World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:12, 2023

On-Road Text Detection Platform for Driver Assistance Systems

Guezouli Larbi, Belkacem Soundes

Abstract—The automation of the text detection process can help the human in his driving task. Its application can be very useful to help drivers to have more information about their environment by facilitating the reading of road signs such as directional signs, events, stores, etc. In this paper, a system consisting of two stages has been proposed. In the first one, we used pseudo-Zernike moments to pinpoint areas of the image that may contain text. The architecture of this part is based on three main steps, region of interest (ROI) detection, text localization, and non-text region filtering. Then, in the second step, we present a convolutional neural network architecture (On-Road Text Detection Network - ORTDN) which is considered as a classification phase. The results show that the proposed framework achieved ≈ 35 fps and an mAP of ≈ 90%, thus a low computational time with competitive accuracy.

Keywords—Text detection, CNN, PZM, deep learning.

I. Introduction

NOWADAYS, most of the work on object detection is based on CNNs. Many applications now use technologies such as text extraction and recognition from scene images, text-to-speech converters etc. Scene images are complex due to their non-uniform background. In addition, other complex problems arise with this type of image, such as shadows, light reflection, etc.

In this contribution, an efficient two-step model is presented to reduce computation time and increase accuracy. We summarize our contributions as follows:

- A proposal for a textual area segmentation method based on Pseudo Zernike moments. PZMs are invariant to noise, scale, rotation, perspective distortion and translation, making the method robust to camera motion, perspective distortion, and luminance conditions.
- The use of PZMs allowing both the detection and the segmentation of textual information on low-resolution frames directly and without any pre-processing step. This will also allow its use for all types of videos.
- Proposal of a scheme based on deep learning and its corresponding design methodologies.

Our proposal achieves state-of-the-art results at a real-time speed on the RoadText-1K dataset with 98.97 % mAP running at 35 FPS on a High Performance Computing cluster with a 2 CPU's Intel Xeon Gold 14 Cores, DDR4 128 Gb, and 4 GPU's Nvidia V100, HBM2 16 Gb, CUDA Cores 5120, Tensor Cores 640 (HPC cluster machine).

The rest of the paper is organized as follows: Section II covers the most widely used techniques in literature for text detection. Section III focuses on the proposed approach, namely Two Steps On-Road Text Detection (2SORTD). This section presents how to use PZMs to extract candidate text regions. Then, a proposed OnRoad Text Detection Network (ORTDN) is detailed and used to classify candidate text regions into two classes: text and non-text. Section IV presents the performance measures used to evaluate our model. Finally, the conclusion is given in Section V.

L. Guezouli is with LEREESI, Laboratory of Renewable Energy, Energy Efficiency and Smart Systems, of the Higher National School of Renewable Energies, Environment and Sustainable Development, Batna, Algeria (e-mail: larbi.guezouli@hns-re2sd.dz).

S. Belkacem is with University of Batna2.

II. Related Work

Recently, there has been a growing interest in DL in the field of text detection [1], [2]. In [3], authors applied the You Only Look Once (YOLO v2) object detection system to text detection in natural scenes with the optimal parameters. They carry out the regression analysis of the coordinate parameters and categories of bounding boxes. Wang et al. [4] propose a novel intentional Detection Network (ADN) to extend the semantic information of feature maps. They propose to add a branch of layers to the classical detection network.

In [5], the authors present an improvement in the results of Devanagari ancient characters recognition using SIFT and Gabor filter feature extraction techniques. Support vector machine classifier is used for the classification task in this work.

However, the above-mentioned detection algorithms do not consider the tilt of the camera and the direction of the text which can be horizontal, vertical or diagonal. To detect the text in different directions and with different sizes, we thought of Pseudo-Zernike Moments. In the next section, we present our PZM-based proposal, where we combined PZMs and CNN to complement the failures of CNNs with the benefits offered by PZMs.

III. Two Steps On-Road Text Detection (2SORTD)

In order to solve the above mentioned issues, we propose in this paper a convolutional neural network fed by segmented images based on Pseudo-Zernike Moments. Our proposal, named 2SORTD, is a two-stage solution. The first stage is a filtering phase based on PZMs to extract candidate text regions. The second stage is a decision phase based on CNN. This stage makes the final decision on whether a candidate region is textual or not.

