

A Robust Wavelet Transform Based Technique for Video Text Detection

Palaiahnakote Shivakumara, Trung Quy Phan and Chew Lim Tan
School of Computing, National University of Singapore
{shiva, phanquyt, tancl}@comp.nus.edu.sg

Abstract

In this paper, we propose a new method based on wavelet transform, statistical features and central moments for both graphics and scene text detection in video images. The method uses wavelet single level decomposition LH, HL and HH subbands for computing features and the computed features are fed to k means clustering to classify the text pixel from the background of the image. The average of wavelet subbands and the output of k means clustering helps in classifying true text pixel in the image. The text blocks are detected based on analysis of projection profiles. Finally, we introduce a few heuristics to eliminate false positives from the image. The robustness of the proposed method is tested by conducting experiments on a variety of images of low contrast, complex background, different fonts, and size of text in the image. The experimental results show that the proposed method outperforms the existing methods in terms of detection rate, false positive rate and misdetection rate.

1. Introduction

With more and more digital devices, video has now become the most popular media type in our daily life. Since text embedded in video contains much semantic information related the video content, it plays an important role in content-based multimedia indexing and retrieval systems. Many text detection approaches have been proposed in the past several years; however, due to low resolution and complex backgrounds of videos and various sizes, colors, styles and alignments of text, text detection and extraction is still challenging [1-3]. Two types of text in video are: (1) *caption/graphics/artificial* text which is artificially superimposed on the video at the time of editing, and (2) *scene* text which naturally occurs in the field of view of the camera during video capture. Clearly, the detection of scene text is a challenging task due to varying lighting, complex movement and transformation [1].

Our literature study shows that connected-component based methods are not robust because they

assume that text pixels belonging to the same connected region share some common features such as color or grey intensity [4]. On the other hand, texture based methods tend to be computationally expensive for large databases [5,6] as they involve expensive operations such as DCT for text detection in images. The edge and gradient based methods have been developed to reduce the number of computations in detecting text in the images. However, these methods are not robust to complex background as they give more false positives. In addition, selection of threshold values to classify text pixel from non text pixel is another major problem [7-10]. To overcome these problems, the method in [11] introduced the segmentation of text portion with the help of candidate text block identification. A method based on uniform colors in $L^* a^* b^*$ space is also proposed in [12] to locate uniform colored text in video frames. Obviously, this method fails when text in video contains multiple colors in a text line or in a word. Recently, the wavelet and the SVM combination is used for text detection in the images, which performs well but they include large number of features and extensive training with the classifier [13,14]. On the other hand, unsupervised method without training for text detection in video images is not yet explored so far. The above observation shows a gap in developing a robust technique to give a better detection rate with fewer false alarms without any constraints for text detection in video images

This paper hence proposes a new robust wavelet and statistical features and central moments based method with k means clustering for detecting text in video images. We have noticed the text regions in subbands of wavelets located on or near the edge yield large wavelet coefficients, making text regions detectable in the high frequency subband images. Hence, we use statistical and central moment features to exploit that property in wavelet domain for detecting text in video.

2. Proposed Text Detection Algorithm

For the gray image of size 256×256 as shown in Figure 1(a), we use single level 2D Haar wavelet decomposition for detecting text in the video image.

Experiments on Haar wavelet basis proved that it has good ability to characterize texture features for the text in video image and is computationally efficient [15]. Therefore, we use three high frequency subband images LH, HL and HH for text detection purpose. The reconstructed LH, HL and HH by inverses 2D wavelet transform are shown in Figure 1(b)-(d) respectively where we can see large wavelet coefficients near or on the edge as thick white color compared to its background.

