Effective Text Localization in Natural Scene Images with MSER, Geometry-based Grouping and AdaBoost

Xuwang Yin\textsuperscript{1}, Xu-Cheng Yin\textsuperscript{1} *, Hong-Wei Hao\textsuperscript{2}, and Khalid Iqbal\textsuperscript{1}

\textsuperscript{1} Department of Computer Science, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
\textsuperscript{2} Institute of Automation, Chinese Academy of Sciences, Beijing 100090, China
xuwangyin@gmail.com, xuchengyin@ustb.edu.cn, hongwei.hao@ia.ac.cn, kik.ustb@gmail.com

Abstract

Text localization in natural scene images is an important prerequisite for many content-based image analysis tasks. In this paper, we proposed a novel and effective approach to accurately localize scene texts. Firstly, Maximally stable extremal regions (MSER) are extracted as letter candidates. Secondly, after elimination of non-letter candidates by using geometric information, candidate regions are constructed by grouping similar letter candidates using disjoint set. Candidate region features based on horizontal and vertical variances, stroke width, color and geometry are extracted. An AdaBoost classifier is built from these features and text regions are identified. The overall system is evaluated on the ICDAR 2011 competition dataset and the experimental results show that the proposed algorithm yields high precision and recall compared with the latest published algorithms.

1. Introduction

Among all the content in image, text information is quite useful and can be used for many content-based image applications, such as text-based image search, license plate reading, product recognition and landmark recognition. To extract text information from natural scene images, text in images has to be robustly located and extracted. Text localization is the process of determining the location of text in the image, which is a challenge due to the complex background, different lighting conditions and wide variety of text appearance.

Existing techniques for scene text localization can be categorized into two groups: region-based and connected component (CC)-based [11]. Region-based methods use sliding windows to search for all possible text regions in an image, extract region features and use a classifier to identify text regions. Lee et al. [6] extracted six different classes of region features and used Modest AdaBoost classifier to identify text regions. While in CC-based methods, connected components in images are extracted as letter candidates and grouped into text regions if they have similar geometric properties, additional checks may be performed to remove non-text region. Epshtern et al. [3] measured stroke width for each pixel and group neighboring pixels with approximately similar stroke width into CCs. These CCs were filtered and aggregated into text regions.

In this work, we proposed a novel approach to localize scene text. Firstly, MSER are extracted as letter candidates. Some non-letter candidates are filtered out based on geometric information. An adjacency relation is defined between letter candidates and an undirected graph is derived. Candidate regions are constructed as connected components of the undirected graph using disjoint set. Region features based on horizontal and vertical variances, stroke width, color and geometry are extracted. An AdaBoost classifier is built from these features and used to identify text regions. In order to use the ICDAR 2011 [12] competition dataset, text regions are later separated into words. The result of each step is shown in Figure 1. Our method is trained on the compe-
tion training dataset and tested on the test dataset. The experimental results show the excellent performance of our method. Additionally, our algorithm runs rather fast on the competition dataset.

The main contribution of this article is the introduction of a new candidate regions construction method. The proposed scheme defines an adjacency relation constrained by different features of letter candidates and treats letter candidates of the same region as connected components of undirected graph. As connected components extraction is a classic graph algorithm, our approach is easy to understood yet very powerful.

The remainder of this paper is organized as follows. The proposed text localization algorithm is described in section 2 and the experimental results are presented in section 3. Section 4 concludes the paper.

![Figure 1. Results of each step in our text localization algorithm. (a) Original image. (b) MSER extracted as letter candidates. (c) Candidate regions with green bounding box after region construction. (d) Text regions after region classification.](image)

2. The Text localization Algorithm

The proposed algorithm is divided into three parts: letter candidates extraction, candidate regions construction and candidate regions classification.

2.1. Letter Candidates Extraction with MSER

Considering distortions in camera-based (document) images [14] and inspired by Kim’s work [12], we extract MSER from image as letter candidates. MSER was proposed by Matas et al. [9] to find correspondences between images of different viewpoints. MSER has been reported as one of the best region detector [10], and is robust against to scale and light change. After letter candidates extraction, non-letter candidates elimination is performed based on geometric information.

2.2. Candidate Regions Construction with Geometry-Based Grouping

Candidate regions can be effectively constructed from letter candidates by using the technique for determining connected components of undirected graphs.

Given letter candidates set $V$, Adjacency relation $A \subset V \times V$ is defined. Letter candidates $u, v \in V$ are adjacent if $d(u, v) \leq \epsilon$, where $d(u, v)$ denotes the distance between $u$ and $v$, and $\epsilon$ is the threshold. An undirected graph is defined $G = (V, E)$, where $V$ is the set of letter candidates, $E$ is the set of pairs of adjacent letter candidates.