A. Candidate Text Region Extraction

On the road, reading the text is of great importance. It allows us to get more information about the environment around us.

Our proposed solution faces a significant challenge. To make things easier, we first need to segment the image to

filter out non-text regions to minimize the text search area [6], [7]. Nevertheless, the acquired videos are usually of average quality. At first, we need to locate and extract the candidate text area and eliminate all other parts of the frame. To locate the text area in an RGB frame, we compute the local PZMs features on a sliding window, and then we use the k-means algorithm to perform the clustering. Finally, these local characteristics are grouped together according to their similarities to form global characteristics. The orthogonality of PZMs allows obtaining values close to zero, which gives a small vector dimension and a reduced computation time.

PZMs have several interesting characteristics:

1) Invariance to rotation;
2) Less sensitivity to noise [8];
3) Expressiveness: Minimal information redundancy. The orthogonality property of PZ polynomials leads to almost zero information redundancy;
4) Multi-level representation: Moments of different orders refer to different characteristics. Low-order moments capture general details, while higher-order moments capture more local information;
5) Efficiency: The number of moments calculated for order and more than that of other orthogonal moments, such as Zernike moments.
6) Image reconstruction.

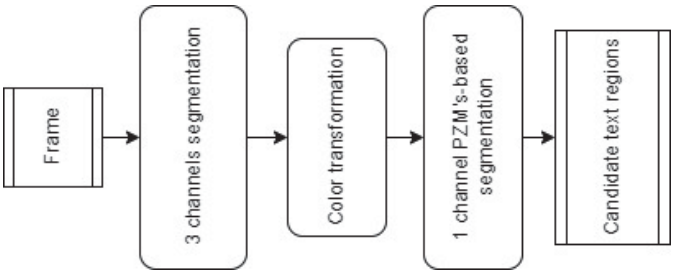

Fig. 1 PZM-based candidate text region extraction

To extract the text zone, several steps are performed:

1) Image partitioning and normalization: The frame is decomposed into its three components R, G and B. Each image I of size $N \times M$ will be divided into parts of size $W \times W$. The number of partitions $NBblock$ is given by:
   - $NBwidth = \frac{N}{W}$
   - $NBlength = \frac{M}{W}$
   - $NBblock = NBwidth \times NBlength$

   Each partition is located by the coordinates (x, y) of its upper left corner where $x \in [0, NBlength - 1]$ and $y \in [0, NBwidth - 1]$. The intensity of each pixel (x_i, y_j) is given by the equation:

$$f^{x,y}(x_i, y_j) = f(W_x + x_i, W_y + y_j) \quad (1)$$

   The parameter W is fixed experimentally. The value $W = 6$ gives the best ratio (quality of description / execution time). The coordinates of the pixels are converted to polar space, where each partition is represented in a unit circle in order to avoid the loss of information since all pixels are taken into account for the calculation of the moment.
2) Features extraction: for each part resulting from the first step, the PZMs are calculated. Among the existing algorithms, we use the recursive algorithm of Papakostas et al. [9] for its speed of calculation. Fig. 1 explains the details of the process.

   The image is put in HSV mode in order to use the V channel. Then, segmentation will be applied on this channel, according to the PZMs (see Fig. 2) to locate the regions that have a high probability of containing text as shown in Fig. 3.
3) Candidate regions are used as input for our convolutional neural network.

