2010 International Conference on Pattern Recognition

New Wavelet and Color Features for Text Detection in Video

Palaiahnakote Shivakumara,
School of Computing
National University of Singapore
Singapore
shiva@comp.nus.edu.sg

Trung Quy Phan
School of Computing
National University of Singapore,
Singapore
phanquyt@comp.nus.edu.sg

Chew Lim Tan
School of Computing
National University of Singapore
Singapore
tancl@comp.nus.edu.sg

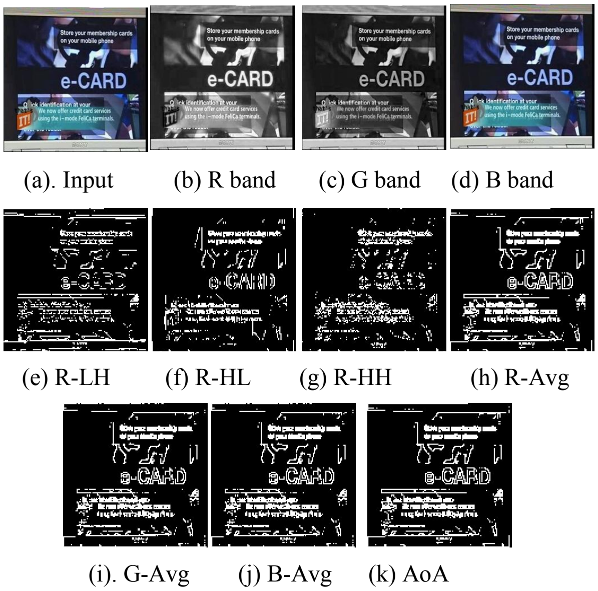
***Abstract*-Automatic text detection in video is an important task for efficient and accurate indexing and retrieval of multimedia data such as events identification, events boundary identification etc. This paper presents a new method comprising of wavelet decomposition and color features namely R, G and B. The wavelet decomposition is applied on three color bands separately to obtain three high frequency sub bands (LH, HL and HH) and then the average of the three sub bands for each color band is computed further to enhance the text pixels in video frame. To take advantage of wavelet and color information, we again take the average of the three average images (AoA) obtained by the former step to increase the gap between text and non text pixels. Our previous Laplacian method is employed on AoA for text detection. The proposed method is evaluated by testing on a large dataset which includes publicly available data, non text data and ICDAR-03 data. Comparative study with existing methods shows that the results of the proposed method are encouraging and useful.**

I. INTRODUCTION

Due to enormous video data of daily activities, efficient indexing and retrieving relevant information from multimedia databases becomes hard and challenging problem for the researchers [1]. Therefore, for the past decades, several ways based on annotation and content have been introduced to meet real challenges of the retrieval. But, to the best of our knowledge, none of the methods achieve good accuracy in filling the semantic gap between low level and high level features to understand the video [1]. This is because of unexpected and undesirable properties of video such as low resolution, complex background, different font and size and text moments. Hence, an alternate way to fill the semantic gap is text detection, extraction and recognition to understand the video content. Text detection and recognition is quite familiar work for document analysis community but due to the above properties of video, document analysis based methods may fail to give satisfactory results [2-3].Text detection and extraction in video is usually addressed by three main approaches, namely, connected component based [4-5], texture based [6-7], edge and gradient based [8-13]. These methods solve the problem to some extent but still there is room for improvements especially for large datasets containing both graphics and scene text. Recently, integrating wavelet and color features is new way for text detection in video. Thus, in this paper, we take advantage of color and wavelet decomposition as color of text component usually will have uniform color but not in contrast [13]. Wavelet decomposition generally enhances the high contrast pixels by suppressing low contrast pixels [4]. These factors motivated us to propose a hybrid method for text detection in video.

II. PROPOSED METHODOLOGY

Proposed wavelet and color features based method assumes text lines in video are in horizontal direction.


