IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 1, 2020, accepted March 17, 2020, date of publication April 14, 2020, date of current version May 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2987869

On the Fusion of Text Detection Results: A Genetic Programming Approach

JOSE L. FLORES CAMPANA[1], ALLAN PINTO[1], (Member, IEEE), MANUEL ALBERTO CÓRDOVA NEIRA[1], LUIS GUSTAVO LORGUS DECKER[1], ANDREZA SANTOS[1], JHONATAS S. CONCEIÇÃO[1], AND RICARDO DA SILVA TORRES[2], (Member, IEEE)

[1]Institute of Computing, University of Campinas, Campinas, 13083-852, Brazil
[2]Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), 6009 Ålesund, Norway

Corresponding author: Ricardo Da Silva Torres (ricardo.torres@ntnu.no)

This work was supported in part by the CNPq under Grant #307560/2016-3, in part by the São Paulo Research Foundation–FAPESP under Grant #2014/12236-1, Grant #2015/24494-8, Grant #2016/50250-1, Grant #2017/20945-0, and Grant #2019/16253-1, in part by the FAPESP-Microsoft Virtual Institute under Grant #2013/50155-0 and Grant #2014/50715-9, in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brazil (CAPES) through a Finance Code 001, and in part by the Samsung Eletrônica da Amazônia Ltd., through the Algoritmos para Detecção e Reconhecimento de Texto Multilíngue (MLTSR) Project, within the scope of the Brazilian Informatics Law No. 8248/91.

ABSTRACT Hundreds of text detection methods have been proposed, motivated by their widespread use in several applications. Despite the huge progress in the area, which includes even the use of sophisticated learning schemes, ad-hoc post-processing procedures are often employed to improve the text detection rate, by removing both false positives and negatives. Another issue refers to the lack of the use of the complementary views provided by different text detection methods. This paper aims to fill these gaps. We propose the use of a soft computing framework, based on genetic programming (GP), to guide the definition of suitable post-processing procedures through the combination of basic operators, which may be applied to improve detection results provided by multiple methods at the same time. Performed experiments in the widely used ICDAR 2011, ICDAR 2013, and ICDAR 2015 datasets demonstrate that our GP-based approach leads to F1 effectiveness gains up to 5.1 percentage points, when compared to several baselines.

INDEX TERMS Scene text detection, multi-oriented text, convolutional neural network, data fusion, genetic programming.

I. INTRODUCTION

Texts are essential elements for effective communication in our daily life. Texts and words are everywhere, being used to guide us in specific activities or even to label objects. In both scenarios, textual elements can play an important role in the semantic understanding of scenes. Similarly, in several computer vision tasks, the understanding of textual elements in a scene may be paramount for machines to be able to recognize important events in multimedia data. In light of this, several researchers are striving towards devising applications that aim at understanding textual elements present in scenes [1]–[3].

Different from the classic optical character recognition problem, the task of localizing and recognizing text in real scenes introduces some research challenges that are still associated with open problems. The variability in the way a textual element can appear in a scene leads to failures in the recognition of texts within images, considering the algorithms and techniques available in the literature. This variability is given mainly due to differences in font style, texture, color, size, contrast, and perspective distortions.

To deal with these challenges, the research community has been making efforts towards proposing new algorithms and techniques for localizing and recognizing texts within scenes effectively by adopting deep learning-based solutions. Those solution demand high computational costs in terms of energy consumption, memory, and storage footprints. Compared to methods proposed before the deep learning "era" [4], [5], state-of-the-art solutions are associated with high effectiveness localization and recognition results. However, at the same time, those recent methods often need

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

more computational resources. In fact, for some scenarios (e.g., mobile-oriented applications), the high costs, in terms of memory consumption, of some effective deep learning solutions may prevent their use in real-word applications. On the other hand, the ability of devices with constrained resources (e.g., embedded devices and smartphones) of running several applications in parallel[1] enables the design of methods that take advantage of complementary views from different text localization approaches.

In this work, we focus on finding complementary information from lighter text localization methods for devising applications that require low memory consumption, without losing sight the idea of taking advantage of sophisticated methods towards enabling effective client-server applications, which allow an off-line processing.

Moreover, despite the use of sophisticated segmentation and even learning procedures, often ad-hoc post-processing procedures are used to improve localization results even considering powerful deep learning-based approaches to design solutions for the localization task. Common procedures include the analysis of a set of rectangular and multi-oriented bounding boxes in order to: (i) remove overlapped bounding boxes; (ii) keep bounding boxes that contain regions of interest; and (iii) remove all bounding boxes that do not contain any region of interest. Performing such simple yet effective procedures often lead to the increase of both the recall and precision of the final text detection results.

In order to address the aforementioned gaps in the literature, this paper introduces a novel method to combine localization results from different text localization methods aiming to exploit the complementary information of different methods for text detection. We model the bounding box fusion problem as an optimization problem, whose solution takes advantage of a soft computing solution based on genetic programming (GP), as illustrated in Figure 1. GP is an artificial intelligence apparatus often successfully used to find near-optimal solutions by using evolution-like solution search procedures. In the GP framework, individuals of a population are possible solutions to a target problem, which evolve over generations, subject to various genetic operators such as cross-over and mutation [6]. Figure 2 shows complementary results from different methods for text localization algorithms such as TextBoxes++, PixelLink, and Pelee-Text networks [7]–[9]. In the example, our GP-based fusion approach is used to combine effectively those complementary views provided by the different algorithms.

The main contributions of this paper are threefold: (i) an algorithm capable of combining the detection results of two or more algorithms towards capturing their complementary views; (ii) an algorithm able to filter out bounding boxes of a given algorithm towards removing overlapped bounding boxes and false positive cases; and (iii) a method for filtering bounding boxes that can be adapted to different operating scenarios and datasets (e.g., (near)-horizontal, vertical, and multi-oriented texts).

[1]https://developer.android.com/training/multiple-threads (As of April 2020).

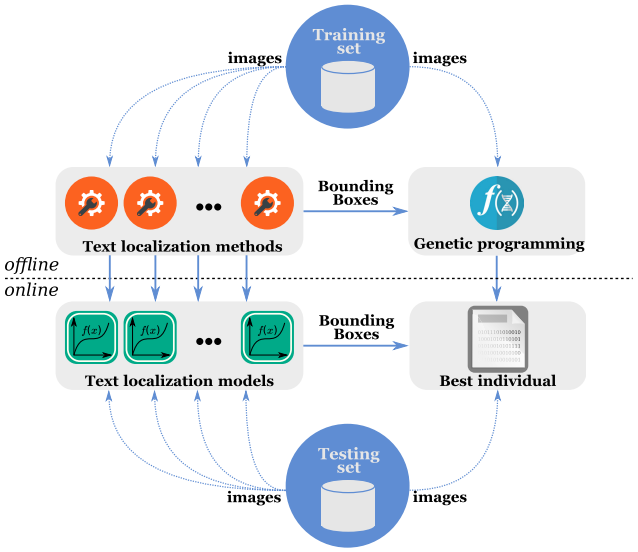


FIGURE 1. Overview of the proposed method for fusing bounding boxes from different text localization methods. Given a training set, we use part of images for training the text localization methods, and the remaining images for training the GP-based algorithm, aiming to select the best individual for fusing bounding boxes. Next, we use the text localization models for predicting the bounding boxes from the test set and the best individual, found during the training phase of our GP-based method, to fuse the predictions of text localization methods.

