An Efficient Hybrid Scheme for Key Frame Extraction and Text Localization in Video

*Monika Singh and Amanpreet Kaur
Department of Electrical Electronics and Communication Engineering
ITM University, Sector-23A,
Gurgaon, India
*monikasingh9@gmail.com

***Abstract*— Efficient algorithms for caption text and scene text detection in video sequences are highly in-demand in the area of multimedia indexing and data retrieval. Due to challenges like, low resolution, low contrast, complex background and texts with multiple orientation/style/color/alignment, scene text extraction from video images is undoubtedly more challenging task. In this paper, a method has been proposed to efficiently extract the key frames from the videos based on color moments and then text localization is done only on the key frames. Since the text information does not change with each frame, text extraction is performed only on key frames which help in reducing the computational/processing time of the algorithm. Further, this paper proposes a hybrid robust method to localize scene and graphic text in the video frames using 2-D haar discrete wavelet transform (DWT), Laplacian of Gaussian filter and maximum gradient difference method. DWT provides a fast decomposition of the images into an approximate and three detail components. The three detail components contain the information about the vertical, horizontal and diagonal edges of the image which are used to easily differentiate texts from image. Maximum gradient difference method is used to further refine the text localization process and the gradient difference magnitude is used in the thresholding process. A dynamic thresholding technique has been used to convert the images into binary form. Since this thresholding technique obtains different threshold values for different images, it can be used for automatic text localization in video sequences. Two mask operators has been employed to obtain an equation which when applied on each pixel provides the intended threshold value. False positives are eliminated using morphological operations and connected component analysis is done to finally localize the text. The comparison metrics in the results show that the proposed method gives a good performance of detection rate, false alarm rate and misdetection rate.**

Keywords— Shot detection; Color Moments; Key Frame Extraction; Discrete Wavelet Transform; Laplacian of Gaussian Filter; Gradient Difference.

I. INTRODUCTION

With the advancement in multimedia technology, there has been a tremendous increase in the multimedia database comprising of digital images and videos. Consequently, this has led to a requirement for efficient multimedia indexing and retrieval techniques. Video labeling and annotation based on content has been an emerging area of research in recent past. Video content can be categorized as: (i) Perceptual content - based on attributes like shape, intensity, color, texture and temporal changes and (ii) Semantic content - based on objects present in the video. Embedded text in the videos contains valuable information and could be easily used for semantic content based video annotation. Unfortunately, there isn't any one robust and consistent approach that can extract text from all kinds of videos. Video texts can be classified as: (i) Scene text are naturally captured in the video and (ii) Caption text, are embedded superficially in the video. Certain undesirable characteristics of videos like: complex background, low resolution, low intensity and various sizes, styles, colors and orientation of the video text pose a challenge to the researchers in this field. Amongst scene text and caption text, it is quite apparent that extraction of scene text is more difficult task to achieve.

In this paper, an efficient scheme is proposed to first extract key frames from the video using color moments and then discrete wavelet transform (DWT), maximum gradient difference and morphological operations are used to localize the text in the video key frames. Rest of the paper is organized as follows. Section II, provides an overview of the methodologies used and related work in this field. Proposed method is elaborated in Section III. Experimental results and comparison metrics are discussed in Section IV. Finally, the conclusions are summarized in Section V.

II. PREVIOUS WORK

Innumerous algorithms for localization, extraction and recognition of text in video image sequence had been proposed in past few years. Text detection and localization techniques can be broadly classified into: (i) Region based and (ii) Texture based techniques.

Region based methodology focuses on the region properties to extract the text, based on the verity that there is a considerable distinction in video/image text properties and its neighboring background. Edge features, color features, and connected component methods are most commonly used in the implementation of this technique. Region based methods work in downside-up fashion. Firstly, the image is segmented into small candidate character regions; these regions are further grouped/segregated to form text lines. Last step is to classify text and non-text regions.

