International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Localizing Text in Images and Videos based on Morphology

Mohamed Amin Ben Atitallah, Rostom Kachouri, Hassene Mnif

Abstract*: Many multifaceted images comprise observable text. If the occurrences of this text can be identified, segmented, and recognized automatically, they will be a prized source of high-level semantics; for retrieval and indexing. In this paper, we will propose a novel method for localizing and detecting text in complex images and video frames based on morphology. A morphological Gardient is generated by computing the variance between the dilation and the erosion image. Then the candidate of regions are connected via a morphological closing operation and every text areas are determined used the occurrence of text in each candidate. The identified text regions are localized perfectly via the projection of the text pixels in the morphological Gardient map. This method is sturdy to different position, character size, color and contrast. The updating of the text region between images is also used to minimize the processing time. Tests are realized on divers images to confirm the good efficient of our method.

***Keywords* : object detection, object segmentation, morphology Gardient operator, text detection, text segmentation, video processing.**

I. INTRODUCTION

Videos and images in databases and on webs are growing. Newscasters are explaining interest in construction large digital records of their assets to use again the record materials for the programs of TV, on line disponibility to the general public and other companies. To answer this demand there is necessity to systems that are capable to supply an efficient retrieval and indexing by content of image segments using the extraction of the information of content level related with visual-data. Although efficient content based retrieval of visual data of video frames is determined by backing content impersonation through low level frame features; the similar doesn't apply to content based retrieval of frames, expecting a limited application settings. In fact, the videos effective retrieval must be based on high level content descriptor [1].

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author

Mohamed Amin Ben Atitallah, Laboratory of Electronics and Information Technology (E.N.I.S.), University of Sfax, PHD student at National Engineering School of Gabes (ENIG), University of Gabes, TUNISIA. E-mail: benatitallahmohamedamin@yahoo.fr.

Rostom Kachouri, LIGM, Univ Gustave Eiffel, CNRS, ESIEE Paris, F-77454 Marne-la-Vallée, France. E-mail: rostom.kachouri@esiee.fr.

Hassene Mnif, National school of electronics and telecommunications of Sfax, Laboratory of Electronics and Information Technology (E.N.I.S.), University of Sfax, TUNISIA. E-mail: hassene.mnif@enetcom.usf.tn.

Most video tend to rise the utilization of text to transmit more direct outline of deliver and semantics improved viewing experience. In this context, headlines resume the reports in subtitles and news videos in drama documentary to help viewers to comprehend the content. Also, the sport videos include text indicating the player names, team and scores [2]. Generally, we can classify the text in video frames into overlay text and scene text [3].

The text presented in scene occurs by nature in the contextual as a part of the video scene. For example, the banners, advertising boards, etc. But on the contrary, overlay text in the scene of video used to aid viewers to comprehending, since it is highly structured and compact, it can be used also for frame video retrieval and indexing [4]. In addition, the extraction of text from frame video for OCR (Optical Character Recognition) will be more defying compared to extraction of text from document images for OCR, because of many difficulties resultant from size, unknown color of text and complex background.

The remainder of our paper is prepared as follows. The section 2 presents the related work, followed by the detection region of the text, which is indicated in section 3. Finally, experiments are given in section 4 followed by conclusion in section 5.

II. RELATED WORK

A Many existing methods of the detection text from video have been presented based on texture based feature, edge and color. Color based approaches suppose that the text of video is created of a unified color. Agnihotri and Dimitrova [5] proposed a method to detect and binarize horizontal yellow, black and white caption text in the frames of video. Afterward; the edge detector find the edge pixels with an immovable threshold. The regions of frame with high density of edge are considered very noisy for the extraction of text. The analyze of the linked component is accomplished on the edge pixel of residual regions. The components of edge are integrated based on structural heuristics to localize the regions of the text. The binarization is completed by the application of thresholding at the average value of pixel of every localized region of text. Xuemei Zhao and al [6] cluster color using an Euclidean-distance in the RGB-space and based on 64 clustered color channel for the detection of text. Nonetheless, the video text is rarely true that it consists of a uniform-color because of the degradation resulting low contrast and compression coding between background and text. Since the regions of text contain many edge information, the edge based approaches are considered useful for the detection of text from video.

Retrieval Number: A2598059120/2020©BEIESP
DOI:10.35940/ijrte.A2598.059120
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This method is applied to detect the edge from the frame of video and to identify the regions that have a high edge of strength and density. The proposed method is efficacy just for the simple background but in the complex background it becomes less reliable. Sumathi and al [7] utilize a modified edge-map to detect the region of text and to localize it utilizing coarse to fine projection. In addition, they extract also the strings of text used an inward filling generality and local thresholding.