(a). Input (b) R band (c) G band (d) B band

(e) R-LH (f) R-HL (g) R-HH (h) R-Avg

(i). G-Avg (j) B-Avg (k) AoA

Figure 1. Wavelet and Color features

A. Wavelet and Color Features

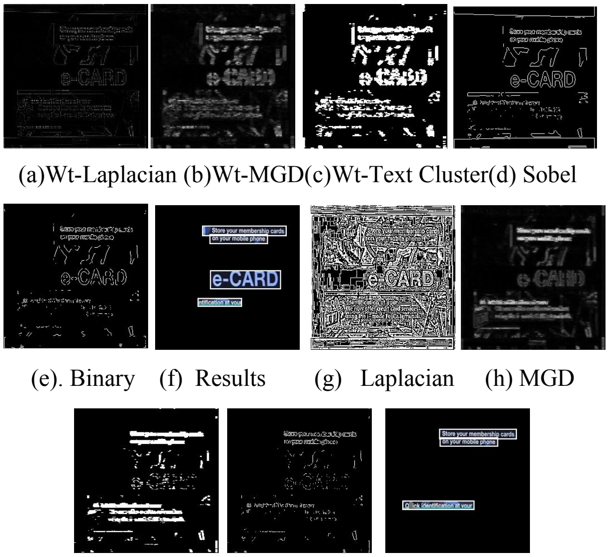
For each set of R, G and B bands of a color frame in Figure 1(a) as shown in Figure 1(b)-(d) respectively, wavelet (Haar) is applied to obtain high frequency subbands such as LH (horizontal), HL (vertical) and HH (diagonal) for the frame shown in Figure 1(b), are shown respectively in Figure 1(e)-(g). The averages of high frequency subbands of R, G and B bands denoted as R-avg, G-Avg and B-Avg are computed and the results are shown respectively in Figure 1(h)-(j). Further, the average of R-Avg, G-Avg and B-Avg (AoA) is computed and the result is shown in Figure 1(k). This result serves as the input for the ensuing text detection method.

B. Laplacian Method for Text Detection

We employ our previous Laplacian method [8] on AoA for text detection. According to our literature review, Laplacian method gives better results than other exiting methods. However, we notice that the method gives poor

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.972

results for low contrast text as it depends on Laplacian mask operation in addition to noise introduction. In this work, we use the enhanced image obtained by the previous section as input to our Laplacian method for accurate text detection and it is called Wavelet-Laplacian method. The steps of the method are illustrated in Figure 2, where (a) shows the result of Laplacian 3×3 mask over AoA, (b) shows the result of Maximum Gradient Difference (MGD), (c) shows the text cluster given by K-means clustering (K=2), (d) shows the Sobel edges of input image in Figure 1(a), (e) shows the result of intersection of Sobel and the text cluster image, (f) shows the text extraction result after eliminating false positives. Figure 2(g)-(k) shows the steps of existing Laplacian method, where one can notice from Figure 2(g) that Laplacian mask operation produces more noisy edges when compared to Figure 2(a). It is also observed from Figure 2(i) and (j) that the text line “e-Card” is missing in Figure 2(i) due to noise introduction and small fonts text below “e-Card” line is missing in Figure 2(j) because text cluster intersects with Sobel edge map which usually detects only high contrast text pixels. Thus Laplacian method does not detect text line “e-Card” and small font text as shown in Figure 2(k) while the proposed method restores the line “e-Card” because of the combination of wavelet and color features as shown in Figure 2(c). But it fails to restore the small font text lines since the obtained text cluster image intersects with Sobel edge map as shown in Figure 2(e). Thus the proposed method detects “e-Card” text line and misses low contrast text as in Figure 2(k) as compared to Laplacian method results. Therefore, performance of the proposed method improves over existing Laplacian method. More details of the Laplacian method can be found in [8]. Further, the clear flow of the Laplacian method can be seen in Figure 3.