978-1-4799-8792-4/15/$31.00 ©2015 IEEE

Texture based methods utilize the quantitative measure of the arrangement of intensities/color sub elements in a region of the video text to segregate it from the background. These techniques mostly use Gabor filters, wavelet decomposition, discrete cosine transform, FFT, spatial variance etc. to achieve the task. Firstly, the texture features are extracted from the video/image and then the text regions are located in this technique. Although texture based techniques are quite robust for complex background images as compared to region based techniques, they have a high computational complexity.

Chung-Wei Liang [1] used DWT and morphological operations in his work, for text region extraction in static images or video sequences using DWT and morphological operation. Wei et al. [2] defined a pyramidal scheme to detect text in images. D. Chen et al. [3] presented a twin-step method for detecting and recognizing text in complex images and video frames. The method consists of: (i) Fast text localization process which enables normalization of text size and (ii) a vigorous machine learning text verification process which is applied on the independent features in the background. Shivakumara et al. [4] proposed a method which used gradient difference method to segregate the candidate text regions.

The proposed method, reduces the computation time by extracting text from only the selected key frames instead of all the frames which would be rather time consuming. Further, the video texts are usually multi-oriented and using DWT helps to find the edge details in the horizontal, vertical and diagonal simultaneously.

III. PROPOSED METHODOLOGY

The flowchart of the proposed framework for text localization video is shown in fig. 1. The details of each processing block are discussed below.

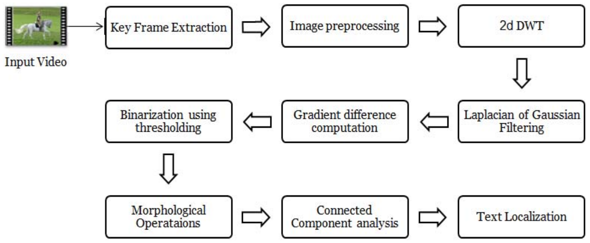

Input Video
Key Frame Extraction
Image preprocessing
2d DWT
Laplacian of Gaussian Filtering
Gradient difference computation
Binarization using thresholding
Morphological Operataions
Connected Component analysis
Text Localization

Fig. 1 Flowchart of the proposed method

A. Key Frame Extraction

A video comprises of number of frames which run in succession to appear as a seamless video. When a video is shot the pictures are usually taken contiguously by a single camera and represent a continuous action in time and space. These basic temporal units of a video sequence are called shots. Shots are separated from one another by gradual and abrupt transitions. These transitions can be detected by extracting the characteristic features of each frame and if the dissimilarity index is very high for consecutive frames then it marks the presence of a shot transition. In the proposed model, a shot is detected based on the color moments [5] of the frames and Euclidean distance measure is used to measure the congruity between the frames.

To compute the color moments, first the RGB values of each frame are converted into YIQ color model. YIQ color map isolates the luminance Y and chrominance I & Q components. The Intensity image Y is given by the equation below:

$$Y = 0.299R + 0.587G + 0.114B \qquad (1)$$

Formula for the calculation of h^{th}; $h = 1,2,3,\ldots,$ color moment of the i^{th} color component is given by [5] :

$$M_i^h = \left(\frac{1}{N}\sum_{k=1}^{M}(p_{i,k} - M_i^1)^h\right)^{\frac{1}{h}} \qquad (2)$$

where, $p_{i,k}$ is the intensity of the i^{th} color component of the k^{th} pixel of a frame and N is the total number of pixels in the frame. However for the ease in computation, here we have used the color moments for only the Y channel of the image and only first 2 color moments (mean and standard deviation) are calculated. Based on the color moments, a feature vector is created for each frame which is given as :

$$F_j = [\alpha_1 M_1^1, \alpha_1 M_1^2, \ldots, \alpha_1 M_1^H, \alpha_1 M_1^1, \alpha_2 M_2^2, \ldots, \alpha_2 M_2^H] \quad (3)$$