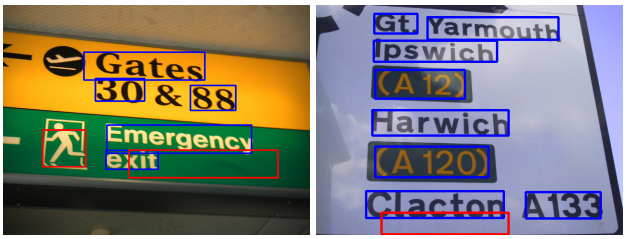
The remaining of this paper is organized as follow: Section II provides an overview of related work; Section III introduces the proposed GP-based fusion approach; Section IV presents the adopted experimental protocol, while Section V presents and discusses achieved results; finally, Section VI provides our conclusions and points out possible future research venues.

II. RELATED WORK

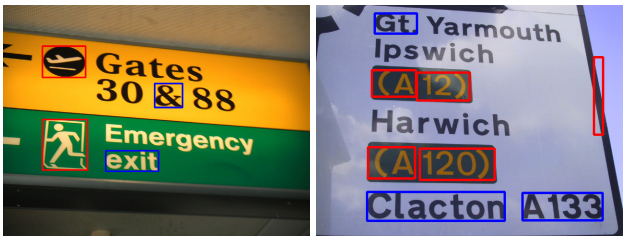
The literature available for the text localization problem is vast and covers a wide range of approaches that exploit the problem from different views. Existing methods can be divided into two main categories: bottom-up and top-down. The first approach tries to localize words by exploiting character-level patterns and grouping all detected characters towards forming words. On the other hand, the bottom-up approaches seek to detect patterns found in text lines that are more stable than patterns found at character-levels, which are more sensitive to variations such as font size and style, and disconnected stroke. This section aims to cover the main methods from three distinct groups — character-, word-, and text- line-based methods — towards emphasizing the wide variability of methods.

A. CHARACTER-BASED METHODS

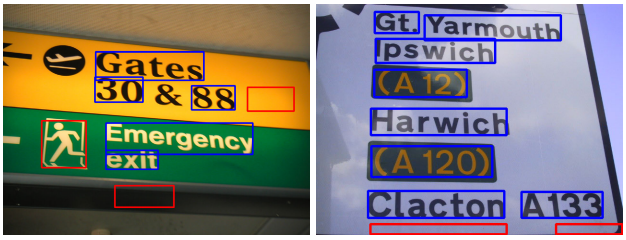
Character-based methods comprise the approaches which seek to detect characters present in a scene and, after applying grouping methods, to detect words or text lines.



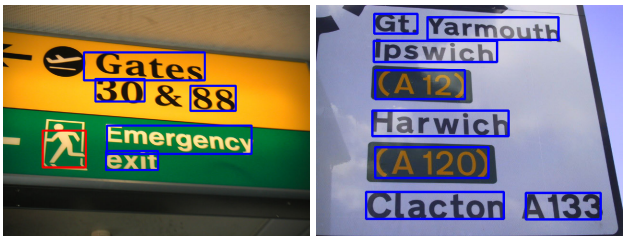
(a) TextBoxes++ results.



(b) PixelLink results.



(c) Pelee-Text results.



(d) Fusion results.

FIGURE 2. Example of fusion of some state-of-the-art methods for text detection algorithms. The last row shows the results of our proposed GP-based fusion method taking as input the bounding boxes produced by those detectors.

Character detection is a challenging task due to variability in terms of how this element can appear in a scene (e.g., different size, color, or style). For this reason, classification methods based on stages are preferable and they are designed to achieve high recall rates in the first stages by using weakly classifiers and to increase the precision, in the final stages, by removing false positives using strong classifiers.

Minetto *et al.* [10] proposed a text localization method composed of four main steps: image segmentation, character filtering, character grouping, and text region filtering. Initially, it locates candidate characters on images by means of a segmentation and a character/non-character binary classification system. The segmentation approach takes advantages of morphological operations for local contrast enhancement and thresholding. The classification system relies on shape descriptors (e.g., Fourier descriptors, Pseudo-Zernike moments, and Polar descriptor) and an ensemble of SVM classifiers. The candidate characters, represented by their bounding boxes, are then grouped according to a geometric criteria. The resulting groups, i.e., candidate text regions, are validated by means of another texture-based classification system, which exploits a multi-cell histogram of oriented gradients (named T-HOG) [11], and another SVM classifier.

Zhang and Kasturi [12] proposed a solution for the text detection problem based on character and link energies. In their text model, each character is a part and two connecting parts are connected by a link. In this method, closed boundaries in the edge map are used to detect text objects. The energies associated with characters and links are used to compute the probability that a candidate text model is really a text object. The character energy is computed based on the fact that each character stroke forms two edges with high similarity in terms of length, curvature, and orientation. Link energy, in turn, depends on the similarity of characters in terms of color, size, stroke width, and spacing. Text units, whose energy is greater than a threshold, are considered valid text objects.

Neumann and Matas [13] proposed an end-to-end real-time text localization and recognition method, where the real-time performance is achieved by posing the character detection problem as an efficient sequential selection from the set of Extremal Regions (ER), which can be summarized in four steps. Firstly, different channels are used to be processed independently: Hue, Saturation, intensity, and gradient magnitude, as well as their complements. In the second step, a component tree [14] is extracted from each channel. Later, shape-based features (e.g., aspect ratio, compactness, number of holes and number of horizontal crossings) are computed for each ER, and used as input of a classifier, which estimates the class-conditional probability $p(ER|character)$ of each ER being a character. Next, the ERs that survive to the first-stage classification, are submitted to a second-stage classifier that exploits more computationally expensive features, such as hole area ratio, convex hull ratio, among others. In the fourth step, the final set of ERs is used to find all possible text line or words.

B. WORD-BASED METHODS

These methods aim to detect words based on shared features among the characters with certain spatial proximity. In this context, He *et al.* [15] proposed a method for detecting text in natural scenes, which directly outputs word-level bounding boxes without post-processing, except for the NMS method. In short, the method can be decomposed into three parts: a convolutional component, a text-specific component, and a box prediction component. The convolutional and box prediction components are inherited from the SSD detector [16], while the a text-specific component was specifically designed for the text localization problem, which comprises two modules: a text attention module and a hierarchical inception module. The text attention module aims to automatically

learn rough spatial regions of text from the convolutional features with the goal of improving the performance with regard to three aspects: reduction of false alarms, detection of ambiguous text, and improvement of the word-level detection accuracy. The hierarchical inception module is used to aggregate multi-scale inception features in order to identify very small-scale text and working reliably on the multi-scale text.

Liao *et al.* [7] proposed a text localization and recognition solution able to predict arbitrary orientation word bounding boxes. The proposed method consists of a Fully Convolutional Network (FCN) that inherits from the popular VGG-16 architecture and is adapted to detect arbitrary-oriented words. In short, the main modifications to reach these goals are: the proposal of default boxes with vertical offsets, which enable better detection in regions with many textual elements; the use of default boxes with aspect ratios more adaptable to detect ''long words;'' and adaptation in the training stage to detect quadrilateral bounding boxes, instead of rectangular bounding boxes.

C. TEXT LINE-BASED METHODS

The main idea behind these approaches consists of detecting text lines, whose patterns associated with these elements present a better regularity in comparison with patterns extracted from individual characters since characters are more sensitive to several conditions such as blur, low-resolution, disconnected stroke, among others. He *et al.* [17] introduced a method based on Fully Convolutional Network (FCN) for the scene text detection problem. This network seeks to find text center lines of each word, which are segmented in order to came up to word-level detection. More precisely, FCN contains three branches with shared convolutional parameters and a per-scale loss function that learns features from multiple scales. In each branch, the FCN detects the center lines of words and after performs a segmentation towards detecting word instance, considering words with more than two characters.