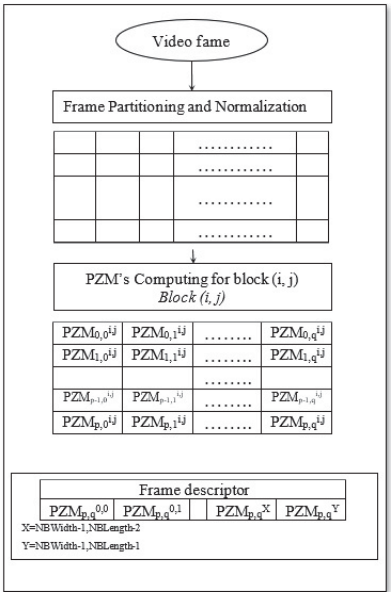

Fig. 2 Extracting features of an image from the V channel based on PZMs

Fig. 3 Example of results: (a) Original frame, (b) Segmented frame

B. OnRoad Text Detection Network (ORTDN)

It is worth mentioning that a state-of-the-art model based on neural networks has been proven in the field of computer vision, especially in object detection, shape detection, etc. To this end, we propose a convolutional neural network model comprising different types of layers: Conv2D layers, BatchNormalization layers, ReLu layers, MaxPooling2D layers, Flatten layers, and finally Dense and Sigmoid layers.

In this section, we will introduce our OnRoad Text Detection Network (ORTDN) in detail. It is the nub of our proposal. After the stage of filtering frames to extract candidate text zones, we will use the results of this stage as input to our neural network. We closely looked at existing neural networks used in this field and inspired a network that fits road text detection. It consists of three bottleneck residual blocks and two dense layers.

The proposed model architecture (see Fig. 4) contains the initial fully convolution layer with 32 filters using a 5×5 size kernel and stride = 2, followed by 2 residual bottleneck layers. The output of the map filters is then flattened and followed by a dense layer with ReLU activation function, then a second dense layer with a sigmoid activation function for classification. The output of this network is two classes: text or non-text. In our bottleneck residual blocks, we use the ReLU activation function as the non-linearity due to its robustness and a kernel size of 3×3 as is standard for modern networks. We use batch normalization during training.

IV. Performance Measures

A. RoadText-1K Dataset

There are not many datasets intended for road text detection. The best one is RoadText-1K [10] which is 20 times larger than the largest existing dataset for text in videos. It includes 1000 video clips of 10 seconds duration with annotations for text bounding boxes and transcriptions in each frame. These video clips are collected with a variety of weather conditions, including sunny, overcast, rainy, as well as at different times of the day, including day and night.

B. Experimental Results

Among the datasets cited in the previous section, we have chosen RoadText-1K dataset and COCO-Text dataset for their large contents. We carry out our experiments on an HPC cluster with a 2 CPU's Intel Xeon Gold 14 Cores, DDR4 128 Gb, and 4 GPU's Nvidia V100, HBM2 16 Gb, CUDA Cores 5120, Tensor Cores 640. Detailed experimental results on the RoadText-1K dataset are shown in Table I where we find that our model achieves state-of-the-art performance with an mAP of 98.97% after 200 epochs.

TABLE I
Comparison of 2SORTD with Other Models by RoadText-1K Dataset

Models	Precision	Recall	Accuracy (MAP)	F1-Measure
Fast Text Detection for Road Scenes (M1) [11]	62.43	61.54	52.13	61.99
Natural scene text detection based on YOLO V2 (M2) [3]	61.24	73.14	60.98	66.67
Single Shot Text Detector with Regional Attention (M3) [12]	88.27	**86.48**	86.72	**87.37**
Temporally-aware Convolutional Block Attention Module for Video Text Detection (M4) [13]	64.10	41.28	22.04	50.22
FREE Video Text Spotter (M5) [14]	63.20	43.39	26.37	51.39
2SORTD	**96.96**	78.49	**98.97**	86.75

The performance comparison of the training accuracy and the test accuracy is given in Table II.

TABLE II
The Performance Comparison of the Training Accuracy and the Test Accuracy by RoadText-1K Using 200 Epochs

Models	Training accuracy	Training time (hours)	Test accuracy
Fast Text Detection for Road Scenes [11]	0.909	-	0.675
Natural scene text detection based on YOLO V2 [3]	0.914	-	0.701
Single Shot Text Detector with Regional Attention [12]	0.902	-	0.681
Temporally-aware Convolutional Block Attention Module for Video Text Detection [13]	0.756	-	0.621
FREE Video Text Spotter [14]	0.710	-	0.592
2SORTD	**0.983**	23	**0.857**

Experimental results on the RoadText-1K dataset are compared in Table I in terms of precision, recall and f1-measure. The precision represents the number of correctly classified objects out of the number of positively classified objects [15] For unbalanced learning, recall is typically used to measure the coverage of the minority class. F1-measure, which weights precision and recall equally, is the most commonly used variant when learning from unbalanced data. To analyze this result, we plot them together in Fig. 5 as a Bar Chart to get a full picture of the performance variation between the different models. The classification performance of the proposed model is summarized in the form of a confusion matrix shown in Fig. 6.