Yu Liu and al [8] present a novel method using an edge map proposed by Sobel continued by geometrical constraints, smoothing filter and the morphological operation.

The wavelet transform and the salient point detection; have been utilized to detect the regions of text. Lamberto Ballan and al [9] propose to detect the corner points from the scene of video and after that they detect the region of text utilizing similarity of points of corner between frames. Baisakhi Sur Phadikar and al [10] utilize the features of texture from the coefficients of DCT to detect the text from MPEG\JPEG compressed domain. Firstable, they reveal the blocks of the candidates of text that have a high horizontal of spatial-intensity-variation, and after that, they refine the detected candidates into different areas by spatial limitations. The vertical spectrum of energy checks the potential caption of regions of text. In any case, its strength in complex-background may its not be fulfilling for the constraint of a spatial features domain.

The next step of the recognition of text from video frame is the extraction of text which be applied before the application of the OCR. The methods of the extraction of text can be categorized into method based on color [11] and method based on stroke [12], since there is a different color between the background of the frame and the text; the application of thresholding allows to extract the text from the video frame. The Otsu method[11] is a robust method for the extraction of text due to the efficiency and the simplicity of its algorithm, it utilized the color based method for the extraction of the text. Nevertheless, the method of Otsu isn't efficient for the extraction of text from video frame which have same color with the background causing the application of thresholding in all frame. To tackle this issue, the identified text areas are splitted into a different squares and afterward Otsu strategy is applied locally to each square, for example, in [7] the author defined a dam point using the adaptive thresholding to separate the background from the text. There are a few of channels that have been utilized to extract the text in the strategies based on stroke, this channels dependent essentially on the course of strokes. To improve the stroke as the shapes and to suppress others, the author utilized the character extraction filters [12] for four directions. Nonetheless, a few of characters or words without evident stripe shape can likewise be stifled, since the filtre of stroke is language reliant.

In the rest of this work, we present a novel method to detect the text from video frame utilizing the transition of the area between the background and text. Firstable, we create the morphological Gardient dependent on our perception that there is a transient hues among the adjacent background and text. After that, the regions of text are generally recognized by calculating the texture on every side of the transition of pixel and transition density of pixel.

III. DETECTION REGION OF TEXT

Our method is based-on our observation that the colors of contrast is exist between the adjacent background and the text. The proportional contrast between the text and its background is a very important feature to detect the text from the region. The Fig.1 shows the overall procedure of our proposed method to detect the text.

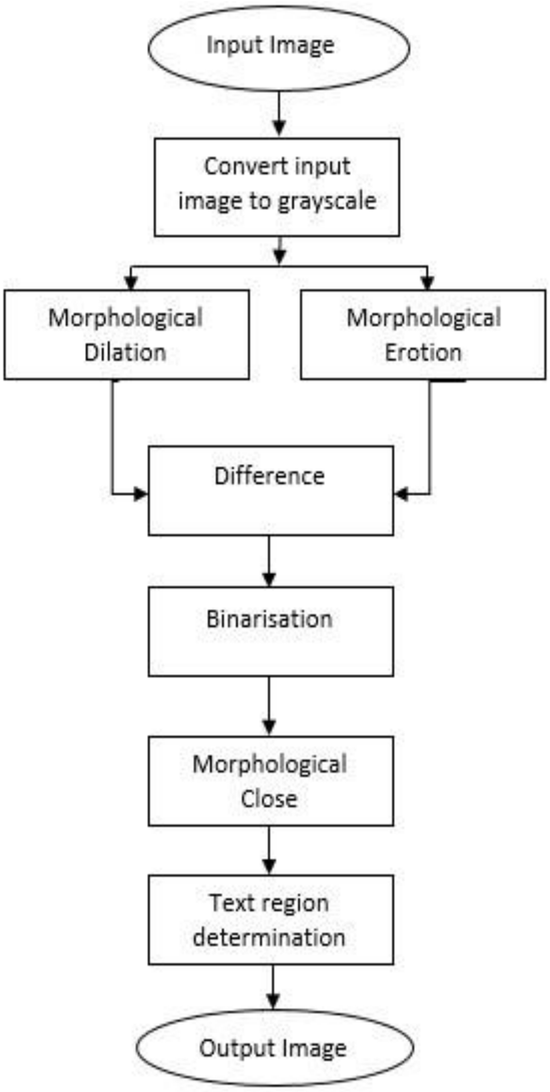


Fig. 1. Overall procedure of the proposed detection method.