Zhang *et al.* introduced an approach based on two fully Convolution Network (FCN) architectures for predicting a salience map of text regions in a holistic manner (named as *Text-Block FCN*), and also for predicting the centroid of each character (named *Character-Centroid FCN*) in order to eliminate false text line candidates [18]. The *Text-Block FCN* network inherits five convolutional layers from the VGG-16 network. These layers are followed by deconvolutional and up-sampling layers. The goal is to get feature maps from the intermediate representations, which are fused to generate a single salient map of text region candidates. Similarly, the Character-Centroid FCN network also inherits three layers of the VGG-16 network, which is adjusted during the training stage to remove non-text line regions considering a character-level detection.

Finally, He *et al.* [19] proposed a scene text detection method by using a FCN [20] with mechanisms for locating text line boundaries. The first step of this approach consists of extracting several visual features by using deep CNNs such as S-VGG, VGG-16, and ResNet-50. Here the CNN's outputs were redesigned such that the maximum receptive field was larger than input image size, to get long texts and then find more accurate bounding boxes. Next, these visual features are combined via another CNN able to produce feature maps of finer-resolution with multi-level features fusion since such architecture can perform a multi-scale detection, which can benefit both classification and regression of bounding boxes locations. Next, a multi-task learning stage is used to classify segments into text and non-text and to predict an oriented text boundary. The authors considered a post-processing step, named as Recalled Non-Maximum Suppression, to avoid redundancy.

III. PROPOSED METHOD

This section introduces the proposed method for fusing bounding boxes, which is based on Genetic Programming (GP) [6]. The fusion is guided by analyzing some properties of bounding boxes, such as localization and geometric aspects, that might reveal false localization, redundant localization or complementary ones. Our fusion approach was designed to learn in which case we should fuse, keep, or remove bounding boxes in order to maximize the precision and recall rates of the final results.

A. BACKGROUND ON GP

Genetic Programming (GP) comprises a set of artificial intelligence solutions, which was inspired on the theory of evolution. GP is commonly used in optimization problems, whose solutions are modeled as individuals of a population that evolves over generations, subject to genetic operations (reproduction, mutation, crossover). The objective is to discover near-optimal solutions (individuals with the best performance) to the target problem, as illustrated in Figure 3.

Algorithm 1 outlines the main GP evolutionary steps. First, a population of randomly generated individuals is created (line 1). In the following, this population is evolved over generations (lines 3 – 9). The fitness of each individual is computed (line 4) and then individuals are selected (line 5) according to their fitness to be sent to the next generations. After this step, individuals are subjected to genetic operations in order to define the next population generation (lines 6–8). At the end of the process, the best performing individual is returned (line 10).

A common application of GP is related to the evolution of programs. In this case, the goal is to find a program that best performs a particular task, based on the combination of basic fusion operators. We exploit this research venue in this work.

B. GP-FRAMEWORK FOR BOUNDING BOX FUSION

Let $\mathcal{B} = \{b_1, b_2, \ldots b_n\}$ be a set of n *candidate* bounding boxes, which are expected to be associated with text regions within images. Set $\mathcal{B}$ may be associated, for example, with the results of one or more text detection algorithms. Let $\mathcal{F}$ be a function that maps $\mathcal{B}$ to a set $\mathcal{B}' = \{b'_1, b'_2, \ldots, b'_m\}$,

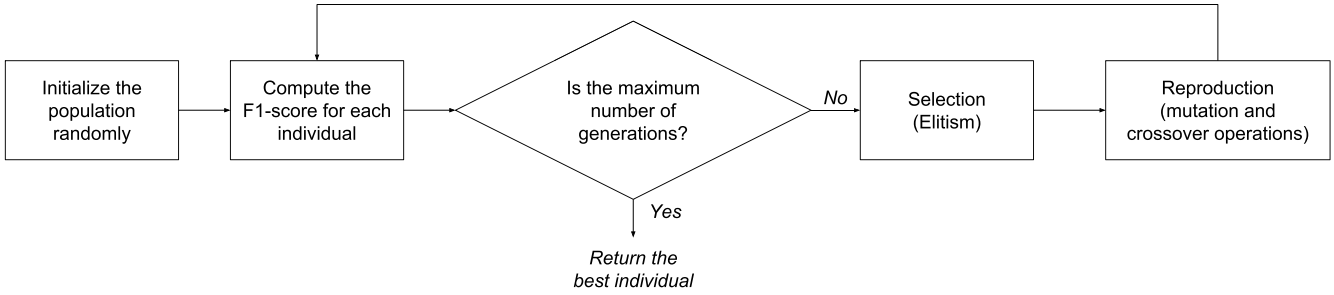


FIGURE 3. Pipeline of GP-based fusion method. After the initialization of the population. Elitism mechanisms select the best individuals (e.g., 20% best) of the population to the next generation, without any changing, while the reproduction step produces the remaining (e.g., 80%) of the population by applying mutation and crossover operations upon the entire population.

Algorithm 1 Basic GP Evolution Algorithm

```
procedure GP Evolution
    Generate an initial population of individuals
    for N generations do
        Calculate the fitness of each individual
        Select the individuals to genetic operations
        Apply reproduction
        Apply crossover
        Apply mutation
    end for
    return the best performing individual
end procedure
```

with m bounding boxes, which are expected to be associated with *all* text regions within images. Our goal is take advantage of the GP framework to find a solution that implements the most effective function $\mathcal{F}$, i.e., the one that leads to the most effective text detection results.

1) INDIVIDUAL REPRESENTATION

In our formulation, a GP individual is a program comprised of a sequence of binary and unary fusion operators, which in turn, are formed by a *condition* and a *method* (image-based operator). A binary fusion operator acts upon two overlapping bounding boxes and aims to remove redundant localization, which is performed by fusing bounding boxes or by keeping the best one, according to the condition and method of the fusion operator.

In this work, we consider four methods to build binary fusion operators as follow:

1) *non-maximum suppression (NMS)* [21], which removes the bounding boxes with the lower confidence;
2) *mean*, which fuses two bounding boxes based on the mean value of their (x, y) coordinates;
3) *union*, which merges a pair of bounding boxes using a minimum rectangle; and
4) *nothing*, which returns the bounding boxes without any transformation.

In turn, a unary fusion operator acts upon an isolated bounding box and aims to remove false positive localization. For this, we consider two methods to build unary fusion operators:

1) *remove*, which removes a bounding box according to operator's condition or
2) *nothing*, which returns the bounding boxes without any transformation.

The conditions, proposed in this work, were defined in terms of properties of the bounding boxes aiming to explore their possible complementary views. Let b_i be a bounding box defined in terms of its upper-left (x_{min}, y_{min}) and bottom-right corners (x_{max}, y_{max}). Let $\mathcal{A}_{b_i}$ and $\mathcal{C}_{b_i}$ be the area and the confidence score of b_i, respectively. Let h_{b_i}, and w_{b_i} be the height and the width of b_i. Let IoU_{b_i,b_j} be the intersection over union of bounding boxes b_i and b_j, and $\mathcal{A}_{\cap_{b_i,b_j}}$ be the area of intersection of b_i and b_j. The following conditions are used to build a GP population:

$$\left(\frac{b_i \cdot y_{max} - b_j \cdot y_{min}}{H} > \mathcal{T}_Y\right) \text{ and } \left(\frac{b_j \cdot y_{max} - b_i \cdot y_{min}}{H} > \mathcal{T}_Y\right) \tag{1}$$

$$\left(\frac{b_i \cdot x_{max} - b_j \cdot x_{min}}{W} > \mathcal{T}_X\right) \text{ and } \left(\frac{b_j \cdot x_{max} - b_i \cdot x_{min}}{W} > \mathcal{T}_X\right) \tag{2}$$

$$(\mathcal{A}_{\cap_{b_i,b_j}} > \mathcal{T}_\cap \cdot \mathcal{A}_{b_i}) \text{ or } (\mathcal{A}_{\cap_{b_i,b_j}} > \mathcal{T}_\cap \cdot \mathcal{A}_{b_j}) \tag{3}$$

