab *et al.* [26] compared the performance of four different saliency detection models at scene text detection. Meng and Song [27] also adopted the saliency framework of [11] for scene text detection.

While the aforementioned approaches have demonstrated that saliency detection models facilitate scene text detection, they share a common inherent limitation, which is that they are distracted by other salient objects in the scene. The approach we propose here differs from these existing methods in that we propose a text-specific saliency detection model (*i.e.* a charcterness model) and demonstrate its robustness when applied to scene text detection.

Measures of ‘objectness’ [13] have built upon the saliency detection in order to identify windows within an image that are likely to contain an object of interest. Applying an objectness measure in a sliding-window approach thus allows the identification of interesting objects, rather than regions. This approach has been shown to be very useful as a pre-processing step for a wide range of problems including occlusion boundary detection [29], semantic segmentation [30], and training object class detectors [31].

We propose here a similar approach to text detection, in that we seek to develop a method which is capable of identifying individual, bounded units of text, rather than areas with text-like characteristics. The unit in the case of text is the character, and much like the ‘object’, it has a particular set of characteristics, including a closed boundary. In contrast to the objects of [13], however, text is made up of a set of inter-related characters. Therefore, effective text detection should be able to compensate for, and exploit these dependencies between characters. The object detection method of [13] is similar to that proposed here in as much as it is based on a Bayesian framework combining a number of visual cues,

including one which represents the boundary of the object, and one which measures the degree to which a putative object differs from the background.

In contrast to saliency detection algorithms which either attempt to identify pixels or rectangular image windows that attract the eye, our focus here is instead on identifying individual characters within non-rectangular regions. As characters represent the basic units of text, this renders our method applicable in a wider variety of circumstances than saliency-based paragraph detection, yet more specific. When integrating the three new characterness cues developed, instead of simple linear combination, we use a Bayesian approach to model the joint probability that a candidate region represents a character. The probability distribution of cues on both characters and non-characters are obtained from training samples. In order to model and exploit the inter-dependencies between characters we use the graph cuts [32] algorithm to carry out inference over an MRF designed for the purpose.

To the best of our knowledge, we are the first to present a saliency detection model which measures the characterness of image regions. This text-specific saliency detection model is less likely to be distracted by other objects which are usually considered as salient in general saliency detection models. Promising experimental results on benchmark datasets demonstrate that our characterness approach outperforms the state-of-the-art.

II. RELATED WORK

A. Scene text detection

Existing scene text detection approaches generally fall into one of three categories, namely, texture-based approaches, region-based approaches, and hybrid approaches.

Texture-based approaches [33], [34] extract distinct texture properties from sliding windows, and use a classifier (*e.g.*, AdaBoost [33], [34]) to detect individual instances. Some widely used features include Histograms of Gradients (HOGs), Local Binary Patterns (LBP), Gabor filters and wavelets. Foreground regions on various scales are then merged to generate final text regions. Yet, there is something profoundly unsatisfying about texture-based approaches. Firstly, the brute force nature of window classification is not particularly appealing. Its computational complexity is proportional to the product of the number of scales. Secondly, it is unclear which features have contributed most for removing various interfering backgrounds. Thirdly, an extra post processing step, *i.e.* text extraction, is needed before text recognition.

Region-based approaches [35]–[41], on the other hand, first extract potential characters through edge detection [35]–[37], color clustering [38] or Maximally Stable Extremal Region (MSER) detection [39], [40]. After that, low level features based on geometric and shape constraints are used to reject non-characters. As a final step, remaining regions are clustered into lines through measuring the similarities between them. A typical example of region-based approaches was proposed by Epshtain *et al.* [35], where a local image operator, called Stroke Width Transform (SWT), assigned each pixel with the most likely stroke width value, followed by a series of rules in

order to remove non-characters. All region-based approaches, although not particularly computationally demanding, involve many parameters that need to be tuned manually. An advantage of region-based approaches is that their results can be sent to Optical Character Recognition (OCR) software for recognition directly, without the extra text extraction step.

Hybrid approaches [42]–[47] are a combination of texture-based and region-based approaches. Usually, the initial step is to extract potential characters, which is the same as region-based approaches. Instead of utilizing low level cues, these methods extract features from regions and exploit classifiers to decide whether a particular region contains text or not, and this is considered as a texture-based step. In [45], potential characters were extracted by the SWT initially. To reject non-characters two random forest classifiers were trained using two groups of features (component and chain level) respectively.

B. Saliency detection

The underlying hypothesis of existing saliency detection algorithms is the same: *the contrast between salient object and background is high*. Contrast can be computed via various features, such as intensity [12], edge density [13], orientation [12], and most commonly color [8]–[19]. The measurement of contrast also varies, including discrete form of Kullback-Leibler divergence [12], intersection distance [9], χ^2 distance [11], [13], [16], [18], Euclidean distance [17]. As no prior knowledge about the size of salient objects is available, contrast is computed at multiple scales in some methods [11], [12], [16], [17], [19]. To make the final saliency map smoother, spatial information is also commonly adopted in the computation of contrast [9], [10], [14], [16], [17].

The large amount of literature on saliency detection can be broadly classified into two classes in terms of the scope of contrast computed. *Local methods* [12], [16], [19] estimate saliency of an image patch according to its contrast against its surrounding patches. They assume that patches whose contrast values are higher against their surrounding counterparts should be salient. As computing local contrast at one scale tends to only highlight boundaries of the salient object rather than the whole object, local methods are always performed at multiple scales

Global methods, *e.g.*, [4], [7], [9], [10], [14], [15], [17], on the other hand, take the entire image into account when estimating saliency of a particular patch. The underlying hypothesis is that globally rare features correspond to high saliency. Color contrast over the entire image is computed in [10]. Shen and Wu [15] stacked features extracted from all patches into a matrix and then solved a low-rank matrix recovery problem. Perazzi *et al.* [14] showed that global saliency can be estimated by high-dimensional Gaussian filters.