From Table I and Fig. 5, we can easily see the variation of results. For Recall, 'Single Shot Text Detector with Regional Attention' model achieves the best result with 86.48 followed by 2SORTD, which achieves the second-best result with 78.49. For Precision, 2SORTD model achieves the best result with 96.96 followed by 'Single Shot Text Detector with Regional Attention' which reached a good score of 88.27. For Accuracy, 2SORTD model achieves the finest result with 89.97 followed by 'Single Shot Text Detector with Regional Attention' with 86.72. For F1-Measure, 'Single Shot Text Detector with Regional Attention' model achieves the greatest result with 87.37 followed by 2SORTD with 86.75. 2SORTD has achieved the best results in terms of Accuracy and the F-Score, which is considered the metric indicating the harmony between Precision and Recall.

Results are summarized in the form of a confusion matrix shown in Fig. 6 where we can see that true negatives and true positives are the best values.

To understand the progress of our Neural Network, we use the Accuracy curve shown in Fig. 7 (a) and Loss curve shown in Fig. 7 (b). Overfitting is the difference between learning precision and validation precision. A large variance indicates a large overfitting. [16]. Images are trained 200 times. The calculation of the training accuracy and the test accuracy is given in Fig. 7 (a). As we can see in Fig. 7 (a), the training results have converged.

Our model obtains significant performance improvements and outperforms recent work such as the work of Wu et al. [17] whose solution does not converge after 30000 epochs.