$$IoU_{b_i,b_j} > \mathcal{T}_{IoU} \tag{4}$$

$$\frac{|\mathcal{A}_{b_i} - \mathcal{A}_{b_j}|}{H \cdot W} > \mathcal{T}_\mathcal{A} \tag{5}$$

$$|\mathcal{C}_{b_i} - \mathcal{C}_{b_j}| > \mathcal{T}_\mathcal{C} \tag{6}$$

$$\frac{\mathcal{A}_{b_i}}{H \cdot W} < \mathcal{T}_{\mathcal{A}_{b_i}} \tag{7}$$

$$\mathcal{C}_{b_i} < \mathcal{T}_{\mathcal{C}_{b_i}} \tag{8}$$

$$\frac{h_{b_i}}{w_{b_i}} > \mathcal{T}_{ar} \tag{9}$$

$$label_{b_i} == \mathcal{T}_l \tag{10}$$

where $\mathcal{T}_Y$, $\mathcal{T}_X$, $\mathcal{T}_\cap$, $\mathcal{T}_{IoU}$, $\mathcal{T}_\mathcal{A}$, $\mathcal{T}_\mathcal{C}$, $\mathcal{T}_{\mathcal{A}_{b_i}}$, $\mathcal{T}_{\mathcal{C}_{b_i}}$, $\mathcal{T}_{ar}$, and $\mathcal{T}_l$ are thresholds learned during the training phase, and H and W are the height and the width of an image within which the

bounding boxes are located. The conditions defined in Equations 1 and 2 verify the alignment of text bounding boxes. Equation 4, in turn, checks if the intersection over union of the two input bounding boxes are enough. The methods might be applied conditioned to differences of the input bounding boxes in terms of their areas (Equation 5) and confidence scores (Equation 6). Finally, Equations 7, 8, 9, and 10 are specifically designed to build unary fusion operators, which analyze an input bounding box in terms of its area, confidence, aspect ratio, and method used (label), respectively. Figure 4 illustrates a population of individuals composed of a sequence of *fusion operators*, which are applied if a *condition* satisfies.

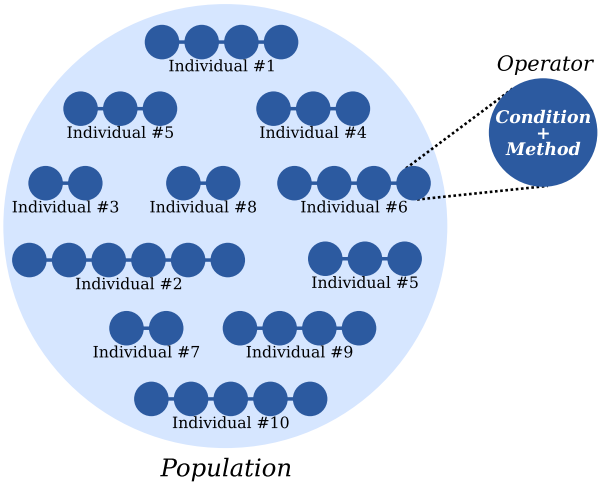


FIGURE 4. Example of a population with 10 individuals. In our formulation, each individual represents a sequence of fusion operators formed by a condition and a method. Our fusion method learns from data which sequence of fusion operators, and their configurations, that maximize the detection results when we fuse bounding boxes from different detectors.

In this work, we consider the fusion of (near)-horizontal, vertical, and multi-oriented texts. Different from (near)-horizontal and vertical texts, the fusion of multi-oriented text needs to deal with bounding boxes with different angles or orientations. For this reason, both mean and union methods used to fuse two bounding boxes were adapted to deal with multi-oriented texts. The "mean" method might not work properly for merging two multi-oriented bounding boxes, as the mean value of their coordinates may produce a bounding box that does not fit well a multi-oriented texts. Therefore, this fusion operator was not considered in text detection tasks related to multi-oriented texts. The "union" method, in turn, is suitable for handling multi-oriented texts. In its implementation, we used the convex hull algorithm to merge two oriented bounding boxes, instead of finding the minimum bounding rectangle as we do for horizontal and vertical texts (see Figure 5). In both cases, the adaptations lead us to find a tight-fitting convex boundary that encloses all points of bounding boxes. Similarly, the conditions presented in Equations 1 and 2 were adapted to deal with multi-oriented

(a) Horizontal texts

(b) Multi-oriented texts

FIGURE 5. Example of union of horizontal (a) and multi-oriented (b) texts. The first column and second columns illustrate the final bounding boxes before and after applying the union operator, respectively.

bounding boxes. Let θ_{b_i}, be the angle of a bounding box. The following angle condition is defined to be used as a condition in GP to multi-oriented text:

$$|\theta_{b_i} - \theta_{b_j}| \leq \mathcal{T}_\theta \tag{11}$$

where $\mathcal{T}_\theta$, is the angle threshold and used as a binary fusion operator. For unary operators, the condition presented in Equation 9 does not work correctly with multi-oriented text.

2) GENETIC OPERATORS

We implement two genetic operators: mutation and crossover. Mutation aims to change an operator by modifying its conditions and methods, randomly. The crossover selects two individuals as *parents* and, for each individual, a crossover position is determined. Next, operators from that positions are exchanged, leading to new individuals. Finally, that *reproduction* refers to the copy of the most effective individuals from one generation to another.

3) FITNESS FUNCTION

Let $\mathcal{S}$ be a set of images for training, $\mathcal{G}$ their respective ground truth defined in terms of the coordinate of bounding boxes associated with text regions, and $\mathcal{B}$ a set of candidate bounding boxes from different text detection algorithms. An individual $\mathcal{H}$ aims to find a subset $\mathcal{B}' \subseteq \mathcal{B}$ that maximize the fitness function defined in Equation 12:

$$F1 = \frac{1}{N}\sum_{n=1}^{N}\left(2 \times \frac{P_n \times R_n}{P_n + R_n}\right) \tag{12}$$

where N refers to the total number of examples in the training set $\mathcal{S}$, while P_n and R_n are the precision and recall computed for n-th example in $\mathcal{S}$, respectively.

In this work, we use the average F1-score as fitness function to guide the optimization process to sub-optimum solutions. However, other measures could be used according to a target application. Algorithm 2 outlines the main steps of

Algorithm 2 Fitness Computation

Input: Individual $\mathcal{H}$, a set of bounding boxes $\mathcal{B} = \{b_1, b_2, \ldots b_n\}$ to fuse, and a set of ground-truth bounding boxes $\mathcal{G} = \{b_1^{\mathcal{G}}, b_2^{\mathcal{G}}, \ldots b_{|\mathcal{G}|}^{\mathcal{G}}\}$

Output: F1-score

```
1: function Fitness(H,B,G)
2:     B' ← ∅
3:     for b_i ∈ B do
4:         for b_j ∈ B do
5:             B' ← B' ∪ FusionOperators(H, b_i, b_j)
6:         end for
7:     end for
8:     return compute_F1Score(B', G)
9: end function
```

the fitness function. The function *FusionOperators* (line 5) applies an individual $\mathcal{H}$ in all possible pairs of bounding boxes in $\mathcal{B}$ toward adding the fused bounding box in $\mathcal{B}'$ or discard them. Finally, the F1-Score (line 8) is used to measure the effectiveness of the individual $\mathcal{H}$ by comparing the bounding boxes in $\mathcal{B}'$ with bounding boxes available in $\mathcal{G}$.

