IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 4, Issue 2, April - May, 2016
ISSN: 2320 – 8791 (Impact Factor: 2.317)
www.ijreat.org

Text Extraction From A Video Image Using MSF Based Approach

Mrs.S.Aruna[1], V. Sesha Sriteja[2], V.S.M.Durga Prasad[3], S.V.Vamsi[4]

**[1]Assistant professor, [2,3,4]Student in Electronics & Communication, Andhra University,
Visakhapatnam, Andhra Pradesh 530003, India**

Abstract

Scene text detection from video as well as natural scene images is challenging due to the variations in background, contrast, text type, font type, font size, and so on. Besides, arbitrary orientations of texts with multi-scripts add more complexity to the problem. The proposed approach introduces a new idea of convolving Laplacian with wavelet sub-bands at different levels in the frequency domain for enhancing low resolution text pixels. Then, the results obtained from different sub-bands (spectral) are fused for detecting candidate text pixels. We explore maxima stable extreme regions along with stroke width transform for detecting candidate text regions. Text alignment is done based on the distance between the nearest neighbour clusters of candidate text regions. In addition, the approach presents a new symmetry driven nearest neighbour for restoring full text lines. We conduct experiments on our collected video data as well as several benchmark data sets, such as ICDAR 2011, ICDAR 2013, and MSRA-TD500 to evaluate the proposed method. The proposed approach is compared with the state-of-the-art methods to show its superiority to the existing methods.

Keywords: **laplacian wavelet, multispectral fusion, maxima stable extreme regions, stroke width transform.**

1. Introduction

The recent advances in multimedia and network technologies combined with the rapid decline in hardware prices, the contents of digital video and images are growing at a tremendous speed. As the statistics data of 2010 shows, more than 35 hours of video contents were uploaded to video sharing sites (e.g., YouTube) every minute and more than 35 million photos had been uploaded to social networking sites like Facebook every month [1]. This results in huge databases and thus requires approaches which work at high level semantics. The conventional approaches that use low level features may not be sufficient for handling such large databases due to the gaps between low level features and high level semantics. To alleviate this problem, text detection and recognition has become popular as it provides meaningful cues which are close to the content of video or image [2]–[4]. So, it has been widely used in video summarisation, content based image indexing and video sequence retrieval. On top of these applications, text detection and recognition has also been used for real time surveillance applications, such as assisting a blind person to walk freely on roads, assisting tourists to reach their destinations, enhancing safe driving, navigating vehicles based on license plate information, exciting event extraction from sports video, identifying athletes in marathon events, etc [5]. Video consists of two types of texts, namely, caption text and scene text. Caption text is manually edited, which has good clarity and visibility and hence is easy to process. Scene text exists naturally in video frames, the detection of which suffers from colour bleeding, low contrasts, low quality due to distortion, different orientations, backgrounds, etc. Hence, scene text is hard to process compared to caption text [4], [6], [7]. Scene images captured through a high resolution camera usually contain only scene texts with high contrast and complex background, while video contains both caption and scene texts with low resolution and complex background. Achieving good accuracy for text detection from both video and natural scene images is still an open issue in the field of image processing and pattern recognition because most of the existing approaches [8], [9] either focus on caption text in video or scene text in natural scene images but not both video and natural images. The problem of text detection and recognition from scanned document images is not new for the document analysis community because for different scripts we can find several Optical Character Recognisers (OCR engines) that are available publicly.

2. Multispectral fusion based approach

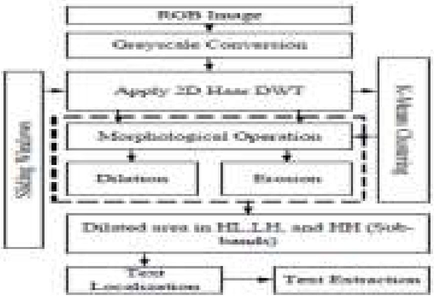


Fig 1. Overall Architecture of the proposed system

The proposed work is designed to accept the input as an image where the final effective output is obtained as extracted text using k-means clustering algorithm and mathematical morphological operations. For contrast in the results, discrete wavelet transform is applied for decomposing the image to sub-bands at various scales with diversified resolution. The text area is considered as special texture with unbalanced texture characteristics. Various statistical features like mean, standard deviation, and energy is estimated when the image with text is subjected to discrete wavelet transformation algorithm. After the image is subjected to wavelet transform, classification based on region is applied for compacting the text area within the scope of image. A specific sliding window is designed which reads the high frequency sub bands by sliding steps. The application can be considered that the dimension of each sub-band is M×N after

, d1if d1 and d2 are not equal to zero, than it fails to superimpose all the area of every sub-band when sliding window reads the high frequency sub-bands by the step l1 subjecting one-level wavelet transform, and we have= mod (M-W, l1), d2= mod(N-H, l2)x l2. The work also rejects all the contents which do not belong to the region. The statistical characteristics of every sub-band are estimated. The process achieves 12 features by evaluating the characteristics of three high frequency sub bands. Finally 12-dimension text feature vector is constructed.

