Analysis of an Automatic Text Content Extraction Approach in Noisy Video Images

C.P. Sumathi
Department of Computer Science
SDNB Vaishnav College
For Women, Chennai

N. Priya
Department of Computer Science
SDNB Vaishnav College
For Women, Chennai

ABSTRACT

Text either embedded or superimposed within video frames is very useful for describing the contents of the frames, as it enables both keyword and free-text based search, automatic video logging, and video cataloging. Low contrast, noise and poor quality are the main problems of text extraction in video images. This article explores a novel approach for text extraction from video frames, which can handle complex image backgrounds with different font sizes, font styles, and font appearances such as normal and noisy video. The pre processing is done to de-noise the images through wavelet based approach by removing noise in the frequency field and reducing by the soft-threshold method. Then, the enhanced image is obtained through the inverse wavelet transform. The Morphological operators are applied to sharpen the image for clear edges and to detect the connected components accurately. Lastly, features are extracted and fed into an artificial neural network to classify the text pixel from that of the background of the image. A quantitative measure of comparison and analysis is provided by the different parameters with different noises.

Keywords

Morphological Operators, Neural Network, Text Extraction, Wavelet

1. INTRODUCTION

Text objects embedded in videos contain much semantic information related to the video content. Therefore, the extraction of text objects plays an important role in content-based information indexing and retrieval systems. Manual annotation of video is extremely time consuming, expensive, and unscalable in the face of ever growing video databases. Therefore, automatic extraction of video descriptions is desirable in order to annotate and search large video databases often superimposed on the frames in textual form. Video commercials ensure that the product and other shopping information are presented as readable text. When video text is automatically extracted, it not only provides keywords for annotation for search of image and video libraries but also aids in highlighting events which can then be used for summarizing a video. Text extracted can also be used in video categorization, cataloging of commercials, logging of key events, and efficient video digest construction.

Text in video can be divided into Scene text and caption text. Scene text is text that occurs naturally in the 3-D scene being recorded and is distorted by perspective projection. Caption text comprises 2-D strings that are composited on to the video frame during the editing stage of production. Video frames are typically noisy, low–resolution, full–color with interlace artifacts. The text in a video frame can be multi–colored, multi–font, and be transparent, with the background showing through. Since caption text is usually composited on to the video frame, the background behind and around the text characters can be changing even if the text is stationary. In this paper the proposed method tends to provide an efficient and effective approach to the issue of text content extraction for a wider range of noisy video images.

2. PREVIOUS WORK

Many text detection approaches have been proposed since several decades. However, due to low resolution and complex backgrounds of videos and various sizes, colors, styles and alignments of text, text detection and extraction are still challenging [1][2][3]. Smith [4] used vertical edge information for localizing caption text in images while Jung [5] used a neural network based filtering scheme to classify the pixels of input image as belonging to text or non-text regions. Jiang et al. [6] applied merging bounding blocks using special color features, edge features and morphology operator. These features are used to eliminate the false text candidates. However, this method is script dependent and is reported to be working well for Chinese documents. Yuan & Tan [7] used edge information to extract textual blocks in Manhattan layout. Ariki and Teranishi [8] assumed text pixels to have a minimum intensity and use frame subtraction to isolate the characters in video. Yeo [9] proposed a method for the detection of text caption events in video by modified scene change detection. The authors attributed any changes in the video frame which are not due to shot changes to cause of abrupt caption appearances or disappearances. Hauptmann and Smith [10] performed text localization in video and used the heuristic that text regions consist of a large number of horizontal and vertical edges in spatial proximity. 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

However, there are many differences between the proposed method and the previous approaches such as de-noising the image using wavelet transformation and soft thresholding technique, sharpening the edges using morphological operators, features of text extracted using statistical measures and the classification of text and background using neural network. Fig 1 shows the flow of the complete process of the proposed technique and the rest of the paper is organized as follows. In section 3, the methodology of text extraction technique is described. Analysis and comparison with proposed algorithm are presented in section 4 followed by conclusion in Section 5.