(a)Wt-Laplacian (b)Wt-MGD(c)Wt-Text Cluster(d) Sobel

(e). Binary (f) Results (g) Laplacian (h) MGD

(i). Text cluster (j) Binary (k) Results

Figure 2. Wt-RGB vs Laplacian method

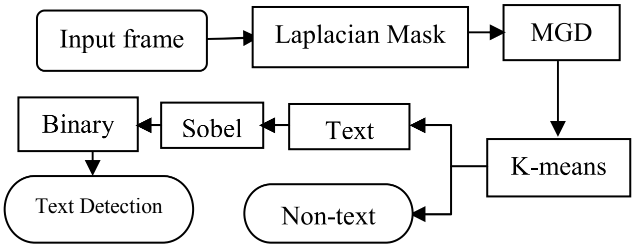


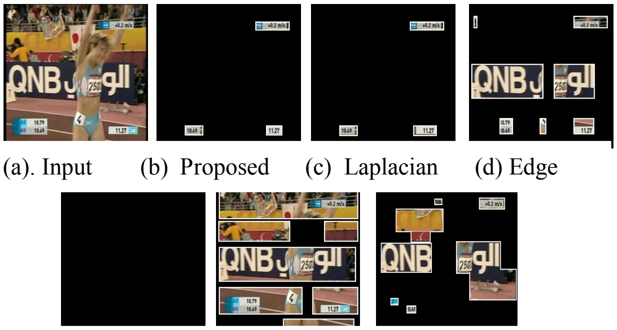
Figure 3. Flow diagram for Laplacain method

III. EXPERIMENTAL RESULTS

We create our own database as there is no benchmark data for text detection in video. The created database consists of (1) 800 frames of different text size, fonts and graphics, scene text etc, (2) a publicly available dataset of 45 images, (3) 251 ICDAR 03 data of camera images and (4) 300 non text frames. In total, 1396 frames are used to test the proposed method in comparison with several existing methods.

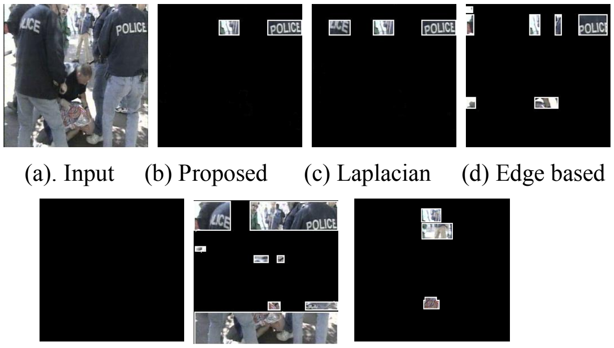
A. *Experiments on Large dataset*

The performance of the proposed and existing methods is illustrated in Figure 4 where the proposed and Laplacian methods detect even small font but they miss scene text compared to results of other existing methods. However other existing methods do not detect text properly as detected text blocks includes background information. Gradient based method fails to detect text while edge based method detects text to some extent but Sobel-Color and Uniform-Color methods detects text lines with inaccurate boundary with more false positives.


(a). Input (b) Proposed (c) Laplacian (d) Edge

(e). Gradient (f) Sobel-Color (g) Uniform-Color

Figure 4. Sample results for sports news frame


(a). Input (b) Proposed (c) Laplacian (d) Edge based

(e). Gradient (f) Sobel-Color (g) Uniform-Color

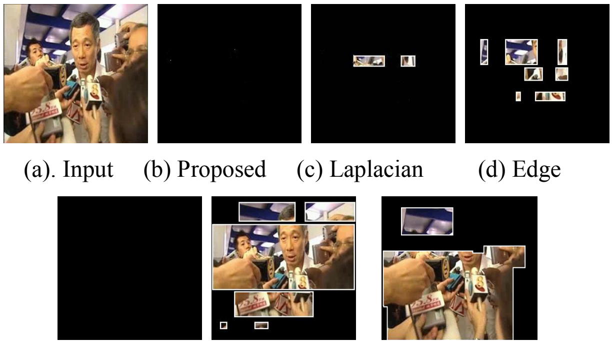
Figure 5. Experiment on Hua’s data

B. Experiments on Hua's data [14]

We test the proposed and existing methods on this public dataset to show that the proposed method is applicable to an independent set of video data. Figure 5 shows that the Laplacian method gives better results than the proposed method as the proposed method considers true text blocks as non text blocks due to the background enhancement and hence conditions used in Laplacian method eliminates true text blocks as false positives in this work because we use the same condition in this work. Thus there is a room to think tradeoff between false positive elimination and true text blocks identification [15]. This concludes that we need a mechanism that eliminates false positives without tuning the parameters and changing the conditions. However, the proposed method is competitive in comparison with the results of existing methods.

