I J C T A, 9(20), 2016, pp. 403-410

Text Detection and Speech Synthesis from Natural Scene Image and Video Frame Using Smartphone

Lalita Kumari*, J. L. Raheja and Vidyut Dey*****

ABSTRACT

Detection and extraction of text from natural scene image or video stream is an important task for intelligent text detection system. In this paper, a text detection algorithm is proposed for natural scene image and complex video stream. The algorithm detects text from the input video frames / scene images, taken from smart phone, and converts the detected text into speech. Our proposed algorithm uses a built in dictionary to validate and auto correct detected text for possible error. Error detection and correction is performed using minimum error distance mapping. The effectiveness of the proposed method is demonstrated by experimental results.

Keywords: text detection, video stream, natural scene image, Local Binary Patterns, language dictionary

1. INTRODUCTION

Gone are the days when videography used to be the job a specialized person. Now-a-days, with the market flooded with high end smart phones, videography is now a Everyman's passion. These videos apart from their scenic contents contain vital information in the form of text. Grabbing such texts from the video is a challenge as such text is difficult to detect due to their orientation, alignment, contrast, complex colored, textured background etc. The texts have to be detected, tracked, extracted, and enhanced before they can be recognized for further processing. These texts may be street signs, labels license plates etc. Text extraction involves detection, localization, tracking, binarization, extraction, enhancement and recognition of the text from the images. Scene texts can be characterized as street signs, business cards, bottle labels, and license plates text regions. Text extraction consists of two parts: scene text extraction and scene text recognition. Scene text detection can be categorized into three: region based methods, connected component based methods and hybrid based methods. Connected component based methods search for possible text in image or video then apply machine learning techniques to identify texts. The purpose of the extraction process is to separate text regions from the natural scene image. The purpose of the recognition process is to determine labels from the extracted text regions. Image embedded in video is exploited in many ways such as content-based web image retrieval, video information retrieval, mobile based text analysis and recognition.

Rest of the paper is organized as follows. Related work and state of the art is discussed in section II. Proposed algorithm and its flow chart are explained in section III. Testing and experimental result is discussed in section IV. Conclusion remarks are given in section V.

2. RELATED WORK

This section discusses state of the art work on text detection from natural scene images. Following are the different approaches/methods to detect text from images.

* Department of Electronics & Communication Engineering, NIT Agartala, Tripura, India, *Email: kumaril2003@yahoo.co.in*

** Digital System group CSIR/ CEERI, Pilani, Rajasthan, India, *Email: jagdish.raheja.ceeri@gmail.com*

*** Department of Production Engineering, NIT Agartala, Tripura, India, *Email: vidyut.pe@nita.ac.in*

2.1. Texture-based method

Texture-based method is used for background subtraction. It is based on the local binary pattern texture measure, through which we get texture description for gray scale image .The textual description number is then converted to a decimal label. The histogram of the labels can be used as a texture descriptor. The binary number is converted into a decimal Local Binary Pattern (LBP) number. The resulted decimal LBP number is regarded as a label corresponding to the center pixel. In this way, a gray-scale image is converted into a LBP labeled image and the histogram of which will be used to describe the texture of the gray-scale image.

In 2005 Zhixin Shi,proposed an algorithm to convert a gray scale document image into an adaptive local connectivity map (ALCM), which is also a gray scale image. Second, applying a thresholding algorithm on the ALCM to reveal the text line patterns in terms of connected components. Third, using a straightforward grouping algorithm easily group the connected components into location masks for each text line. Finally, the text lines from a binarized version of the document image (using any standard thresholding algorithm) can be extracted by mapping the location masks back onto the binary image to collect the text line components. For those components touching multiple lines, a splitting algorithm is applied.

2.2. Connected component based methods

Connected component analysis is done when region boundary is detected on the edge pixels, which is not separated by a boundary. Shim[1] et al. proposed a method in which connected components are filtered geometrically by using texture property to exclude CCs. It measures stroke width of each pixel and merges each neighboring pixel into a connected component which may form a character. Constant stroke width feature is used to separate texts from other components of the scene. In this method logical operators are used with geometrical reasoning to recognize stroke width in scenes containing text. Yi [2] proposed a method to extract the text string from a video text string by using color-based partition and adjacent character grouping. But this method fails to locate the text information if texts are in small size, overexposed, characters with non-uniform colors, or strings with less than three character.