A. Morphological Gardient

So as to recognize the regions of text from complex background a morphology based methodology is utilized to remove high contrast feature [13].

Let I(x,y) signify a gray level of the input of image. Let Sm,n mean an organizing element with size m×n; where m,n are chances and bigger than zero. In addition, let ⊕ signify a dilation operation, and □ signify an erosion operation.

Dilation Operation:

$$I(x,y) \bullet S_{m,n} = I(x,y) \oplus S_{m,n} \qquad (1)$$

Erosion Operation:

$$I(x,y) \circ S_{m,n}= I(x,y) \square S_{m,n} \qquad (2)$$

Difference:

$$D(I_1,I_2)= I_1(x,y)\text{-}I_2(x,y) \qquad (3)$$

To acquire the morphological map, dilation (1) and erosion (2) operations are performed utilizing a disk auxiliary element $S_{3,3}$.

The difference (3) got from subtracting the two images are the consequence of the next step.

The Fig 2 presents the entire strategy of our morphology based procedure to extract the features of the contrast.

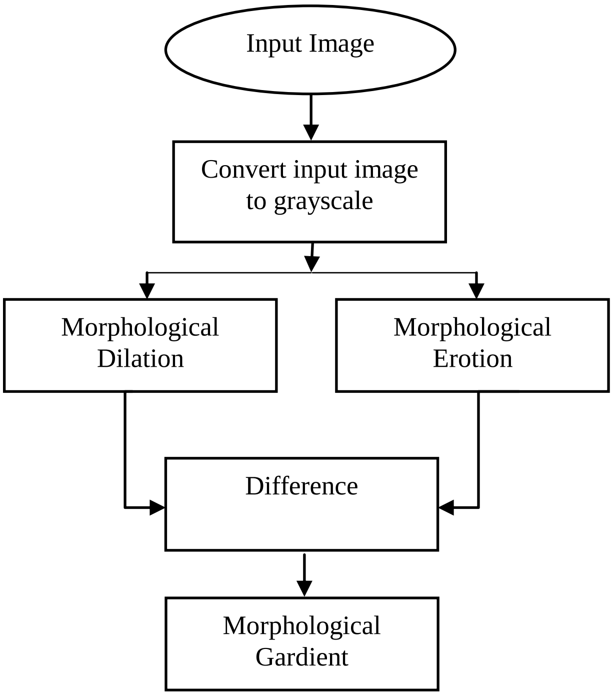


Fig. 2. Flowchart to extract the features of contrast for detection regions of text.

The Fig 3(b) shows an experiment result of the current process.

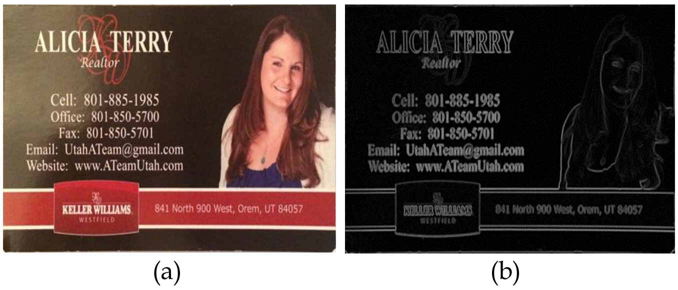


Fig. 3. Generation of morphological Gardient (a) Input image (b) Morphological Gardient image.

B. Thresholding

A threshold Otsu operation (4) is applied trailed by a marking procedure to extract the segments of text. In the threshold the background of the image. This parameter is dependable to decide the limit value of the binarization operation.

Thresholding:

$$T(I(x,y))= \begin{cases} 255, \text{ if } I(x,y) > T \\ 0\ ,\ otherwise \end{cases} \qquad (4)$$

The Fig 4(b) shows an experiment result of the current process.

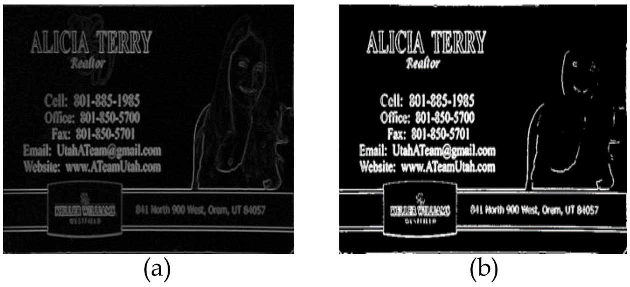


Fig. 4. Generation of binarized image (a) Morphological Gardient image (b) Binarized image.

