LEVERAGING SURROUNDING CONTEXT FOR SCENE TEXT DETECTION

Yao Li¹, Chunhua Shen¹, Wenjing Jia², Anton van den Hengel¹

¹The University of Adelaide, Australia  ²University of Technology, Sydney, Australia

ABSTRACT
Finding text in natural images has been a challenging task in vision. At the core of state-of-the-art scene text detection algorithms are a set of text-specific features within extracted regions. In this paper, we attempt to solve this problem from a different prospective. We show that characters and non-character interferences are separable by leveraging the surrounding context. Surrounding context, in our work, is composed of two components which are computed in an information-theoretic fashion. Minimization of an energy cost function yields a binary label for each region, which indicates the category it belongs to. The proposed algorithm is fast, discriminative and tolerant to character variations and involves minimal parameter tuning.

Index Terms—Surrounding context, energy minimization, scene text detection

1. INTRODUCTION
As a rule of thumb, human beings can detect text in natural scene with surprising speed and accuracy before performing recognition. Stimulating such an ability in machine vision is a crucial prerequisite for many tasks such as content-based image analysis, image retrieval and robot navigation. Though scene text detection has received rapid growth of concern and some breakthroughs have emerged in recent decades, it still remains a challenging task: the performance of state-of-the-art algorithms are well below human performance.

Previous approaches to scene text detection can be broadly classified into three categories: texture-based, connected component (CC)-based and hybrid.

Texture-based approaches [1, 2] perform text detection from a top-down scheme, which includes feature extraction, classification and multi-scale merging. In contrast, CC-based approaches [3–7] are actually bottom-up, whose main steps are candidate extraction, verification and CC grouping. More recently, hybrid approaches [8–10] are enjoying more popularity. Intuitively, they are the combination of the two aforementioned kinds. The approach we propose here can be considered as a CC-based kind, in as much as no supervised classifiers are utilized. Though heavy training is not included, we still achieve the state-of-the-art performance. In addition, for both region-based and hybrid approaches (based on training), the expandability becomes the bottleneck in generalized scene text detection tasks, however, our algorithm survives from the problem.

One of the crucial sub-problems of scene text detection is how to remove non-character interferences as much as possible. To address this problem, texture-based and hybrid approaches extract features to train a classifier (e.g. SVM, AdaBoost). On the other hand, CC-based approaches rely on hand-crafted heuristic rules or geometric constrains. While most of the work in scene text detection has examined text-specific features within extracted regions, little work has looked at whether it is possible to distinguish characters from interferences outside those regions.

Our work differs from the previous works mentioned above mainly because we focus on leveraging surrounding context. Surrounding context, in our definition, refers to unary and pairwise features in the surrounding background of those regions, rather than the region themselves. We not only consider surrounding context of a particular region (unary), interaction between regions are also taken into account (pairwise). To leverage the surrounding context, we propose an energy minimization framework. Characters and non-character interferences are separated by minimizing an energy cost function designed on the surrounding context. Therefore, the two main contributions of our work are:

1. Novel surrounding context for scene text detection. In contrast to the general text-specific features which are extracted within a region, we propose to depict the surrounding context of a region and use that for text detection. To the best of our knowledge, we are the first to report this finding.

2. An energy minimization formulation for labelling. To reject non-character interferences, most previous works rely on either hand-designed heuristic rules or classifiers. In this work, however, we address the problem from an energy minimization perspective.

The rest of this paper is organized as follows. In Sec. 2 we introduce an edge-preserving MSER preprocessing step for extracting candidate regions. We give details of surrounding context and energy formulation in Sec. 3 and 4 respectively, followed by region grouping in Sec. 5. Comprehensive experiments are presented in Sec. 6. Finally, we conclude the paper in Sec. 7.
2. EDGE-PRESERVING REGION EXTRACTION

Maximally Stable Extremal Region (MSER) [11] has been shown suitable for extracting perceptually homogeneous characters in the literature [8, 3, 12]. However, as pointed out in existing work [3, 10, 12], MSER is sensitive to blur, which may lead to incorrect character boundaries. To address this problem, Chen et al. [3] pruned out MSER pixels which are located outside the boundary detected by Canny edge detector. Tsai et al. [12] performed judicious parameter selection and multi-scale analysis of MSERs. Neumann and Matas [10] extended MSER to Extremal Region (ER). In our work, we incorporate the gradient information of images so as to produce edge-preserving MSERs.