2.3. Region-based methods

Region-based methods consists of two stages First is detection of text detection and second text localization. This approach was proposed by Chen et al. [3]. In this technique text is extracted from video images by detecting horizontal and vertical edges with a Canny filter and thereby smoothens the edges. Wolf et al. [37] enhanced Otsu's method to binarize text regions from background to reduce noise and correct classification error by using a sequence of morphological processing. This method is based on morphological operations which extracts text regions from scene. In region based method and edge based method is applied to observe text edges in character and background pixels.

2.4. Hybrid based methods

Hybrid methods is combination of region based and texture based methods. In 2011 Yi proposed a hybrid approach to robustly detect and localize texts in natural scene images. A text region detector is designed to estimate the text existing confidence and scale information in image pyramid, which help segment candidate text components by local binarization. To efficiently filter out the non-text components, a conditional random field (CRF) model considering unary component properties and binary contextual component relationships with supervised parameter learning is proposed.

Phan et al. [11] proposed an algorithm for text detection in video, in which combination of Laplacian and k-means clustering is used. Shivakumara et al. [15] proposed Bayesian classifier method for multi-

oriented video scene text detection. Roy et al. [39] proposed a method to recognize text through binarization, which is based on the fusion concept .It is mainly work for video text.

3. PROPOSED ALGORITHM

In this paper a new framework to extract text from natural scene images, captured by a smart phone has been proposed. In this approach text is extracted from complex and cluttered background, validated or corrected automatically from stored dictionary and finally the algorithm synthesizes voice for the text. Figure 1 shows generalized flow chart of the proposed algorithm and then preprocessing. This algorithm consists of six generalized steps. Image is captured by a mobile phone for text detection. This captured image undergoes through the preprocessing process to remove the defects such as uneven light exposure, brightness/contrast adjusting, etc. This preprocessed image is used for text region detection by Text Localization using Local Binary Patterns (LBP) and Stroke width Transformation (SWT). Next step is to extract text from the identified text regions. Extracted text may contain spelling mistake in word (due to error in text extraction process). The extracted text is compared with words from a stored dictionary to identify the most likely word for the text. Finally the extracted word is passed to the speech synthesizer to convert the text to speech. A ddetailed flowchart of proposed algorithm is shown in figure 2.

3.1. Preprocessing of Captured Image

Scene image captured through mobile phone is not directly usable for text extraction as it contains various imperfections. e.g. uneven light exposure, non uniform or out of focus etc. Therefore, before going to actual process of text extraction, preprocessing of captured scene image is required. First step under this block is image normalization. This is used to modify the contrast in a color image. Histogram equalization transforms the values of color in an image in such a way that the color of the output image is uniform. Second step is conversion of normalized image into binary image. One dimensional adaptive thresholding of wavelet coefficients is used to convert into a binary image. Third step under this block is to restore the binary image using morphological operation. Since binary image conversion may cause text stroke erosion at the time of thresholding, image dilation is used for restoration of text region obtained after applying thresholding.

3.2. Text area detection

The detection of text area from a scene image starts from a binary image. Then the connected components are determined. Thereafter, components are filtered out using various geometric properties. First step under this block is determination of connected components from a binary image which is obtained from the scene image. Two-dimensional connectivity is considered for this purpose with 8-connected neighborhood. flood-