The second phase of the design uses k-means clustering protocol where clustering is deployed by analysing the texture characteristic vector. The clustering factors selected are primary point of text, normal background, and complex background. Care should be taken to update the point of cluster in every processing of k- iterations. The image is segregated into three categories for textual area, simple and complex background area. Binarization technique is applied to the image depending on the results of classification and then mathematical morphological operations are deployed to take out the text details from the image.

2.1 Text Extraction Pipeline

A text extraction system typically consists of five steps: (1) Localisation, (2) Tracking, (3) Enhancement, (4) Binarization and (5) Recognition. The first step (Localisation) aims to detect and accurately locate all the text lines in an image or a video frame. The second step (Tracking) helps to track the movement of the text lines over multiple frames, e.g., a text line moving from bottom to top in a movie credits scene. In the third step (Enhancement), the localised and tracked text lines are enhanced in terms of contrast and resolution to improve their readability. The fourth step (Binarization) converts the text lines into black and white images so that they can be used in the last step (Recognition), which recognises the characters by using either an existing OCR engine or a custom-built OCR engine with its own feature extraction scheme.

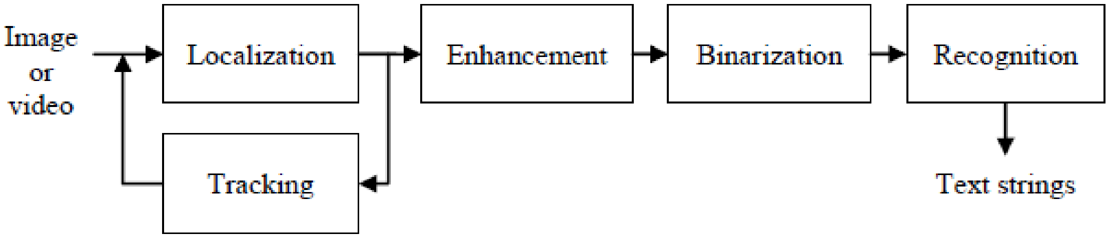


Fig 2. Text extraction pipeline architecture

Some text extraction systems may slightly change the order of the steps or omit certain steps. For example, Binarization is not needed if the Recognition step can work on grayscale or colour images directly. As another example, because temporal information is not available in natural scene images, the Tracking step is omitted for these images. The next section discusses Localisation, the first step in the pipeline.

2.2 Text Localisation

The goal of text localisation is to locate all the text lines in an input image or a video frame. A text line‘s position is usually represented by a rectangular bounding box .Some methods may provide additional information about a localised text line, e.g., a

―text mask‖, which indicates whether a particular pixel in the bounding box is a text pixel or a background pixel. Depending on the application, localisation can also be performed at the word level, instead of at the text line level

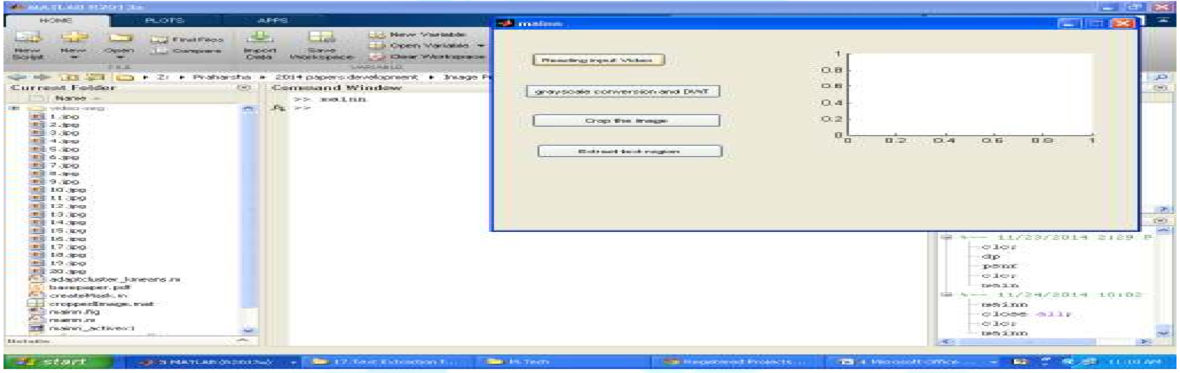
Fig 3. The(white) bounding box the localised text lines