C. Contributions

Although previous work [25]–[28] has demonstrated that existing saliency detection models can facilitate scene text detection, none of them has designed a saliency detection model tailored for scene text detection. We argue that adopting

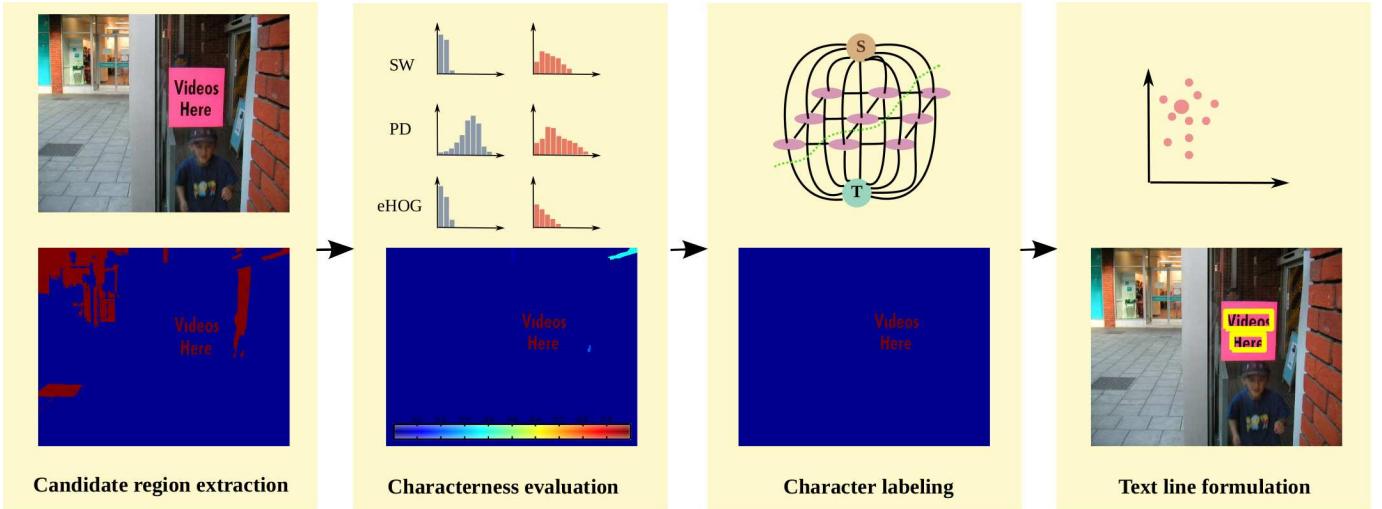


Fig. 1. Overview of our scene text detection approach. The characterness model consists of the first two phases.

existing saliency detection models directly to scene text detection [25]–[28] is inappropriate, as general saliency detection models are likely to be distracted by non-character objects in the scene that are also salient. In summary, contributions of our work comprise the following.

- 1) We propose a text detection model which reflects the ‘characterness’ (*i.e.* the probability of representing a character) of image regions.
- 2) We design an energy-minimization approach to character labeling, which encodes both individual characterness and pairwise similarity in a unified framework.
- 3) We evaluate ten state-of-the-art saliency detection models for the measurement of ‘characterness’. To the best of our knowledge, we are the first to evaluate state-of-the-art saliency detection models for reflecting ‘characterness’ in this large quantity.

III. OVERVIEW

Fig. 1 shows an overview of the proposed scene text detection approach. Specifically, Sec. IV describes the characterness model, in which perceptually homogeneous regions are extracted by a modified MSER-based region detector. Three novel characterness cues are then computed, each of which independently models the probability of the region forming a character (Sec. IV-B). These cues are then fused in a Bayesian framework, where Naive Bayes is used to model the joint probability. The posterior probability reflects the ‘characterness’ of the corresponding image patch.

In order to consolidate the characterness responses we design a character labeling method in Sec. V-A. An MRF, minimized by graph cuts [32], is used to combine evidence from multiple per-patch characterness estimates into evidence for a single character or compact group of characters. Finally, verified characters are grouped to readable text lines via a clustering scheme (Sec. V-B).

Two phases of experiments are conducted separately in order to evaluate the characterness model and scene text detection approach as a whole. In the first phase (Sec. VI),

we compare the proposed characterness model with ten state-of-the-art saliency detection algorithms on the characterness evaluation task, using evaluation criteria typically adopted in saliency detection. In the second phase (Sec. VII), as in conventional scene text detection algorithms, we use the bounding boxes of detected text lines in order to compare against state-of-the-art scene text detection approaches.

IV. THE CHARACTERNES MODEL

A. Candidate region extraction

MSER [48] is an effective region detector which has been applied in various vision tasks, such as tracking [49], image matching [50], and scene text detection [39], [42], [47], [51], [52] amongst others. Roughly speaking, for a gray-scale image, MSERs are those which have a boundary which remains relatively unchanged over a set of different intensity thresholds. The MSER detector is thus particularly well suited to identifying regions with almost uniform intensity surrounded by contrasting background.

For the task of scene text detection, although the original MSER algorithm is able to detect characters in most cases, there are some characters that are either missed or incorrectly connected (Fig. 2 (b)). This tends to degrade the performance of the following steps in the scene text detection algorithms. To address this problem, Chen *et al.* [39] proposed to prune out MSER pixels which were located outside the boundary detected by Canny edge detector. Tsai *et al.* [52] performed judicious parameter selection and multi-scale analysis of MSERs. Neumann and Matas extended MSER to MSER++ [51] and later Extremal Region (ER) [43]. In this paper, we use the edge-preserving MSER algorithm from our earlier work [41] (*c.f.* Algorithm 1).

Motivation. As illustrated in some previous work [50], [53], the MSER detector is sensitive to blur. We have observed through empirical testing that this may be attributed to the large quantities of *mixed pixels* (pixels lie between dark background and bright regions, and *vice versa*) present along character boundaries. We notice that these mixed pixels usually



(a) Original text (b) Original MSER (c) Our results

Fig. 2. Cases that the original MSER fails to extract the characters while the modified eMSER succeeds.

have larger gradient amplitude than other pixels. We thus propose to incorporate the gradient amplitude so as to produce edge-preserving MSERs (see Fig. 2(c)).

Algorithm 1: Edge-preserving MSER (eMSER)

- Input:** A color image, and required parameters
Output: Regions contain characters and non-characters
- 1 Convert the color image to an intensity image I .
 - 2 Smooth I using a guided filter [54].
 - 3 Compute the gradient amplitude map ∇I , then normalize it to $[0, 255]$.
 - 4 Get a new image $I^* = I + \gamma \nabla I$ (resp. $I^* = I - \gamma \nabla I$).
 - 5 Perform MSER algorithm on I^* to extract dark regions on the bright background (resp. bright regions on the dark background).
-

B. Characterness evaluation

1) *Characterness cues*: Characters attract human attention because their appearance differs from that of their surroundings. Here, we propose three novel cues to measure the unique properties of characters.

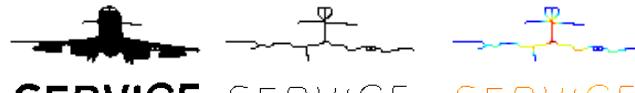
Stroke Width (SW). Stroke width has been a widely exploited feature for text detection [35], [37], [45], [46]. In particular, SWT [35] computes the length of a straight line between two edge pixels in the perpendicular direction, which is used as a preprocessing step for later algorithms [45], [55], [56]. In [46], a stroke is defined as a connected image region with uniform color and half-closed boundary. Although this assumption is not supported by many uncommon typefaces, stroke width remains a valuable cue.