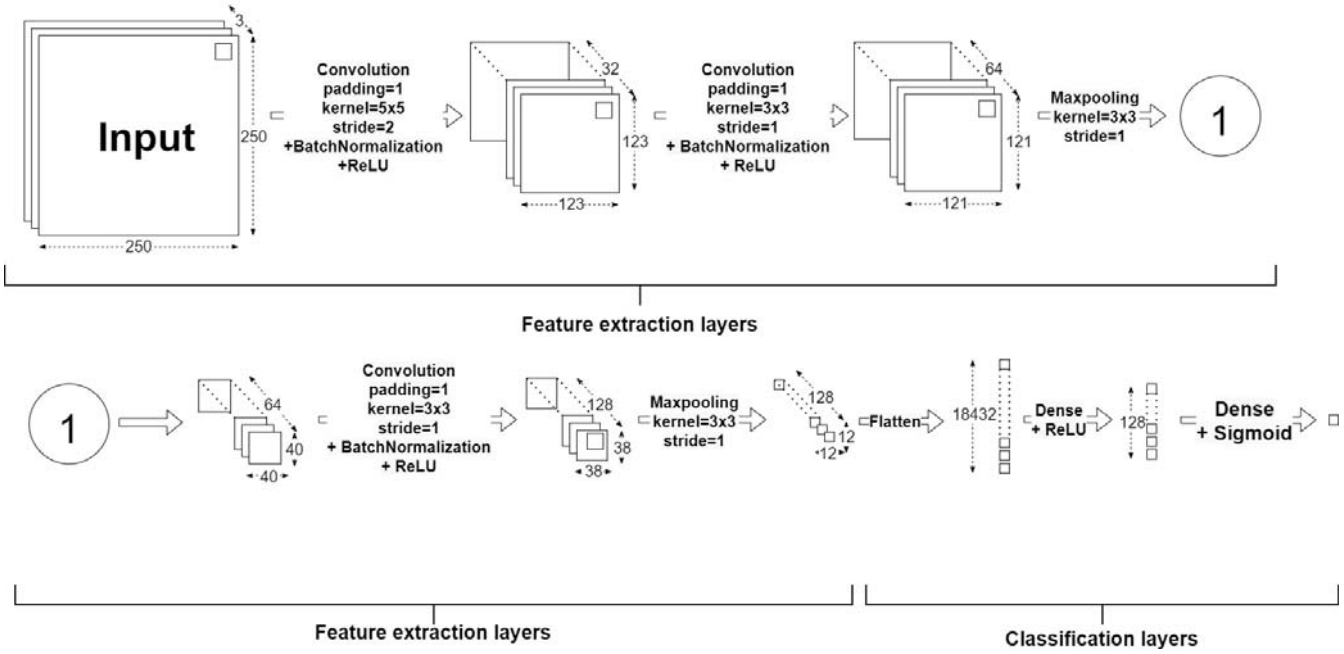

Fig. 4 Proposed CNN model (ORTDN)

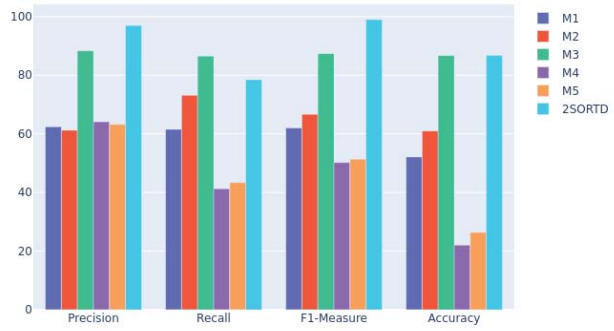

Fig. 5 Comparison with existing models

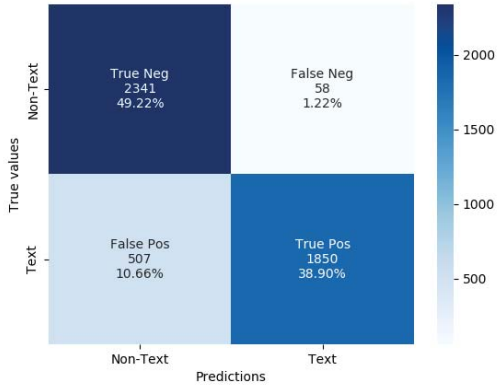

Fig. 6 Classification results presented in a confusion matrix

C. Detection Speed

To calculate the detection speed, we use a set of 1000 images, which we submit to our trained system. We calculate the time of 'Candidate text region extraction' phase (presented in Fig. 1). On GPU, our system reaches an average of 0.025 seconds per image which is equivalent to *35 fps*.

V. Conclusion

In this paper, we proposed an efficient Two Steps On-Road Text Detection called 2SORTD. We used Pseudo Zernike Moments in the filtering phase before the OnRoad Text Detection Network (ORTDN). Our neural network (ORTDN) contains layers for feature extraction and layers for classification. To address the problem of text detection which tends to be time-consuming, we present a lightweight convolutional neural network. We improve the feature maps in the backbone network, which effectively improves the detection performance. On the RoadText-1K Road Text benchmark, we achieved state-of-the-art performance and outperformed previous methods. Our model can be applied to other types of detection architectures, such as object detection, and in existing environmental monitoring tasks, such as meta-learning for few-shot soot density recognition, monitoring and forecast, image-based smoke detection, etc. It may be useful for future research.