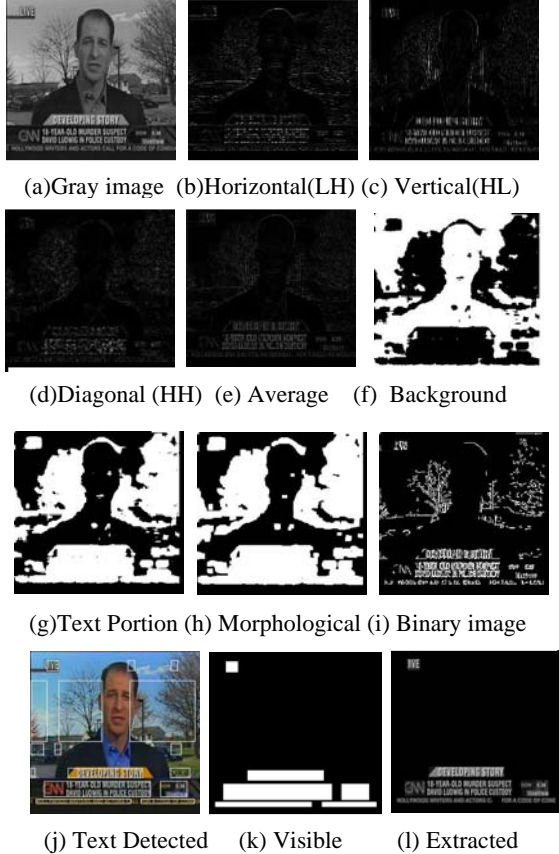


Figure 1. Intermediate steps in text detection

We employ statistical features in the three subband images to capture the texture property of the text in video image. More specifically, the features include energy, entropy, inertia, local homogeneity, mean, second-order (μ_2) and third-order (μ_3) central moments of subband images, are computed. Firstly, a sliding window of size $N \times N$ ($N=8$) pixels is moved over each subband image. For each window position and for each subband image, the features are computed using the formula as follows.

$$E = \sum_{i,j} W^2(i, j) \quad (1)$$

$$Et = \sum_{i,j} W(i, j) \cdot \log W(i, j) \quad (2),$$

$$I = \sum_{i,j} (i - j)^2 W(i, j) \quad (3),$$

$$Hm = \sum_{i,j} \frac{1}{1 + (i - j)^2} W(i, j) \quad (4),$$

$$M = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N W(i, j) \quad (5),$$

$$\mu_2 = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (W(i, j) - M)^2 \quad (6),$$

$$\mu_3 = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (W(i, j) - M)^3 \quad (7),$$

Here, $W(i, j)$ is the subband image, at pixel position (i, j) in the window of size $N \times N$. After feature computation, we obtain 21 features, which are the 7 features above multiplied by three subband images. These features are normalized in the range from 0 to 1, which form the feature vector representation for each pixel. i.e, let FV be the feature vector representing text pixel, obtain the minimum and the maximum element in the feature vector as

$$Min = \min(FV) \quad (8), \quad Max = \max(FV) \quad (9)$$

Then the features in FV are normalized as

$$NFV = \frac{(FV(i) - Min)}{(Max - Min)} \quad (10)$$

The k-means algorithm is applied to classify the feature vector into two clusters: background and text candidates. Area of the two clusters is computed to separate text and background. A cluster is classified as text if its area is less than to area of other cluster. The sample output of the k-means algorithm is shown by separating the background and the text in Figure 1(f) and Figure 1(g), respectively. After this step, we get the initial text candidates which are binary ($Bin(i, j)$). The method uses morphological operations such as opening and dilation to get connected components and to discard too small objects as background. The final result of morphological operations is shown in Figure 1(h). The position of every connected component of the text candidates is projected on the average subband image shown in Figure 1(e) to get the corresponding subband text candidates. The average subband image is calculated as

$$AVG(i, j) = \frac{1}{3} \sum_{k=1}^3 WS_k(i, j) \quad (11), \text{ where}$$

$WS_k(i, j)$ denotes horizontal, vertical and diagonal subband images as shown in Figure 1(b)-(d). The true text pixel is classified using average subband image ($AVG(i, j)$) and binarized image ($Bin(i, j)$) which is the output of morphological operation. i.e

$$(i, j) = \begin{cases} 1, & \text{if } (AVG(i, j) > T \ \& \ (Bin(i, j) = 1)) \\ 0, & \text{Otherwise} \end{cases} \quad (12)$$

Using equation (12) we get text candidates which are in binary as shown in Figure 1(i).

The threshold T is determined automatically using the average subband image as follows. Let FC be the frequency coefficients in the average subband, which are greater than or equal to 0.05. The mean frequency coefficient (MFC) is calculated as

$$MFC = \frac{1}{m} \sum_{i=1}^m FC_i \quad (13), m \text{ is the}$$

number of elements in FC . The number of high frequency coefficients (NHF) are counted if $AVG(i,j) \geq MFC$. Then the threshold T is defined as

$$T = \frac{SFC}{(m + NHF)} \quad (14), \text{ where } SFC$$

is the sum of values in FC .