Edge-preserving MSER (eMSER). Given a grayscale image \( I \) smoothed by guided filter [13], a new image \( I^* \) is computed based on its gradient amplitude, denoted by \( \nabla I \):

\[
I^* = I \pm \gamma \nabla I, \tag{1}
\]

where ‘+’ is for detecting dark characters on bright background, whereas ‘−’ is chosen to detect bright characters in dark background, and the factor \( \gamma \) weighs the contribution of the image gradient versus the original grayscale value. Note that both \( I \) and \( \nabla I \) are in the range of \([0, 255]\). Then, we perform MSER detector on the resultant image \( I^* \). We obtain a set of candidate regions \( \mathcal{X} \) after this procedure, which contain both characters and non-character interferences.

3. SURROUNDING CONTEXT

From our observation, we found two important properties of the surrounding background that characters have (Fig. 1):

1. They tend to be perceptually homogeneous.
2. They are similar for neighbouring characters in the same word.

Intuitively, these two findings can be utilized to discriminate between characters and non-character interferences. Specifically, we propose two novel surrounding contexts, i.e., unary surrounding context and pairwise surrounding context, to depict these two properties respectively. Both the unary and pairwise surrounding contexts are defined in an information-theoretic fashion.

3.1. Unary surrounding context (USC)

The unary surrounding context, in our work, is defined as the perceptual homogeneity in the surrounding background of a region.

Given a candidate region \( x \ (x \in \mathcal{X}) \), we define its surrounding background \( S_x \), as the relative complement of \( x \) in a rectangle \( R_x \) that encloses \( x \) (i.e. \( S_x = x \setminus R_x \)). To measure how perceptually homogeneous \( S_x \) is, we compute the sum of entropy of \( S_x \) with respect to color histograms:

\[
USC(x) = -\sum_{R,G,B} \sum_{i=1}^{b} h_i(S_x) \log h_i(S_x), \tag{2}
\]

where \( h_i(S_x) \) is the \( i \)th bin of the histogram of \( S_x \) in a color channel, and \( b \) denotes the number of bins (26 here). In information theory, entropy measures the average information of transmitting a signal given its probability distribution. The less predictable the occurrence of a signal is, the higher the entropy. In our scenario, for a true character, typically there is a dominating color in \( S_x \), therefore, histogram of \( S_x \) should have one sharp peak. However, for a non-character interference, its color values in \( S_x \) span the histogram as result of color variation. This corresponds to the entropy of the former is usually smaller than that of the latter. Note that since we treat the histogram in each color channel as a probability distribution function (PDF), it needs to be normalized before computing the entropy. In practice, the rectangle \( R_x \) is the bounding box of the region \( x \).

3.2. Pairwise surrounding context (PSC)

We define the pairwise surrounding context as the perceptual divergence of the surrounding backgrounds of two regions. This definition is motivated by the center-surround computation in saliency detection algorithms [14, 15]. In information theory, the Kullback-Leibler divergence (KLD) is interpreted as a measure of the dissimilarity of two distributions [16]. Here, we take advantage of its discrete form [15] to measure the dissimilarity between the histograms of the surrounding backgrounds (denoted as \( S_x \) and \( S_y \) respectively) of two re-
gions $x$ and $y$ as:
\[
PSC(x, y) = \sum_{R,G,B} \sum_{i=1}^{b} h_i(S_x) \log \frac{h_i(S_x)}{h_i(S_y)}. \tag{3}
\]
Note that the more perceptually divergent the two regions are, the higher the corresponding value $PSC(x, y)$ is.

## 4. ENERGY FORMULATION

We cast the task of separating characters from background as a labelling problem. Let $\mathcal{L}$ denotes the set of labels (e.g. $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$). Each candidate $x_i \in \mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ should be labelled as either character $l_x = 1$ or background $l_x = 0$. The optimal labelling $\mathcal{L}^*$ is found by minimizing an energy cost function that is a sum of unary and pairwise terms:
\[
E(l) = \sum_{x \in \mathcal{X}} D_x(l_x) + \sum_{(x, y) \in \mathcal{N}} V_{x,y}(l_x, l_y), \tag{4}
\]
where the first term, $D_x(l_x)$, is the data term that determines the cost of assigning the label $l_x$ to candidate $x$, the second term, $V_{x,y}(l_x, l_y)$, is the smoothness term that reflects the cost of assigning different labels to neighbouring candidates $x$ and $y$. $\mathcal{N}$ is the set of neighbouring candidate pairs. In our work, we consider two candidates $x$ and $y$ as neighbours only if the Euclidean distance between their centroids is smaller than the minimum of their characteristic scales\(^1\).

### 4.1. The design of $D_x(l_x)$

The unary surrounding context defined in Sec. 3.1 is encoded in the design of $D_x(l_x)$ using a sigmoid function as:
\[
D_x(l_x) = \frac{1}{1 + e^{-USC(x) + t}}, \tag{5}
\]
where $t$ is a constant that takes positive values.