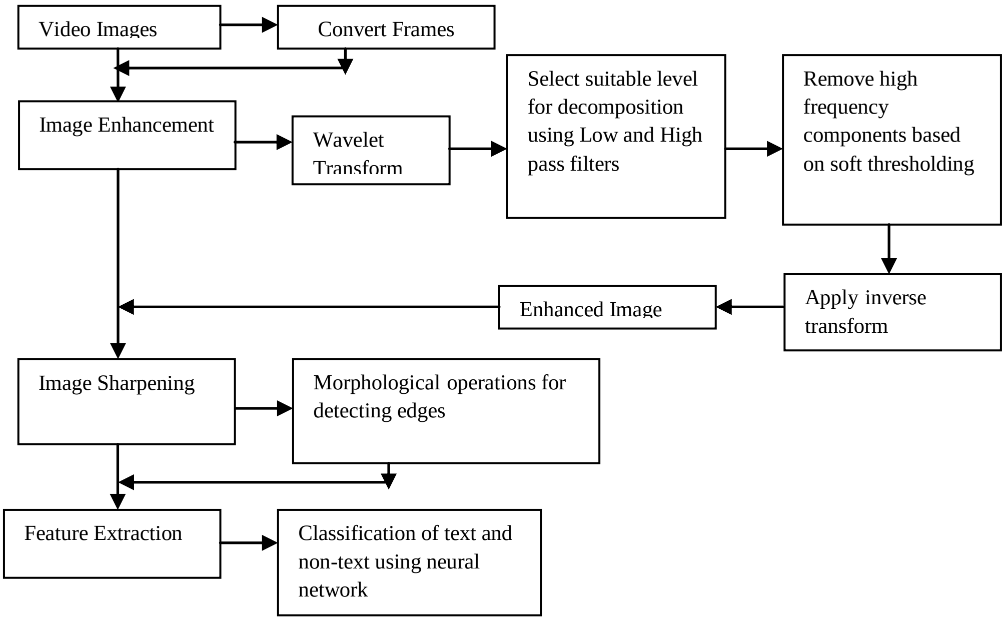
Video Images
Convert Frames
Image Enhancement
Wavelet Transform
Select suitable level for decomposition using Low and High pass filters
Remove high frequency components based on soft thresholding
Apply inverse transform
Enhanced Image
Image Sharpening
Morphological operations for detecting edges
Feature Extraction
Classification of text and non-text using neural network

Fig 1: Flow of Proposed Technique

3. METHODOLOGY

3.1 Image Enhancement

Image enhancement is a process which principally focuses on processing an image in such a way that the processed image is more suitable than the original one for the specific application. Image noise is unwanted fluctuations. There are various types of image noises present in the image like gaussian noise, salt & pepper noise, speckle noise, shot noise, white noise[11] and there are various noise reduction techniques which are used for removing them. Most of the standard algorithms are used to de-noise the noisy image and perform the individual filtering process. The result is that it generally reduces the noise level. But the image is either blurred or over smoothed due to losses like edges or lines. Noise reduction should be done to remove the noise without losing much detail present in an image [12]. To achieve this goal, we use the mathematical function known as the wavelet transform to localize an image into different frequency components or useful sub-bands and effectively reduce the noise in the sub-bands.

3.1.1 Wavelet Transform

The wavelet transform has become a useful computational tool for noise reduction in signal. For many signals, the low-frequency content is the most important part. It is what gives the signal its identity. On the other hand, the high frequency content imparts flavor or nuance. Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. The basis of the wavelet transform is to decompose an image into approximations and details by down sampling through the low and high pass decomposition filters(L and H). The approximations are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components. The type of wavelet transform is designed to be easily reversible which means that the original signal can be easily recovered by upsampling with reconsruction filters (L' and H') after it has been transformed. Decomposition and Reconstruction are shown in the fig 2.

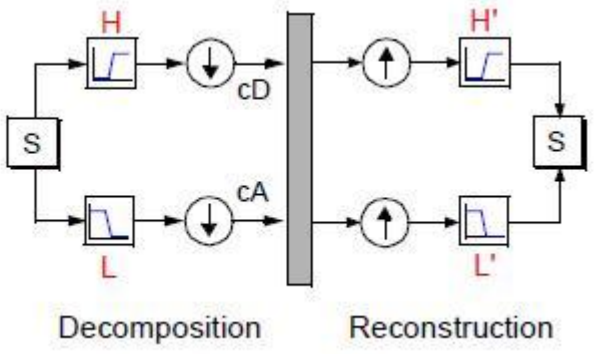

Fig 2: Decomposition and Reconstruction of wavelet transform

Depending on the application, the choice of the suitable wavelets and the corresponding decomposition levels are based on different criteria. In this experiment, the image decomposition is implemented by Daubechies wavelet with two levels (fig 3) and the noise is removed in the frequency field based on soft-threshold method. Then the image is reconstructed to the original signal with no loss of information.

