International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 2, April 2021, pp. 1634~1640
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i2.pp1634-1640

Real-time Arabic scene text detection using fully convolutional neural networks

Rajae Moumen, Raddouane Chiheb, Rdouan Faizi
National School of Computing and System Analysis (ENSIAS), Mohammed V University, Morocco

Article Info

Article history:

Received Nov 28, 2019
Revised Aug 18, 2020
Accepted Oct 13, 2020

Keywords:

Arabic text detection
Convolutional neural networks
Natural language processing
Scene text detection

ABSTRACT

The aim of this research is to propose a fully convolutional approach to address the problem of real-time scene text detection for Arabic language. Text detection is performed using a two-steps multi-scale approach. The first step uses light-weighted fully convolutional network: TextBlockDetector FCN, an adaptation of VGG-16 to eliminate non-textual elements, localize wide scale text and give text scale estimation. The second step determines narrow scale range of text using fully convolutional network for maximum performance. To evaluate the system, we confront the results of the framework to the results obtained with single VGG-16 fully deployed for text detection in one-shot; in addition to previous results in the state-of-the-art. For training and testing, we initiate a dataset of 575 images manually processed along with data augmentation to enrich training process. The system scores a precision of 0.651 vs 0.64 in the state-of-the-art and an FPS of 24.3 vs 31.7 for a VGG-16 fully deployed.

Corresponding Author:

Rajae Moumen
National School of Computing and System Analysis
Mohammed V University
Avenue Mohamed Ben Abdellah Regragui, Rabat, Morocco
Email: rajae.moumen@um5s.net.ma

1. INTRODUCTION

Scene text detection stands for localizing in an accurate way existing text area in a scene in natural environment with natural conditions. Technically, it consists on localizing the bounding box of each textual element. There are many challenges in this task; first of all, the heterogeneity of the scene, that may contain, several texts in different scale range, multi-oriented and curved; Image quality that may present blur, lack of light, etc.

Scene text detection has received an increasing amount of interest through last years, as it is a mandatory step for scenes comprehension and analysis used in many applications such as video and image understanding, visually impaired assistance, automatic driving, robot sensing, etc. The focus of research in this area is to produce a transcription of scene texts with a high accuracy and a reasonable velocity in order to evolve towards a real-time accurate text detection and recognition.

Main studies concern English language, and few researches focus on Arabic scene text processing. Actually, for many years, the focus of research around Arabic was on text detection in close material, such as documents, printed or handwritten at a hand-held distance range. The prominent paradigm in this area is segmentation of texts, then words and identification of single characters, making the process a classification problem. This model showed its limits first because of the lengthy process, and the multi-ambiguous characteristic of Arabic representation. Other segmentation free models showed up later, such as hidden markov models (HMM) and neural networks resulting in speeding up processing time and achieving better accuracy. More recently, the spectrum of text analysis expanded to computer vision applications such as text detection and recognition from images, scenes and videos. However, real-time scene text processing requires a high capacity approach in terms of processing functionality, efficiency and latency. In addition to memory efficient and fast processing hardware, making it difficult for mobile devices to host such applications.

For many years, the traditional sliding window-based methods were widely used [1, 2]. It consists on moving a multi-scale window over an image and determine if the content is text or non-text. Another popular approach is the connected component (CC) as used in [3, 4], an algorithm designed to extract maximally stable extremal regions (MSERs) as character candidates. The higher probabilities of text candidates corresponding to non-text are estimated with a character classifier. These methods obtain modest results because the task is divided to several steps, text line detection, and character classification. Each step with an error rate, the inferred over-all error rate is significant. Recently, text detection is mainly undertaken with deep learning approaches. These methods can be divided into three sections: Regression-based methods, segmentation-based methods and hybrid methods.

Regression-based methods use the bounding box concept considering text elements as object and treating text detection as an object detection problem. For example, work in [5] and TextBoxes [6] that predicts the text box by applying a fully convolutional network. TextBoxes++ [7] uses quadrilateral regression for text detection. EAST [8] use pixel-level regression for detection multi-oriented texts. These methods present the inconvenient to render modest results when it comes to curved text. Regression-based methods work in two methods, one-stage method and two-stage methods; the two-stage method considers a second step for refining the results of first step. In literature, the two-steps stage achieves better results than the one-step methods.