### 4.2. The design of $V_{x,y}(l_x, l_y)$

The pairwise potential favours neighbouring regions having the same label unless their pairwise surrounding context is large. This is defined as:
\[
V_{x,y}(l_x, l_y) = [l_x \neq l_y](1 - \tanh(PSC(x, y))), \tag{6}
\]
where $[\cdot]$ is the Iverson bracket.

### 4.3. Optimization

In this work, we leverage the standard graph cuts [17] to minimize Eqn. (4). It casts the problem as a graph-partitioning one using the mincut/maxflow graph algorithm. Regions with labels $l_x = 0$ are rejected as non-character interferences.

\(^1\)Characteristic scale is defined as the sum of the length of major axis and minor axis as in [9].

## 5. REGION GROUPING

Ground truth in state-of-the-art datasets are in the form of text lines. For the sake of better evaluation of the proposed algorithm, we group remaining regions into readable lines of text. Two regions are aggregated only if they are close and their stroke width [7] is similar.

## 6. EXPERIMENTS

### 6.1. Experimental setup

**Datasets.** The proposed algorithm is evaluated on two publicly available benchmark scene text detection datasets: ICDAR 2003 and 2011 datasets [18, 19]. ICDAR 2003 dataset [18] contains 509 images, 258 of which are used for training with the rest for testing. ICDAR 2011 dataset [19] is composed of 229 training images and 255 testing images.

**Evaluation metric.** Same as most previous work, precision, recall and f-measure criteria are adopted to quantify the performance and compare with others. Note that f-measure is the harmonic mean of precision and recall.

### 6.2. Implementation details

Since we have no prior knowledge about text color, eMSER algorithm is performed twice on each test image so as to extract bright characters on dark background and vice versa. The final detection result is the combination of the two above results. Unless otherwise specified, parameters in Eqn. (1) (5) were set as follows: $\gamma = 0.5$, $t = 7.5$. We experimentally found that the above values work well for different datasets.

### 6.3. Quantitative evaluation

**ICDAR 2003 dataset [18].** Our algorithm achieves the precision of 62%, the recall of 65% and the f-measure of 63%. Table 1 shows our result along with results of other methods. Compared with existing methods, our method has achieved the state-of-the-art performance.

**ICDAR 2011 dataset [19].** Our result on this dataset is 63%
Table 2: Results on ICDAR 2011 dataset [19].

<table>
<thead>
<tr>
<th>method</th>
<th>precision</th>
<th>recall</th>
<th>f-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim [23]</td>
<td>0.81</td>
<td>0.69</td>
<td>0.75</td>
</tr>
<tr>
<td>Yi [24]</td>
<td>0.81</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>Ours</td>
<td>0.63</td>
<td>0.68</td>
<td>0.65</td>
</tr>
<tr>
<td>Yi</td>
<td>0.67</td>
<td>0.58</td>
<td>0.62</td>
</tr>
<tr>
<td>TH-TextLoc</td>
<td>0.67</td>
<td>0.58</td>
<td>0.62</td>
</tr>
<tr>
<td>Li [7]</td>
<td>0.59</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>Neumann</td>
<td>0.69</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>TDM-JACS</td>
<td>0.64</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td>LIP6-Rev</td>
<td>0.63</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>KAIST AIPR</td>
<td>0.60</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>ECNU-CCG</td>
<td>0.35</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>Text Hunter</td>
<td>0.50</td>
<td>0.26</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table 2 shows our results along with the results of other state-of-the-art methods. Notice that methods without reference are those presented in the competition. We report the best performance amongst all the participating CC-based methods.

6.4. Qualitative evaluation

Some sample outputs of the proposed algorithm are presented in Fig. 2. The detected text regions are bounded by yellow rectangles. As it shows, our algorithm is robust to text variations including color, size, font. Also, it detects text successfully in different scenarios, such as business board, book cover, road signs, etc.

In terms of the failure cases, false positives mainly arise from non-character interferences embedded in a perceptually homogeneous surrounding background. In that circumstance, unary surrounding context of those interferences can be low which leads to wrong labels.

7. CONCLUSION

An effective method to detect text in natural scene images is presented. In contrast with previous work, the proposed approach is based on encoding surrounding context in an energy minimization framework. The novelty of this paper is twofold: surrounding context and energy minimization formulation. Surrounding context refers to the text-specific features in the surrounding background of an extracted region, which is rarely considered in existing works. The energy minimization formulation renders the unary and pairwise surrounding contexts are combined in the same framework. The proposed approach has shown promising results on benchmark scene text detection datasets.

References