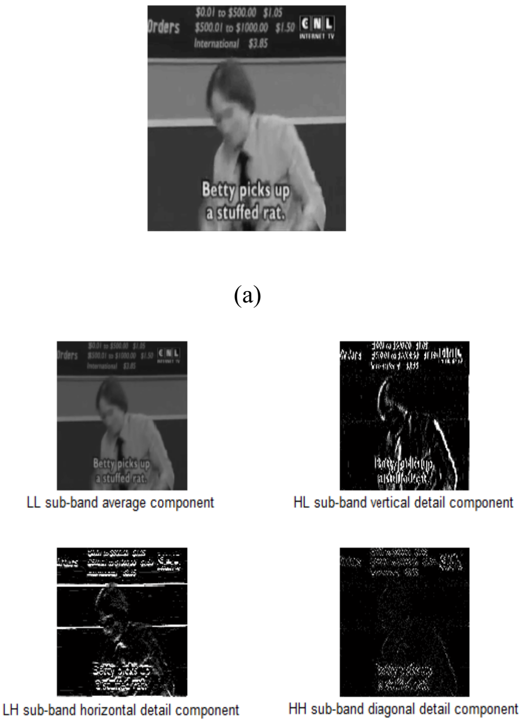
Then the Euclidean distance is calculated between the consequential frames k_j and k_{j-1} as follows:

$$D = d\left(F_j - F_{j-1}\right) - d(F_{j-1} - F_{j-2}) \qquad (4)$$

where $d\left(F_j - F_{j-1}\right) = \sum_{k=1}^{z}\left|F_j(k) - F_{j-1}(k)\right|$ and q is set to 2. If $D > T$ then a key frame is detected. Threshold T is set as equal to mean of all such Euclidean distances.

B. Edge Map Extraction

1) *2-d Discrete Wavelet Transform:* Once the key frames are extracted, each frame is converted into a gray-scale image if it is a colored video frame. Discrete wavelet transform (DWT) [6] is one of the useful transforms for multi-resolution analysis of images. In two-dimensional DWT, the input image signal is decomposed into four subbands, viz. low-low (LL), low-high (LH), high-low (HL), and high-high (HH). This is accomplished by first filtering the image along the row and is decimated by two. The sub-image thus obtained is filtered along the column and decimate it by two. LL sub-band provides an average component of the image while LH, HL and HH components provide horizontal, vertical and diagonal detail components respectively. Usually, the text contains all three types of the edges provided by the detail components. The primary reason for using 2-d DWT for edge extraction is that it can detect all three kinds of edges simultaneously as compared to traditional edge detectors, as a result this reduces the processing time. Another advantage of using DWT is that it can remove noise while other edge detectors falsely identify noise pixels as edge pixels.



(a)

LL sub-band average component

HL sub-band vertical detail component

LH sub-band horizontal detail component

HH sub-band diagonal detail component

(b)

Fig. 2. (a) Gray image of the original image (b) DWT coefficients

2) *Laplacian of Gaussian Mask Filtering:* The three detail edge components separated using DWT are filtered using 5x5 Laplacian of Gaussian (LoG) mask operator to localize the text blocks in each detail component. Low pass filters are used for smoothing of the image by removing the noise. These filters usually employ mask operations i.e., moving window operator which modifies each pixel of the image at a time, according to their neighborhood contained in the window region. For edge detection the Laplacian of an image is taken since it highlights the regions of quick intensity change. The resultant image after applying Laplacian filter image contains negative and positive value for every edge and the changeover between these values (zero junctions) are used to identify the text and background transitions. The Gaussian smoothing operation reduces the noise sensitivity problem while detecting zero crossing using Laplacian operator by band-limiting the image to a small range of frequencies. Let be the Gaussian standard deviation then the 2-d LoG function centered at zero is given as follows:

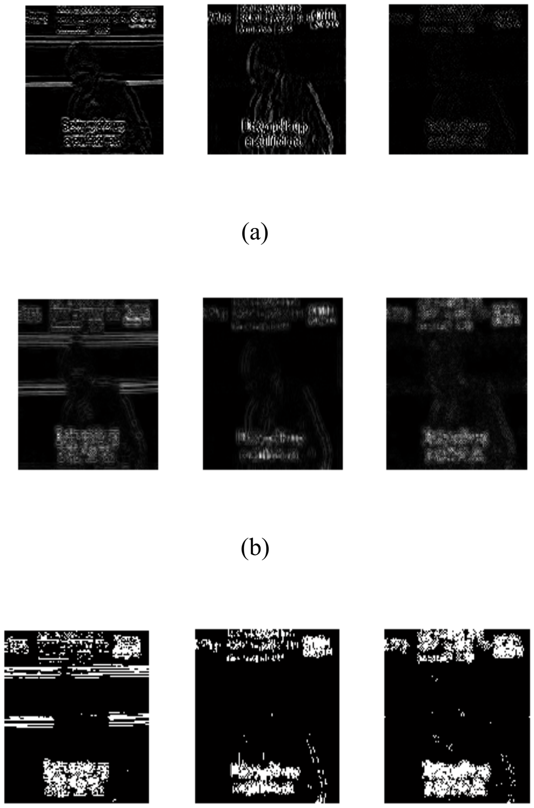
$$LoG(x,y) = -\frac{1}{\pi\sigma^4}\left[1-\frac{x^2+y^2}{2\sigma^2}\right]e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

C. *Maximum Gradient Difference*

The gradient information of the text areas in an image is considerably different from the non-text background. The positive and negative values in the LoG-filtered detail components of the image are captured using the maximum gradient difference (MGD) method [7]. MGD is represented as the difference of the maximum and minimum values within a local $1 \times N$ window [8]. For the LoG-filtered image F the MGD at pixel (x, y) are computed as follows [10]:

$$MGD(x,y) = \max\big(F(x,y)\big) - \min\,(F(x,y)) \tag{6}$$

Typically text regions have larger MGD value as compared to the background and hence text regions appear brighter in the gradient map which is obtained by sliding a window over the image.



(a)

(b)

(c)

Fig.3. (a) Laplacian of Gaussian filtered detail components (b) Gradient images of the detail components (c) Results after binarization of LH, HL and HH respectively.

D. *Binarization*

After obtaining the gradient map, each of the detail components of the image is converted into binary form. Here we have employed dynamic thresholding technique [1] for binarization. For setting the threshold value, two mask operators are used to obtain an equation. This equation when performed on each pixel with its neighboring pixels gives the value of threshold. Threshold value is computed for the gradient map of all the three detail components. This is a dynamic thresholding method because it provides different values for different images. Let G be the gradient map of a detail component then threshold T_G for G is given by;

$$T_G = \frac{\sum(G(i,k) \times h(i,k))}{\sum h(i,k)} \qquad (7)$$

where, $h(i,k) = Max(|m_1 ** G(i,k)|)$ and $m1 = [-1 \quad 0 \quad 1], m2 = [-1 \quad 0 \quad 1]^t$ are the mask operators.

E. Text Localization

The basic steps used for text localization in this method are:

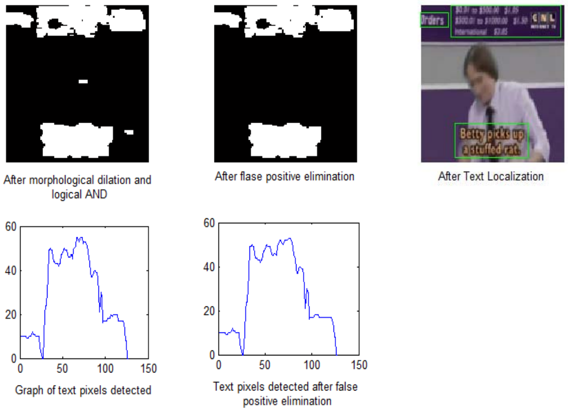
1. Morphological operations: Morphological dilation is performed on binary images of the 3 detail components using different structuring elements for each component. In this case, 3×8 rectangle is used as a structuring element for horizontal and diagonal components and a rectangle of 5×8 is used for vertical component.

2. Logical AND: Since the texts contain vertical, horizontal and diagonal components, these three components were then combined together using logical AND operation to further separate text regions.