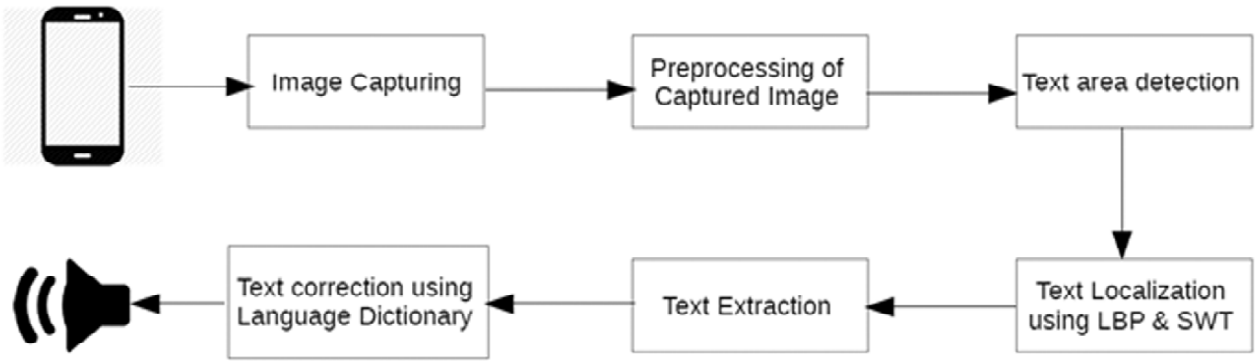

Figure 1: Generalized flowchart of proposed framework

fill algorithm is used to label all the pixels in the connected component. Second step is to filter out these components by matching geometric properties of text and that of the obtained connected components. For filtering out the connected components, geometric properties such as: aspect ratio, area, solidity, eccentricity, extent, Euler number are used. Based on these features, connected components are filtered out for possible text region.

3.3. Text Localization

Text localization process is performed on the filtered connected components obtained from text area detection block of the proposed algorithm. As described in the detailed flowchart, in figure 2, features are extracted using Local Binary Patterns (LBP). In the second stage filtering, only connected components of possible test area are allowed. The final filtering is done with the help of constant stroke width on the previously filtered regions. At the end of this process block, text localization is finalized for further text extraction

3.4. Text Extraction

After text localization, text extraction process is performed to extract text from the captured scene image. Text is extracted by a series of operations such as background subtraction, text alignment using Hough Transform. finally text is extracted using OCR.

3.5. Text correction using Dictionary

Generally extracted text are not correct enough to produce any meaningful word. Even any mistake in detecting a single character may produce null word. Therefore the proposed algorithm uses a dictionary to validate