Finally, the projection profile analysis is done to find horizontal and vertical boundary for the text lines in the image and then the bounding boxes are fixed for the text lines. The sample text detection result is shown in Figure 1(j). The detected text blocks are filled by the white color to make visible in Figure 1(k) and the true text blocks are extracted as shown in Figure 1(l)

2.1. False positive elimination

As it is noted from [2] that elimination of false positives without missing any text is challenging and interesting as it helps in improving the performance of the method in terms of detection rate. Hence, we introduce some heuristics based on height, width, area of the text blocks detected by the proposed method. Let $T(i,j)$, H and W be text block, height and width of the detected text block respectively. Aspect ratio of the $T(i,j)$ is computed as $A_R = H/W$. Let $B(i,j)$ be the binary information corresponding to detected text block in $Bin(i,j)$ image. Area of the $B(i,j)$ is computed as $A_HW = Area/(H*W)$. Let $G(i,j)$ be the gray information corresponding to $T(i,j)$ in gray image. The number of Sobel edge components (NS) and the number of Canny edge components (NC) for the text block is computed. The number of transitions from 0 to 1 and 1 to 0 in Canny edge block of $T(i,j)$ is also noted. The difference between the number of Sobel and Canny edge components is calculated as $S_D = NS - NC$. Initially, we eliminate the text blocks which has too height or too small or a few number of transitions.

- (i) $If(A_R > 2)$
- (ii) $If(S_D < t) \& if(A_HW < 0.2)$
- (iii) $If(S_D > t) \& if(A_HW < 0.1)$
- (iv) $If(H > W) \& if(A_HW > 0.2)$

If the detected text blocks satisfy any of the above heuristic then we treat them as false positives and hence we eliminate them.

3. Experimental Results

3.1 Dataset and Methods for Comparison

For experimentation, we have created our own dataset as there is no standard dataset available in literature, which includes a variety of video images, such as images taken from movies, news clips containing some scene texts and sports videos. In this experiment, we have selected 101 video images from the above said sources which give 491 actual number of text blocks. The method implemented using MATLAB software is run on a PC with Pentium IV 2.33 GHz processor. The approximate processing time for each video image of size 256×256 is about 15.2 seconds for text detection.

We have chosen three existing methods [9, 10, 12] for comparison. Method [9] is based on Sobel edge information for text detection. Method [10] is also based on gradient for text detection. Method [12] makes use of uniform color for text location.

3.2 Sample Test Results

Figures 2-4 show the text detection results of the proposed method in (a)-(c) and the above three existing methods in (d)-(f) for a variety of sample video images. In (a)-(c), we show the original image, the text detection and the final text extraction results using the proposed method. In (d)-(f), we show the text detection results of the three existing methods.

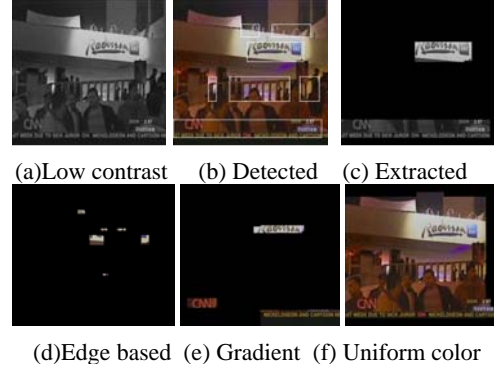


Figure 2. Text detection for low contrast image

Figure 2 shows that the edge based method fails to detect text in low contrast image because Sobel operator works for high contrast images. Therefore the detection rate of the method is quite low compared to the proposed method. The gradient based method detects text in low contrast image with inaccurate boundaries because the method suffers from many ad hoc threshold values used in detection. The uniform color method also fails to detect text in the image because of its limitation. On the other hand, the proposed method detects the text in the image correctly

including scene text. Hence the detection rate is high compared to the existing methods.

Figure 3 shows that the proposed method gives better results even for scene text images compared to the three existing methods.