Segmentation-based methods like [9-11] usually proceed by segmenting the background and using pixel-level prediction. Example in PSENet [12], authors propose a novel approach to detect text with arbitrary shapes, PSENet generates different scales of kernels for each text instance and gradually expands the kernel to the complete shape of the text instance, study in [13] proposes a novel segmentation based on cascaded convolution neural networks. Authors in [14] propose to perform text detection with a deep approach using connected components.

Arabic text detection research works are limited; Authors in [15] propose a hybrid approach for Farsi/Arabic text detection and localization in video frames, where the image is divided into macro blocks and fed into support vector machine (SVM) classifier to categorize them into text and non-text group. In [16], authors use a CNN-RNN hybrid architecture by transcribing convolutional features from the input image to a sequence of target labels. In [17], a convolutional neural network is used as a deep classifier to detect scene characters; the network is trained with distinct learning rates. In [18], a deep fully convolutional networks (FCN) multi-oriented system for real-time text detection. In [19], authors propose a deep scene text detector for Arabic text detection.

In this paper, we apply a multiscale approach to perform Arabic scene text detection. We use the idea behind the approach proposed in [20] for Arabic language, a two-steps approach; first a novel network, scale-based region network (SPRN) a multi scale framework, aiming to eliminate non-textual elements of the scene image, providing textual areas and estimation of scale range of each element. A fully convolutional network is then used to determine narrow range text areas. Text detection is processed in two-steps for multi textual areas scenes. The advantage of this method is its speed in comparison with other multi-scale approaches making it more suitable with real-time processing. The second step uses a fully convolutional neural network to localize text in an accurate way since the output of the first step contains no noise. We show that the two-steps method scores better runtime results than one-step approaches in literature related to Arabic processing, comparable results to English scene text detection and better results than a one-step approach based on VGG-16.

2. RESEARCH METHOD

To our knowledge, Arabic text detection has never been undertaken with two-steps approaches. Therefore, we propose to segment text detection into two steps. First, detecting wide scale range text, which means eliminating non-textual areas, estimate the scale of each text instance, we will call this first step text block detection and the first network TextBlockLocalizer; we feed the resulting images to a text detector we name TextDetector, which role will be detecting narrow scale range text. Figure 1 describes the overall process of text detection. In this work, we assume that:

- Elements of a text block have same orientation, size and font
- Bounding boxes are quadrilateral, but not necessarily rectangle

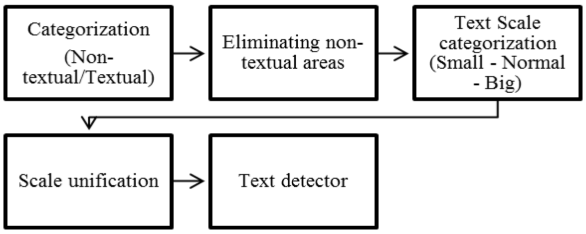

Categorization (Non-textual/Textual)
Eliminating non-textual areas
Text Scale categorization (Small - Normal - Big)
Scale unification
Text detector

Figure 1. Process of text detection framework

2.1. Implementation details

First step is performed with a fully convolutional network (FCN) named TextBlockLocalizer, an adaptation of VGG-16 with a series of modifications to keep the network light-weighted with a regard on efficient runtime. VGG is used for many reasons, first it considers local and global information, it is trained end-to-end and it has shown its efficiency in pixel labeling. First, TextBlockLocalizer feature extraction is derived from VGG-16 with 3*3 convolution layers stacked straightforward. A serie of modifications are processed to the network in order to optimize runtime, such as equal channel width to increase forward speed. Slowing operations such as excessive group convolution, element-wise operations and network fragmentation are optimized to increase the network speed. Number of kernels is drastically reduced to 16 kernels, the quarter in VGG-16. We rely on experimental studies in [21], that use less parameters and performs random initialization for scene feature extraction. For feature fusion, a 1*1 convolution is used to normalize channel width, and a deconvolution layer to up-sample spacial resolution. TextBlockLocalizer FCN performs two tasks:

First task is localizing text blocks, the network deals it as a classification problem by performing categorization of textual and non-textual regions; this step aims to filter out non-textual areas, that consists noisy and slowing elements for the detection tool. Text localization is performed as a classification problem at a pixel-level. Non-textual areas are filtered out and this step outputs wide scale range textual instances. We make the presumption that pixels within the bounding box are positive text instance; first, because the rendered output is to be refined in the second step and in either ways, regions between characters are different in contrast to non-textual instances.