Based on the efficient stroke width computation method we have developed earlier [40] (*c.f.* Algorithm 2), the stroke width cue of region r is defined as:

$$\text{SW}(r) = \frac{\text{Var}(l)}{\text{E}(l)^2}, \quad (1)$$

where $\text{E}(l)$ and $\text{Var}(l)$ are stroke width mean and variance (*c.f.* Algorithm 2). In Fig. 3 (c), we use color to visualize the stroke width of exemplar characters and non-characters, where larger color variation indicates larger stroke width variance and *vice versa*.

Perceptual Divergence (PD). As stated in Sec. II, color contrast is a widely adopted measurement of saliency. For



(a) Detected regions (b) Skeleton (c) Distance transform
Fig. 3. Efficient stroke width computation [40] (best viewed in color). Note the color variation of non-characters and characters on (c). Larger color variation indicates larger stroke width variance.

Algorithm 2: Efficient computation of stroke width

- Input:** A region r
Output: Stroke width mean $\text{E}(l)$ and variance $\text{Var}(l)$
- 1 Extract the skeleton S of the region.
 - 2 For each pixel $p \in S$, find its shortest path to the region boundary via distance transform. The corresponding length l of the path is defined as the stroke width.
 - 3 Compute the mean $\text{E}(l)$ and variance $\text{Var}(l)$.
-

the task of scene text detection, we observed that, in order to ensure reasonable readability of text to a human, the color of text in natural scenes is typically distinct from that of the surrounding area. Thus, we propose the PD cue to measure the perceptual divergence of a region r against its surroundings, which is defined as:

$$\text{PD}(r) = \sum_{R,G,B} \sum_{j=1}^b h_j(r) \log \frac{h_j(r)}{h_j(r^*)}, \quad (2)$$

where the term $\int_x p(x) \log \frac{p(x)}{q(x)}$ is the Kullback-Leibler divergence (KLD) measuring the dissimilarity of two probability distributions in the information theory. Here we take advantage of its discrete form [12], and replace the probability distributions $p(x)$, $q(x)$ by the color histograms of two regions $h(r)$ and $h(r^*)$ (r^* denotes the region outside r but within its bounding box) in a sub-channel respectively. $\{j\}_1^b$ is the index of histogram bins. Note that the more different the two histograms are, the higher the PD is.

In [57], the authors quantified the perceptual divergence as the overlapping areas between the normalized intensity histograms. However, using the intensity channel only ignores valuable color information, which will lead to a reduction in the measured perceptual divergence between distinct colors with the same intensity. In contrast, all three sub-channels (*i.e.*, R, G, B) are utilized in the computation of perceptual divergence in our approach.

Histogram of Gradients at Edges (eHOG). The Histogram of Gradients (HOGs) [58] is an effective feature descriptor which captures the distribution of gradient magnitude and orientation. Inspired by [36], we propose a characterness cue based on the gradient orientation at edges of a region, denoted by eHOG. This cue aims to exploit the fact that the edge pixels of characters typically appear in pairs with opposing gradient

Videos

Fig. 4. Sample text (left) and four types of edge points represented in four different colors (right). Note that the number of edge points in blue is roughly equal to that in orange, and so for green and crimson.

directions [36]¹.

Firstly, edge pixels of a region r are extracted by the Canny edge detector. Then, gradient orientations θ of those pixels are quantized into four types, *i.e.*, Type 1: $0 < \theta \leq \pi/4$ or $7\pi/4 < \theta \leq 2\pi$, Type 2: $\pi/4 < \theta \leq 3\pi/4$, Type 3: $3\pi/4 < \theta \leq 5\pi/4$, and Type 4: $5\pi/4 < \theta \leq 7\pi/4$. An example demonstrating the four types of edge pixels for text is shown in Fig. 4 (right), where four different colors are used to depict the four types of edge pixels. As it shows, we can expect that the number of edge pixels in Type 1 should be close to that in Type 3, and so for Type 2 and Type 4.

Based on this observation, we define the eHOG cue as:

$$\text{eHOG}(r) = \frac{\sqrt{(w_1(r) - w_3(r))^2 + (w_2(r) - w_4(r))^2}}{\sum_{i=1}^4 w_i(r)}, \quad (3)$$

where $w_i(r)$ denotes the number of edge pixels in Type i within region r , and the denominator $\sum_{i=1}^4 w_i(r)$ is for the sake of scale invariance.

2) *Bayesian multi-cue integration*: The aforementioned cues measure the characterness of a region r from different perspectives. SW and eHOG distinguish characters from non-characters on the basis of their differing intrinsic structures. PD exploits surrounding color information. Since they are complementary and obtained independently of each other, we argue that combining them in the same framework outperforms any of the cues individually.

Following the Naive Bayes model, we assume that each cue is conditionally independent. Therefore, according to Bayes' theorem, the posterior probability that a region is a character (its characterness score) can be computed as:

$$\begin{aligned} p(c|\Omega) &= \frac{p(\Omega|c)p(c)}{p(\Omega)} \\ &= \frac{p(c)\prod_{cue \in \Omega} p(cue|c)}{\sum_{k \in \{c,b\}} p(k)\prod_{cue \in \Omega} p(cue|k)}, \end{aligned}$$

where $\Omega = \{\text{SW}, \text{PD}, \text{eHOG}\}$, and $p(c)$ and $p(b)$ denote the prior probability of characters and background respectively, which we determine on the basis of relative frequency. We model the observation likelihood $p(cue|c)$ and $p(cue|b)$ via distribution of cues on positive and negative samples respectively, with details provided as follows.

Learning the Distribution of Cues. In order to learn the distribution of the proposed cues, we use the training set of text segmentation task in the ICDAR 2013 robust reading

¹Let us assume the gradient orientation of an edge pixel p is θ_p . If we follow the ray along this direction or its inverse direction, we would possibly find another edge pixel q , whose gradient orientation, denoted by θ_q , is approximately opposite to p , *i.e.*, $|\theta_p - \theta_q| \approx \pi$, as edges of a character are typically closed.