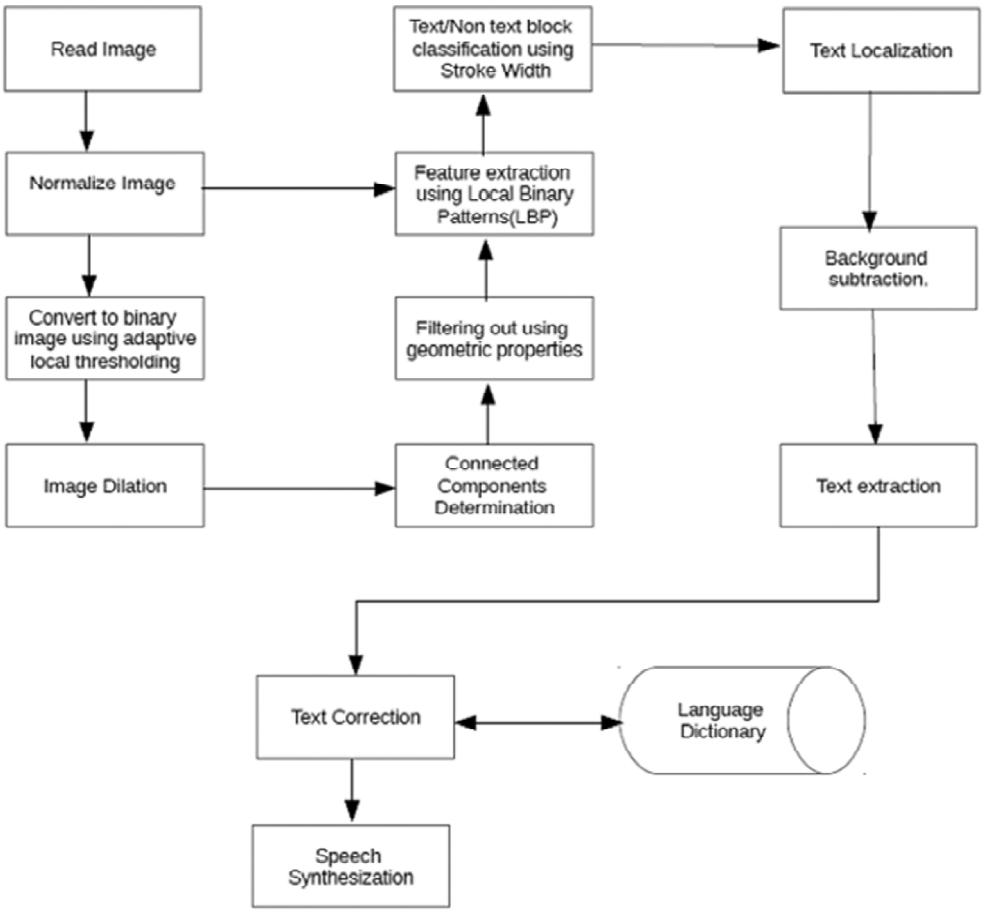

Figure 2: Detailed flowchart of proposed framework

extracted word. Error correction is carried out in this step, using minimum hamming distance concept. Possible word with minimum character error count is chosen intelligently to identify the correct text word

4. EXPERIMENTAL RESULT

In order to carry out trial a standard database, ICDAR 2015, was used. Along with the above mentioned standard database a small database was also created from the images captured by a mobile phone. Figure 3,

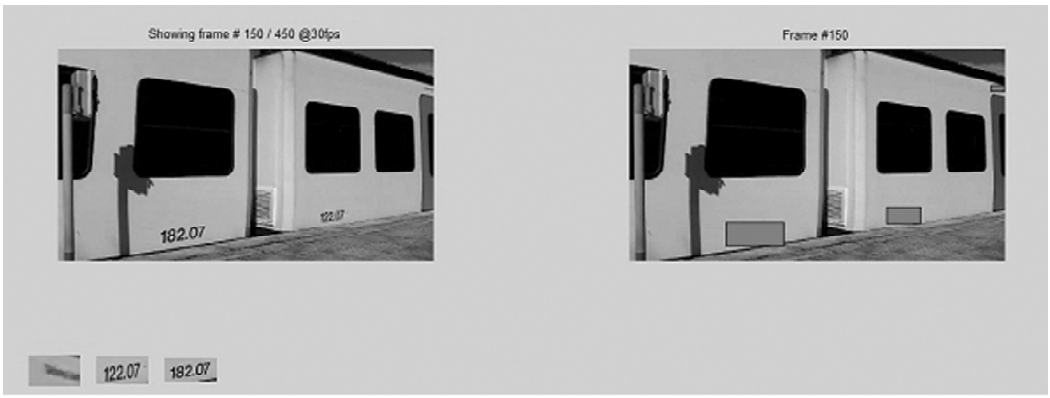

Figure 3: showing text areas in frame number 150 of Video_51_7_4.mp4

Figure 4: showing text areas in frame number 80 of Video_33_2_3.mp4

Figure 5: showing text areas in frame number 368 of Video_19_5_1.mp4

Figure 6: showing text areas in frame number 10 of Video_16_3_2.mp4

Figure 7: Text extraction steps. (a) Shows cropped text area from the video frame, (b) binary image of text area after background subtraction and histogram matching. (c) Extracted text from the binary text blocks, (d) text correction using minimum distance matching from available dictionary.

figure 4, figure 5, and figure 6 show four video frames showing detected text areas. These frames represent different scenario for testing purpose.

After detection of text area, text was extracted from the cropped text areas. After these operations, binary image consisting of text only is obtained for further text extraction. Figure 7 shows images of intermediate steps in the process of text extraction. It display cropped image, text only binary image (after background extraction), and extracted text.

