A Pixel-based Evaluation Method for Text Detection in Color Images

Marios Anthimopoulos, Nikolaos Vlissidis, Basilis Gatos
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, NCSR “Demokritos” GR-153 10 Agia Paraskevi, Athens,Greece
{anthimop, nvlissid, bgat}@ iit.demokritos.gr

Abstract—This paper proposes a performance evaluation method for text detection in color images. The method, contrary to previous approaches, is not based on the inexplicitly defined text bounding boxes for the evaluation of the text detection result but considers only the text pixels detected by binarizing the image and applying a color inversion if needed. Moreover, in order to gain independence from the chosen binarization algorithm, the method uses the skeleton of the binarized image. The results produced by the proposed evaluation protocol proved to be quite representative and reasonable compared to the corresponding optical result.

Keywords—performance evaluation, text detection, color images

I. INTRODUCTION

The tremendous increase of multimedia content has raised the need for automatic semantic information indexing and retrieval systems. Textual information in images and videos proves to be an important source of high-level semantics. Among all steps of a textual information extraction system, text detection is the most crucial. Although text detection has been extensively studied in the past decade presenting quite promising results, the absence of a reliable evaluation protocol deprived the researchers of the capability to optimize their algorithms and compare to the state-of-the-art methods.

II. RELATED WORK

The evaluation of a text detection system is an aspect not as trivial as it might seem. It resembles the generic problem of object detection evaluation having additionally its own issues. Most researchers use for their experimentation simple methods while very few works have focused on the specific problem of evaluation. Moreover these works propose evaluation strategies with several drawbacks or require great effort for the generation of the ground truth ([1], [2], [3]). In [2] and [3] Kasturi, Manohar et al. propose as overall measure of text detection in a frame, a box-based measure called Frame Detection Accuracy (FDA):

\[ FDA = \frac{\text{Overlap}_\text{Ratio}}{N_G + N_D} \]  \hspace{1cm} (1)

where, \( N_G \) is the number of the ground truth objects, \( N_D \) the number of detected objects and

\[ \text{Overlap}_\text{Ratio} = \frac{\sum_{i=1}^{N_{mapped}} G_i \cap D_i}{\sum_{i=1}^{N_{mapped}} |G_i \cup D_i|} \]  \hspace{1cm} (2)

\( N_{mapped} \) is the number of mapped object pairs, where the correspondence is established between objects which have the best spatial overlap. Kasturi et al. also proposed a thresholding of the overlap ratio in order to forgive minor inconsistencies between the boundaries of the system’s output and the ground truth boxes.

The evaluation methods of this kind are based on the mapping between ground truth and detected objects. Especially for the text detection problem, text lines are considered to be the objects where a text line is usually defined as an aligned series of characters with a small intermediate distance relative to their height. However, this subjectively small distance can result arbitrarily in bounding box splits or merges among annotators and detectors making the object mapping inappropriate. In addition, the number of correctly retrieved boxes is not generally a measure of the retrieved textual information since the number of characters in different boxes may vary considerably.

Wolf et al. [4] proposed the creation of match score matrices with the overlap between every possible pair of blocks, in order to evaluate document structure extraction algorithms. The benefit of this kind of algorithms is their ability to consider the possible splits or merges of the bounding boxes besides one-to-one matching. However, in order to match two ground truth boxes with one resulting box, the total overlap threshold has to be very low (~40%). This will have as a result accepting as correct, a box with size even higher than the double size of the ground truth box.

Many researchers have used the overall overlap to compute area-based recall and precision measures ([5], [6]).
Recall\textsubscript{pixel} = \frac{|G \cap D|}{|G|} \quad (3)

Precision\textsubscript{pixel} = \frac{|G \cap D|}{|D|} \quad (4)

where $G$ is the ground truth area and $D$ the detected area. However, the main drawback here similarly to the box-based approaches is the fact that the number of retrieved pixels does not correspond to proportional textual information since different text lines may contain characters of various sizes.

Anthimopoulos et al. [7] proposed a method based on the recall and precision of the area coverage, normalised by the approximation of the number of characters for every box (see Eq. 5, 6). The number of characters in a bounding box was approximated by the ratio width/height of the box, assuming that this ratio is invariable for every character, the spaces between different words in a text line are proportional to its height and each textline contains characters of the same size.

\[
\text{Precision} = \frac{\sum_i \frac{EGD_i}{hd_i^2}}{\sum_i \frac{ED_i}{hd_i^2}} \quad (5)
\]

\[
\text{Recall} = \frac{\sum_i \frac{EGD_i}{hg_i^2}}{\sum_i \frac{EG_i}{hg_i^2}} \quad (6)
\]

where $hg_i$ is the height of the $i^{th}$ ground truth bounding box, $EG_i$ is its number of pixels, $EGD_i$ is the number of pixels of the intersection that belong to $i^{th}$ ground truth bounding box, $hd_i$ is the height of the $i^{th}$ detection bounding box, $ED_i$ is its number of pixels, $EDG_i$ is the number of pixels of the intersection that belong to $i^{th}$ detection bounding box, $N$ is the number of ground truth bounding boxes and $M$ is the number of detected bounding boxes.

The previous approach, based on the estimated character number of each text line, managed to overcome some problems that box-based and overlap-based methods have faced. However, the method still depends on the subjective ground truth annotation due to the text object ambiguous definition.

In this work we strive towards a more accurate evaluation methodology considering as text only the pixels that belong to characters. In that way we overcome all the problems caused by the ambiguously defined text bounding boxes. However, distinguishing text from non-text pixels actually refers to binarizing the initial detection result and deciding between normal or inverse text. Normal is denoted any text whose characters have lower intensity values than the background while inverse text is the opposite.