Second task is to render scale estimation of each text instance; the framework affects to each text instance three possible values (Big–Normal–Big). In a regressive way, category Cat of a bounding text box β is determined with the following equation taking into consideration bounding box dimensions:

$$Cat(\beta) = \begin{cases} Small & if\ L(\beta) < T_s \\ Normal & if\ L(\beta) \in [T_s, T_n] \\ Big & if\ L(\beta) > T_n \end{cases}$$

T_s Standing for the superior threshhold to text of small scale
T_n Standing for the superior threshold to text of normal scale

Second step is the TextDetector; it is also an FCN with the task of detecting text in an accurate way. It is based on VGG-16 fully deployed following the design described in Figure 2.

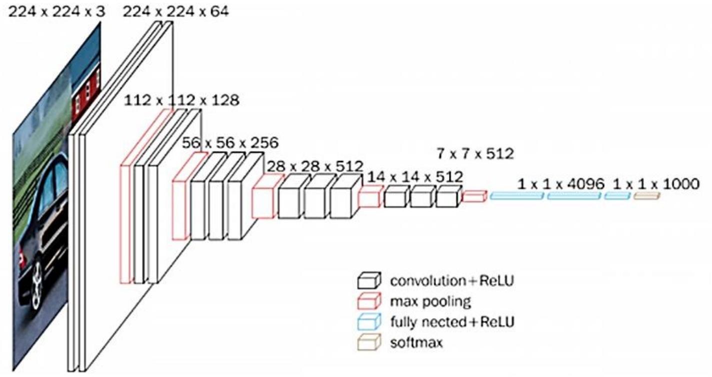


Figure 2. VGG-16 architecture

2.2. Dataset

We count several datasets aimed for handwritten Arabic or off-line printed documents such as APTI [22] and few datasets for Arabic text detection and recognition in videos i.e. AcTiv [23], ARASTI [24]. To our knowledge, there is a single dataset EASTR [25] for Arabic scene text detection, unfortunalty it is not available publically. For this reason, we initiated a dataset for the needs of this research work. In Table 1, statistics of the dataset:

Table 1. Dataset statistics

Attribute	Number
Images	575
Textual instances	762
Words	1120
Percentage of curved text	20.86%
Percentage of images presenting blur	10.08%
Percentage of images presenting missing/hidden characters	11.82%
Number of fonts	41
Percentage of images with no text	14.8%

2.3. Technical environment

We built manually the associated ground truth corresponding to images, and determined the bounding box of each word in semi-autmatic mode. The scale category is determined as described in 2.1 (small, normal, big). Because of the limited number of images in the dataset, we exploit data augmentation by using online augmentation for two reasons: first created images are not stored, hence do not have additional memory requirements; and second system does not go through same image twice. For this purpose, we use Keras library image data generator [26] with specific configuration to avoid unnecessary transformations in order to maintain realistic natural images, for example no vertical or horizontal flip needed, zoom range limited to twice the initial size.

The two networks are trained on the dataset, TextBlocklocalizer to categorize textual and non-textual images and give scale estimation, and TextDetector to detect text in an accurate way. The databases are divided into training and test sub-sets. We exploit ADAM optimizer [27] to update network weights instead of stochastic gradient for many reasons; first, Adam optimizes computation, requires light memory, and optimizes training in the case of noisy data problems, as it is the case in text detection. All development is made with Python on Keras.