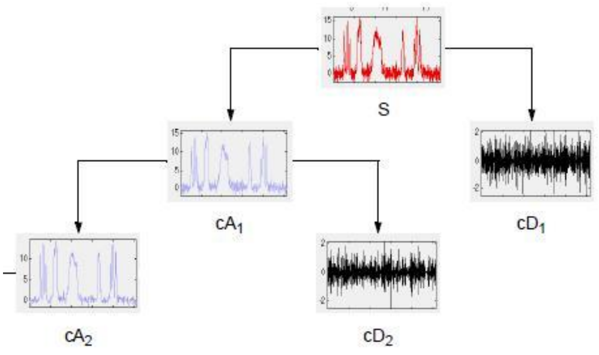

Fig 3: Two Levels of Decomposition

The decomposition of the wavelet is given by (1) and (2)

$$c(n) = h_0 x(2n) + h_1 x(2n+1) + h_2 x(2n+2) + h_3 (2n+3) \quad (1)$$

$$d(n) = h_3 x(2n) - h_2 x(2n+1) + h_1 x(2n+2) - h_0 (2n+3) \quad (2)$$

where the multipliers are:

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}, \quad h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}, \quad h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}, \quad h_1 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

The inverse block that reverses this decomposition uses the same multipliers and is given by (3) and (4)

$$y(2n) = h_0 c(n) + h_2 c(n-1) + h_3 d(n) + h_1 d(n-1) \quad (3)$$

$$y(2n+1) = h_1 c(n) + h_3 c(n-1) - h_2 d(n) - h_0 d(n-1) \quad (4)$$

The addition of noise to the original image is shown in fig 4(a) and de-noise the image using wavelet transformation is shown in fig 4(b).

3.2 Image Sharpening

Sharpening techniques improve the clarity of digital images by enhancing the objects which are present in the scene. This improves their edges and their details, giving greater neatness and depth to the images. The edges are considered as a very important portion of the perceptual information content in an image. Edges can be found when the difference between luminance intensity from one point to the other appears. Practically, the more the difference of light luminance, the edges are easier to define. In contrast the lesser the difference in the intensity harder the edges to be defined. Edge detection evaluates the brightness of each area with difference luminance. Mathematical morphology is a topological and geometrical based approach for image analysis. It provides powerful tools for extracting geometrical structures and for representing shapes in many applications. Morphological operations are very effective in the detection of boundaries in a binary image. The most basic morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. These two operations can be combined with opening and closing operations for boundary detection. Opening operation is used to smoothen the inner object contour to break narrow strips and eliminates thin portions of the image. It is also used to remove the noise. Closing operations fills the small holes and gaps in a single-pixel object.

Let $S_{m,n}$ denote a structure element with the size m x n, where m and n are odds and larger than zero and $I_{x,y}$ denote a gray-level input image. According to the definition of $S_{m,n}$ the smoothing, dilation, erosion, closing, opening, and other operations are mathematically represented as

Smoothing operation:

$$E_{S_{m,n}}(I(x,y)) = \frac{1}{mn} \sum_{i=-m/2}^{m/2} I(x+i, y+j) S_{m,n}(i,j), \quad (5)$$

dilation operation:

$$I(x,y) \oplus S_{m,n} = \max_{|i| \le m/2, |j| \le n/2} I(x-i, y-j) S_{m,n}(i,j), \quad (6)$$

erosion operation:

$$I(x,y) \Theta S_{m,n} = \min_{|i| \le m/2, |j| \le n/2} I(x-i, y-j) S_{m,n}(i,j), \quad (7)$$

closing operation:

$$I(x,y) \bullet S_{m,n} = (I(x,y) \oplus S_{m,n}) \Theta S_{m,n}, \quad (8)$$

opening operation:

$$I(x,y) \circ S_{m,n} = (I(x,y) \Theta S_{m,n}) \oplus S_{m,n}, \quad (9)$$

and differencing operation:

$$I(x,y) \circ S_{m,n} = (I(x,y) \Theta S_{m,n}) \oplus S_{m,n}, \quad (10)$$

The sharpening model in this research begins with brightness enhancement and then followed by edge detection, using the above operations. The output image appears sharpened. But, this sharpening operation is selective. Edges of big objects, which are preserved with distance, appear enhanced. On the other hand, small details, which are smoothed with distance, are not sharpened. The edge detection evaluation $S_{(x,y)}$ and position (x, y) may need a suitable threshold level to recover the loss of edge from the small connected components in an image. The threshold value is obtained using the function given below.

thresholding operation:

$$T(I(x,y)) = \begin{cases} 255, & if\,(I(x,y) > T \\ 0, & otherwise \end{cases} \quad (11)$$

The technique used in this research separates the thin layers of the area and determines the differences of each pixel accurately as shown in fig 5.