4) COMPUTATIONAL COMPLEXITY

The GP training procedure takes $\mathcal{O}(N_g \times N_i \times F)$, where N_g is the number of generations considered in the evolution process, N_i is the number of individuals in the population, and F is the cost for evaluating the fitness function [22]. The costs for computing F depends on the size of individuals (i.e., the number of operations), the cost associated with operators, and the number of training samples. Recall that the training process is performed offline. On average, for the ICDAR 2011 dataset, a typical GP training takes 960 s. For the ICDAR 2013 dataset, it takes 120 s, while for the ICDAR 2015 dataset, 17,040 s. The training process for ICDAR 2011, ICDAR 2013, and ICDAR 2015 datasets considered around 1000, 1000, and 5000 bounding boxes, respectively.

IV. EXPERIMENTAL SETUP

In this section, we present datasets (Section IV-A), along with their respective protocols (Section IV-B) used to validate our method. We also present the metrics (Section IV-C) adopted for measuring the effectiveness of the proposed method.

A. DATASETS

We evaluated the proposed methods in three datasets widely used for evaluating text localization methods, the ICDAR 2011, ICDAR 2013, and ICDAR 2015.

1) ICDAR 2011

This dataset was introduced in *ICDAR 2011 Robust Reading Competition* and it was built for evaluating text localization and recognition algorithms. The ICDAR'11 dataset provides images found in Web pages and emails, which typically contain text born-digital images, i.e., text created digitally. Usually, these multimedia objects present a low-resolution and several compression artifacts since they are generated to be transmitted over the Internet at a minimum cost.

In the official evaluation protocol of this dataset, the 551 images were divided into two subsets: training and test sets. The training set contains 410 images and it was used to estimate the parameters of the proposed method. The test set comprises of 141 images, which was used only to report the performance results of the proposed method.

2) ICDAR 2013

This dataset was introduced in *ICDAR 2013 – "Focused Scene Text challenge competition"* and it is composed of scene text images. In scene text images, the textual elements appear in real scenes, which were captured by a camera in an indoor or outdoor environment. For this reason, the text localization and recognition in scenes are usually a challenging scenario due to mainly the variability in which the text appear in real scenes, such as font style and sizes, color, texture, among others. In total, this dataset provides 462 images whose annotations were built in terms of rectangle word bounding boxes, totaling 1,943 words. All the text lines are horizontal or near horizontal.

The official evaluation protocol defined for this dataset divides the 462 images into two subsets, training and testing sets, which contain 229 and 233 images, respectively [23]. In this work, the training set was used to estimate the parameters of the proposed method, while the test set was used only to report the performance results of our approach.

3) ICDAR 2015

This dataset was introduced in *ICDAR 2015 – "Incidental Scene Text challenge competition"* and it is composed of scene text images. This dataset provides images that were captured by Google glasses in an indoor or outdoor environment where the user of the camera does not take any action before captured the image, causing that the image captured has poor quality and text positioning. In total, this dataset comprises 1500 images whose annotations were built in terms of multi-oriented word bounding boxes, totaling 6545 words. All the text lines are arbitrary. The official evaluation protocol defined for this dataset divides the 1500 images into two subsets, training and testing sets, which contain 1000 and 500 images, respectively [24]. In this work, the training set was used to estimate the parameters of the proposed method, while the test set was used only to report the performance results of our approach.

B. EVALUATION PROTOCOL

This section describes the evaluation protocol adopted to validate the GP-based method for fusing the detection results from different text localization methods. For all datasets used in this work, we split the training set into two subsets with equal size, hereafter named as training and validation sets. The training set was used to train the text localization methods, and the validation set was used in the GP-based

fusion function discovery process. To have a more generalized method for fusing the bounding boxes detected by the text localization methods, we split the validation set again into two subsets, also with equal size. The first subset was used to train the GP-based method, and the second subset was used to select the best individual considering a set containing the best individuals (e.g., 100 best), which were tracked during the training stage of the GP-based method. Finally, we used the official test set to measure the efficacy of the proposed methods and the baseline methods. Table 1 summarize the number of images considered on each subset.

TABLE 1. Number of images used for training the text localization methods and the GP-based fusion methods, after split the official training set of datasets considered in this work.

Dataset	Text Localization		GP-based Fusion	
	Training set	Validation set	Training set	Validation set
ICDAR 2011	205	205	102	103
ICDAR 2013	114	115	57	58
ICDAR 2015	500	500	250	250

C. EVALUATION METRICS

We evaluated the effectiveness of the proposed methods in terms of recall, precision, and f-measure. Here, we consider a correct detection (true positive) if the overlap between the ground-truth annotation and detected bounding box, which is measured by computing the intersection over union, is greater than 50% (similar to standard practice in object recognition [25]). Otherwise, the detected bounding box is considered an incorrect detection (false positive). For a fair comparison with other methods available in the literature, we use the evaluation tools provided by the *"ICDAR Robust Reading Competition"* organizers. All experiments were performed considering a Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz with 12 cores, and 64GB of RAM.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the performance results of our proposed approach for fusing bounding boxes from different text localization methods. The experimental protocol considered two scenarios: a restrictive computing scenario, which requires low-cost solutions, such as detectors designed with classical machine learning techniques; and a nonrestrictive scenario that allows the use of high-cost solutions, such as deep learning approaches. For this, we select from literature effective text localization methods based on classical machine learning techniques such as Scene Text Recognition [13], SnooperText [10], and MSER-SWT Text Detection [26], [27], hereinafter, referred to non-deep learning methods. Although these methods were not proposed recently, they are, in fact, among the most effective text localization methods based on fundamental feature engineering techniques. For the experiments related to nonrestrictive computing scenario, we select two effective and efficient methods based on Convolutional Neural Network (CNN), the TextBoxes++ [7], Pelee-Text [9], and PixelLink [8] methods. We also consider the PSENet [28] network for the experiments in the ICDAR 2015 dataset.

A. WOULD THE FUSION LEADS TO IMPROVED RESULTS, IN COMPARISON WITH PERFORMANCE OF INDIVIDUAL ALGORITHMS FOR TEXT LOCALIZATION?

This section evaluate our GP-based solution toward fusing bounding boxes from different text localization approaches. The next sections, we present the performance results for the three datasets considered in this work.

1) ICDAR 2011: BORN-DIGITAL IMAGES

This section presents the performance results of our fusion approach for the ICDAR 2011 dataset, which provides born-digital images with low-quality and with considerable amount of JPEG artifacts. Table 2 shows the results considering the fusion of non-deep methods. Our GP solution for fusing bounding boxes achieved the best results, in terms of precision, recall, and F1, in comparison with individual performance of methods for text localization methods. Our fusion method was able to bring a maximum percentage increase of 16.5% and 89.4%, in terms of precision and recall, respectively, in comparison with MSER-SWT method. Considering the best text detection approach, SceneText method, the percentage increase was over 6.0% for all metrics.

TABLE 2. Performance results of the GP-based fusion method considering the non-deep methods and the ICDAR 2011 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
SceneText	83.7	70.4	76.5
SnooperText	79.2	63.9	70.7
MSER-SWT	76.9	39.5	52.2
GP-based Fusion	**89.6**	**74.8**	**81.5**

TABLE 3. Performance results of the GP-based fusion method considering the deep learning-based methods and the ICDAR 2011 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
TextBoxes++	90.0	**90.7**	90.3
Pelee-Text	85.9	88.4	87.2
PixelLink	89.0	52.8	66.3
GP-based Fusion	**93.7**	89.0	**91.3**

Our GP solution also presented a better precision, recall and F1 values, in comparison with the deep learning-based methods used during the fusion step (see Table 3). We could observe percentage increases of 68.6% and 5.3% of precision and recall, in comparison with PixelLink network. Considering the best CNN architecture available in our baseline (TextBoxes++), the the percentage increase in terms of precision reached a value of 4.1%.