Now the candidate region set $R$ is the set of connected components of $G$, which can be effectively computed using disjoint set [2]. The detailed procedure for candidate region construction is illustrated in Algorithm 1. In this algorithm, procedure $\text{make\_region}(x)$ creates a new region containing only one letter candidate $x$, $\text{find\_region}(x)$ returns the address of the region containing $x$ and $\text{region\_union}(x, y)$ unites the region containing $x$ and $y$ into a new region; original regions containing $x$ and $y$ are destroyed after the union operation. After the region construction algorithm ended, regions containing only one letter candidate are removed.

```
Algorithm 1 Candidate regions construction algorithm.
for each letter candidate $v \in V$ do
  $\text{make\_region}(v)$
end for

for each pair of letter candidates $(u, v) \in E$ do
  if $\text{find\_region}(u) \neq \text{find\_region}(v)$ then
    $\text{region\_union}(u, v)$
  end if
end for
```

The distance between two letter candidates is defined as weighted sum of distances of different features $d(u, v) = \sum_{f \in F} d_f(u, v)w_f$, where $F$ is the set of features, include letter candidates’ widths, heights, centroids, colors, stroke widths, top and bottom alignments, $d_f(u, v)$ is the distance between $u, v$ under feature $f$ and $w_f$ is the weight assigned to feature $f$. Weights of features and threshold $\epsilon$ are set empirically.

Letter candidate’s stroke width is computed by averaging over its pixels’ stroke widths. As is shown in Figure 2, the calculation of a pixel’s stroke width con-
Figure 2. Calculation of a pixel’s stroke width. Four line segments’ lengths are considered. In this case, length of line segment b is selected as the pixel stroke width as it is the minimum.

Figure 3. Horizontal and vertical variances of sub-regions. Horizontal variance is large in the central while vertical variance is large in the top and bottom. (a) Candidate region with three sub-regions. (b) Candidate letter with three sub-regions. (c) and (d) Horizontal and vertical variances of sub-regions.

Figure 4. Calculation of exterior edge pixels’ derivatives. Filters in (b) and (c) are applied to exterior edge pixels in (a) with red color. (b) Filter to calculate a pixel’s $x$ derivative. (c) Filter for the $y$ derivative.
Table 1. Performance (%) comparison of text localization algorithms.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim’s Method</td>
<td>62.47</td>
<td>82.98</td>
<td>71.28</td>
</tr>
<tr>
<td>Our Method</td>
<td>62.22</td>
<td>81.53</td>
<td>70.58</td>
</tr>
<tr>
<td>Yi’s Method</td>
<td>58.09</td>
<td>67.22</td>
<td>62.32</td>
</tr>
<tr>
<td>TH-TextLoc System</td>
<td>57.68</td>
<td>66.97</td>
<td>61.98</td>
</tr>
<tr>
<td>Neumann’s Method</td>
<td>52.54</td>
<td>68.93</td>
<td>59.63</td>
</tr>
<tr>
<td>TDM IACS</td>
<td>53.52</td>
<td>63.52</td>
<td>58.09</td>
</tr>
</tbody>
</table>

*aThis method is not published.*

The training and test dataset of our algorithm are from ICDAR 2011 Robust Reading Competition [12]. The whole dataset consisted of 485 images containing text of various fonts and colors with different backgrounds. Performances of our system and some of the published algorithms in ICDAR 2011 competition are presented in Table 1.

The proposed method shows a competitive f score of 70.58, only a little bit lower than that of Kim’s method, which is not published yet. In its very simple description [12], Kim’s method is also a CC-based method and employs MSER as letter candidates, and the features used by Kim’s method to identify text regions are gradient features from oriented gradient images.

Additionally, our system runs rather fast on the test dataset. The average processing time for an image (with an average resolution of $1000 \times 790$) on a Linux laptop with a dual-core 1.83GHz CPU is 0.59 seconds, which is much faster than most of the methods reported in ICDAR 2005 competition [7].

4. Conclusion

We have proposed a novel approach for text localization in natural scene images. The algorithm employs Maximally stable extremal regions as letter candidates and constructs candidate regions from letter candidates using disjoint set. An AdaBoost classifier is trained from candidate region features and used to identify text regions. Several important features, includes letter candidate’s stroke width, candidate region’s horizontal and vertical variances are exploited. The experimental results demonstrated the excellent performance of the proposed method.

Acknowledgments

The research was supported by National Natural Science Foundation of China (61105018, 61175020), and the R&D Special Fund for Public Welfare Industry (Meteorology) of China (GYHY201106039, GYHY201106047).

References