Text in images often has the following characteristics, which makes it distinguishable from the background:

Text has sufficient contrast to the local background (to be readable). The strokes of a character are in four main directions: horizontal, vertical, left diagonal and right diagonal. The pixels of a single character have almost uniform intensity values or colours. Characters of the same text line are aligned on a straight line. Characters of the same text line have similar widths and heights. Characters of the same text line are spaced regularly.

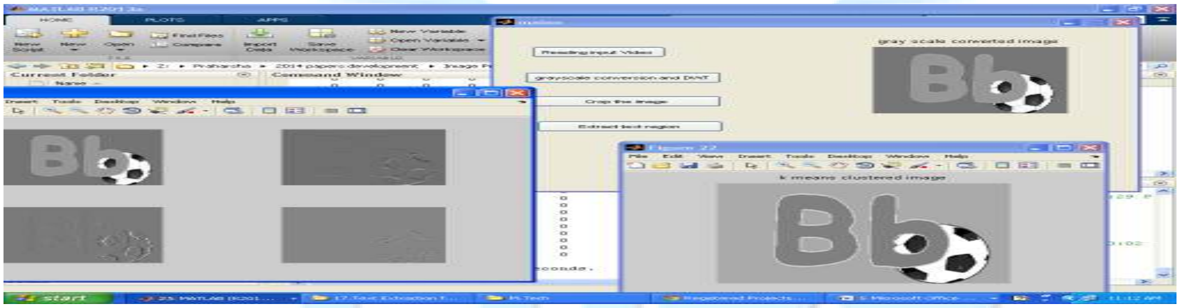
Different methods make use of different properties to localise the text lines. They can be classified into three main approaches: gradient-based, intensity/colour-based and texture-based. As its name suggests, the first approach relies on the first two properties of text and often performs edge detection to identify regions in the input image with those properties. Similarly, the second approach analyses regions in which the pixels have similar intensity values or colours (the third text property). Different from the previous two approaches, the last approach considers text as a special texture and applies techniques such as Discrete Cosine Transform and wavelet decomposition for feature extraction. For text/non-text classification, this approach typically employs machine learning techniques such as neural networks and Support Vector Machines (SVM).

It is worth mentioning that unlike the first three properties, the last three properties of text are usually used at a later stage in a localisation method (rather than as the main feature). For example, these properties can be used to remove false positives.

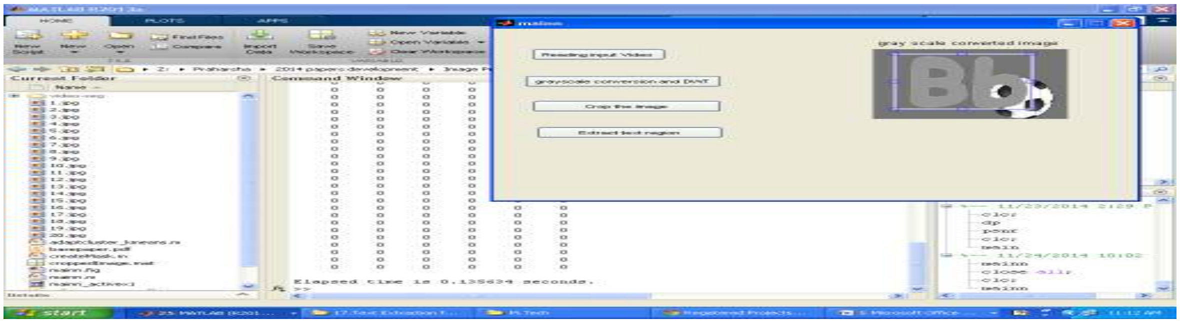


```
>> mainn
>>
```

3. Results



Reading Input Video
Gray Scale Conversion



```
Elapsed time is 0.135634 seconds.
>>
```