C. Extraction region candidate

The close regions can be easily associate based on the morphological closing operator (5) but for the regions whose its positions are far away to one another leave confined. In this part of the section, we apply the morphological dilation operator [14] on the image of the previous step. The application of this step is done for square of 9 x 1 of the image to get joint zone-referred to as writings blobs. On the off chance that a hole of back to back pixels among two nonzero points in a similar column is shorter than 5% of the picture width, they are loaded up with 1s. On the off chance that the associated components are littler than the value of the threshold, they are deleted. The value of the threshold is experimentally chosen by watching the minimum size of the region of the text. At that point, each associated component is reshaped to have a smooth limits. Since it is reasonable to acknowledge that the districts of content are generally in rectangular shapes, a rectangular ricocheting box is made by associating four focuses, which compare to (minmum_i, minmum_j), (maximum_i, minmum_j), (minmum_i, maximum_j), (maximum_i, maximum_j) taken from the text-blobs. The refined regions of candidate are appeared in Fig. 5(b).

Let I(x,y) signify a gray level of the input of image. Let Sm,n mean an organizing element with size m×n; where m,n are chances and bigger than zero. In addition, let ⊕ signify a dilation operation, and □ signify an erosion operation.

Closing operation:

$$I(x,y) \bullet Sm,n=(I(x,y) \oplus Sm,n) \square Sm,n \qquad (5)$$

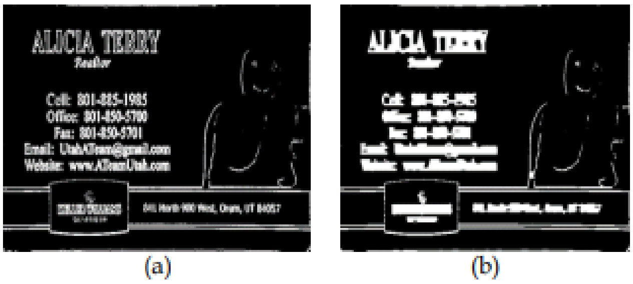


Fig. 5. Extraction of the regions of candidate (a) Binarized image (b) Connected components through closing.

D. Extraction region candidate

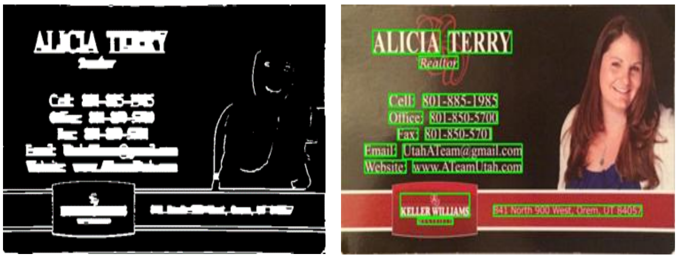
The subsequent stage is to decide the real area of text between the limit smoothed candidate areas by a few valuable clues. The vertically longer candidates can be eliminate easily, since a large portion of writings are put horizontally in the frames of video. In view of the perception that variation of intensity around the change pixel is huge because of the structure of the content of the text is very complex, so we utilize the prevailing nearby double example (DLBP) acquainted in [15] with portray the surface around the progress pixel. DLBP adequately catch the ruling examples in surface pictures. Dissimilar to the customary LBP approach, which just adventures the uniform LBP [16], given a surface picture, the DLBP approach figures the event frequencies of all revolution invariant examples characterized in the LBP gatherings. These examples are then arranged in plummeting request. The initial a few most as often as possible happening examples ought to contain ruling examples in the picture and, accordingly, are the prevailing examples.

LBP is a proficient and basic device to speak to the consistency of surface utilizing just the power design. LBP structures the twofold example utilizing current pixel and its all round neighbor pixels and can be changed over into a decimal number as shows the equation (6):

$$LBP_{P,R} = \sum_{i=0}^{P-1} s(g_i - g_c)2^i, \text{ where } s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (6)$$

Where, P and R signify the user's picked number of round neighbor pixels of a particular pixel and the radius of circle respectively. gc and gi mean the intensity of the current-pixel and circular neighbor-pixels respectively.

In figure 6 we can see that the region of text is well identified from other candidates.