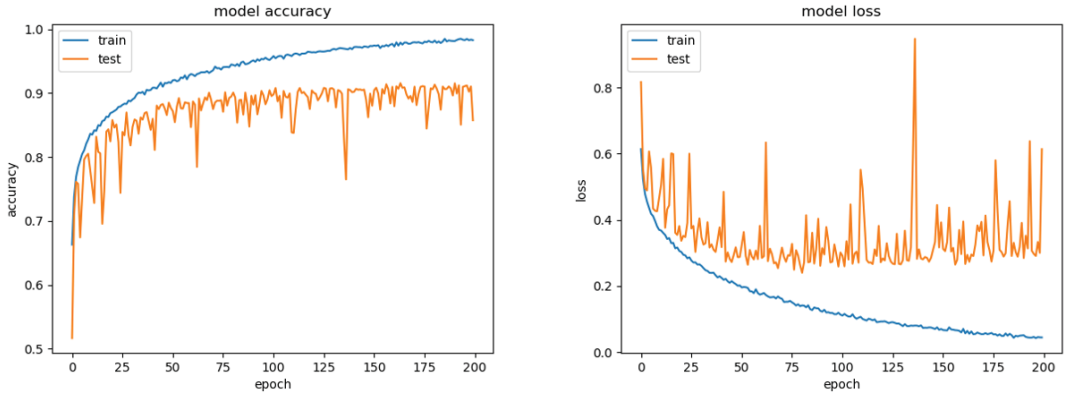

Fig. 7 Calculation of training and testing accuracy and loss

REFERENCES

[1] N. Anwar, T. Khan, and A. F. Mollah, "Text detection from scene and born images," in *Recent Trends in Communication and Intelligent Systems*, A. K. S. undir, N. Yadav, H. Sharma, and S. Das, Eds. Singapore: Springer Nature Singapore, 2022, pp. 115–122.

[2] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[3] D. Haifeng and H. Siqi, "Natural scene text detection based on yolo v2 network model," *Journal of Physics: Conference Series*, vol. 1634, p. 012013, sep 2020.

[4] J. Wang, H. Hu, and X. Lu, "Adn for object detection," *IET Computer Vision*, vol. 14, no. 2, pp. 65–72, 2020.

[5] S. R. Narang, M. K. Jindal, S. Ahuja, and M. Kumar, "On the recognition of devanagari ancient handwritten characters using sift and gabor features," *Soft Computing*, vol. 27, no. 22, pp. 17 279–17 289, 2020.

[6] M. Sravani, A. Maheswararao, and M. K. Murthy, "Robust detection of video text using an efficient hybrid method via key frame extraction and text localization," *Multimedia Tools and Applications*, 2020. [Online]. Available: https://doi.org/10.1007/s11042-020-10113-2

[7] Z. Liu, W. Zhou, and H. Li, "Scene text detection with fully convolutional neural networks," *Multimedia Tools and Applications*, vol. 78, no. 13, pp. 18 205–18 227, 2019. [Online]. Available: https://doi.org/10.1007/s11042-019-7177-4

[8] X. Wang and L. min Hou, "A new robust digital image watermarking based on pseudo-zernike moments," *Multidimens. Syst. Signal Process*, vol. 21, no. 2, pp. 179–196, 2010.

[9] G. A. Papakostas, Y. S. Boutalis, D. A. Karras, and B. G. Mertzios, "Efficient computation of zernike and pseudo-zernike moments for pattern classification applications," *Pattern Recognition and Image Analysis*, vol. 20, pp. 56–64, 3 2010.

[10] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas, and C. V. Jawahar, "Roadtext-1k: Text detection & recognition dataset for driving videos," in *2020 IEEE International Conference on Robotics and Automation, (ICRA) 2020, Paris, France, May 31 - August 31, 2020*. IEEE, 2020, pp. 11 074–11 080.

[11] V. Toro and M. Alejandro, "Fast text detection for road scenes," Master's thesis, Department of Computer Science, University of Applied Sciences Bonn-Rhein-Sieg, Bonn-Rhein-Sieg, 5 2015.

[12] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3066–3074.

[13] M. Fujitake and H. Ge, "Temporally-aware convolutional block attention module for video text detection," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021, pp. 220–225.

[14] Z. Cheng, J. Lu, B. Zou, L. Qiao, Y. Xu, S. Pu, Y. Niu, F. Wu, and S. Zhou, "Free: A fast and robust end-to-end video text spotter," *IEEE Transactions on Image Processing*, vol. 30, pp. 822–837, 2021.

[15] C. Fernandez, *Learning from Imbalanced Data Sets*, 1st ed. Springer; 1st ed. 2018 edition, 11 2018.

[16] G. V. Jose, "Useful plots to diagnose your neural network," https://towardsdatascience.com/useful-plots-to-diagnose-your-neural-network-521907fa2f45, 10 2019, accessed by: 26-12-2020.

[17] Z. Wu and S. He, "Improvement of the alexnet networks for large-scale recognition applications," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 2020.