Multi-Lingual Text Localization via Language-Specific Convolutional Neural Networks

Jhonatas S. Conceição*, Allan Pinto*, Luis Decker*, Jose Luis Flores Campana*,
Manuel Cordova Neira*, Andreza A. dos Santos*, Helio Pedrini*, Ricardo Torres*†,

*Institute of Computing, University of Campinas (UNICAMP), Campinas, SP, Brazil, 13083-852
†Department of ICT and Natural Sciences, NTNU, Norwegian University of Science and Technology, Trondheim, Norway

***Abstract*—Scene text localization and recognition is a topic in computer vision that aims to delimit candidate regions in an input image containing incidental scene text elements. The challenge of this research consists in devising detectors capable of dealing with a wide range of variability, such as font size, font style, color, complex background, text in different languages, among others. This work presents a comparison between two strategies of building classification models, based on a Convolution Neural Network method, to detect textual elements in multiple languages in images: (i) classification model built on a multi-lingual training scenario; and (ii) classification model built on a language-specific training scenario. The experiments designed in this work indicate that language-specific model outperforms the classification model trained over a multi-lingual scenario, with an improvement of 14.79%, 8.94%, and 11.43%, in terms of precision, recall, and F-measure values, respectively.**

I. Introduction

Text localization and recognition are challenging problems in computer vision, which consist in identifying characters and words in images or videos. While text localization consists in finding delimited candidate regions that contain textual information, text recognition aims to transform a scene text into machine-encoded text. Fig. 1 shows examples of challenging scenes containing text for both tasks with a variability of factors, such as font sizes and styles, color, and distortions.

Fig. 1: Examples of textual elements with different font sizes and styles.

Besides the complicating factors mentioned previously, the presence of multiple languages in a scene demands for language independent methods capable of localizing words accordingly. In this context, the specificity of languages must be taken into account in order to reach proper detection rates such as correct split of text line into words, language-specific symbols, accents, and orientation (see Fig. 2). Several methods available in the literature present large variations in accuracy when we consider the detection rates from different languages. FOTS method [1] achieved an F-measure of 73.31% considering six languages and symbols. However, when we analyze the F-measure per language, we observe a large difference in performance that reaches 36.97%, considering the Latin and Bangla languages. Similarly, CRAFT method [2] presented an overall F-measure of 74.03% and a difference in performance, considering the best and the worst results for each language separately, of 41.66%. The same phenomenon was observed for other approaches, such as the PixelLink network [3].

Fig. 2: Examples of images containing textual elements of different languages.

In this work, we aim to investigate the reasons why the current methods suffer from detecting scene text in different languages. Our hypothesis is that current formulations for text localization methods, i.e., Convolutional Neural Network-based methods work better if we encode the specificity of a language in a single classification model. In this paper, we

design experiments to verify this hypothesis and the phenomena associated with these differences, in terms of accuracy, for each language by creating language-specific classification models for the text localization problem. In our experiments, we use the PixelLink neural network. This approach was chosen based on its outstanding results in text detection and recognition problems [3].

The remainder of the paper is organized as follows. Section II presents the methodology used to contextualize the neural network PixelLink. Section III presents and discusses the experimental results. Finally, Section IV concludes the paper with some final remarks and directions for future work.

II. Methodology

This section presents the methodology adopted in this work to verify the hypothesis raised in this work and to improve the performance results of text localization methods in a multilingual scenario. Next, we describe the PixelLink method, which is a Convolutional Neural Network (CNN) designed to classify the pixels of an image as text/non-text, and also to predict the links among them to come up into a word-based text localization.

A. Method Overview

The PixelLink method addresses the problem of localizing text in a scene based on instance segmentation. In the instance segmentation problem, there are two main tasks involved: prediction of categories for pixels of an image by performing a pixel-wise labelling; and differentiation of objects of a same category (e.g., segment individuals in crowds, cars in heavy traffics). In the context of the text localization problem, PixelLink predicts positive pixels, i.e., pixels belonging to textual elements, and joins them into text instances by predicting positive links. To link a pixel to another one, PixelLink verifies its neighbors, considering an eight-connected neighborhood, to check if there is any neighbor labeled as positive pixel. Thus, positive pixels are grouped to form connected components, where each connected component represents a text instance. Finally, the method computes a minimum-area bounding rectangle from all text instances found and removes all bounding boxes with a shorter side smaller than 10 pixels or an area smaller than 300 pixels.