3. RESULTS AND DISCUSSIONS

Figure 3 shows some results of the proposed framework in various situations and scene configuration. We notice that system performs well and gives satisfying results. To demonstrate the strength on the proposed framework in performance and time cost improvement, we perform text detection with a VGG-16 fully deployed. The network is trained and tested on same dataset and we confront results of both approaches. Table 2 shows precision/Recall/and F-measure and average time cost for the proposed framework in comparison with best-recorded studies (to our knowledge) and the VGG-16 results:

- Best results in English scene text detection
- Best results in Arabic scene detection
- VGG-16 performance

The proposed framework achieves satisfying results in performance, outperforming state-of-the-art performance in Arabic text detection. For time cost, comparison between a VGG-16 time cost and the proposed framework shows a significant improvement equivalent to 23.34%. However, performance is to be improved in some cases. Figure 4 (a-f) shows examples that the framework failed to process, they present respectively problems of calligraphy, small font size, curved text, vertical text, missing/unclear character, high blur.

Table 2. Performance comparison of detection approaches

Algorithm	Precision	Recall	F-measure	Time cost (fps)
Proposed framework	0.651	0.714	0.68	24.3
VGG-16	0.661	0.752	0.69	31.7
Arabic best performance [19]	0.64	0.72	0.70	N.A
English best performance [20]	0.7698	0.8485	0.8072	16.5

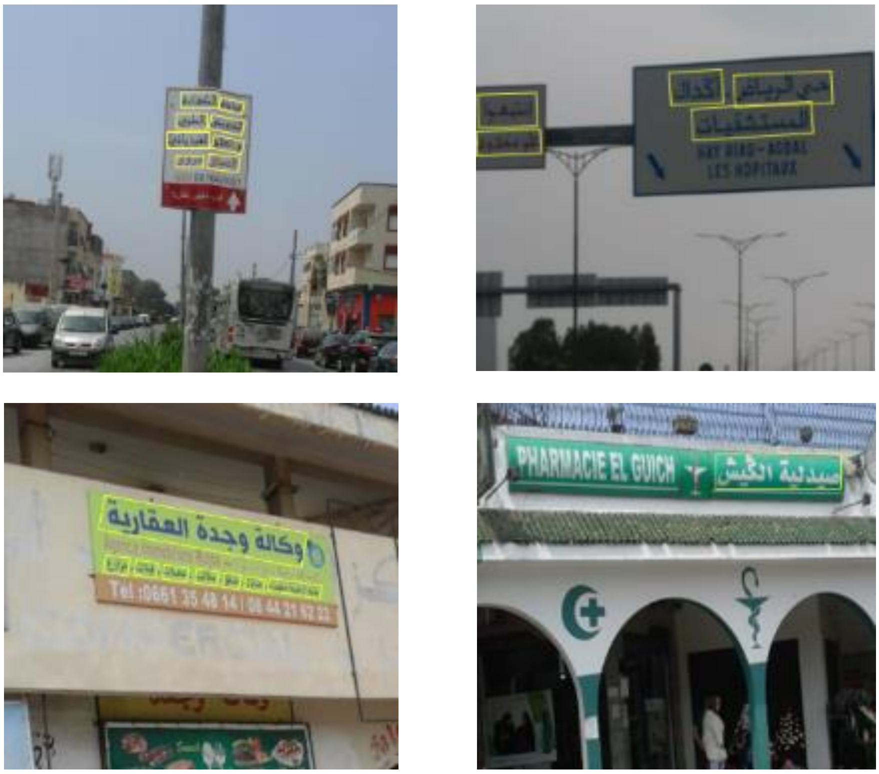
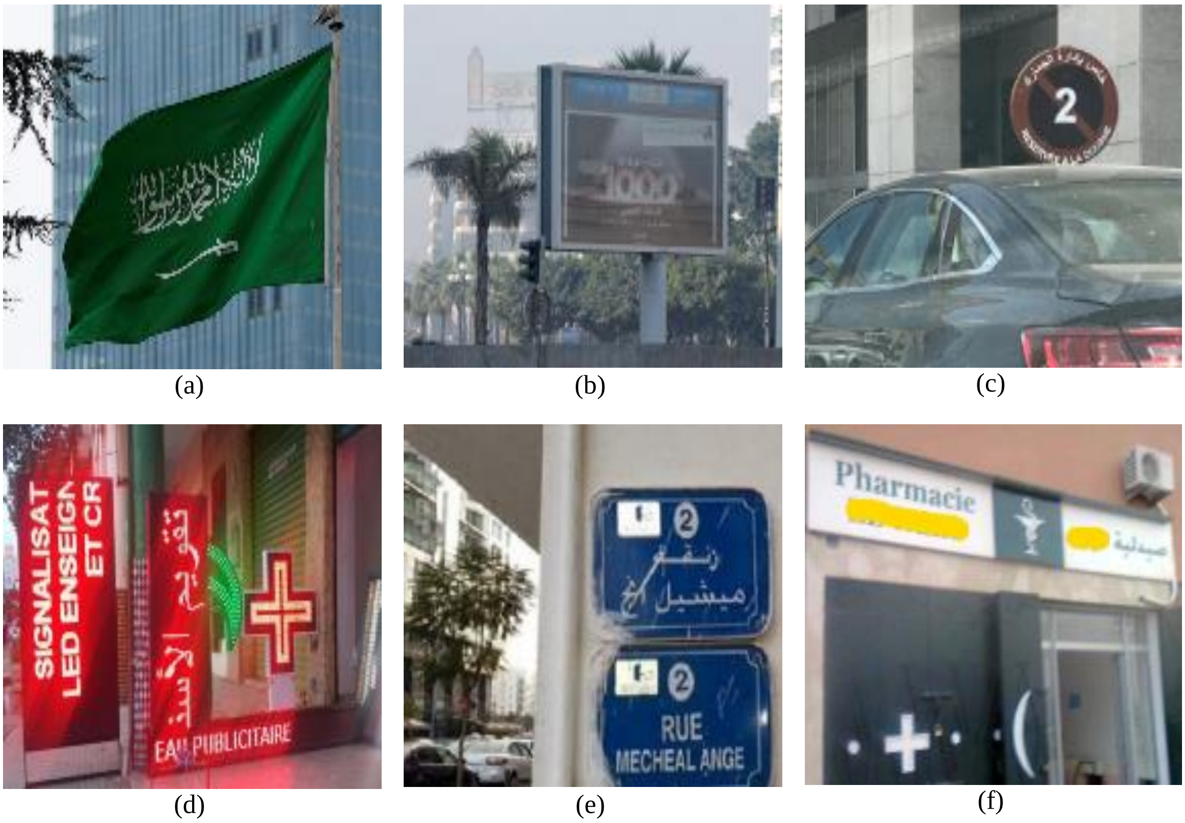