TABLE 4. Performance results of the proposed method considering the fusion of non-deep and deep learning-based methods in the ICDAR 2011 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
SceneText	83.7	70.4	76.5
SnooperText	79.2	63.9	70.7
MSER-SWT	76.9	39.5	52.2
TextBoxes++	90.0	**90.7**	90.3
Pelee-Text	85.9	88.4	87.2
PixelLink	89.0	52.8	66.3
GP-based	**93.0**	89.6	**91.3**

To evaluate the ability of the proposed method to fuse non-deep and deep learning methods, we designed an experiment in which we gather the detection results of all methods, deep and non-deep methods, and then we train our GP-fusion method to find the best fusion operator. Table 4 shows a comparison of performance results between our fusion approach and the results obtained by all methods, deep and non-deep methods. In this scenario, our fusion approach obtained the best results for precision and F1 metrics, whose values were 93.0% and 91.3%, respectively. From this experiment, we could conclude that our approach was still able to capture complementary information among the text detectors.

Finally, in both scenarios, using non-deep and deep learning-based methods, our GP solution was superior than performance of individual methods, which suggest that the proposed method for fusing bounding boxes is able to extract complementary information among different approaches for text localization. That opens new opportunities for further investigations related to development of methods for constrained processing scenarios. Those methods would improve the effectiveness of efficient non-deep methods by combining their complementary views.

TABLE 5. Performance results of the GP-based fusion method considering the non-deep methods and the ICDAR 2013 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
SceneText	72.3	49.1	58.5
SnooperText	80.5	56.9	66.7
MSER-SWT	63.8	34.9	45.1
GP-based Fusion	**81.2**	**61.5**	**70.0**

2) ICDAR 2013: HORIZONTAL AND VERTICAL SCENE TEXTS

This section presents the performance results of our fusion approach for the ICDAR 2013 dataset, which provides (near)-horizontal and vertical scene texts. Tables 5, 6 and 7 show the effectiveness of our approach in combining bounding boxes from different text localization methods. The GP-based fusion achieved a better precision, recall, and F1 values than individual methods and baseline method, considering the non-deep methods. In the Tables 5, we could observer percentage increases of 76.2% and 27.3%, in terms of precision and recall, respectively. Considering the best text localization method in this dataset, our fusion could bring a percentage increase of 4.9%, in terms of F1 value.

For the deep learning-based methods (Tables 6), our GP solution also achieved the best results for precision and F1 metrics. The minimum and maximum percentage increase in terms of F1 value was 1.6% and 53.9%, respectively. We also evaluated the fusion among deep and non-deep methods. Tables 7 presents the performance results after fusing all methods, from which we could observe that our approach obtained the best results for precision and F1 metrics with values of 90.1 and 85.9, respectively. We also could observe an improvement for all metrics in comparison with the fusion results achieved by fusing deep and non-deep methods, separately. Figure 6 illustrates some examples of fusion bounding boxes. As we can observe, our proposed solution was able to properly fuse overlapped bounding box (Figures 6(a), (g), and (f)) and, at the same time, to remove false positive detections (Figures 6(a), (c), and (e)).

TABLE 6. Performance results of the GP-based fusion method considering the deep learning-based methods and the ICDAR 2013 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
TextBoxes++	84.9	82.2	83.5
Pelee-Text	76.4	**82.3**	79.2
PixelLink	48.8	63.3	55.1
GP-based Fusion	**89.5**	80.6	**84.8**

TABLE 7. Performance results of the proposed method considering the fusion of non-deep and deep learning-based methods in the ICDAR 2013 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
SceneText	72.3	49.1	58.5
SnooperText	80.5	56.9	66.7
MSER-SWT	63.8	34.9	45.1
TextBoxes++	84.9	82.2	83.5
Pelee-Text	76.4	**82.3**	79.2
PixelLink	48.8	63.3	55.1
GP-based Fusion	**90.1**	81.8	**85.9**

3) ICDAR 2015: MULTI-ORIENTED SCENE TEXTS

This section presents the performance results of our fusion approach for the ICDAR 2015 dataset, which contains multi-oriented scene texts. Table 8 shows the performance results to fuse deep learning-based methods. We could observe that our GP-based fusion bring improvements for both recall and F1 metrics, with a percentage increases of 14.7% and 9.6%, respectively, in comparison with PSENet network, and a percentage increases of 0.7% and 1.5%, also in terms of recall and F1, in comparison with the best text localization method (TextBoxes++).

Figure 7 illustrates examples of fusion bounding boxes, from which we can confirm the ability of our proposed

FIGURE 6. Example of detection results achieved by non-deep methods (first column) and their fusion (second column), considering horizontal bounding boxes (ICDAR 2011 and ICDAR 2013).

method for learning complementary information from different detectors. For instance, the example illustrated in the second row shows that Pelee-Text and PSENet networks detected

TABLE 8. Performance results of the GP-based fusion method considering the deep learning-based methods and the ICDAR 2015 dataset.

Methods	Precision (%)	Recall (%)	F1 (%)
TextBoxes++	82.0	79.4	80.7
Pelee-Text	**83.5**	77.3	80.3
PixelLink	81.7	80.7	81.2
PSENet	72.0	77.6	74.7
GP-based Fusion	82.6	**81.3**	**81.9**

block of texts instead of words, which increased the false positive rates for these methods. On the other hand, PixelLink network split the word *"MARINA:SQUARE"* into two words, which also increased the false positive rate of this network. However, our fusion method was able to properly fuse the detection results of these methods, filtering out false positive detections and accepting correct bounding boxes detected. Finally, the third example (last row) shows a clear example of spurious bonding boxes removal.

B. WOULD GENETIC PROGRAMMING BE AN EFFECTIVE APPROACH FOR BOUNDING BOX FUSION?

This section presents a comparison of performance of our GP-based solution for fusion and other well-known fusion rule such as union-based fusion, i.e., OR-rule. This experiment aims to verify if GP-framework could find, in training phase, an effectiveness criteria for fusing bounding boxes considering the (near)-horizontal, vertical, and multi-oriented texts.

Tables 9, 10 and 11 show the comparison of performance considering the non-deep and deep learning-based methods, respectively. We could observe that our approach presented better results in terms precision ad F1 values for all scenarios. In comparison with Union-based fusion, the proposed method brings a percentage increase was of 9.7% and 16.3%, for ICDAR 2011 and ICDAR 2013 datasets, respectively, in terms of F1 and considering the fusion of non-deep methods (Table 9). For deep learning-based methods (Table 10), the percentage reaches 3.5%, 18.4%, and 10.8% for ICDAR 2011, ICDAR 2013, and ICDAR 2015 datasets, respectively, also in terms of F1. Finally, our approach also presented better results for precision and F1 metrics considering the fusion of deep and non-deep methods (Table 11). These results suggest that our proposed method was able to find criteria that lead a effective fusion of bounding boxes under different scenarios.

C. WOULD OUR PROPOSED SOLUTION BASED ON GP LEADS TO IMPROVED RESULTS WHEN ACTING AS POST-PROCESSING METHOD?

This section presents experimental results of our proposed method for bounding boxes filtering. For these experiments. In this task, our GP-based solution is expected to remove or fuse overlapped bounding boxes, to remove bounding boxes with low-confidence, and mainly to remove false positive cases. We compare our results with the

FIGURE 7. The fusion of multi-oriented bounding boxes detected by deep learning-based methods. The first four columns illustrate detection results achieved by the Pelee-Text, TextBoxes++, PixelLink, and PSNet networks, respectively, and the last column shows the performance results achieved by our GP-based fusion method.

TABLE 9. Comparison of performance results among our proposed method and union rule-based fusion, considering the non-deep methods.