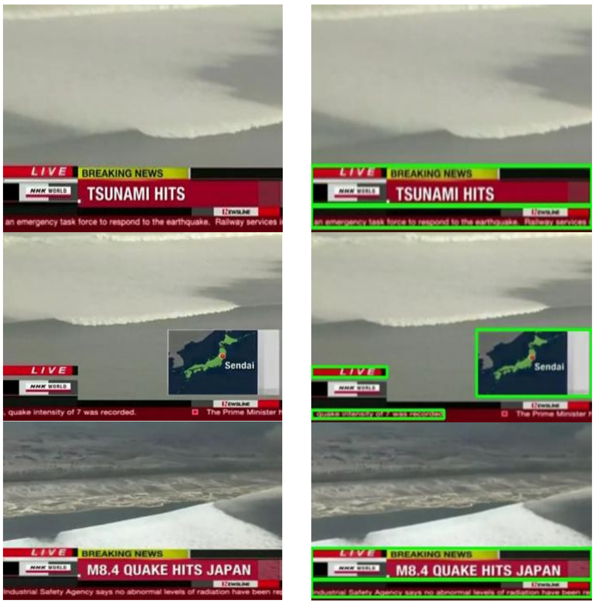


(a) (b)

Fig. 6. Determination of the region of text.

IV. Experimental Results

Our method has been tried on different images video cam and a videos for a real life. Our algorithm have been implemented in visual studio C++ on an Intel center I7 PC with 2.70 GHz CPU. We made our database comprising of various videos arrangements of with size 640 × 480 pixels. The most of content occasions were stationary and all of content had a horizontal orientation. The database consists of videos imported from the TV. These videos include news TV, televised advertisements, televised football matches etc... . The videos chosen are very complex, they are made up of different colors, of different text and of different languages. The sequences of video were caught at 20 frames per second.



(a) (b)

Fig. 7. Experimental results of scene text detection and overlay text (a) original frames (b) Result.

It is difficult to detect the text scene because some type of text have a different size and irregularly aligned. The Fig 7 presents the detection of the scene text in different video frames.

V. Conclusion

Content installed in recordings regularly conveys the most significant data, such.as time, spot, name or subjects, and so on. This data may do incredible assistance to video ordering and video content comprehension. We proposed a new strategy to de detect the text from complex frames of video. Our identification strategy depends on the observation if the colors of contrast is exist between the adjacent background and the text. The generation of the morphological map is obtained by the difference among opening and closing frame. For each region of candidate, the connected components are produced and afterward each linked component is reshaped. The predominant local binary pattern is utilized to discover the variation of intensity around transition of pixel. The limits of the regions of the detected text are localized precisely utilizing the projection of the pixels of text in the morphological map. To prove the good performance of our proposed method to detect the text, different videos and images have been tried. The proposed strategy is valuable for the real time application.

ACKNOWLEDGMENT

We thank our colleague in ESIEE Paris to supported this work.