5. CONCLUSION

This paper presented an algorithm to detect and extract text, from the video stream of natural scene image, captured using a Smartphone. The algorithm is capable to extract text from complex and cluttered background, validate orcorrect the words automatically after comparing with a dictionary, and finally produced, synthesized voice for the text. The presented algorithm used, an improved approach, which was backed by testing result. This algorithm can work well in natural scene image as well as in video stream. Conversion to speech from the extracted text was validated using a dictionary, can be improved further by using a domain specific dictionary. Future plan of this approach is to concentrate on the shortcomings of this method which include text detection from more complex scene, self error correction from broad domain, etc.

REFERENCES

[1] J. C. Shim, C. Dorai, and R. Bolle, Automatic Text Extraction from Video for Content-based Annotation and Retrieval,Proc. of International Conference on Pattern Recognition, Vol. 1, 1998, pp. 618-620

[2] C. Wolf, J. Michel. Jolion and F. Chassaing (2002). Text Localization, Enhancement and Binarization in Multimedia Documents. In Proc. ICPR, 1037-1040.

[3] Y. Liu, H. Lu, X. Xue, and Y. P. Tan, "Effective video text detection using line features," in Proc. Int. Conf. Control, Automation, Robotics and Vision, Dec. 2004, vol. 2, pp. 1528–1532.

[4] K. C. K. Kim et al., "Scene text extraction in natural scene images using hierarchical feature combining and verification," in Proc. Int. Conf. Pattern Recognition, Aug. 2004, vol. 2, pp. 679–682.

[5] J. Liang, D. Doermann, H. Li, "Camera based analysis of text and documents: a survey", Int. Journ. on Doc. Anal. And Recog. (IJDAR) vol. 7, pp. 84-104, 2005.

[6] C. Pal, C. Sutton, A. McCallum, Sparse forward–backward using minimum divergence beams for fast training of conditional random fields, Proceedings of the 2006 International Conference on Acoustics, Speech and Signal Processing, 2006, p. 5.

[7] T. Saoi, H. Goto, H. Kobayashi, "Text Detection in Color Scene Images Based on Unsupervised Clustering of Multihannel Wavelet Features," Proc. of 8 th Int. Conf. On Doc. Anal. and Recog. (ICDAR), pp. 690-694, 2005.

[8] C. Liu, C. Wang, and R. Dai, "Text detection in images based on unsupervised classification of edge-based features," in Proc. Int. Conf. Document Analysis and Recognition, Sep. 2005, vol. 2, pp. 610–614.

[9] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Trans. Circuit and Systems for Video Technology, vol. 15, no. 2, pp. 243–255,Feb. 2005.

[10] V. Kolmogorov, Convergent Tree-Reweighted Message Passing for Energy Minimization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1568–1583.

[11] T. Kasar, J. Kumar, and A. G. Ramakrishnan, "Font and background color independent text binarization," in Proc. 2nd Int. Workshop Camera-Based Document Anal. Recognit., 2007, pp. 3–9.

[12] Sunil Kumar, Rajat Gupta, Nitin Khanna, Santanu Chaudhury, Shiv Dutt Joshi(2007),"Text Extraction And Document Image Segmentation Using Matched Wavelets And MRF Model", IEEE Transactions On Image Processing, Vol. 16, No. 8.

[13] C.-B. Jeong, S.-H. Kim, Word image decomposition from mixed text/graphics images using statistical methods, Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, 2007, pp. 624–628.

[14] X. Liu, H. Fu, Y. Jia, "Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images", Pattern Recognition, vol. 41, pp. 484 –493, 2008.

[15] W. Pan, T. D. Bui, C. Y. Suen, "Text Detection from Scene Images Using Sparse Representation", Proc. of the 19th International Conference on Pattern Recognition, 2008.

[16] J. Liang, D. De Menthon, D. Doermann, Geometric Rectification of Camera-Captured Document Images, PAMI 30 (4) (2008) 591–605.

[17] P. Shivakumara, W. Huang, and C. L. Tan, "An efficient edge based technique for text detection in video frames," in The Eighth IAPR Workshop on Document Analysis Systems, 2008

[18] Jun Ye,Lin-Lin Huang,XiaoLi Hao "Neural Network Based Text Detection in Videos Using Local Binary Patterns", Pattern Recognition, 2009. CCPR 2009. China,IEEE.