Figure 3. Example images form dataset


(a) (b) (c)

(d) (e) (f)

Figure 4. Images where text detection failed, (a) problems of calligraphy, (b) small font size, (c) curved text, (d) vertical text, (e) missing/unclear character, and (f) high blur

4. CONCLUSION

In this paper, we presented a novel approach to deal with real-time Arabic scene text detection using a two-steps approach instead of a single one, we confronted the results to a one-time framework based on VGG-16 and our approach gives satisfying results with optimized runtime. However, the framework is to be improved in following aspects: First, the framework does not resolve curved text detection, second, a single language could be processed, in our case Arabic, information in multilingual scenes is not fully retrieved. Finally, yet importantly, the lack of Arabic datasets for scene text detection and recognition is a big brake to achieving better results thus building a large dataset for this purpose is an urgent matter.

REFERENCES

[1] T. Shangxuan, et al., "Text Flow: A Unified Text Detection System in Natural Scene Images," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1-9.

[2] Z. Zhang, et al., "Symmetry-based text line detection in natural scenes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 2558-2567.

[3] S. Lei, et al., "A robust approach for text detection from natural scene images," *Pattern Recognition*, vol. 48, no. 9, pp. 2906-2920, 2015.

[4] Y. Xu-Cheng, et al., "Robust Text Detection in Natural Scene Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970-983, 2014.

[5] M. Liao, et al., "Real-time Scene Text Detection with Differentiable Binarization," *arXiv: 1911.08947*, 2019.

[6] M. Liao, et al., "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," *arXiv: 1611.06779*, 2016.

[7] M. Liao, et al., "TextBoxes++: A Single-Shot Oriented Scene Text Detector," *IEEE Transactions on Image Processing*, vol. 27, pp. 3676-3690, 2018.

[8] X. Zhou, et al., "EAST: An Efficient and Accurate Scene Text Detector," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551-5560.