B. Loss Functions

The PixelLink defines three loss functions to (i) examine each pixel individually, (ii) examine the predicted links to linkage pixels of the same instance, and (iii) compute the overall error on training phase. Equation 1 shows the training loss, which consists of the weighted sum of pixel (L_{pixel}) and link (L_{link}) losses:

$$L = \lambda L_{pixel} + L_{link} \tag{1}$$

Equation 2 shows the pixel loss function used in this work, where r refers to the positive-negative ratio, S refers to area of the instance, and W is a matrix of weights for all positive pixels that is used to balance the loss computed over a small and large areas (Equation 3), for all N instances. Finally, the L_{pixel_CE} is the matrix of Cross-Entropy loss computed for the text and non-text predictions:

$$L_{pixel} = \frac{1}{(1+r)S} + W L_{pixel\_CE} \tag{2}$$

$$w_i = \frac{B_i}{S_i}$$
$$B_i = \frac{S}{N} \quad S = \sum_{i}^{N} S_i \quad \forall i \in \{1, \cdots, N\} \tag{3}$$

In turn, the link loss is defined as a sum of positive and negative link losses, as shown in Equation 4:

$$L_{link} = \frac{L_{link\_pos}}{rsum(W_{pos\_link})} + \frac{L_{link\_neg}}{rsum(W_{neg\_link})}$$
$$L_{link\_pos} = W_{pos\_link} L_{link\_CE}$$
$$L_{link\_neg} = W_{neg\_link} L_{link\_CE} \tag{4}$$
$$W_{pos\_link}(i,j,k) = W(i,j) \times (Y_{link(i,j,k)==1)}$$
$$W_{neg\_link}(i,j,k) = W(i,j) \times (Y_{link(i,j,k)==0)}$$

where k is the k-th neighbor of pixel (i, j), $rsum$ is a reduce sum function that computes the sum of all elements of a tensor, W is the matrix of weights defined in Equation 2, and Y is the label matrix of links.

III. Results and discussion

This section presents the datasets and protocols used in this work, as well as results achieved with the experiments designed to validate our hypothesis. We report the quality of the results in terms of Recall, Precision, and F-measure. In the text localization problem, recall measures the fraction of correct bounding boxes detected over all bounding boxes present in the ground truth, precision measures the fraction of correct bounding boxes over all bounding boxes detected with the method, and the F-measure is the harmonic mean between recall and precision. All experiments were conducted on an Intel Core i7-8700 @3.20GHz with 62GB of RAM and Nvidia RTX 2080 Ti 11GB running a Linux operating system.

A. Datasets

In this work, we used three datasets widely employed to design and evaluate text localization and recognition methods, the ICDAR 2015 [4], MLT 2017 [5], and MLT 2019 [2] datasets, which are described in this section. Figure 3 illustrates examples from these datasets.

a) ICDAR 2015: This dataset [4] contains $1,500$ images, $1,000$ training images and 500 testing images. The images were captured by Google glasses and contain texts with different orientations, blurred, or with low resolution. The annotations were built in terms of quadrangle word bounding boxes.

Fig. 3: Examples of images from ICDAR 2015 (left), MLT 2017 (center), and MLT 2019 (right) datasets.

b) MLT 2017: This dataset [5] comprises $18,000$ images containing text from nine languages, $2,000$ images per language, including Arabic, Bangla, Chinese, English, French, German, Italian, Japanese, and Korean. In total, this dataset contains $9,000$ training images and $9,000$ testing images.

c) MLT 2019: This dataset [6] contains $10,000$ training images and $10,000$ testing images containing scene images with text in 10 languages, $1,000$ images per language, including Arabic, Bangla, Chinese, Devanagari, English, French, German, Italian, Japanese and Korean.

B. Experimental Setup

The training phase of PixelLink network was performed by using input RGB images with 512×512 pixels, a learning rate of 10^{-3} and a batch size of 8. We also used the Online Hard Example Mining (OHEM) method [7] to select negative pixels in order to have a negative-positive pixels ratio of 3. Finally, we set $\lambda = 2.0$, in Equation 1, in order to explicitly give more importance for the pixel-wise labelling task. PixelLink was implemented in Python using the TensorFlow framework.

C. Experiment 1: Text Detection Considering a Multi-Lingual Training.

This experiment aims to verify the performance results of PixelLink network in a multi-lingual scenario. In [3], the authors conducted experiments considering a dataset with two languages, English and Chinese. In this experiment, we went further in this analysis by using the MLT 2019 dataset. Initially, we performed a pretraining of the network, considering the ICDAR 2015 dataset, with 110 epochs. Next, we performed a fune tuning using the training set of MLT 2019 dataset, also with 110 epochs. Table I shows the performance results considering the testing images of 10 languages.