### III. PROPOSED METHOD

The proposed algorithm (Fig.1) generates two binarized images for the ground truth and the resulted bounding boxes respectively. For each case, the algorithm takes as input the image and a set of bounding boxes. The pixels contained in each box are binarized and then conditionally inverted producing black pixels for text and white pixels for the background. All pixels outside boxes are set to white.

For the binarization of the images, Otsu [8] thresholding was chosen after relative experimentation. Otsu’s method proved to be a very good solution for this kind of text images although the choice of the binarization method does not affect considerably the result of the proposed evaluation because of the skeletonizing that follows.

In order to classify between normal or inverse text we firstly apply a connected component analysis. The numbers of white ($WCC$) and black ($BCC$) connected components are counted, discarding components with height less than 8 pixels or less than the 40% of the box height. If $|WCC-BCC|>1$ then the color that corresponds to the largest number of connected components is regarded as text color. Else if the distance between $WCC$ and $BCC$ is less or equal to 1, the condition for the inversion is based on the pixel values of the borders of the bounding boxes. If the majority of border pixels are black then text is considered inverse.

Then, the two binarized images have to be compared in order to compute the recall and precision rates. However, a straight comparison between the two images would produce evaluation rates strongly depended to the used binarization method. To overcome this problem and focus only on the evaluation of text detection we use the skeleton of each image computed by an iterative skeletonization method presented in [9]. Specifically, the recall and precision rates are defined by the equations (7) and (8).

\[
\text{Recall} = \frac{|\text{GT}_{\text{ske}} \cap \text{RS}_{\text{bin}}|}{|\text{GT}_{\text{ske}}|} \quad (7)
\]

\[
\text{Precision} = \frac{|\text{RS}_{\text{ske}} \cap \text{GT}_{\text{bin}}|}{|\text{RS}_{\text{ske}}|} \quad (8)
\]

where $\text{GT}_{\text{bin}}$ is the binarized ground truth image and $\text{GT}_{\text{ske}}$ is its skeleton while $\text{RS}_{\text{bin}}$ is the binarized image of the detection result and $\text{RS}_{\text{ske}}$ is its skeleton. The operator [⋯] denotes the number of text (black) pixels and $\cap$ is the intersection of the text (black) pixels.
Figure 2 shows an example of video text and two different cases of text detection results while Table 1 presents the evaluation results. For the result shown in fig. 2(b) every evaluation method based on one to one mapping, like FDA, would fail because of the merging of the two ground truth boxes in one. Moreover, the case in fig. 2(c) will produce a relatively low value, since resulting boxes are larger than the corresponding ground truth boxes. However, the actual text that will be recognized by an OCR system after detection and binarization is the same, despite the different values produced. This is why the proposed method will give the same value for both cases. Figure 3 shows the images produced for the proposed evaluation.

IV. EXPERIMENTAL RESULTS

In order to experiment with different algorithms for normal/inverse text classification we created a dataset consisting of 3048 normal and 1827 inverse text images taken from the ground truth or the results of different text detection methods. The method used in this work scored 97% in terms of accuracy. Table 2 presents the results of three different text detection algorithms on a dataset consisting of 203 video frames with a total of 2790 textlines. Table 3 displays the corresponding results using a different binarization method based on [10]. The resemblance between the two result tables indicates the proposed method’s independence from the chosen thresholding technique.

TABLE I. EVALUATION RESULTS FOR FIG.2

<table>
<thead>
<tr>
<th>%</th>
<th>FDA</th>
<th>F-measure of (3) and (4)</th>
<th>F-measure of proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig.2 (b)</td>
<td>33.3</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>Fig.2 (c)</td>
<td>67</td>
<td>74</td>
<td>99</td>
</tr>
</tbody>
</table>

TABLE II. PROPOSED EVALUATION’S RESULTS

<table>
<thead>
<tr>
<th>%</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al. [11]</td>
<td>52.4</td>
<td>50.1</td>
<td>51.7</td>
</tr>
<tr>
<td>Chen et al. [6]</td>
<td>62.4</td>
<td>73.8</td>
<td>67.6</td>
</tr>
<tr>
<td>Anthimopoulos et al. [12]</td>
<td>73.3</td>
<td>68.3</td>
<td>70.7</td>
</tr>
</tbody>
</table>

TABLE III. PROPOSED EVALUATION’S RESULTS USING A BINARIZATION METHOD BASED ON [10]

<table>
<thead>
<tr>
<th>%</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al. [11]</td>
<td>52.7</td>
<td>51.3</td>
<td>52</td>
</tr>
<tr>
<td>Chen et al. [6]</td>
<td>62.3</td>
<td>75.4</td>
<td>68.2</td>
</tr>
<tr>
<td>Anthimopoulos et al. [12]</td>
<td>74.4</td>
<td>69.3</td>
<td>71.7</td>
</tr>
</tbody>
</table>
V. CONCLUSION

In this paper we proposed a novel performance evaluation method for text detection in color images. In order to avoid the inexplicitly defined text bounding boxes, the algorithm considers as text pixels only the ones belonging to characters. The discrimination between text and background pixels in a bounding box is done by binarizing and inverting the image if needed. The algorithm that decides for the inversion proved to be very accurate after relative experimentation so that it does not affect the detection evaluation rates. Moreover, the method uses the skeleton of the binarized image, to gain independence from the chosen binarization algorithm. The results produced by the proposed evaluation protocol proved to be quite representative and reasonable compared to the corresponding optical result.

ACKNOWLEDGMENT

This research is carried out within the framework of the co-funded by the EU project CASAM (FP7-ICT-217061).

REFERENCES


Figure 3. Images produced by the proposed evaluation method. (a) The ground truth binarized image, (b) the skeleton of (a), (c) the resulted binarized image, (d) the skeleton of (c)