Dataset	Methods	Precision (%)	Recall (%)	F1 (%)
ICDAR 2011	Union-Rule + NMS	73.3	**75.2**	74.3
	GP-based Fusion	**89.6**	74.8	**81.5**
ICDAR 2013	Union-Rule + NMS	59.9	60.4	60.2
	GP-based Fusion	**81.2**	**61.5**	**70.0**

TABLE 10. Comparison of performance results among our proposed method and union rule-based fusion, considering the deep learning-based methods.

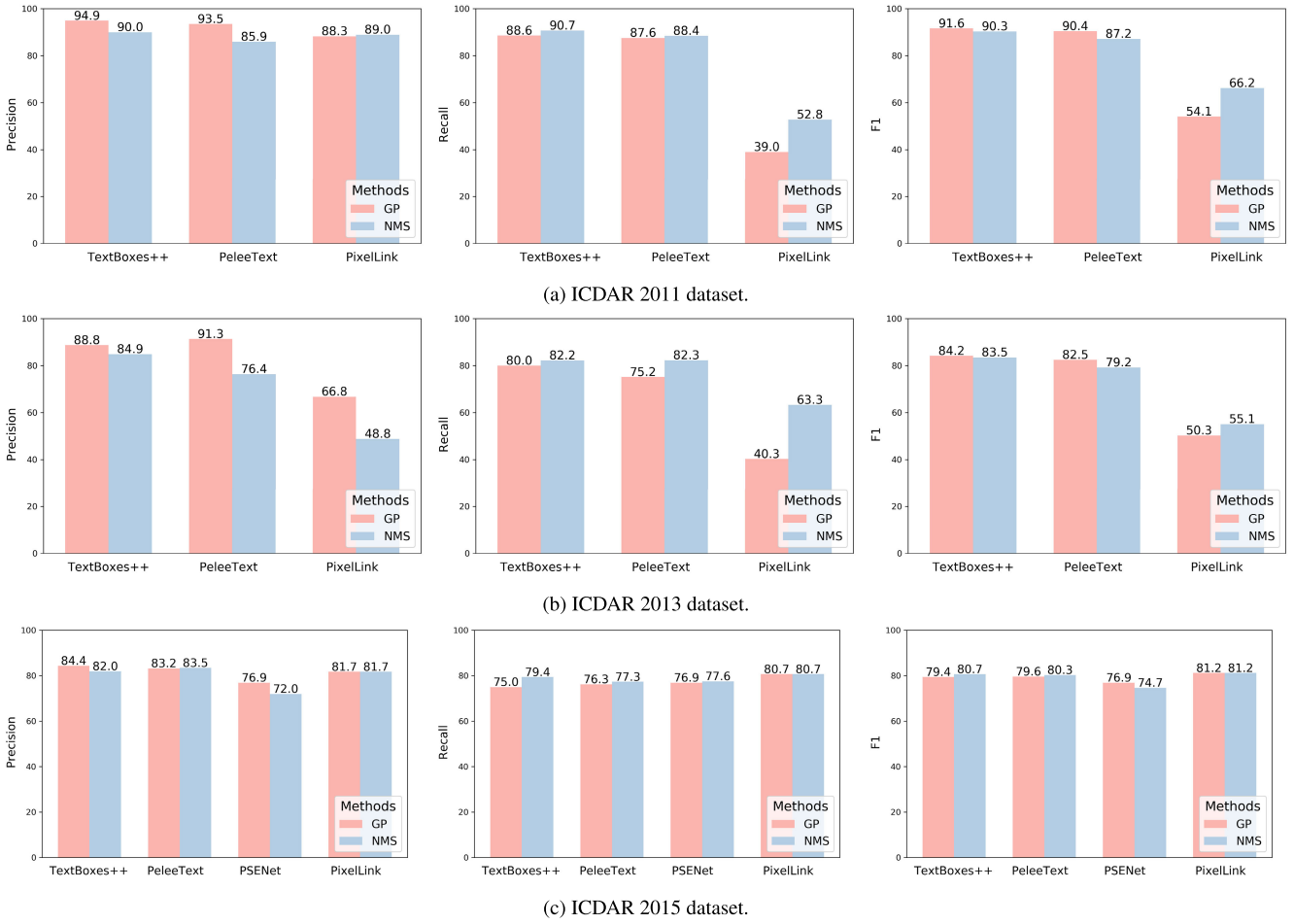
Dataset	Methods	Precision (%)	Recall (%)	F1 (%)
ICDAR 2011	Union-Rule + NMS	85.7	**90.8**	88.2
	GP-based Fusion	**93.7**	89.0	**91.3**
ICDAR 2013	Union-Rule + NMS	62.7	**83.4**	71.6
	GP-based Fusion	**89.5**	80.6	**84.8**
ICDAR 2015	Union-Rule + NMS	65.1	**85.6**	73.9
	GP-based Fusion	**82.6**	81.3	**81.9**

TABLE 11. Comparison of performance results among our proposed method and union rule-based fusion, considering the deep and non-deep methods.

Dataset	Methods	Precision (%)	Recall (%)	F1 (%)
ICDAR 2011	Union-Rule + NMS	78.8	**91.2**	84.5
	GP-based Fusion	**93.0**	89.6	**91.3**
ICDAR 2013	Union-Rule + NMS	55.7	**84.1**	67.0
	GP-based Fusion	**90.1**	81.8	**85.8**

standard method used in the literature to remove overlapped bounding boxes, the non-maximum suppression (NMS) method [21]. We do not consider the non-deep methods in these experiments because such methods already have a post-processing step in their original pipelines, which could lead to biases in our conclusions regarding the use of the proposed method as a post-processing step upon these approaches.

Figure 8 shows the results for the proposed method considering all datasets considered in this work. We could observe that our method was able to improve the precision for all datasets and text localization methods and datasets, except for the PixelLink network in the ICDAR 2011 dataset, which suggest that our proposed method could remove false positive cases, i.e., bounding boxes whose content does not have textual elements. On the other hand, our methods did not lead to improvements in terms of recall, which was expected since the fusion method does not generate bounding boxes in text regions that was not detected for any text localization methods. In fact, our GP-based fusion solution is expected to increase the precision rates and not decrease the recall rates, as much as possible, towards having better results in terms of F1 metric. In this context, we could observe that our proposed method led to improvements in terms of



(a) ICDAR 2011 dataset.

(b) ICDAR 2013 dataset.

(c) ICDAR 2015 dataset.

FIGURE 8. Performance results of the proposed method acting as post-processing method.

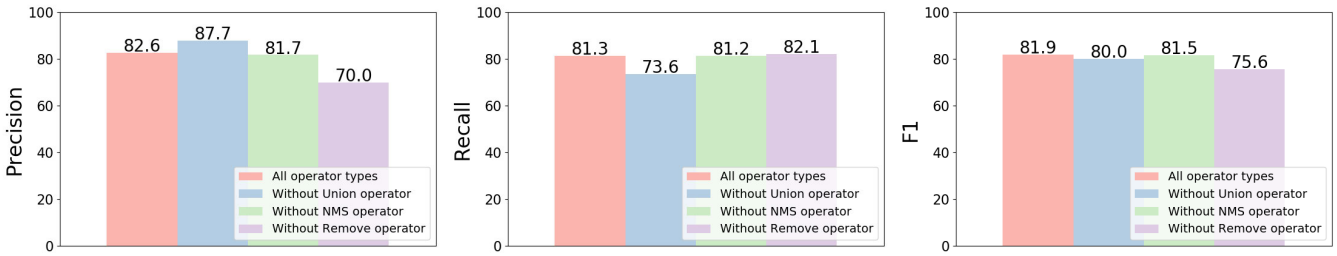


FIGURE 9. Comparison of performance of our proposed method trained in the absence of a particular fusion operator type.

F1 for all text localization methods, except for the PixelLink network.