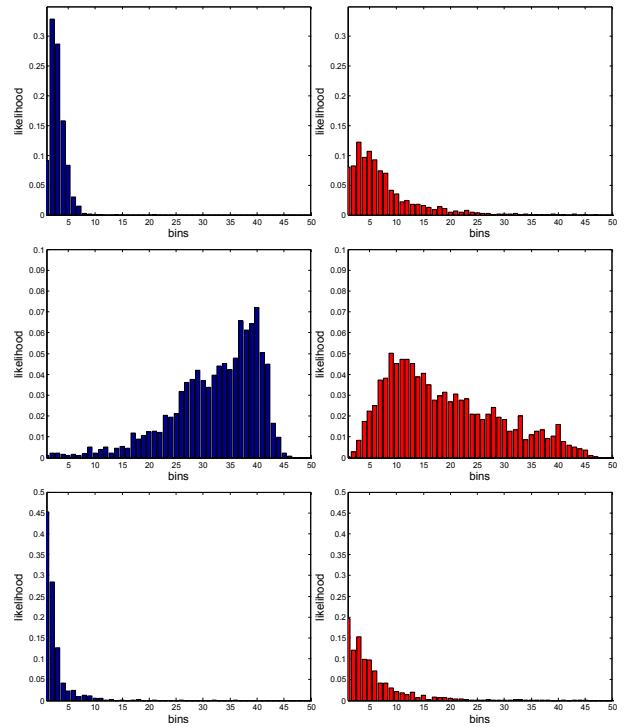


Fig. 5. Observation likelihood of characters (blue) and non-characters (red) on three characterness cues *i.e.*, SW (top row), PD (middle row), and eHOG (bottom row). Clearly, for all three cues, observation likelihoods of characters are quite different from those of non-characters, indicating that the proposed cues are effective in distinguishing them. Notice that 50 bins are adopted.

competition (challenge 2). To our knowledge, this is the only benchmark dataset with pixel-level ground truth so far. This dataset contains 229 images harvested from natural scenes. We randomly selected 119 images as the training set, while the rest 100 images were treated as the test set for characterness evaluation in our experiment (Sec. IV-B).

- To learn the distribution of cues on positive samples, we directly compute the three cues on characters, as pixel-level ground truth is provided.
- To learn the distribution of cues on negative samples, eMSER algorithm is applied twice to each training image. After erasing ground truth characters, the rest of the extracted regions are considered as negative samples on which we compute the three cues.

Fig. 5 shows the distribution of the three cues via normalized histograms. As it shows, for both SW and eHOG, compared with non-characters, characters are more likely to have relatively smaller values (almost within the first 5 bins). For the distribution of PD, it is clear that characters tend to have higher contrast than that of non-characters.

V. LABELING AND GROUPING

A. Character labeling

1) *Labeling model overview*: We cast the task of separating characters from non-characters as a binary labeling problem. To be precise, we construct a standard graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_i\}$ is the vertex set corresponding to the candidate characters, and $\mathcal{E} = \{e_j\}$ is the edge set corresponding

to the interaction between vertexes.² Each $v_i \in \mathcal{V}$ should be labeled as either character, *i.e.*, $l_i = 1$, or non-character, *i.e.*, $l_i = 0$. Therefore, a labeling set $\mathcal{L} = \{l_i\}$ represents the separation of characters from non-characters. The optimal labeling \mathcal{L}^* can be found by minimizing an energy function:

$$\mathcal{L}^* = \arg \min_{\mathcal{L}} E(\mathcal{L}), \quad (4)$$

where $E(\mathcal{L})$ consists of the sum of two potentials:

$$E(\mathcal{L}) = U(\mathcal{L}) + V(\mathcal{L}) \quad (5)$$

$$U(\mathcal{L}) = \sum_i u_i(l_i) \quad (6)$$

$$V(\mathcal{L}) = \sum_{ij \in \mathcal{E}} v_{ij}(l_i, l_j), \quad (7)$$

where $u_i(l_i)$ is the unary potential which determines the cost of assigning the label l_i to v_i . $v_{ij}(l_i, l_j)$ is the pairwise potential which reflects the cost of assigning different labels to v_i and v_j . This model is widely adopted in image segmentation algorithms [59], [60]. The optimal \mathcal{L}^* can be found efficiently using graph-cuts [32] if the pairwise potential is submodular.

2) *The design of unary potential:* characterness score of extracted regions is encoded in the design of unary potential directly:

$$u_i(l_i) = \begin{cases} p(c|\Omega) & l_i = 0 \\ 1 - p(c|\Omega) & l_i = 1. \end{cases} \quad (8)$$

3) *The design of pairwise potential:* As characters typically appear in homogeneous groups, the degree to which properties of a putative character (stroke width and color, for example) match those of its neighbors is an important indicator. This clue plays an important role for human vision to distinguish characters from cluttered background and can be exploited to design the pairwise potential. In this sense, similarity between extracted regions is measured by the following two cues.

Stroke Width Divergence (SWD). To measure the stroke width divergence between two extracted regions r_1 and r_2 , we leverage on stroke width histogram. In contrast with Algorithm 2 where only pixels on the skeleton are taken into account, distance transform is applied to all pixels within the region to find length of shortest path. Therefore, the stroke width histogram is defined as the histogram of shortest length. Then, SWD is measured as the discrete KLD (*c.f.* Equ.2) of two stroke width histograms.

Color Divergence (CD). The color divergence of two regions is the distance of their average color (in the LAB space) measured by L2 norm.

The aforementioned two cues measure divergence between two regions from two distinct perspectives. Here, we combine them efficiently to produce the unified divergence (UD):

$$UD(r_1, r_2) = \beta SWD(r_1, r_2) + (1 - \beta) CD(r_1, r_2), \quad (9)$$

where the coefficient β specifies the relative weighting of the two divergence. Without losing generality, in our experiments

²In our work, we consider the edge between two vertexes (regions) exists only if the Euclidean distance between their centroids is smaller than the minimum of their characteristic scales. Characteristic scale is defined as the sum of the length of major axis and minor axis [45].

we set $\beta = 0.5$ so that the two divergence are equally weighted. We make use of the unified divergence to define the pairwise potential as:

$$v_{ij}(l_i, l_j) = [l_i \neq l_j](1 - \tanh(UD(r_i, r_j))), \quad (10)$$

where $[\cdot]$ is the Iverson bracket. In other words, the more similar the color and stroke width of the two vertexes are, the less likely they are assigned with different labels.

B. Text line formulation

The goal of this step, given a set of characters identified in the previous step, is to group them into readable *lines* of text. A comparable step is carried out in some existing text detection approaches [35], [45], but the fact that these methods have many parameters which must be tuned to adapt to individual data means that the adaptability of these methods to various data sets remains unclear. We thus introduce a mean shift based clustering scheme for text line extraction.

Two features exploited in mean shift clustering are characteristic scale and major orientation [45]. Note that both features are normalized. Clusters with at least two elements are retained for further processing.

Within each cluster a bottom-up grouping method is performed, with the goal that only characters within the same line of text will be assigned the same label. In order to achieve this goal all regions are set as unlabeled initially. For an unlabeled region, if another unlabeled region is nearby (less than the average of their skeleton length), both are given the same label and the angle of the line connecting their centroids is taken as the text line angle. On the other hand, for a labeled region, if another unlabeled region is nearby and the angle between them is similar to that of the text line (less than 30 degrees), the latter is assigned the label of the former, and the angle of the text line is updated.