3. False positive elimination: In the final image the connected components are labeled using 8-connectivity but it might still contain some false positives and following geometrical rules were used to localize candidate text regions. These values have been derived experimentally.

   a) $Area \leq 70$

   b) $Width > 60, Height < 20$



After morphological dilation and logical AND

After flase positive elimination

After Text Localization

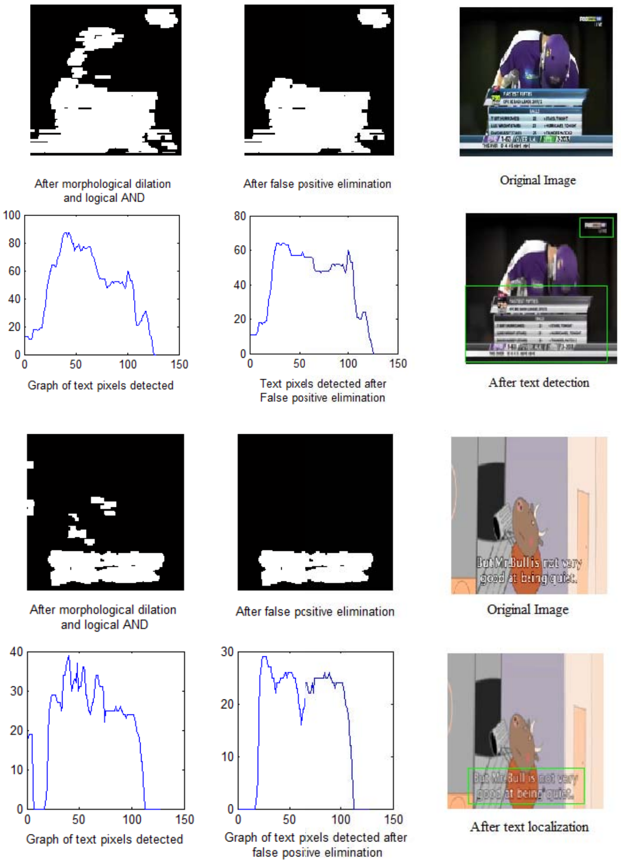
Graph of text pixels detected

Text pixels detected after false positive elimination

Fig.4 Intermediate results for text localization

IV. Experimental Results and Comparison Metrics

The proposed algorithm is implemented using MATLAB software. For the experimental purpose we have created our own video dataset which includes cartoon videos, sports video and educational video.This algorithm has also been tested on some images. The results of text localization in sample video and images are given in fig. 5 and fig. 6.



After morphological dilation and logical AND

After false positive elimination

Original Image

Graph of text pixels detected

Text pixels detected after False positive elimination

After text detection

After morphological dilation and logical AND

After false pcsitive elimination

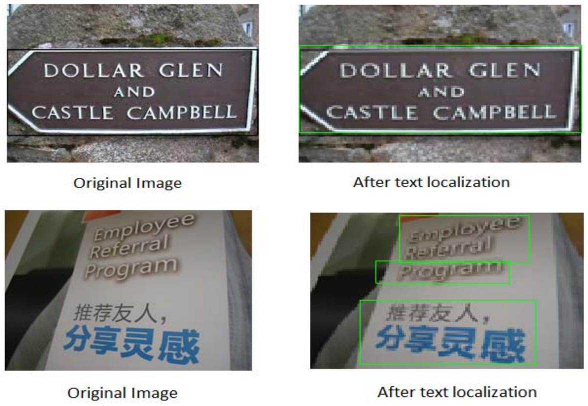
Original Image

Graph of text pixels detected

Graph of text pxels detected after false posiive elimination

After text localization

Fig. 5. Text localization results for sample videos



Original Image

After text localization

Original Image

After text localization

Fig. 6. Text localization result for sample images

The performance evaluation of the method is done using the well-known existing method of 'Precision and Recall' and F-measure.