3.3 Feature Selection

Feature selection is a special form of dimensionality reduction that involves simplifying the amount of resources required to describe a large set of data accurately. Selection criteria usually involve the minimization of a specific measure of predictive error for models fit to different subsets. The criterion for selecting statistical features in the proposed method is to look at geometric object level features. These object level features include Min, Max, area, center of area, axis of least second moment, perimeter, Euler number, projections, thinness ratio, and aspect ratio. Min, Max, Area, center of area, axis of least second moment, perimeter determines about where the object is. The other four conveys the shape of the objects. These features are extracted from the clearly defined labeled components after the image segmentation and transforms. Some of the features are explained below:

The area Ai is measured in pixels and indicates the relative size of the object.

$$Ai = \sum\sum Ii(r,c)(0 <= r,c <= N-1) \qquad (12)$$

Where:

$$Ii(r,c) = 1 \quad if \;\; I(r,c) = i^{th} \; object$$
$$0 \quad otherwise$$

We define the center of area for an object by the pair (ri, ci)

$$ri = \left(\sum\sum rIi(r,c)\right)/\,Ai \qquad (0 <= r,c <= N-1) \qquad (13)$$
$$ci = \left(\sum\sum cIi(r,c)\right)/\,Ai \qquad (0 <= r,c <= N-1) \qquad (14)$$

(sum of the r, c value of all the pixels divides the number of pixels)

Theses correspond to the row coordinate of the center of area for the i^{th} object ri, and the column coordinate of the center of area for the i^{th} object ci . This feature will help to locate an object in the two-dimensional image plan.

The axis of least second moment provides information about the object's orientation. This axis corresponds to the line about which it takes the least amount of energy to spin an object of like shape. If we move the origin to the center of area r, c , the axis of least second moment is defined as:

$$\tan(\,2\theta_i) = 2\sum\sum rcI_i(r,c)/\left(\sum\sum r^2 I_i(r,c) - \sum\sum c^2 I_i(r,c)\right) \qquad (15)$$

The perimeter of the object can help us to locate in space and provide information about the shape of the object. The perimeter can be found by applying edge detector to the object, followed by counting the '1' pixels that have '0' pixels as neighbors.

After feature computation, the above features are normalized and form the feature vector representation for each connected pixel. From that vector the texture properties are determined to classify the text and non-text in the video images. The first ten features in Table 2 are selected for classification.

3.4 Text Extraction Using Classification Technique

The classification module creates a learning pattern and creates a neural network classifier. Using the 10 values of features extracted structurally, it creates the learning pattern. Finally, the neural network classifier is trained with input data using the learning pattern. The feed forward neural network classifier consists of three layers with an input layer, a hidden layer, and an output layer. The input layer has 10 input nodes, the hidden layer has 40 hidden nodes, and the output layer has 1 output node. The neural network is trained and change weights until the minimum error reduces to 0.1. Based on that classification, the text blocks are extracted from the sharpened image.

4. IMPLEMENTATIONS OF THE PROPOSED ALGORITHM

In order to analyze the performance of this proposed approach the own dataset have created as there is no standard dataset available in literature, which includes 300 different images taken from variety of video images such as movies, news and sports video. The method implemented using MATLAB software is run on a PC with Celeron processor. The approximate processing time for each video image of size $256X256$ is about 17.2 seconds for text detection. Analysis and comparisons are based on different noises occurred in digital video images which arise during image acquisition or transmission of the transformed image. Then the performance of text extraction is also analyzed using certain parameters.

4.1 Experimental Results

4.1.1 Enhancement Model :

The objective of the first step in this experiment of video images is to remove the background containing complicated low resolution objects such as Gaussian noise, speckle noise and salt and pepper noise. The idea is to decompose the image with wavelet transform and remove the high frequency components with soft thresholding filtering method and produce different enhancement weight coefficients in different sub-images which are used to enhance an image. The detailed process is as follows.

An image can be seen as a 2D signal, an image's edge feature information in approximation and detail information are distributed in high-frequency sub-images. When an image is decomposed through wavelet transform of k scales, $3k+1$ sub-images can be obtained:

$$\{LL_k, HL_j, LH_j, HH_j\}$$

where $j = 1,2,...k,$ k denotes the image's decomposition scale levels of wavelet transform, LL_k denotes the k^{th} scale level low-frequency subimage, and HL_j, LH_j, HH_j denote the j^{th} scale level high frequency sub-images. There is an abundance of image detail information in high-frequency sub-images. But there are also plenty of noises in these sub-images. The wavelet transform's smooth function can help us to reduce an image's noise, but it cannot meet out the requirements. A non-linear soft thresholding method is used to reduce the noises of high frequency sub-images. The formula of reducing noise is given by

$$G(x,y) = \begin{cases} H(x,y) - T_{jil}, & H(x,y) \geq T_{jil} \\ 0, & -T_{jil} \leq H(x,y) \leq T_{jil} \\ H(x,y) + T_{jil}, & H(x,y) \leq -T_{jil} \end{cases} \qquad (16)$$