C. Experiments on Non-text data

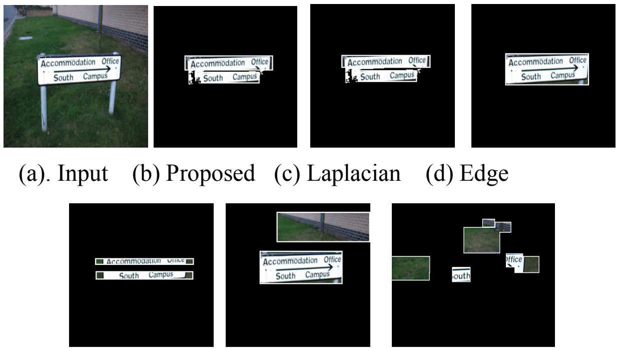
It is often a misconception that text detection methods can be used for text frame classification from the large number of text and non text video frames. However, experimental results on non text data show that text detection methods including the proposed method fails to classify non text frame as non text correctly. This indicates that text frame classification before text detection is another research issue where we need to focus. For this particular example shown in Figure 6, the proposed method and gradient based methods do not produce any false positives while other methods produce false positives. This shows that the proposed and gradient based methods better than other methods in terms of false positives elimination.



(a). Input (b) Proposed (c) Laplacian (d) Edge

(e). Gradient (f) Sobel-Color (g) Uniform-Color

Figure 6. Experiment on non-text data



(a). Input (b) Proposed (c) Laplacian (d) Edge

(e). Gradient (f) Sobel-Color (g) Uniform-Color

Figure 7. Experiment on ICDAR-03 data

D. Experiments on ICDAR- 03 data

If the method works for low resolution and low contrast text images, it should work for high resolution camera based images also. Our experimental results in Figure 7 show definitely text detection methods works well for camera based images. Figure 7 shows the proposed, Laplacian and gradient based methods give better results than other methods as these are good in false positive elimination.

IV. COMPARATIVE STUDY AND DISCUSSION

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 frames in the dataset. Based on the number of blocks, the following metrics are calculated to evaluate the performance of the methods.

Detection rate (DR) = Number of truly detected text blocks / Number of actual text blocks (ground truth). **False positive rate (FPR)** = Number of falsely detected blocks / Number of truly detected blocks + number of falsely detected blocks. **Misdetection rate (MDR)** = Number of Misdetected blocks / Number of truly detected blocks. The performance of the proposed method in comparison with the existing methods is summarized in Tables I-IV respectively for experiments on horizontal data, Hua' data, Non text data and ICDAR-03 data.

To give a comparative study with the existing methods, we have chosen "Laplacian method [8]", which works based on maximum gradient difference in Laplacian values to detect text efficiently. "Edge based [9]", which basically uses different directional maps of Sobel and a set of texture features to detect text", Gradient based [10]", which is based on maximum gradient difference and identifying potential line segments and text lines, "Sobel-Color based [11]" which uses Sobel in color channels and masks to control contrast variation, and "Uniform text color based [13]", which works based on hierarchical clustering. Table I shows that the proposed method gives good detection rate and false positive rate as compared to other existing methods while misdetection rate is lower in the gradient based method. The proposed method outperforms the existing methods in term of false positive rate while detection rate is lower than Laplacian method according to Table II of Hua's data. Table III shows that text detection methods fail when a non-text frame is given as input. The methods including the proposed method produce false positives for non-text frames. However, the proposed method is better than existing methods as it detects 234 non text frames correctly without false positives out of 300 and total number of false positives lower than existing methods. Hence, the proposed method is good in false positive elimination but not good in detection of text blocks. Table IV shows that the performance of the proposed method is better than the existing methods for ICDAR-03 camera images except Laplacian method. However, detection rate is low and false positive rate is high.

This is mainly because of resizing huge image into 256×256 size resulting this, we lose color information. Thus the proposed method has low detection rate and high false positive rate than Laplacian method.

The reasons for the poor performance of the existing methods are as follows. The Laplacian method uses Laplacian mask and Sobel edge map to detect text. Therefore, it expects high contrast for text pixels. Edge based method is good for high contrast text frames but not for low contrast and small font. Gradient based method basically suffers from several thresholds for identifying text segments. Sobel color based method also suffers from thresholds which are used over mask to control the contrast. Uniform text color based method fails because of its assumption that text will have same color in video. On the other hand the proposed method gives better results because of the advantages of wavelet and color features for text enhancement.