[9] D. Deng, et al., "PixelLink: Detection scene text via instance segmentation," in *the thirty-second AAAI Conference of Artificial Intelligence*, 2018.

[10] L. Shangbang, et al., "TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes," *European Conference on Computer Vision ECCV*, 2018, pp. 19-35.

[11] Z. Zhang, et al., "Multi-Oriented Text Detection with Fully Convolutional Networks," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4159-4167.

[12] W. Wang, et al., "Shape Robust Text Detection with Progressive Scale Expansion Network," *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9336-9345.

[13] Tang, Yuan et al., "Scene Text Detection and Segmentation Based on Cascaded Convolution Neural Networks," in *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, vol. 26, no. 3, pp. 1509-1520, 2017.

[14] J. Fan, et al., "Deep Scene Text Detection with Connected Component Proposals," *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1-10.

[15] M. Mohieddin and M. Saeed, "Hybrid approach for Farsi/Arabic text detection and localisation in video frames," *IET Image Processing*, vol. 7, no. 2, pp. 154-164, 2012.

[16] M. Jain, et al., "Unconstrained scene text and video text recognition for Arabic script," in *1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 1-5, 2017.

[17] S. B. Ahmed, et al., "Deep learning based isolated Arabic scene character recognition," in *1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 1-6, 2017.

[18] M. S. H. Sassi, et al., "Multi-Oriented Real-Time Arabic Scene Text Detection with Deep Fully Convolutional Networks," in *16th International Conference on Computer Systems and Applications*, Abu Dhabi, United Arab Emirates, 2019, pp. 1-6.

[19] I. Beltaief and M. B. Halima, "Deep FCN for Arabic Scene Text Detection," in *IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition*, London, pp. 129-134, 2018.

[20] W. He, et al., "Realtime multi-scale scene text detection with scale-based region proposal network," *Pattern Recognition*, vol. 98, 2020.

[21] W. He, et al., "Multi-Oriented and Multi-Lingual Scene Text Detection with Direct Regression," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5409-5419, 2018.

[22] F. Slimane, et al., "A New Arabic Printed Text Image Database and Evaluation Protocols," in *10th International Conference on Document Analysis and Recognition*, Barcelona, 2009, pp. 946-950.

[23] O. Zayene, et al., "A dataset for Arabic text detection, tracking and recognition in news videos-AcTiV," in *13th International Conference on Document Analysis and Recognition*, Tunis, 2015, pp. 996-1000.

[24] M. Tounsi, et al., "ARASTI: A database for Arabic scene text recognition," in *1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 140-144, 2017.

[25] S. B. Ahmed, et al., "A Novel Dataset for English-Arabic Scene Text Recognition (EASTR)-42K and Its Evaluation Using Invariant Feature Extraction on Detected Extremal Regions," *IEEE Access*, vol. 7, pp. 19801-19820, 2019.

[26] "Image data preprocessing," *Keras*, [Online]. Available: https://keras.io/api/preprocessing/image/.

[27] P. K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," *CoRR*, 2014.

BIOGRAPHIES OF AUTHORS

Rajae Moumen is a PhD student at the National School of Computer Science and Systems Analysis at Mohammed V University, Rabat, Morocco. She has a degree of engineer in Computer sciences from the same school and has 10 years experience as IT manager. Her research interests are in the area of neural networks, natural language processing, especially Arabic.

Raddouane Chiheb is a professor of applied mathematics at the National School of Computer Science and Systems Analysis at Mohammed V University, Rabat, Morocco. He obtained his Master from the National Institute of Applied Sciences of Lyon and PhD from the Jean Monnet University of Saint-Etienne. His research interests are in the area of Semantic Analysis, Structural Optimization, Education, Optimization of the logistics chain, Machine learning, and Value Analysis. He supervised over 10 students. Prof. Raddouane Chiheb is President of the Moroccan Association for the Value Analysis

Rdouan Faizi has a Ph.D. in linguistics from Mohammed V Agdal University, Rabat, Morocco. Currently, he teaches English at Ecole Nationale Supérieure d’Informatique et d’Analyse des Systèmes (ENSIAS), Mohammed V Souissi University.