Crop the Image

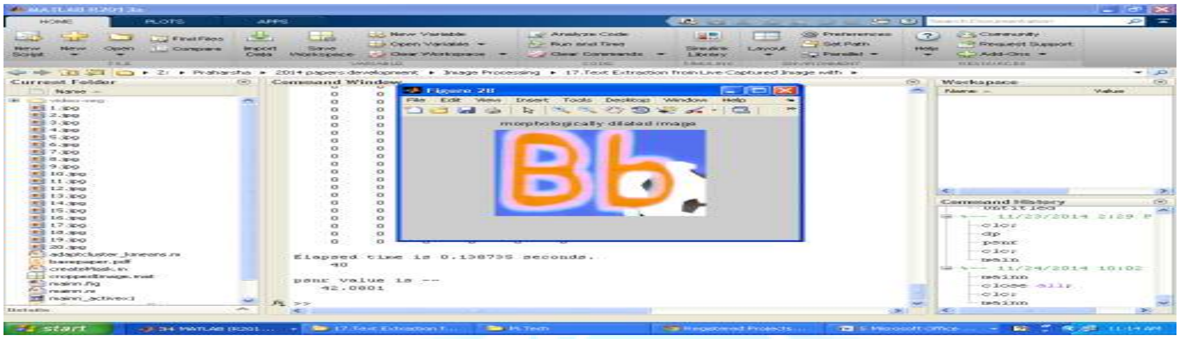


Figure 28
File Edit View Insert Tools Desktop Window Help
morphologically dilated image
Command Window
Current Folder
Workspace
Command History
start

```
Elapsed time is 0.138735 seconds.
    40

psnr value is --
   42.0001
>>
```

Morphological dilated image

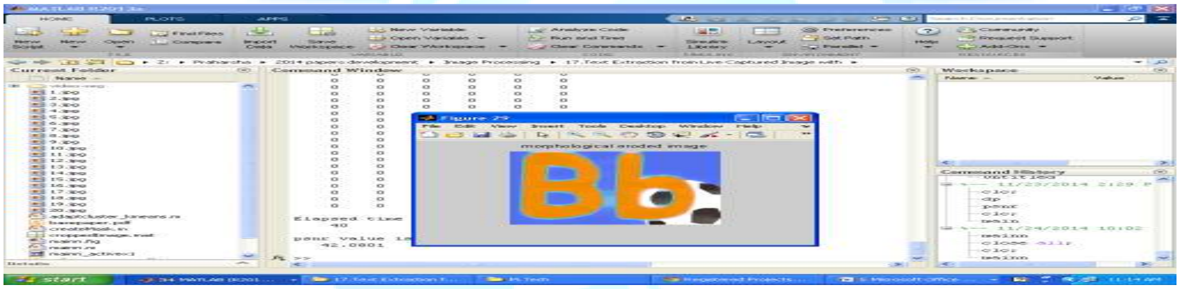


Figure 29
File Edit View Insert Tools Desktop Window Help
morphological eroded image
Command Window
Current Folder
Workspace
Command History
start

```
Elapsed time
    40

psnr value is
   42.0001
>>
```

Morphological eroded image

4. Conclusions

The proposed system has introduced a novel process of text extraction considering multiple cases of image with its textual contents. The system has been implemented using 2D Haar DWT along with k-means clustering algorithm. It also deploys methodology of sliding window for reading sub-bands of high frequency. Morphological operations like dilation and erosion has been introduced finally to refine the text and non-text region appropriately. For more realistic and robust results, the proposed system has been experimented with images with single / multiple text, multiple text of different sizes / style / languages, images with uniform and non-uniform background. The system is also evaluated with major research results in the past for conventional text extraction approach and is found to be potential for more accurately extracting text information. The future work will be to extending the similar concept of extracting text from video with higher accuracy

References

[1] Palumbo, P. W., Srihari, S. N., Soh, J., Sridhar, R. and Dem-janenko, V., 1992. Postal address blocks location in real time. Computer 25(7), pp. 34–42.

[2] Arth, C., Limberger, F. and Bischof, H., 2007. Real- time license plate recognition on an embedded DSP-platform. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR ’07) pp. 1–8.

[3] Wolf, C., Michel Jolion, J. and Chassaing, F., 2002. Text localisation, enhancement and binarization in multimedia documents. In: In Proceedings of the International Conference on Pattern Recognition (ICPR) 2002, pp. 1037–1040.

[4] Kavallieratou, E., Balcan, D., Popa, M. and Fakotakis, N., 2001. Handwritten text localisation in skewed documents. In: International Conference on Image Processing, pp. I: 1102–1105.