D. WHAT IS THE MOST IMPORTANT FUSION OPERATOR TYPE FOR AN EFFECTIVE FUSION?

This section presents the results of experiments designed to find out the importance of fusion operators proposed in this work. Figure 9 shows a comparison of performance of our GP-based fusion method trained in the absence of a particular fusion operator. We could observe that remove operator plays an important role during the fusion, followed by the union operator. We could observe a great drop in the overall performance of our GP-based fusion when we discard these operators. The percentage decreases in terms of F1 values reaches 0.5% and 16.3% when we discard the union and the remove operator, respectively.

VI. CONCLUSION

In this paper, we proposed a solution that learns how to effectively fuse bounding boxes from text localization methods. This work modeled the fusion as an optimization process and takes advantage of a genetic programming framework towards exploiting complementary views provided by results from different text detectors. For this, we designed a set of binary and unary operators capable of merging and removing bounding boxes according to some conditions designed in this

work to explore localization and geometric aspects of text candidate regions. Our GP-based fusion solution was able to learn a sequence of operators and their parameters that analyze both pair of bounding boxes and isolated ones. The goal is to decide which pairs should be merged or kept; or which bounding boxes should be removed towards maximizing the precision and recall rates of the final results.

The experimental results demonstrate that the GP-based fusion approach leads to highly effective results for widely used benchmarks. These results suggest that our approach is promising for improving the effectiveness of text detectors based on the combination of efficient non-deep methods. That opens the opportunity of developing applications for devices with constrained processing capabilities (e.g., mobile devices) based on non-deep approaches. Also, the GP-based fusion scheme was able to fuse and to improve the detection scores of highly effective deep learning methods, which makes it a promising alternative for fusing effective text detectors in operating scenarios that allow off-line processing and also for devising data-driven post-processing strategies.

Future work will be concerned with the inclusion of novel operators to improve the fusion function discovery process. We also plan to develop fusion approaches for arbitrarily shaped texts (e.g., curved text collections).

ACKNOWLEDGMENT

The authors would like to thank Samsung R&D Institute, Brazil.

REFERENCES

[1] C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, and Q. Dai, "Effective uyghur language text detection in complex background images for traffic prompt identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 220–229, Jan. 2018.

[2] L. Wang, Z. Wang, Y. Qiao, and L. Van Gool, "Transferring deep object and scene representations for event recognition in still images," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 390–409, Apr. 2018.

[3] C. Yi, Y. Tian, and A. Arditi, "Portable camera-based assistive text and product label reading from hand-held objects for blind persons," *IEEE/ASME Trans. Mechatronics*, vol. 19, no. 3, pp. 808–817, Jun. 2014.

[4] H. Zhang, K. Zhao, Y.-Z. Song, and J. Guo, "Text extraction from natural scene image: A survey," *Neurocomputing*, vol. 122, pp. 310–323, Dec. 2013.

[5] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[6] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA, USA: MIT Press, 1992.

[7] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[8] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 6773–6780.

[9] M. A. Cordova, L. G. L. Decker, J. L. Flores-Campana, A. A. dos Santos, J. S. Conceicao, A. Pinto, H. Pedrini, and R. da S. Torres, "Pelee-text: A tiny convolutional neural network for multi-oriented scene text detection," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Florida, FL, USA, Dec. 2019, pp. 400–405.

[10] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "SnooperText: A text detection system for automatic indexing of urban scenes," *Comput. Vis. Image Understand.*, vol. 122, pp. 92–104, May 2014.

[11] R. Minetto, N. Thome, M. Cord, N. J. Leite, and J. Stolfi, "T-HOG: An effective gradient-based descriptor for single line text regions," *Pattern Recognit.*, vol. 46, no. 3, pp. 1078–1090, Mar. 2013.

[12] J. Zhang and R. Kasturi, "A novel text detection system based on character and link energies," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4187–4198, Sep. 2014.

[13] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3538–3545.

[14] D. Nistér and H. Stewénius, "Linear time maximally stable extremal regions," in *Computer Vision—ECCV*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Germany: Springer, 2008, pp. 183–196.

[15] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3066–3074.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision–ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[17] D. He, X. Yang, C. Liang, Z. Zhou, A. G. Ororbia, D. Kifer, and C. L. Giles, "Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 474–483.

[18] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4159–4167.

[19] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Multi-oriented and multi-lingual scene text detection with direct regression," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5406–5419, Nov. 2018.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[22] W. Fan, M. D. Gordon, and P. Pathak, "Discovery of context-specific ranking functions for effective information retrieval using genetic programming," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 4, pp. 523–527, Apr. 2004, doi: 10.1109/TKDE.2004.1269663.

[23] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Washington, DC, USA, Aug. 2013, pp. 1484–1493.

[24] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Washington, DC, USA, Aug. 2015, pp. 1156–1160.

[25] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

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

[27] Á. González, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Text location in complex images," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 617–620.

[28] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9336–9345.

JOSE L. FLORES CAMPANA received the B.Sc. degree in informatics engineering from the UNSAAC, Peru, in 2016, and the M.Sc. degree in computer science from the University of Campinas (Unicamp), Brazil, where he is currently pursuing the Ph.D. degree. His research interests include machine learning, deep learning, and image processing, especially in text detection and recognition in images.

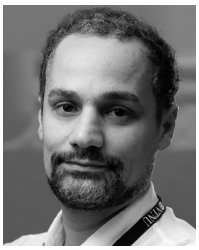
ALLAN PINTO (Member, IEEE) received the B.Sc. degree in computer science from the University of São Paulo (USP), Brazil, in 2011, and the M.Sc. and Ph.D. degrees in computer science from the University of Campinas (Unicamp), Brazil, in 2013 and 2018, respectively. He is a Postdoctoral Researcher with the Unicamp. A part of his doctoral was accomplished at the University of Notre Dame, USA, in which he worked on different topics such as presentation attack detection in biometric systems, content-based image retrieval, and multimedia forensics. He is currently a member of the Editorial Board of the *Forensic Science International: Reports*.

MANUEL ALBERTO CÓRDOVA NEIRA received the B.Sc. degree in systems engineering from the National University of Loja (UNL), Ecuador, in 2010, and the M.Sc. degree in computer science from the University of Campinas (Unicamp), Brazil, in 2015, where he is currently pursuing the Ph.D. degree. He worked as a Professor at the Department of Systems Engineering, UNL, Ecuador, from 2016 to 2018.

LUIS GUSTAVO LORGUS DECKER received the bachelor's degree in computer science from the Universidade Federal de Santa Catarina (UFSC) and the master's degree from the University of Campinas (Unicamp), where he is currently pursuing the Ph.D. degree. He has experience in computer vision and digital holography. He is currently working on deep learning.

ANDREZA SANTOS is currently pursuing the B.Sc. degree in computer science with the University of Campinas (Unicamp). Her research interests include machine learning and deep learning, specifically in detecting objects on images and videos.

JHONATAS S. CONCEIÇÃO is currently pursuing the B.Sc. degree in computer science with the University of Campinas (Unicamp). His research interests include machine learning and deep learning, specifically in exploiting context information.



RICARDO DA SILVA TORRES (Member, IEEE) received the B.Sc. degree in computer engineering and the Ph.D. degree in computer science from the University of Campinas (Unicamp), Brazil, in 2000 and 2004, respectively. He used to hold a position as a Professor at the Unicamp, from 2005 to 2019. He is a Professor in visual computing with the Norwegian University of Science and Technology (NTNU). He has been developing multidisciplinary eScience research projects involving multimedia analysis, multimedia retrieval, machine learning, databases, information visualisation, and digital libraries. He has authored or coauthored more than 200 articles in refereed journals and conferences and serves as a PC member of several international and national conferences. He has been serving as a Senior Associate Editor for the IEEE Signal Processing Letters and an Associate Editor of the *Pattern Recognition Letters*.