TABLE I: Performance results of PixelLink network trained and evaluated with images containing text in 10 languages.

Precision	Recall	F-measure
61.00%	53.69%	57.11%

From the results obtained in this experiment (see Table I) and from an analysis of success and failure cases of this model, we could observe that PixelLink was able to detect several text candidate regions partially correct. In several cases, the methods performed a text-line detection (see Figure 4), which decreased the performance results of the method since the ground truth provided along with datasets used in this work provide a word-based annotation.

D. Experiment 2: Text Localization via Language-Specific Model.

The results obtained in the previous experiment motivated us to investigate a training schema considering a specific language. We hypothesize that PixelLink network was not able to properly encode specificity of languages provided in the MLT 2019 dataset such as symbols, links between characters and word and character spacing. In order to verify our hypothesis, we first chose a language that would be our basis for our implementation. We chose Arabic language since it presented many errors of bounding box adjustment.

In this experiment, we also used the $1,000$ training image from MLT 2017 dataset that contains Arabic text, since MLT 2019 dataset provides only $1,000$ images with Arabic text for training. As in the previous experiment, we also performed a pre-training in the ICDAR 2015 dataset. Table II shows the performance results considering testing images containing only Arabic text and Figure 5 shows visual examples achieved with the classification model obtained from this experiment.

TABLE II: Performance results of PixelLink network trained and evaluated with images containing only Arabic text.

Precision	Recall	F-measure
90.17%	76.88%	83.00%

Table III shows a comparison between models estimated in a multi-language and language-specific training scenarios, considering the testing images containing only Arabic text. We could observe that the language-specific model outperforms the model built on a multi-lingual training scenario, with an increase of 11.43%, in terms of F-measure. Thus, we can state that, to address text detection problems with multiple languages, it is possible to obtain better results using an individual training for detection.

TABLE III: Comparison between two strategies adopted for training the PixelLink network.

Model	Precision	Recall	F-measure
Multi-lingual Model	77.32%	66.19%	71.33%
Language-Specific Model	92.11%	75.13%	82.76%

IV. Conclusions and Future Work

In this work, we investigated strategies to build a Convolutional Neural Network-based approach, referred to as PixelLink network, for the text localization problem in a multi-lingual scenario. We compared two methods for training a classification model: (i) performing a multi-lingual training of PixelLink network; and (ii) performing a language-specific training stage, considering the Arabic language. From the conducted experiments, it is possible to conclude that language-specific models are a proper choice for deploying more accurate models to operate in multi-lingual scenarios.

Fig. 4: Examples of bounding boxes predicted by the method trained over multi-lingual scenario (green) and their respective ground-truth annotation (blue).

Fig. 5: Examples of bounding boxes predicted by a language-specific classification model (green) and their respective ground-truth annotation (blue).

Future research efforts will focus on training language-specific models for the remaining languages available on MLT 2019 dataset to improve the performance results in all languages. In addition, we will investigate methods to integrate the outcomes from multiple CNN-based models to improve the overall results on this dataset.

ACKNOWLEDGEMENTS

The authors are grateful to Samsung R&D Institute Brazil, FAPESP, CNPq and CAPES for their financial support. This research was partially supported by Samsung Eletrônica da Amazônia Ltda., through the project "Algoritmos para Detecção e Reconhecimento de Texto Multilíngue (MLTSR)", within the scope of the Informatics Law No. 8248/91.

REFERENCES

[1] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast Oriented Text Spotting with a Unified Network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5676–5685, 2018.

[2] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character Region Awareness for Text Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374.

[3] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting Scene Text via Instance Segmentation," *ArXiv*, vol. abs/1801.01315, 2018.

[4] 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 *13th International Conference on Document Analysis and Recognition*, Aug. 2015, pp. 1156–1160.

[5] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, W. Khlif, M. Luqman, J.-C. Burie, C.-L. Liu, and J.-M. Ogier, "ICDAR 2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT," in *14th IAPR International Conference on Document Analysis and Recognition*, 11 2017, pp. 1454–1459.

[6] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J.-C. Burie, C. lin Liu, and J.-M. Ogier, "ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition - RRC-MLT-2019," *ArXiv*, vol. abs/1907.00945, 2019.

[7] A. Shrivastava, A. Gupta, and R. Girshick, "Training Region-Based Object Detectors with Online Hard Example Mining," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp. 761–769.