Where T_{jil} are soft threshold values of the jil^{th} sub-image. $H(x, y)$ denotes the high frequency coefficient of the position (x, y) in the jil^{th} sub-image, and $G(x, y)$ denotes the coefficient of the position (x, y) de-noised. Through the inverse wavelet transform the enhanced image was generated.

Different noises and filters are used for experiments and compared with this enhancement approach. Three objective measures, Peak signal-to-noise ratio(PSNR), Root Mean Squared Error (RMSE), Pearson Correlation Coefficient(PCC) have been used to evaluate the performance and analysis of this technique. These measures are most commonly used as a measure of quality of reconstructed and noisy image. PSNR is most easily defined via the RMSE. Given a noise-free $m \times n$ monochrome image I and its noisy approximation K, RMSE and PSNR are defined as:

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i, j) - K(i, j)]^2} \qquad (17)$$

$$PSNR = 20 \cdot \log_{10}\left(\frac{MAX_I}{RMSE}\right) \qquad (18)$$

PCC is a technique to show the association between the reconstructed and the noisy image with the continuous data ranging between -1 and +1. Table 1 gives a quantitative account of the performance of our system on five sequences with ground truth.

It can be observed from Table 1 that our system performs very well in terms of measures where the image appears much brighter than the other filters and the enhanced image are shown in the fig 4.

Table 1: Performance and Analysis of Proposed Enhancement with different filters

Filters		**Gaussian Noise**			**Speckle Noise**			**Salt & Pepper Noise**		
		PSNR	**RMSE**	**PCC**	**PSNR**	**RMSE**	**PCC**	**PSNR**	**RMSE**	**PCC**
Gaussian	[Video_1]	28.07	9.18	0.87	29.87	8.18	0.90	31.45	6.82	0.92
	[Video_2]	28.24	9.87	0.85	29.23	8.81	0.88	31.06	7.13	0.91
	[Video_3]	30.46	7.65	0.97	32.16	6.29	0.99	34.80	4.6	0.92
	[Video_4]	30.70	7.43	0.90	33.10	5.64	0.98	37.71	3.31	0.99
	[Video_5]	29.86	8.19	0.82	30.84	7.32	0.92	30.78	7.37	0.90
Average	[Video_1]	28.72	9.34	0.82	30.84	7.32	0.91	32.71	5.90	0.91
	[Video_2]	28.91	9.14	0.84	30.59	7.54	0.89	31.89	5.14	0.89
	[Video_3]	29.07	8.96	0.88	32.77	5.86	0.94	34.86	4.60	0.92
	[Video_4]	29.13	8.91	0.89	33.58	5.36	0.92	27.29	5.14	0.81
	[Video_5]	28.88	9.16	0.90	32.12	6.31	0.91	31.10	6.12	0.87
Laplacian	[Video_1]	25.39	13.70	-0.35	24.61	8.18	-0.03	24.52	15.15	-0.20
	[Video_2]	25.06	14.23	-0.38	24.46	15.24	-0.10	24.45	15.27	-0.28
	[Video_3]	26.09	12.64	-0.28	25.27	13.89	0.10	25.16	14.06	-0.12
	[Video_4]	26.17	12.52	-0.33	25.18	14.03	0.06	25.06	14.23	-0.15
	[Video_5]	25.65	13.30	-0.31	24.56	15.07	0.07	25.06	14.23	-0.15
Proposed	[Video_1]	31.44	6.83	0.98	34.85	4.61	0.99	37.37	3.45	0.99
Enhancement	[Video_2]	31.81	6.54	0.98	35.99	4.04	0.98	40.22	2.48	0.99
	[Video_3]	31.94	6.45	0.98	37.73	3.31	0.99	39.71	2.64	0.99
	[Video_4]	32.02	6.38	0.99	38.77	2.93	0.99	40.61	2.37	0.99
	[Video_5]	31.68	6.63	0.98	36.53	3.79	0.99	38.95	2.88	0.99