REFERENCES

1. CeyhunCelik, Hasan SakirBilge, “Content based image retrieval with sparse representations and local feature descriptors : A comparative study”. Pattern Recognition Volume 68, August 2017, Pages 1-13.
2. Jonas NygaardBlom, Kenneth Reinecke Hansen,“ Forward-reference as lure in online news headlines,” Journal of Pragmatics Volume 76, January 2015, Pages 87-100.
3. Haojin Yang, Bernhard Quehl, and Harald Sack, “A framework for improved video text detection and recognition,” Multimedia Tools and Applications volume 69, pages217–245(2014).
4. Weiming Hu ; Nianhua Xie ; Li Li ; Xianglin Zeng and Stephen Maybank, “A Survey on Visual Content-Based Video Indexing and Retrieval,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) (Volume: 41 , Issue: 6 , Nov. 2011).
5. Mauro Barbieri, Lalitha Agnihotri and Nevenka Dimitrova, “Method and device for automatic generation of summary of a plurality of images,” in United States, Jun. 19, 2012.
6. Xuemei Zhao, Yu Li and Quanhua Zhao, “Mahalanobis distance based on fuzzy clustering algorithm for image segmentation,” in Digital Signal Processing Volume 43, August 2015, Pages 8-16.
7. C.P. Sumathi1, T. Santhanam and G.Gayathri Devi3, “A SURVEY ON VARIOUS APPROACHES OF TEXT EXTRACTION IN IMAGES,” International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.4, August 2012.
8. Yu Liu, Changwen Zheng, Quan Zheng and Hongliang Yuan, “Removing Monte Carlo noise using a Sobel operator and a guided image filter”, The Visual Computer volume 34, pages589–601(2018).
9. Lamberto Ballan, Marco Bertini, and Alberto Del Bimbo, “Event detection and recognition for semantic annotation of video,” in Multimedia Tools and Applications volume 51, pages279–302(2011).
10. Baisakhi Sur Phadikar, Amit Phadikar, Goutam Kumar Maity, “Content-based image retrieval in DCT compressed domain with MPEG-7 edge descriptor and genetic algorithm”, Pattern Analysis and Applications volume 21, pages469–489(2018).
11. N. Otsu, “A threshold selection method from gray-levl histograms,”IEEE Trans. Syst., Man, Cybern., vol. 9, no. 1, pp. 62–66, Mar. 1979.
12. T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, “Video OCR for digital news archive,” in Proc. IEEE International Workshop on Content Based Access of Image and Video Libraries, Jan. 1998, pp. 52–60.
13. Jui-Chen Wu, Jen-Wei Hsieh and Yung-Sheng Chen“, Morphology based text line extraction”, pp.1195-1200, 2006.
14. Jagath Samarabandu and Xiaoqing Liu,” An Edge-based Text Region Extraction Algorithm for Indoor Mobile Robot Navigation”, pp.273-280. 2007.
15. S. Liao, Max W. K. Law, and Albert C. S. Chung,” “Dominant Local Binary Patterns for Texture Classification”, IEEE Transactions on Image Processing, Vol. 18, No. 5, May 2009.
16. Abdenour Hadid, Juha Ylioinas, and Messaoud Bengherabi, “Gender and texture classification: A comparative analysis using 13 variants of local binary patterns,” Pattern Recognition Letters Volume 68, Part 2, 15 December 2015, Pages 231-238.

AUTHORS PROFILE

Mohamed Amin Ben Atitallah received his Applied License in Computer and Master degrees from the National School of Electronics and Telecommunications of Sfax (ENET’Com) respectively in 2013 and 2016. Since 2016, he is a Ph.D. student in Image Processing System in the National School of Engineers of Gabes (ENIG) in collaboration with ESIEE Paris. From 2016 to 2020, Mohamed Amin Ben Atitallah was a temporary teacher at the Higher Institute of Industrial Management of Sfax (ISGIS).

Rostom Kachouri received his Engineer and Master degrees from the National Engineering School of Sfax (ENIS) respectively in 2003 and 2004. In 2010, he received his Ph.D. in Image and Signal Processing from the University of Evry Val d’Essonne. From 2005 to 2010, Dr. Kachouri was an Assistant Professor at the National Engineering School of Sfax (ENIS) and then at the University of Evry Val d’Essonne. From 2010 to 2012, He held a post-doctoral position as part of a project with the high technology group SAGEMCOM. Dr. Kachouri is currently Associate Professor in the Computer Science Department and Head of apprenticeship computer and application engineering at ESIEE, Paris. He is member of the Institut Gaspard-Monge, unité mixte de recherche CNRS-UMLPE-ESIEE, UMR 8049. His main research interests include pattern recognition, machine learning, clustering and Algorithm-Architecture Matching.

Hassene Mnif was born in Sfax, Tunisia, in 1975. He received the Dip Ing and Master in electrical engineering from the University of Sfax (ENIS) in 1999 and 2000, respectively, the Ph. D. degree in electronics from the University of Bordeaux I, France, in 2004 and the HDR degree from the University of Sfax in 2011. He is currently Professor and Director of the National School of Electronics and Telecommunications of Sfax, University of Sfax, where he has multiple innovative engineering education initiatives. He is a member of Electronic Communication Group in the Electronic and Information Technology Laboratory. His research interests include Energy Harvesting, Design of Radio-Frequency Integrated Circuits and Characterization and compact modeling of both high frequency devices and future emerging technologies like Carbon Nanotube Field Effect Transistor (CNTFET). He participates also in researches for real time image and video text extraction. He has authored and co-authored more than 80 journal publications and conference papers and has gathered significant scientific coordination experience within national and international collaborative research projects. He participated in the organization of several IEEE conferences and workshops, in particular ICECS 2009 and MELECON 2012. He is actually the IEEE Tunisia CAS chapter chair.

Retrieval Number: A2598059120/2020©BEIESP
DOI:10.35940/ijrte.A2598.059120
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication