Scene to Text Conversion and Pronunciation for Visually Impaired People

Saeed Mian Qaisar[†], Raviha Khan, Noofa Hammad
Electrical and Computer Engineering Department, Effat University, Jeddah, KSA
[†]sqaisar@effatuniversity.edu.sa

Abstract

The recent technological advancements are focusing on developing smart systems to improve the quality of life. Machine learning algorithms and artificial intelligence are becoming elementary tools, which are used in the establishment of modern smart systems across the globe. In this context, an effective approach is suggested for automated text detection and recognition for the natural scenes. The incoming image is firstly enhanced by employing Contrast Limited Adaptive Histogram Equalization (CLAHE). Afterward, the text regions of the enhanced image are detected by employing the Maximally Stable External Regions (MSER) feature detector. The non-text MSERs are removed by employing appropriate filters. The remaining MSERs are grouped into words. The text recognition is performed by employing an Optical Character Recognition (OCR) function. The extracted text is pronounced by using a suitable speech synthesizer. The proposed system prototype is realized. The system functionality is verified with the help of an experimental setup. Results prove the concept and working principle of the devised system. It shows the potential of employing the suggested method for the development of modern devices for visually impaired people.

Keywords *–Image Processing, Text Detection and Recognition, MSER Features Detector, OCR, Speech synthesizer.*

I. Introduction

Languages are the oldest way of communication between human beings whether they are in spoken or written forms. In the recent era, visual text in natural or manmade scenes might carry very important and useful information. Therefore, the scientists have started to digitize these images, extract and interpret the data by using specific techniques, and then perform text-to-speech synthesis (TTS). It is done in order to read the information aloud for the benefit and ease of the user. Text extraction and TTS can be utilized together to help people with reading disabilities and visual impairment to listen to written information by a computer system [1].

Image enhancement is an important stage to enhance the quality of the images that could be degraded due to the imperfections in the resolutions of the camera or the scanner used. The term enhancement involves denoising, de-blurring and enhancing the contrast of the image [2], [3].

In [4], [5], Authors have proposed systems that perform two types of detections which are edge detection and region detection. MSER is a region detector that is used to distinguish text regions from non-text regions by comparing the intensity of the text to the backgrounds.

According to [6], the OCR is one of the most widely used technique in recognition of text of natural scene images and videos. OCR works precisely if the image is free of noise and the background is clean.

TTS is also called speech syntheses. It is an artificial form of human beings speech, and a speech synthesizer is a computer system used for this purpose where the implementation is done on a software [7].

This paper proposes a system that takes an image as an input, then performs CLAHE to enhance the image quality. MSER is used to extract the text regions, and the non-texts regions are discarded by applying contour filters. MSER regions are grouped into words, recognized by the OCR. Finally, these words are pronounced by an effective text to speech synthesizer.

II. Materials and Methods

The proposed system block diagram is shown in Fig. 1. A system description is presented in the following.

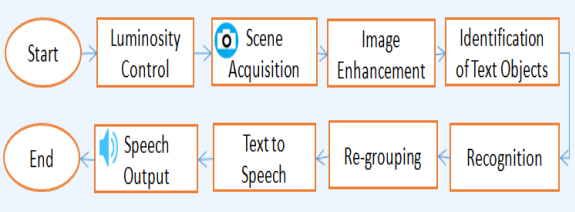

Fig. 1. The proposed system block diagram.

A) The Luminosity control and Image Acquisition

The luminosity of the environment is controlled with the help of an embedded controller. Arduino UNO is employed in this regard to control the motor based daylight saving mechanism [8]. If the current luminosity is inadequate then based on the time in the computer system a curtain should open. The circuit is realized by using a L298 Motor controller interfaced between the Arduino UNO board and a DC motor.

In case of night time or a dark day, an array of Light Emitting Diodes (LEDS) should simply turn on, again controlled by the Arduino UNO board. The images will be taken using a 2MP USB camera integrated with MATLAB by using the webcam toolbox [9].

B) Image Enhancement

The CLAHE is used to bypass the over-exaggeration of noise. CLAHE separates the subjected image into blocks that do not overlap; these blocks are called tiles and are enhanced individually rather than subjecting the whole

image to the change. it combines the excellencies of histogram equalization (HE) and histogram specification, hence the produced histogram closely mimics the probability distribution specified by the user. These are then combined to evade the inter tile edge by using the bilinear interpolation. These artifacts are artificially produced as a result of the enhancement. Contrast enhancement is kept constrained to keep the noise from amplifying [3], [10]. Let us consider that a certain bright image have its pixel values be limited in the high values, to make it fit the criteria of a good image, the Histogram Equalization (HE) is applied. In simple words, what Histogram Equalization does is that it tweaks the contrast of the image. The process is clear from Fig. 2.

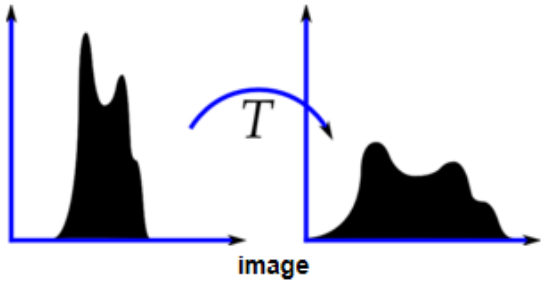

Fig. 2. Histogram Equalization transformation.

C) Text Detection using MSER

The MSER regions are areas that have a relatively distinct intensity compared to their background contrast. They are isolated through a process of attempting numerous thresholds. The regions that preserve constant shapes over a wide range of thresholds are selected (cf. Fig. 3) [11].

Segmenting the text from a scene via MSER intensively helps as a preprocessing step for optical character recognition (OCR). Once the text regions are detected, the other non text regions are dropped. MSER is compatible with text due to the constant color and high contrast with the background, which together give us stable intensity profiles. However it is highly likely that a number of non text regions that are stable are also selected. To remove these non-text regions the stroke-width is considered. Text characters tend to have little variation when it comes to stroke widths of the lines and curves, whereas non text areas display a high stroke width variance [12].

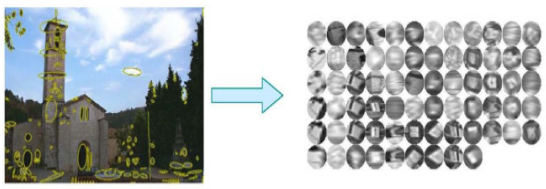

Fig. 3. An example of MSER regions.

D) Classification and Grouping

After using the MSER the individual alphabetical characters are identified. Afterwards, these are regrouped into words and sentences to complete the text recognition task. For example the string “word” must be recognized for it to be useful instead of unordered individual characters such as (‘o’,’w’,’d’,’r’). An approach to achieve this is to first determine nearby regions of text and then enclose them with a box. To find these nearby regions we expended the obtained bounding boxes. The result is that the bounding boxes of all such ad joint text regions overlap such that regions that make up one word or phrase make a series of overlapping bounding boxes. After that, all such bounding boxes are combined together to form complete words or phrases. A graph is employed to observe this overlapping ratio. A MATLAB based specifically developed application is used to calculate the pair overlap ratios for all the existing expanded bounding boxes and then by using the obtained graph all the joint regions are identified. Indices of the overlapping text regions are produced that further have bounded boxes. These indices are used to combine several nearby bounding boxes into one bounding box by determining the smallest and largest values of the separate bounding boxes that the connected component consists of. Lastly, before displaying the final detected text, the bogus text detections are removed as bounding boxes consisting of lone text regions which are probably not text regions since usually text is found as a group of characters. Thenceforth, we use the OCR to identify the text inside all the bounding box. If we have not first detected the text regions using MSER, the output of the OCR would have been much erroneous [12].

F) Speech Synthesis

The MATLAB based Text-to-Speech synthesizer, proposed by W. Garn is employed to pronounce the detected text [13].

III. RESULTS

The implementation of luminosity control hardware setup is shown in Fig. 4.

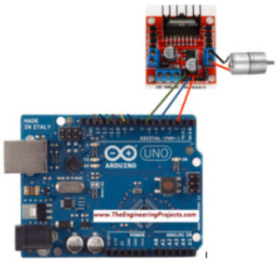

Fig. 4. The hardware setup

The system is tested by using different types of test images, each with a certain variation from others. Like simple binary images with text, simple colored images, complex colored images, images with single line text, images with multi lines text, etc. Certain results are shown in Fig. 5 and 6.

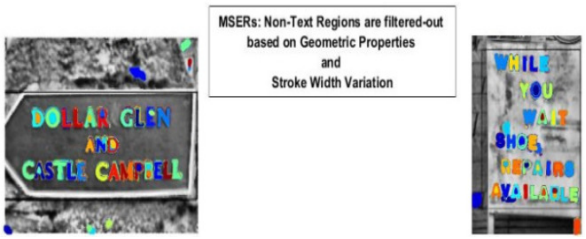

Fig. 5: Filtered MSERs: Most of the non-text regions are filtered-out based on the geometrical properties and stroke width variations.

Fig. 6: Examples of segmented and recognized texts from natural scenes

A real-time system operation is also verified. In this context, a USB camera is integrated in the system to capture scenes which are onwards enhanced by the proposed system (cf. Fig. 7). Later on the text is detected, identified and pronounced respectively by using MSER, OCR and TTS.

An example of text image, acquired by the camera, and a successful text detection mechanism is shown in Fig. 8. The detected text is also successfully pronounced via the employed TTS.

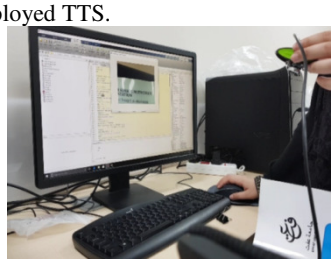

Fig. 7. Test Image acquisition via USB camera

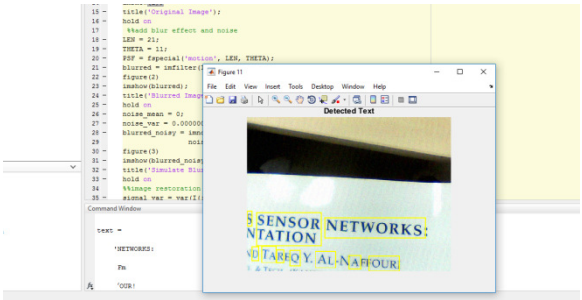

Fig. 8. Test Image acquired from camera and the text is successfully detected

IV. Conclusion

An effective scene to text conversion and pronunciation method has been proposed. The system is helpful for people with reading disabilities such as visual impaired people to be able to understand the written text during their daily life like medicine boxes, elementary food products, caution notices, etc.

The system acquires images by using integrated camera. It enhances the acquired image by employing CLAHE. In next step, it detects text regions via MSER and identifies characters via OCR and then regroups these characters to form meaningful words. Finally the detected words are pronounced by using the TTS. A system prototype is successfully realized and tested.

A detailed system performance evaluation for standard natural scenes database is a future work. Moreover, addition of event-driven features can augment the system performance in terms of processing and power efficiency [14]-[23]. Studying the feasibility of these features integration in the proposed solution is another prospect.

V. Acknowledgement

Authors are thankful to anonymous reviewers for their valuable feedback. This paper is sponsored by the Effat University.

References

[1] Mohd Bilal Ganai1 and Erjyoti Arora. (2015) "Implementation of Text to Speech Conversion Technique".International Journal of Innovative Research in Computer and Communication Engineering. 3(9): DOI: 10.15680/IJIRCCE.2015. 0309075

[2] Kumar, S., Kumar,P., Gupta, M. and Nagawat, A.K.(2010). "Performance Comparison of Median and Wiener Filter in Image De-noising". International Journal of Computer Applications. 12(4):0975 – 8887.

[3] Setiawan, A. W., Mengko, T. R., Santoso, O. S., &Suksmono, A. B. (2013, June). Color retinal image enhancement using CLAHE. In ICT for Smart Society (ICISS), 2013 International Conference on (pp. 1-3). IEEE.

[4] Kumar, A., & Gupta, S. (2017). Detection and recognition of text from image using contrast and edge enhanced mser segmentation and ocr. IJOSCIENCE (INTERNATIONAL JOURNAL ONLINE OF SCIENCE) Impact Factor, 3(3), 3.

[5] Chidiac, N., Damien, P. andYaacoub C. (2016) "A robust algorithm for text extraction from images," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), Vienna, 2016, pp. 493-497.

[6] Mathur1, G. and Rikhari, S. (2017) " Text Detection in Document Images: Highlight on using FAST algorithm." International Journal of Advanced Engineering Research and Science (IJAERS). 4(3):2456-1908.

[7]Khilari, P. and Bhope V. P.(2015) " A REVIEW ON SPEECH TO TEXT CONVERSION METHODS." International Journal of Advanced Research in Computer Engineering & Technology. 4(7).

[8] Arduino, S. A. (2015). Arduino. Arduino LLC.

[9] Ranjini, S., &Sundaresan, M. (2013). Extraction and recognition of text from digital english comic image using median filter. International Journal on Computer Science and Engineering, 5(4), 238.

[10] Joseph, J., Sivaraman, J., Periyasamy, R., & Simi, V. R. (2017). An objective method to identify optimum clip-limit and histogram specification of contrast limited adaptive histogram equalization for MR images. Biocybernetics and Biomedical Engineering, 37(3), 489-497.
[11] Huang, W., Qiao, Y., & Tang, X. (2014, September). Robust scene text detection with convolution neural network induced mser trees. In European Conference on Computer Vision (pp. 497-511). Springer, Cham.
[12] Zhu, Y., Yao, C., & Bai, X. (2016). Scene text detection and recognition: Recent advances and future trends. Frontiers of Computer Science, 10(1), 19-36.
[13] Griebe, T., Hesenius, M., Gesthüsen, M., &Gruhn, V. (2016, August). Test Automation for Speech-Based Applications. In SoMeT (pp. 85-100).
[14] Qaisar, S. M., Fesquet, L., &Renaudin, M. (2008). Computationally efficient adaptive resolution short-time Fourier transform. EURASIP, RLSP.
[15] Qaisar, S. M. (2014). Event Driven Filtering an Intelligent Technique for Activity and Power Consumption Reduction. International Journal of Circuits Systems and Signal Processing, 8.
[16] Qaisar, S. M., Simatic, J., &Fesquet, L. (2017, May). High-level synthesis of an event-driven windowing process. In 2017 3rd International Conference on Event-Based Control, Communication and Signal Processing (EBCCSP) (pp. 1-8). IEEE.
[17] Qaisar, S. M., Dallet, D., Benjamin, S., Desprez, P., &Yahiaoui, R. (2013, May). Power efficient analog to digital conversion for the Li-ion battery voltage monitoring and measurement. In 2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1522-1525). IEEE.
[18] Qaisar, S. M., Fesquet, L., &Renaudin, M. (2009). A signal driven adaptive resolution short-time Fourier transform. International Journal of Signal Processing, 5(3), 180-188.
[19] Sabo, A., &Qaisar, S. M. (2018, July). The Event-Driven Power Efficient Wireless Sensor Nodes for Monitoring of Insects and Health of Plants. In 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP) (pp. 478-483). IEEE.
[20] Jambi, L., Alsubaie, S., &Qaisar, S. M. (2018, July). An Event-Driven Power Efficient Surveillance and Lighting System in the Saudi Arabia Perspective. In 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP) (pp. 423-427). IEEE.
[21Qaisar, S. M., &Subasi, A. (2018, July). An Adaptive Rate ECG Acquisition and Analysis for Efficient Diagnosis of the Cardiovascular Diseases. In 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP) (pp. 177-181). IEEE..
[22] Mina Qaisar, S., Sidiya, D., Akbar, M., &Subasi, A. (2018). An Event-Driven Multiple Objects Surveillance System. International journal of electrical and computer engineering systems, 9(1), 35-44.
[23] Qaisar, S. M. (2018, July). A Computationally Efficient EEG Signals Segmentation and De-noising Based on an Adaptive Rate Acquisition and Processing. In 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP) (pp. 182-186). IEEE.