VI. PROPOSED CHARACTERNESS MODEL EVALUATION

To demonstrate the effectiveness of the proposed characterness model, we follow the evaluation of salient object detection algorithm. Our characterness map is normalized to [0,1], thus treated as saliency map. Pixels with high saliency value (*i.e.*, intensity) are likely to belong to salient objects (characters in our scenario) which catch human attention.

We qualitatively and quantitatively compare the proposed ‘characterness’ approach with ten existing saliency detection models: the classical Itti’s model (IT) [1], the spectral residual approach (SR) [4], the frequency-tuned approach (FT) [7], context-aware saliency (CA) [17], Zhai’s method (LC) [6], histogram-based saliency (HC) [10], region-based saliency (RC) [10], Jiang’s method (CB) [16], Rahtu’s method (RA) [19] and more recently low-rank matrix decomposition (LR) [15]. Note that CB, RC and CA are considered as the best salient object detection models in the benchmark work [61]. For SR and LC, we use the implementation from [10]. For the rest approaches, we use the publicly available implementations from the original authors. To the best of our knowledge, we are the first to evaluate the state-of-the-art saliency detection models for reflecting characterness in this large quantity.

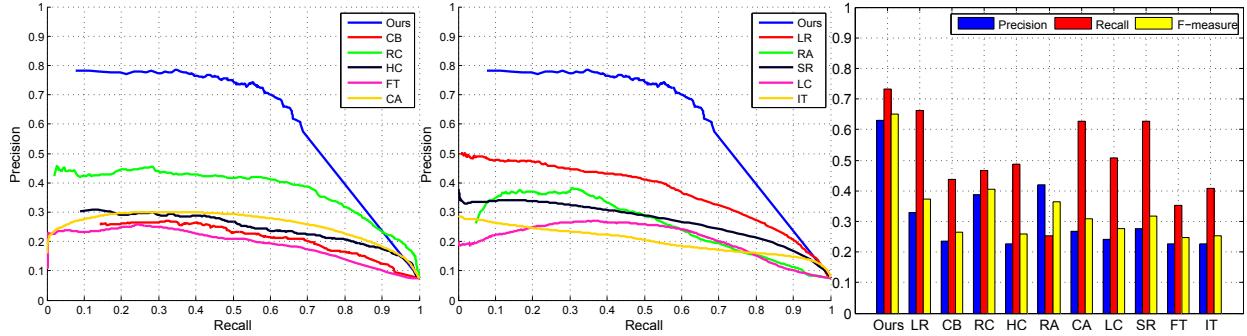


Fig. 6. Quantitative precision-recall curves (left, middle) and F-measure (right) performance of all the eleven approaches. Clearly, our approach achieves significant improvement compared with state-of-the-art saliency detection models for the measurement of ‘characterness’.

TABLE I
QUANTITATIVE PERFORMANCE OF ALL THE ELEVEN APPROACHES IN VOC OVERLAP SCORES.

	Ours	LR [15]	CB [16]	RC [10]	HC [10]	RA [19]	CA [17]	LC [6]	SR [4]	FT [7]	IT [1]
0.5143	0.2766	0.1667	0.2717	0.2032	0.1854	0.2179	0.2112	0.2242	0.1739	0.1556	

Unless otherwise specified, three parameters in Algorithm 1 were set as follows: the delta value (Δ) in the MSER was set to 10, and the local window radius in the guided filter was set to 1, $\gamma = 0.5$. We empirically found that these parameters work well for different datasets.

A. Datasets

For the sake of more precise evaluation of ‘characterness’, we need pixel-level ground truth of characters.³ As mentioned in Sec. IV-B, to date, the only benchmark dataset with pixel-level ground truth is the training set of text segmentation task in the ICDAR 2013 robust reading competition (challenge 2) which consists of 229 images. Therefore, we randomly selected 100 images of this dataset here for evaluation (the other 119 images have been used for learning the distribution of cues in the Bayesian framework).

B. Evaluation criteria

For a given saliency map, three criteria are adopted to evaluate the quantitative performance of different approaches: precision-recall (PR) curve, F-measure and VOC overlap score. In all three cases we generate a binary segmentation mask of the saliency map at a threshold T .

To obtain the PR curve, we first get 256 binary segmentation masks from the saliency map using threshold T ranging from 0 to 255, as in [7], [10], [14], [15]. For each segmentation mask, precision and recall rate are obtained by comparing it with the ground truth mask. Therefore, in total 256 pairs of precision and recall rates are utilized to plot the PR curve.

In contrast with the computation of the PR curve, to get F-measure [7], T is fixed as the twice of the mean saliency value of the image to get precision rate P and recall rate R . F-measure is computed as $((\beta^2 + 1)P \cdot R)/(\beta^2 P + R)$. We set $\beta^2 = 0.3$ as that in [7].

The VOC overlap score [62] is defined as $\frac{|S \cap S'|}{|S \cup S'|}$. Here, S is the ground truth mask, and S' is the our segmentation

³Dataset of pixel-level ground truth is also adopted in [26]. However, it is not publicly available.

mask obtained by binarizing the saliency map using the same adaptive threshold T during the calculation of F-measure.

The resultant PR curve (resp. F-measure, VOC overlap score) of a dataset is generated by averaging PR curves (resp. F-measure, VOC overlap score) of all images in the dataset.

C. Comparison with other approaches

As it shows in Fig. 6, in terms of the PR curve, all existing saliency detection models, including three best saliency detection models in [61] achieve low precision rate (below 0.5) in most cases when the recall rate is fixed. However, the proposed characterness model produces significantly better results, indicating that our model is more suitable for the measurement of characterness. The straight line segment of our PR curve (when recall rates ranging from 0.67 to 1) is attributed to the fact only foreground regions extracted by eMSER are considered as character candidates, thus background regions always have a zero saliency value. It can also be observed from the PR curve that in our scenario, two best existing saliency detection models are RC and LR.

Precision, recall and F-measure computed via adaptive threshold are illustrated in the right sub-figure of Fig. 6. Our result significantly outperforms other saliency detection models in all three criteria, which indicates that our approach consistently produces results closer to ground truth.

Table I illustrates the performance of all approaches measured by VOC overlap score. As it shows, our result is almost twice that of the best saliency detection model LR on this task.

Fig. 7 shows some saliency maps of all approaches. It is observed that our approach has obtained visually more feasible results than other approaches have: it usually gives high saliency values to characters while suppressing non-characters, whereas the state-of-the-art saliency detection models may be attracted by other objects in the natural scene (e.g., sign boards are also considered as salient objects in CB).

In summary, both quantitative and qualitative evaluation demonstrate that the proposed characterness model significantly outperforms saliency detection approaches on this task.

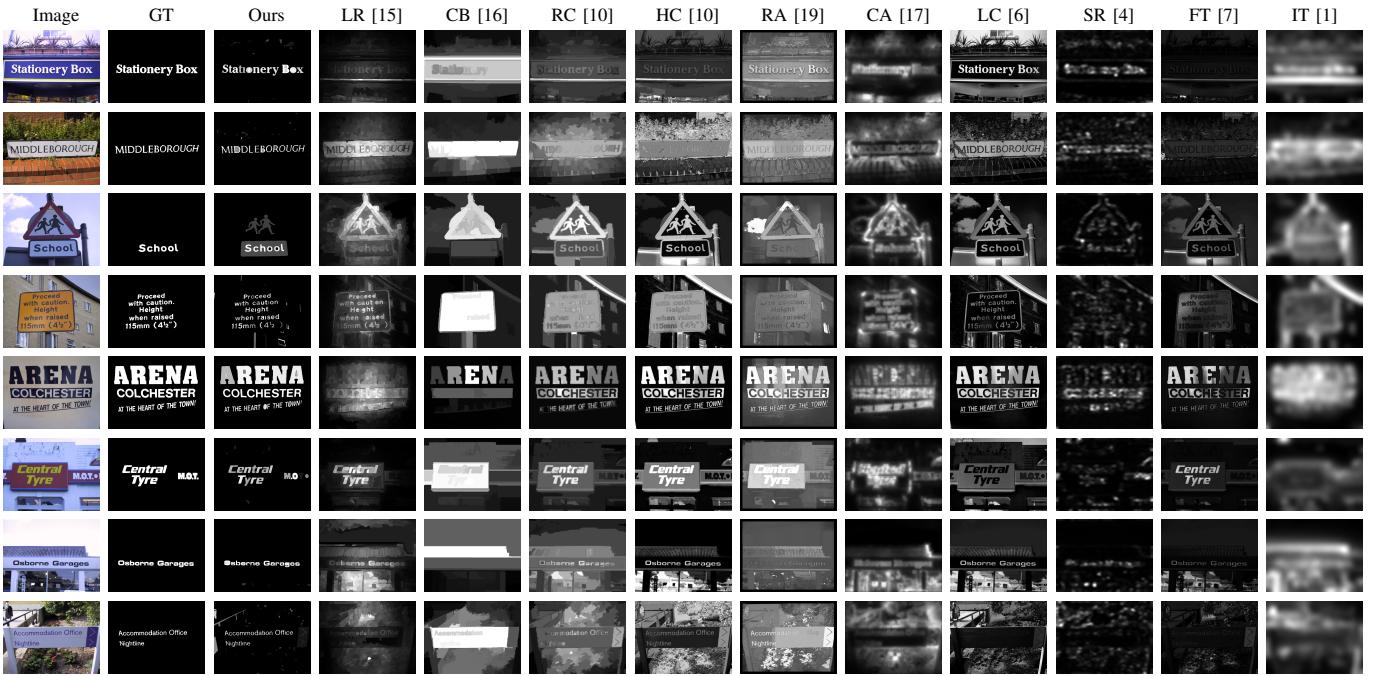


Fig. 7. Visual comparison of saliency maps. Clearly, the proposed method highlights characters as salient regions whereas state-of-the-art saliency detection algorithms may be attracted by other stuff in the scene.

VII. PROPOSED SCENE TEXT DETECTION APPROACH EVALUATION

In this section, we evaluate our scene text detection approach as a whole. Same as previous work on scene text detection, we use the detected bounding boxes to evaluate performance and compare with state-of-the-art approaches. Compared with Sec. VI in which only 119 images are utilized to learn the distribution of cues, all images with pixel-level ground truth (229 images) are adopted here, thus the distribution is closer to the real scene.

From the large body of work on scene text detection, we compare our result with some state-of-the-art approaches, including TD method [45], Epshtein's method [35], Li's method [40], [41], Yi's method [38], Meng's method [27], Neumann's method [42], [43], Zhang's method [36] and some approaches presented in the ICDAR competitions. Note that the bandwidth of mean shift clustering in the text line formulation step was set to 2.2 in all experiments.

A. Datasets

We have conducted comprehensive experimental evaluation on three publicly available datasets. Two of them are from the benchmark ICDAR robust reading competition held in different years, namely ICDAR 2003 [63] and ICDAR 2011 [64]. ICDAR 2003 dataset contains 258 images for training and 251 images for testing. This dataset was also adopted in the ICDAR 2005 [65] competition. ICDAR 2011 dataset contains two folds of data, one for training with 229 images, and the other one for testing with 255 images. To evaluate the effectiveness of the proposed algorithm on text in arbitrary orientations, we also adopted the Oriented Scene Text Database (OSTD) [38]

in our experiments. The dataset set contains 89 images with text lies in in arbitrary orientations.

B. Evaluation criteria

According to literature review, precision, recall and f-measure are the most popularly adopted criteria used to evaluate scene text detection approaches. However, definition of the three criteria are slightly different across datasets.

In the ICDAR 2003 and 2005 competition, precision and recall are computed by finding the best match between each detected bounding boxes $|D|$ and each ground truth $|G|$. In this sense, only one-to-one matches are taken into account. To overcome this unfavorable fact, ICDAR 2011 competition adopts the DetEval software [66] which supports one-to-one matches, one-to-many matches and many-to-one matches. For the OSTD dataset, we use the original definition of precision and recall from the authors [38], which are based on computing the size of overlapping areas between $|D|$ and $|G|$. In all three datasets, f-measure is always defined as the harmonic mean of precision and recall.

C. eMSER versus MSER

Since the proposed characterness cues are computed on regions, the extraction of informative regions is a prerequisite for the robustness of our approach. To demonstrate that the modified eMSER algorithm improves the performance, we compare it with the original MSER algorithm on the ICDAR 2011 dataset. For fair comparison, when learning the distribution of cues on negative samples, we use MSER rather than eMSER to harvest negative samples and then compute the three cues. Other parts of our approach remain fixed.

TABLE II
EVALUATION OF CHARACTERNESS CUES ON THE ICDAR 2011 DATASET.

Cues	precision	recall	f-measure
SW	0.71	0.63	0.67
PD	0.64	0.63	0.63
eHOG	0.58	0.65	0.61
SW+PD	0.78	0.63	0.68
SW+eHOG	0.74	0.63	0.68
PD+eHOG	0.73	0.63	0.67
SW+PD+eHOG	0.80	0.62	0.70
Baseline	0.53	0.67	0.59

Using the MSER algorithm achieves a recall of 66%, a precision of 67% and an f-measure of 66%. In comparison, when the eMSER is adopted, the precision rate is boosted significantly (80%), leading to an improved f-measure (70%). This is owing to that eMSER is capable of preserving shape of regions, whereas regions extracted by MSER are more likely to be blurred which makes cues less effective.