$$P = \frac{CorrectlyDetectedBlocks}{CorrectlyDetectedBlocks+FalsePositives} \quad (8)$$

$$R = \frac{CorrectlyDetectedBlocks}{CorrectlyDetectedBlocks+FalseNegatives} \quad (9)$$

where, false positives are those regions in the image which have been detected as text regions, but in actual are not text regions. False Negatives are those candidate regions in the image which are actually text regions, but have not been detected. F-measure is the combined measure that expresses precision and recall trade off. It is given as follows:

$$F = \frac{2PR}{P+R} \quad (10)$$

The final number of blocks detected for text localization are manually counted and categorized as follows:

1) Candidate text blocks (CTB): is the total number of blocks detected.
2) True text blocks (TTB): is the number of the candidate blocks which contain text.
3) False text blocks (FTB): is the number of falsely detected text blocks i.e., blocks without text characters.
4) Missing text blocks (MTB): is the number of incomplete detected text blocks i.e., blocks with missing text characters.

Using above parameter values, precision, recall and F-measure were calculated for different types of videos and the results thus obtained are tabulated below:

TABLE I.

COMPARISON METRICS FOR SAMPLE VIDEOS

Test videos	**Key Frames**	**CTB**	**Precision (%)**	**Recall (%)**	**F-measure (%)**
Cartoon Video	9	9	100	100	100
Sports Clip	21	41	87.8	85.7	86
Educational Video	7	21	72	90	80

The results show that the suggested method has a good detection rate and recall for both, video frames and still images.

V. Conclusions

In this paper, we have suggested a simple and efficient hybrid algorithm to localize the text contained in the video frames. The method gives good result for still images also. The key frame extraction and use of 2-D DWT increases the efficiency of the algorithm by decreasing the computational time. The thresholding technique used for binarization is dynamic and is based on the maximum gradient difference magnitude. It gives different thresholds for different images. Limitation of this method is that it can only partially detect the text in the video frames with complex backgrounds and high illumination. The method fails to localize the text when the background and text intensity are comparable. Future work in this method can focus on first categorize the images into low resolution and high resolution images and then perform different algorithms based on the image category.

References

[1] Chung-Wei Liang and Po-Yueh Chen, “DWT Based Text Localization”, International Journal of Applied Science and Engineering, 2004, pp. 105-116.

[2] Wei, Y. C., & Lin, C. H., “A robust video text detection approach using SVM”, Expert Systems with Applications, 39(12), 2012, pp. 10832-10840.

[3] Chen, D., Odobez, J. M., and Bourlard, H., “Text detection and recognition in images and video frames”, Pattern Recognition, 37(3),2004, pp. 595-608.

[4] Shivakumara, P., Basavaraju, H. T., Guru, D. S., & Tan, C. L., “Detection of Curved Text in Video: Quad Tree Based Method”, Document Analysis and Recognition (ICDAR), IEEE, August, 2013, pp. 594-598.

[5] Shekar, B., Kumari, M. S., and Holla, R., “An efficient and accurate shot boundary detection technique based on colour moments”, International Journal of Artificial Intelligence and Knowledge Discovery, vol. 1, no. 1, 2011, pp. 77–80.

[6] Mallat, S. G., “A theory for multiresolution signal decomposition: the wavelet representation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11, 7, 1989, pp. 674-693.

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

[8] Phan, T. Q., Shivakumara, P., and Tan, C., “A laplacian method for video text detection,” in Document Analysis and Recognition (ICDAR), 2009, pp. 66–70.

[9] Shivakumara, P., Phan, T. Q., and Tan, C., “A gradient difference based technique for video text detection,” in Document Analysis and Recognition, ICDAR, 2009, pp. 156–160.

[10] Shivakumara, P., Phan, T. Q., & Tan, C. L., “A laplacian approach to multi-oriented text detection in video”, Pattern Analysis and Machine Intelligence, IEEE, 33(2), 2011, pp. 412-419.

[11] B.H.Shekhar, Smitha M.L, Shivkumara, P., “Discrete wavelet transform and gradient difference based approach for text localization in videos”, IEEE , 2014, pp.280-284.