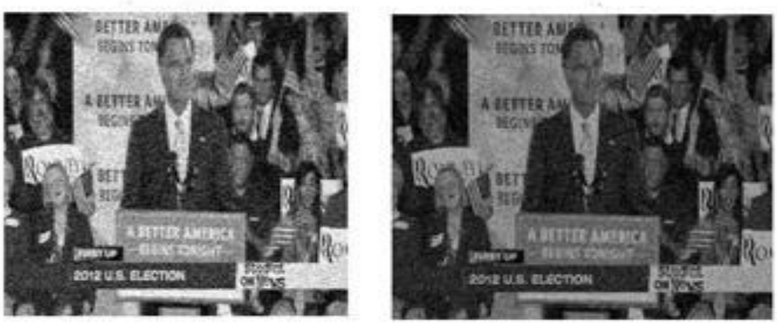

Fig 4: (a) Noisy Image (b) Enhanced Image

4.1.2 Sharpening and Feature Extraction Model:

In the next step of the experiment, the sharpening technique is implemented from the above morphological operations such as opening, closing, differencing and thresholding with a 3×3 structure element. This method separates the thin layers of the area and determines the differences of each pixel accurately and detect the boundaries clearly. Then the algorithm is tried to find out the connected components in a binary image using Connected Component Labelling method. These components have been labeled to identify the candidate text of the image. It can operate on the resulting binary image from a thresholding step.

From the threshold image, a connected component algorithm is performed to find the labelled regions of connected pixels which have the same value. Fig 5 shows an experimental result for this stage.

In the next stage, we employ the statistical measures in the connected component to capture the text property in the video image. More specifically, the above 10 object features mentioned in section 3.3 are extracted and computed. Table 2 lists the samples of each feature. Three conclusions can be drawn from Table2. First, the objects between the $mean-std$ and $mean+std$ in area then second, the objects between the $mean-std$ and $mean+std$ in the center of area and third based on the axis of least second moment i.e. object orientation gives more text property of an image. In order to verify the above conclusions, we trained and tested the data in a neural network classifier to obtain the trustworthy result.

Table 2: Samples of statistical measures

Array1	Amean	Astd	CArea1	CArea2	Cmean1	Cmean2	Cstd1	Cstd2	Orientation
693	47	243	32	160	169	105	92	64	8
14	47	243	2	167	169	105	92	64	107
28	47	243	8	229	169	105	92	64	14
3	47	243	6	21	169	105	92	64	90
466	47	243	28	28	169	105	92	64	7
8	86	639	11	61	174	99	102	64	5
8	86	639	15	27	174	99	102	64	90
3	86	639	15	67	174	99	102	64	90
1200	86	639	44	92	174	99	102	64	74
2	86	639	20	18	174	99	102	64	90
4	86	639	22	39	174	99	102	64	18
7	86	639	23	10	174	99	102	64	-15
24	86	639	25	90	174	99	102	64	113
3	86	639	24	76	174	99	102	64	90
3	86	639	48	227	174	99	102	64	0
2	86	639	49	124	174	99	102	64	0

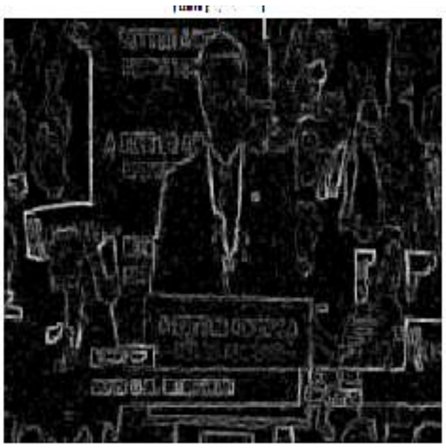

Fig 5: Edge Detections using Morphological Operations

4.1.3 Training and Testing Model:

After selecting the features, we train and test the text and non-text samples with neural network classifier. The idea is that 300 samples are collected from the different enhanced video images for training and testing to classify as text and non-text. For training, 200 samples were taken and 100 samples for testing which includes a complete set of text samples and a portion of non-text samples. After training, the neural network maps each block into a real value between 0 and 1 for non-text and text respectively. In this experiment, the text and non-text classes are well separated when using the neural network as a classifier. Fig 6 shows the distribution of neural network outputs when testing on text and non-text blocks.

4.1.4 Text Extraction Performance:

After training the neural network, text components are extracted based on the statistical measures. To evaluate the performance of this method, the precision, recall, Fragmentation rate and error rate were used. Recall is the ratio of the number $Num_{Correct}$ of correct text detected by the algorithm to the total number Num_{Actual} of actual text appearing in the test images, i.e.,

$$\mathrm{Re}call = Num_{Correct} \; / \, Num_{Actual}$$

In addition, precision is the ratio of the number of text correctly detected by the algorithm to the total number $Num_{Detected}$ of detected text; i.e.,

$$\mathrm{Pr}ecision = Num_{Correct} / Num_{Detected}$$

Fragmentation rate (FR) are those regions in the image which are actually not characters of a text, but have been detected by the algorithm as text.