D. Evaluation of characterness cues

The proposed characterness cues (*i.e.* SW, PD and eHOG) are critical to the characterness model and the final text detection result. In order to show that they are effective in distinguishing characters and non-characters, we evaluate different combinations of the cues on the ICDAR 2011 dataset. Table II shows the evaluation via precision, recall and f-measure. Clearly, the table shows an upward trend in performance with increasing number of cues. Note that the baseline method in Table II corresponds to the result obtained by directly performing text line formulation after candidate region extraction.

As it shows in Table II, the performance of the proposed approach is generally poorer when only one cue is adopted. However, the f-measures are still much higher than the baseline method, which indicates that individual cues are effective. We also notice that the SW cue shows the best f-measure when individual cue is considered. This indicates that characters and non-characters are much easier to be separated by using the SW cue. From Table II, we can easily conclude that the order of discriminability of individual cues (from high to low) is: SW, PD and eHOG.

The performance of the proposed approach is boosted by a large extent (about 5% in f-measure) when two cues are combined, which attributes to the significant increase in precision.

As expected, the best performance is achieved when all cues are combined. Although there is a slightly drop in recall rate, precision rate (80%) is significantly higher than all other combinations, thus the f-measure is the best.

E. Comparison with other approaches

Table III and Table IV show the performance of our approach on two benchmark datasets (*i.e.* ICDAR 2003 and 2011), along with the performance of other state-of-the-art scene text detection algorithms. Note that methods without reference correspond to those presented in each competition.

TABLE III
RESULTS ON ICDAR 2003 DATASET.

method	precision	recall	f-measure
Ours	0.79	0.64	0.71
Kim [47]	0.78	0.65	0.71
TD-Mixture [45]	0.69	0.66	0.67
Yi [46]	0.73	0.67	0.66
Epshtain [35]	0.73	0.60	0.66
Li [41]	0.62	0.65	0.63
Yi [38]	0.71	0.62	0.62
Becker [65]	0.62	0.67	0.62
Meng [27]	0.66	0.57	0.61
Li [40]	0.59	0.59	0.59
Chen [65]	0.60	0.60	0.58
Neumann [42]	0.59	0.55	0.57
Zhang [36]	0.67	0.46	0.55
Ashida	0.55	0.46	0.50

TABLE IV
RESULTS ON ICDAR 2011 DATASET.

method	precision	recall	f-measure
Kim [47]	0.81	0.69	0.75
Ours	0.80	0.62	0.70
Neumann [43]	0.73	0.65	0.69
Li [41]	0.63	0.68	0.65
Yi	0.67	0.58	0.62
TH-TextLoc	0.67	0.58	0.62
Li [40]	0.59	0.62	0.61
Neumann	0.69	0.53	0.60
TDM_IACS	0.64	0.54	0.58
LIP6-Retin	0.63	0.50	0.56
KAIST AIPR	0.60	0.45	0.51
ECNU-CCG	0.35	0.38	0.37
Text Hunter	0.50	0.26	0.34

On the ICDAR 2003 dataset, our method achieves significantly better precision (79%) than all other approaches. Besides, our recall rate (64%) is above the average, thus our f-measure (71%) is superior than others. Although supervised learning (random forest) is adopted in TD-Mixture [45], its precision (69%) is much lower than ours (79%), which indicates the strong discriminability of the Bayesian classifier which is based on fusion of characterness cues.

On the ICDAR 2011 dataset, our method achieves a precision of 80%, a recall of 62%, and an f-measure of 70%. In terms of precision, our rate (80%) is only 1% lower than that of Kim's method [47] (81%) which is based on two times of supervised learning. Besides, we report the best performance amongst all region-based approaches.

Our method achieves a precision of 72%, a recall of 60% and an f-measure of 61% on the OSTD dataset [38] which outperforms Yi's method [38], with an improvement of 6% in f-measure.

Fig. 8 shows some sample outputs of our method with detected text bounded by yellow rectangles. As it shows, our method can handle several text variations, including color, orientation and size. The proposed method also works well in a wide range of challenging conditions, such as strong light,



Fig. 8. Sample outputs of our method on the ICDAR datasets (top two rows) and OSTD dataset (bottom two rows). Detected text are in yellow rectangles.

cluttered scenes, flexible surfaces and so forth.

In terms of failure cases (see Fig. 9), there are two culprits of false negatives. Firstly, the candidate region extraction step misses some characters with very low resolution. Furthermore, some characters in uncommon fonts are likely to have low characterness score, thus likely to be labeled as non-characters in the character labeling model. This problem may be solved by enlarging the training sets to get more accurate distribution of characterness cues. On the other hand, most false positives stem from non-characters whose distribution of cues is similar to that of characters.

VIII. CONCLUSION

In this work, we have proposed a scene text detection approach based on measuring ‘characterness’. The proposed characterness model reflects the probability of extracted regions belonging to character, which is constructed via fusion of novel characterness cues in the Bayesian framework. We have demonstrated that this model significantly outperforms the state-of-the-art saliency detection approaches in the task of measuring the ‘characterness’ of text. In the character labeling model, by constructing a standard graph, not only characterness score of individual regions is considered, similarity between regions is also adopted as the pairwise potential. Compared with state-of-the-art scene text detection approaches, we have shown that our method is able to achieve more accurate and robust results of scene text detection.

ACKNOWLEDGMENT

This work is in part supported by ARC Grants FT120100969 and LP130100156; the UTS FEIT Industry and Innovation Project Scheme.

REFERENCES

- [1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [2] N. D. B. Bruce and J. K. Tsotsos, “Saliency based on information maximization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2005.
- [3] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.
- [4] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2007, pp. 1–8.
- [5] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2009, pp. 2106–2113.
- [6] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *ACM Multimedia*, 2006, pp. 815–824.
- [7] R. Achanta, S. Hemami, F. Estrada, and S. Sussstrunk, “Frequency-tuned salient region detection,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2009, pp. 1597–1604.
- [8] Y. Wei, F. Wen, W. Zhu, and J. Sun, “Geodesic saliency using background priors,” in *Proc. Eur. Conf. Comp. Vis.*, 2012, pp. 29–42.
- [9] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, “Salient object detection by composition,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 1028–1035.
- [10] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “Global contrast based salient region detection,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2011, pp. 409–416.
- [11] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011.
- [12] D. A. Klein and S. Frintrop, “Center-surround divergence of feature statistics for salient object detection,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 2214–2219.
- [13] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object?” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2010, pp. 73–80.