TABLE I. PERFORMANCE ON LARGE DATA (IN %)

Methods	DR	FPR	MDR
Edge based [9]	58.2	32.4	22.1
Gradient based [10]	65.6	16.8	**3.0**
Sobel-Color based [11]	58.1	61.3	12.3
Uniform text color [13]	54.5	54.9	35.4
Laplacian [8]	84.9	26.8	16.3
Proposed	**85.3**	**10.4**	4.2

TABLE II. PERFORMANCE ON HUA'S DATA (IN %)

Methods	DR	FPR	MDR
Edge based [9]	75.4	45.8	16.3
Gradient based [10]	50.8	25.3	12.9
Sobel-Color based [11]	68.8	57.1	13.0
Uniform text color [13]	46.7	56.1	43.8
Laplacian [8]	**94.2**	8.0	**0.86**
Proposed	86.0	**4.5**	1.9

TABLE III. PERFORMANCE ON NON-TEXT DATA (IN %)

Methods	No. of Frames	No. of False Positives
Edge based [9]	62	953
Gradient based [10]	193	196
Sobel-Color based [11]	10	290
Uniform text color [13]	22	278
Laplacian [8]	44	855
Proposed	**234**	**122**

TABLE IV. PERFORMANCE ON ICDAR-03 DATA (IN %)

Methods	DR	FPR	MDR
Edge based [9]	52.7	38.7	24.4
Gradient based [10]	51.6	16.5	8.2
Sobel-Color based [11]	66.5	66.9	42.5
Uniform text color [13]	59.8	55.9	44.5
Laplacian [8]	**70.9**	**6.8**	27.2
Proposed	54.0	16.4	**6.5**

V. CONCLUSION AND FUTURE WORK

This paper presents a new method based on combination of wavelet and color features for text detection in video. Experimental results and comparative study with existing methods have shown that the proposed method outperforms the existing methods for the large dataset in terms of detection rate and false positive rate. Based on experimental results on Hua's and ICDAR-03 dataset, it is also concluded that discriminating false positives and true text blocks is not easy and we need to investigate tradeoff between false positive elimination and true text blocks detection. Future work would be improving results of the proposed method and multi-oriented text extraction.

ACKNOWLEDGMENT

This research is supported in part by the MDA grant R252-000-325-279 and A*STAR grant R252-000-402-305

REFERENCES

[1] J. Zang and R. Kasturi. "Extraction of Text Objects in Video Documents: Recent Progress". DAS, 2008, pp 5-17

[2] D. Crandall and R. Kasturi, "Robust Detection of Stylized Text Events in Digital Video", ICDAR 2001, pp 865-869.

[3] K. Jung, "Neural network-based text location in color images", Pattern Recognition Letters, 2001, pp 1503-1515.

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

[5] A.K. Jain and B. Yu. "Automatic Text Location in Images and Video Frames". Pattern Recognition, 1998, pp. 2055-2076.

[6] H. Li, D. Doermann and O. Kia. "Automatic Text Detection and Tracking in Digital Video". IEEE Transactions on Image Processing, 2000, pp 147-156.

[7] P. Shivakumara, T. Q. Phan and C. L Tan, "A Robust Wavelet Transform Based Technique for Video Text Detection", ICDAR, 2009, pp 1285-1289.

[8] T. Q. Phan, P. Shivakumara and C. L Tan,"A Laplacian Method for Video Text Detection", ICDAR, 2009, pp 66-70.

[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, 2003, pp. 1397-1406.

[11] M. Cai, J. Song and M. R. Lyu, "A New Approach for Video Text Detection", ICIP, 2002, pp 117-120.

[12] P. Shivakumara, W. Huang and C. L. Tan. "An Efficient Edge based Technique for Text Detection in Video Frames". DAS, 2008, pp 307-314.

[13] V. Y. Marinano and R. Kasturi, "Locating Uniform-Colored Text in Video Frames", ICPR, 2000, pp 539-542.

[14] X. S. Hua, L. Wenyin and H. J. Zhang, "Automatic Performance Evaluation for Video Text Detection", ICDAR 2001, pp 545-550.

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