Error rate (ER) Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

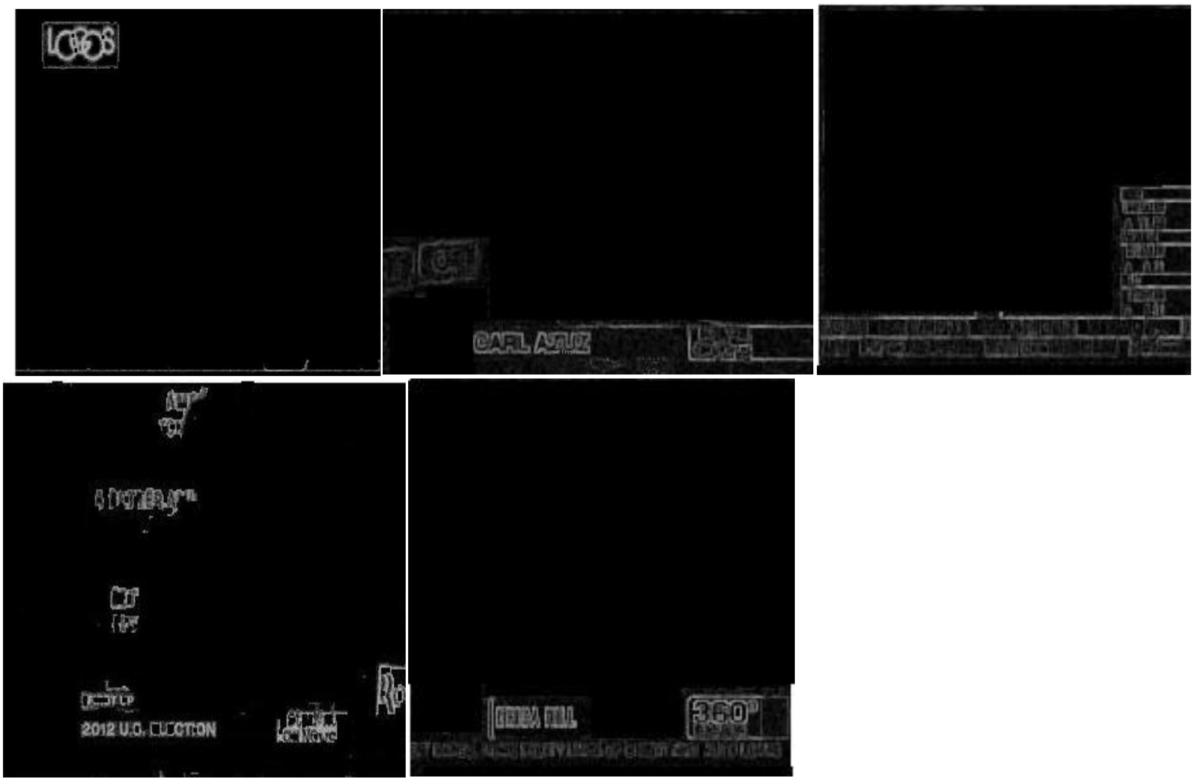

Fig 6: Outputs of Text Content Extraction

Table 3: Performance analysis of the proposed method

Videos	Text/Non-text Blocks in Frames	Recall Rate (%)	Precision Rate (%)	Fragmentation Rate (%)	Error Rate (%)
Video_1	45/100	81.8	86.4	.08	.4
Video_2	60/180	83.6	84.7	.07	.03
Video_3	50/125	95.6	93.5	.02	.01
Video_4	55/140	82.4	80.5	.06	.04
Video_5	140/220	91.5	92.0	.02	.7

Table 3 shows the performance analysis of the proposed method and average error rate is 0.11. Thus, the proposed scheme has high tolerances to noise. The average accuracy of the proposed system is 87%. Fig 7(a) and Fig 7(b) shows the samples of text extraction in video images. Clearly, the proposed method has good abilities to detect all kinds of text regions even under different backgrounds.