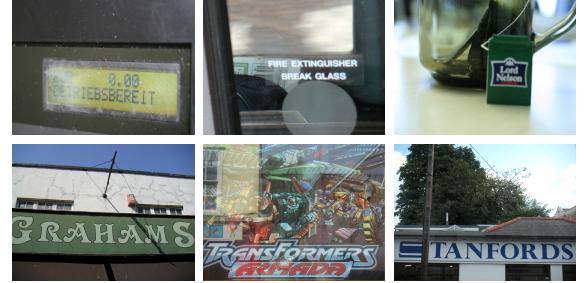


Fig. 9. False negatives of our approach. Clearly, we show two types of characters that our approach fails to detect: (i) characters in extremely blur and low resolutions (top row), (ii) characters in uncommon fonts (bottom row).

- [14] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 733–740.
- [15] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 853–860.
- [16] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Mach. Vis. Conf.*, vol. 3, no. 4, 2011, p. 7.
- [17] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2010, pp. 2376–2383.
- [18] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 914–921.
- [19] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comp. Vis.*, 2010, pp. 366–379.
- [20] Y. Ding, J. Xiao, and J. Yu, "Importance filtering for image retargeting," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2011, pp. 89–96.
- [21] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 1511–1518.
- [22] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 3506–3513.
- [23] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 105–112.
- [24] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, 2009.
- [25] Q. Sun, Y. Lu, and S. Sun, "A visual attention based approach to text extraction," in *Proc. IEEE Int. Conf. Patt. Recogn.*, 2010, pp. 3991–3995.
- [26] A. Shahab, F. Shafait, A. Dengel, and S. Uchida, "How salient is scene text?" in *Proc. IEEE Int. Workshop. Doc. Anal. Syst.*, 2012, pp. 317–321.
- [27] Q. Meng and Y. Song, "Text detection in natural scenes with salient region," in *Proc. IEEE Int. Workshop. Doc. Anal. Syst.*, 2012, pp. 384–388.
- [28] S. Uchida, Y. Shigeyoshi, Y. Kunishige, and Y. Feng, "A keypoint-based approach toward scenery character detection," in *Proc. IEEE Int. Conf. Doc. Anal. and Recogn.*, 2011, pp. 819–823.
- [29] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from an image," *Int. J. Comp. Vis.*, vol. 91, no. 3, pp. 328–346, 2011.
- [30] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. D. Bourdev, and J. Malik, "Semantic segmentation using regions and parts," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012.
- [31] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 3282–3289.
- [32] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.
- [33] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. L. Yuille, and C. Koch, "Adaboost for text detection in natural scene," in *Proc. IEEE Int. Conf. Doc. Anal. and Recogn.*, 2011, pp. 429–434.
- [34] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2004, pp. 366–373.
- [35] B. Epshtain, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2010, pp. 2963–2970.
- [36] J. Zhang and R. Kasturi, "Text detection using edge gradient and graph spectrum," in *Proc. IEEE Int. Conf. Patt. Recogn.*, 2010, pp. 3979–3982.
- [37] ——, "Character energy and link energy-based text extraction in scene images," in *Proc. Asian Conf. Comp. Vis.*, 2010, pp. 308–320.
- [38] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Proc.*, vol. 20, no. 9, pp. 2594–2605, 2011.
- [39] 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.
- [40] Y. Li and H. Lu, "Scene text detection via stroke width," in *Proc. IEEE Int. Conf. Patt. Recogn.*, 2012, pp. 681–684.
- [41] Y. Li, C. Shen, W. Jia, and A. van den Hengel, "Leveraging surrounding context for scene text detection," in *Proc. IEEE Int. Conf. Image Process.*, 2013.
- [42] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. Comp. Vis.*, 2010, pp. 770–783.
- [43] ——, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 3538–3545.
- [44] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. Image Proc.*, vol. 20, no. 3, pp. 800–813, 2011.
- [45] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 1083–1090.
- [46] C. Yi and Y. Tian, "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification," *IEEE Trans. Image Proc.*, vol. 21, no. 9, pp. 4256–4268, 2012.
- [47] H. Koo and D. Kim, "Scene text detection via connected component clustering and non-text filtering," *IEEE Trans. Image Proc.*, vol. 22, no. 6, pp. 2296–2305, 2013.
- [48] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. Brit. Mach. Vis. Conf.*, 2002, pp. 384–393.
- [49] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (mser) tracking," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2006, pp. 553–560.
- [50] P.-E. Forssén and D. G. Lowe, "Shape descriptors for maximally stable extremal regions," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2007, pp. 1–8.
- [51] L. Neumann and J. Matas, "Text localization in real-world images using efficiently pruned exhaustive search," in *Proc. IEEE Int. Conf. Doc. Anal. and Recogn.*, 2011, pp. 687–691.
- [52] S. Tsai, V. Parameswaran, J. Berclaz, R. Vedantham, R. Grzeszczuk, and B. Girod, "Design of a text detection system via hypothesis generation and verification," in *Proc. Asian Conf. Comp. Vis.*, 2012.
- [53] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. V. Gool, "A comparison of affine region detectors," *Int. J. Comp. Vis.*, vol. 65, no. 1-2, pp. 43–72, 2005.
- [54] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. Eur. Conf. Comp. Vis.*, 2010, pp. 1–14.
- [55] N. B. Ali Mosleh and A. B. Hamza, "Image text detection using a bandlet-based edge detector and stroke width transform," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–12.
- [56] J. Pan, Y. Chen, B. Anderson, P. Berkin, and T. Kanade, "Effectively leveraging visual context to detect texts in natural scenes," in *Proc. Asian Conf. Comp. Vis.*, 2012.
- [57] Z. Liu and S. Sarkar, "Robust outdoor text detection using text intensity and shape features," in *Proc. IEEE Int. Conf. Patt. Recogn.*, 2008, pp. 1–4.
- [58] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2005, pp. 886–893.
- [59] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2001, pp. 105–112.
- [60] D. Küttel and V. Ferrari, "Figure-ground segmentation by transferring window masks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 558–565.
- [61] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. Eur. Conf. Comp. Vis.*, 2012, pp. 414–429.
- [62] A. Rosenfeld and D. Weinshall, "Extracting foreground masks towards object recognition," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2011, pp. 1371–1378.
- [63] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," in *Proc. IEEE Int. Conf. Doc. Anal. and Recogn.*, 2003, pp. 682–687.
- [64] 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. and Recogn.*, 2011, pp. 1491–1496.
- [65] S. M. Lucas, "Text locating competition results," in *Proc. IEEE Int. Conf. Doc. Anal. and Recogn.*, 2005, pp. 80–85.
- [66] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *Int. J. Doc. Anal. Recogn.*, vol. 8, no. 4, pp. 280–296, 2006.