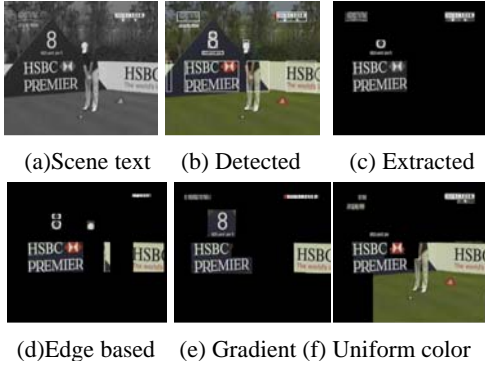


Figure 3. Scene text detection

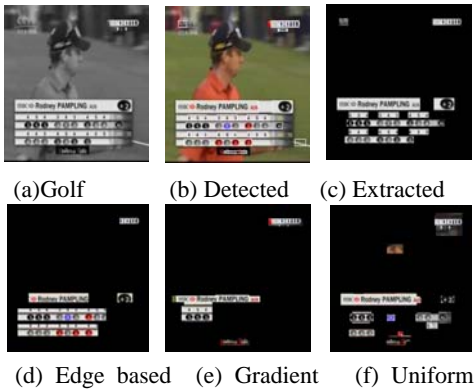


Figure 4. Text detection for Golf image

Figure 4 shows that the proposed method detects text in the golf image including small font and low contrast text in image. While the edge and gradient based methods miss some text and fail to detect small font and low contrast text, the uniform text color method tends to include additional non text information in the bounding boxes and it fails.

We have also conducted experiments to choose a proper window size for detecting text lines in the image as shown in Figure 5 where (c) with 8×8 gives better results among other sizes as it detects all the text lines in the (a) correctly whereas (b) and (d) miss text lines.

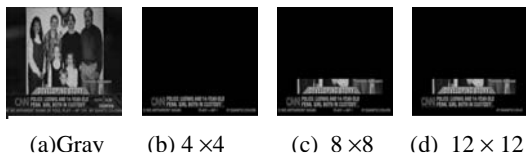


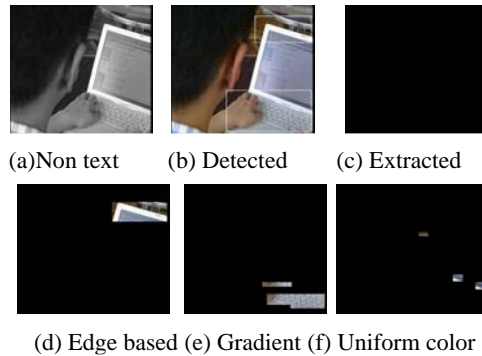
Figure 5. Window size selection

3.3 Experiment on non text images

In order to show the strength of the proposed method, we have conducted experiments on 100 non text images with the intention of checking false positive detection by the methods, along with the three existing methods. The results of the methods reported in Table 1 show that the proposed method produces 32 false text blocks positive for 28 images out of 100 images, while the existing methods perform much worse. The sample of results is shown in Figure 6 where the proposed method does not give any false text blocks positive but the existing methods give at least one false text block positive. With this experiment, it is realized that text detection methods should be evaluated by testing on non text images also before claiming the method detects text accurately.

Table 1. Experiment on 100 non text images

Method	No. Images	No. False Positives
Edge based [9]	42	56
Gradient based [10]	30	38
Uniform text color [12]	91	257
Proposed	28	32



(d) Edge based (e) Gradient (f) Uniform color

Figure 6. Results on non text image

3.4 Comparison Metrics

To give an objective comparison of all the above methods, we use detection rate, false positive rate and misdetection rate as decision parameters and metrics in this work. The detected text blocks are represented by their bounding boxes. To judge the correctness of the text blocks detected, we manually count Actual Text Blocks (ATB) in the images in the dataset. Also we manually label each of the detected blocks as one of the following categories:

Truly detected text block (TDB): a detected block that contains text fully or partially. **Falsely detected text block (FDB):** a detected block that does not contain text. **Text block with missing data (MDB):** a truly detected text block that misses some characters

Based on the number of blocks in each of the categories mentioned above, the following metrics are calculated to evaluate the performance of the methods:

Detection rate (DR) = Number of TDB / Number of ATB. **False positive rate (FPR)** = Number of FDB / Number of (TDB + FDB). **Misdetection rate (MDR)** = Number of MDB/ Number of TDB. The performance of the proposed method in comparison with the existing methods is summarized in Tables 2 and 3. Table 3 shows that the detection rate, false positive and misdetection rates of the proposed method are higher than the three existing methods.