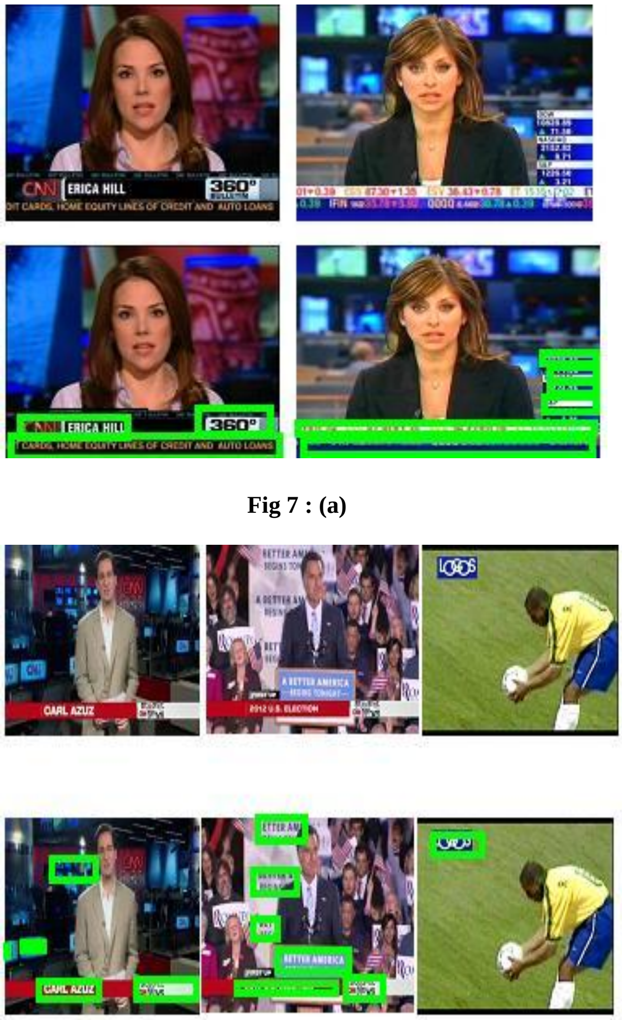

Fig 7 : (a)

Fig 7: (b)

Fig 7: (a) & (b) Text Extraction Samples

5. Conclusion and Future Work

In this paper, an effective combined approach using wavelet transformation, morphological operations, feature extraction and neural network classifier has been proposed for text extraction in noisy video images. Experimental results (Table 3) and comparisons (Table 1) showed that the proposed technique outperforms in different noises using different filters in terms of metrics and gives a good detection rate, low fragmentation and low error rate. In future work, text extraction can be extended to apply any one of the applications like content-based information indexing, video categorization, cataloging of commercials, logging of key events and retrieval systems.

6. References

[1] Ohya.J, Shio.A, and Akamatsu.S," Recognizing Characters in Scene Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 16:214–224, 1994.

[2] Zhong.Y, Karu.K, and Jain.A.K, "Locating Text in Complex Color Images," Pattern Recognition,28(10):1523–1536,October 1995.

[3] Lopresti. D and Zhou.J, " Document Analysis and the World Wide Web," In International Workshop on Document Analysis Systems, Malvern, PA, USA, pages 651–669, 1996.

[4] Smith.M.A and Kanade.T, "Video skimming for quick browsing based on audio and image characterization", CMU-CS-95-186, Technical report, Carnegie Mellon University, 1995.

[5] Jung.K, "Neural network-based text location in color images", Pattern Recognition Letters ,22:1503-1515, 2001.

[6] JiangWu, Shao-Lin Qu, Qing Zhuo,Wen-YuanWang, "Automatic text detection in complex color images", Proc. of Intl. Conf. on Machine Learning and Cybernetics, 2002.Text Localization and Extraction from Complex Gray Images 785

[7] Yuan.Q and Tan.C.L, "Text Extraction from Gray Scale Document Images Using Edge Information", Proc. of Sixth Intl. Conf. on Document Analysis and Recognition, 2001

[8] Ariki.Y and Teranishi.T, " Indexing and Classification of TV News Articles based on Telop Recognition," International Conference on Document Analysis and Recognition, pages 422–427, 1997.

[9] Yeo.B.L," Visual Content Highlighting Via Automatic Extraction of Embedded Captions on MPEG Compressed Video", SPIE/IS&T Symposium on Electronic Imaging Science and Technology: Digital Video Compression: Algorithms and Technologies, volume 2668, 1996.

[10] Hauptmann.A and Smith.M, "Text, Speech, and Vision for Video Segmentation: The Informedia Project," AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision, 1995.

[11] Kazubek.M. "Wavelet domain image de-noising by thresholding and Wiener filtering". Signal Processing Letters IEEE, Volume: 10, Issue: 11, Nov. 2003 265 Vol.3.

[12] Donoho, D.L "Wavelet Shrinkage and W.V.D.: A 10-minute Tour" (David L. Donoho's website)