[19] J.J. Weinman, E.G. Learned-Miller, A.R. Hanson, Scene text recognition using similarity and a lexicon with sparse belief propagation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1733–1746.

[20] E. Kim, S. Lee, J. Kim, Scene text extraction using focus of mobile camera, Proceed- ings of the Tenth International Conference on Document Analysis and Recognition, 2009, pp. 166–170.

[21] B. Epshtein, E. Ofek, Y. Wexler, Detecting text in natural scenes with stroke width transform, in: CVPR, 2963–2970, 2010.

[22] Trung Quy Phan, Palaiahnakote, Shivakumara Chew Lim Tan(2010), "A Skeleton-Based Method For Multi-Oriented Video Text Detection", Das '10 Proceedings Of The 9th IAPR International Workshop On Document Analysis Systems,pp :271-278.

[23] Miriam Leon, Veronica Vilaplana, Antoni Gasull, Ferran Marques(2010), "Region-Based Caption Text Extraction", 11th International Workshop On Image Analysis For Multimedia Interactive Services (Wiamis).

[24] K. Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: ICCV'11, Barcelona, Spain, 1457–1464, 2011.

[25] N. Stamatopoulos, B. Gatos, I. Pratikakis, S. Perantonis, Goal-Oriented Rectification of Camera-Based Document Images, IEEE TIP 20 (4) (2011) 910–920.

[26] T. Saba, G. Sulong, A. Rehman, Document image analysis: issues, comparison of methods and remaining problems, AIR 35 (2011) 101–118.

[27] A. B. Cambra, A. Murillo, Towards robust and efficient text sign reading from a mobile phone, in: ICCV Wshps., 64–71, 2011.

[28] L. Neumann, J. Matas, A method for text localization and recognition in real-world images, Proceedings of the 10th Asian Conference on Computer Vision, 2011, pp. 770–783.

[29] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. IJCV, 2012.

[30] A. Mishra, K. Alahari, C.V. Jawahar, Scene text recognition using higher order language priors, in: Proceedings of the Twenty Third International Conference for British Machine Vision Association, BMVC, 2012.

[31] A. Mishra, K. Alahari, and C. Jawahar. Top-down and bottom-up cues for scene text recognition. In CVPR, 2012.

[32] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection?. In BMVC, 2012.

[33] H. Chen, S. S. Tsai, G. Schorth, D. M. Chen, R. Grzeszczuk and B. Girod (2011). Robust text detection in natural scene images with edge-enhaned maximally stable extremal regions. In Proc. ICIP, pp 2609-2612.

[34] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in IEEE CVPR, 2012, pp. 1083–1090.L. Neumann, J. Matas, On combining multiple segmentations in scene text recognition, in: Proceedings of the 2013 International Conference on Document Analysis and Recognition, ICDAR, 2013.

[35] S. Milyaev, O. Barinova, T. Novikova, P. Kohli, V.S. Lempitsky, Image binarization for end-to-end text understanding in natural images, in: Proceedings of the 2013 International Conference on Document Analysis and Recognition, ICDAR, 2013.

[36] Q. Ye, D. Doermann, Text detection and recognition in imagery: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 37 (99) (2014) 1–20

[37] L.G. I Bigorda, D. Karatzas, Scene text recognition: No country for old men? In: Proceedings of the Twelfth Asian Conference on Computer Vision ACCV, 2014, pp. 157–168.

[38] Syed Saqib Bukhari , Thomas M. Breuel,Faisal Shafait(2009), "Textline Information Extraction From Grayscale Camera-Captured Document Images", ICIP Proceedings Of The 16th IEEE International Conference On Image Processing, pp: 2013 – 2016.

[39] S.Roy, P.Shivakumara, Hamid A. Jalab, Rabha W. Ibrahim, Umapada Pal, Tong Lu: "Fractional poisson enhancement model for text detection and recognition in video frames". Pattern Recognition 52: 433-447 (2016)