Table 2. Results of proposed and existing methods

Method	ATB	TDB	FDB	MDB
Edge based [9]	491	393	86	79
Gradient based [10]	491	349	48	35
Uniform text color [12]	491	252	95	94
Proposed	491	475	21	27

Table 3. Performance of the proposed and existing methods in (%)

Method	DR	FPR	MDR
Edge based [9]	80.0	18.3	20.1
Gradient based [10]	71.0	12.0	10.0
Uniform text color [12]	51.3	27.3	37.3
Proposed	96.7	4.2	5.6

4. Conclusion and Future Work

In this paper, we propose a wavelet, statistical features and moments based text detection method for detecting both graphic text and scene text with different fonts, size, contrast and backgrounds. Experimental results showed that the proposed method outperforms the existing methods in terms of metrics.

For our future work, we plan to use temporal information and use more levels of wavelets to reduce the false positive rate. Furthermore, the method can be extended to fix the bounding boxes for text lines with arbitrary direction.

5. Acknowledgment

This research is supported in part by IDM R&D grant R252-000-325-279.

6. References

[1] J. Zang and R. Kasturi, "Extraction of Text Objects in Video Documents: Recent Progress", *The Eighth IAPR Workshop on Document Analysis Systems (DAS2008)*, Nara, Japan, September 2008, pp 5-17.

[2] J. Zhang, D. Goldgof and R. Kasturi, "A New Edge-Based Text Verification Approach for Video", *ICPR 2008*.

[3] K. Jung, K.I. Kim and A.K. Jain, "Text information extraction in images and video: a survey", *Pattern Recognition*, 37, 2004, pp. 977-997.

[4] A.K. Jain and B. Yu, "Automatic Text Location in Images and Video Frames", *Pattern Recognition*, Vol. 31(12), 1998, pp. 2055-2076.

[5] Y. Zhong, H. Zhang and A.K. Jain, "Automatic Caption Localization in Compressed Video", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 4, 2000, pp. 385-392.

[6] K. L. Kim, K. Jung and J. H. Kim, "Texture-Based Approach for Text Detection in Images using Support Vector Machines and Continuous Adaptive Mean Shift Algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 12, December 2003, pp 1631-1639.

[7] M. Anthimopoulos, B. Gatos and I. "Pratikakis. A Hybrid System for Text Detection in Video Frames", *The Eighth IAPR Workshop on Document Analysis Systems (DAS2008)*, Nara, Japan, September 2008, pp 286-293.

[8] M. R. Lyu, J. Song and M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15, No. 2, February 2005, pp 243-255.

[9] C. Liu, C. Wang and R. Dai, "Text Detection in Images Based on Unsupervised Classification of Edge-based Features", *ICDAR 2005*, pp. 610-614.

[10] E. K. Wong and M. Chen, "A new robust algorithm for video text extraction", *Pattern Recognition* 36, 2003, pp. 1397-1406.

[11] P. Shivakumara, W. Huang and C. L. Tan, "An Efficient Edge based Technique for Text Detection in Video Frames", *The Eighth IAPR Workshop on Document Analysis Systems (DAS2008)*, Nara, Japan, September 2008, pp 307-314.

[12] V. Y. Marinano and R. Kasturi, "Locating Uniform-Colored Text in Video Frames", *15th ICPR*, Volume 4, 2000, pp 539-542.

[13] Q. Ye, Q. Huang, W. Gao and D. Zhao, "Fast and robust text detection in images and video frames", *Image and Vision Computing* 23, 2005, pp. 565-576.

[14] H. Li, D. Doermann and O. Kia, "Automatic Text Detection and Tracking in Digital Video", *IEEE Transactions on Image Processing*, Vol. 9, No. 1, January 2000, pp 147-156.

[15] W. Mao, F. Chung, K. K. M. Lam and W. Siu, "Hybrid Chinese/English Text Detection in Images and Video Frames", *ICPR 2002*.