Text Detection and Localization in Low Quality Video Images through Image Resolution Enhancement Technique

P. Rajendra Kumar
Department of Computer Science and Engineering, CBIT
Hyderabad, India.

Y. Rama Devi, PhD.
Department of Computer Science and Engineering, CBIT
Hyderabad, India.

T. Prathima
Department of Information Technology, CBIT
Hyderabad, India.

ABSTRACT
Video text information plays an important role in semantic-based video analysis, indexing, and retrieval. Text embedded in the image or video contains very useful information like the name of the person, title, location and sometimes brief description of the image. Many algorithms have been proposed to detect and localize the text information present in video images. In this paper, we proposed a methodology to enhance the quality of the image and then detect and localize text regions from low quality video images. Experimental results show the proposed method achieves improved precision rate and recall rate.

General Terms
Text detection and Localization, Connected component

Keywords
Text Region Extraction, Connected component, Edge based Localization, Text detection, Texture features, Discrete wavelet transform (DWT), Stationery wavelet transform (SWT), Low Resolution Image (LRI), Resolution Enhancement Image (REI).

1. INTRODUCTION
Text detection and localization in images and videos is a research area which attempts to develop a computer system with the ability to automatically read from images and videos the text content visually embedded in complex backgrounds. The research field of text recognition receives a growing attention due to the proliferation of digital cameras and the great variety of potential applications, as well. Such applications include robotic vision, image retrieval, intelligent navigation systems and application to visual impaired persons.

Earlier version of videos usually suffers from low quality, perspective distortion, blur, shadow and uneven lighting. Text embedded in images and video sequences, especially captions provide brief and important content information, such as the name of the speaker or player, the title, location and date of an event etc. The text contained in the images and videos can be of any color and gray scale value, low resolution, variable size, unknown font and may appear in different orientation and it is not an easy problem to reliably detect and localize text embedded in images and videos. Hence, the automatic detection and extraction is a challenging problem. Reported works have identified a number of approaches are categorized as connected component based, edge based and texture based methods. Connected component based methods use bottom up approach to group smaller components into larger components until all regions are identified in the image. A geometrical analysis is later needed to identify text components and group them to localize text regions. Edge based methods focus on the high contrast between the background and text and the edges of the text boundary are identified and merged. Later several heuristics are required to filter out non-text regions. But, the presence of noise, complex background, and significant degradation in the low resolution image can affect the extraction of connected components and identification of boundary lines, thus making both the approaches inefficient. Texture analysis techniques are good choice for solving such a problem as they give global measure of properties of a region.

In this paper, resolution enhancement technique has been employed on low quality video images by using interpolation which generates sharper high resolution image. Texture based text detection and segmentation is applied on the enhanced image. The proposed method is robust enough to detect text regions from low quality video images, and achieves improved precision rate and recall rate.

2. RELATED WORK
Image enhancement is the process of improving the quality of the digital image without knowledge about the source of degradation. The source may be a low resolution camera or aliasing due to improper selection of sampling rate or poor illumination. These sources affected the resolution of the image. Resolution has been frequently referred as an important aspect of an image. Images are being processed in order to obtain more enhanced resolution.

One of the commonly used techniques for image resolution enhancement is Interpolation. Interpolation has been widely used in many image processing applications such as facial reconstruction, multiple description coding, and super resolution. There are three well known interpolation techniques, namely nearest neighbor interpolation, bilinear interpolation, and bi-cubic interpolation. Image resolution enhancement in the wavelet domain is a relatively new research topic and recently many new algorithms have been proposed [4]–[7]. Discrete wavelet transform (DWT) [8] is one of the recent wavelet transforms used in image processing. DWT decomposes an image into different subband images, namely low-low (LL), low-high (LH), high-low (HL), and high-high (HH). Another recent wavelet transform which has been used in several image processing applications is stationary wavelet transform (SWT) [9]. In short, SWT is similar to DWT but it does not use down-sampling, hence the subbands will have the same size as the input image.

A number of methods for text localization have been published in recent years and are categorized into connected

component based, edge based and texture based methods. The performance of the methods belonging to first two categories is found to be inefficient and computationally expensive for low resolution images due to the presence of noise, complex background and significant degradation. Hence, the techniques based on texture analysis have become a good choice for image analysis, and texture analysis is further investigated in the proposed work.

A few state of the art approaches that use texture features for text localization have been summarized. A methodology that uses frequency features such as the number of edge pixels in horizontal and vertical directions and Fourier spectrum to detect text regions in real scene images is discussed in [13]. The texture-based text localization method using Wavelet transform is proposed in [14]. Another method which uses DCT coefficients to capture textural properties for caption localization is presented in [8].

Out of many works cited in the literature it is generally agreed that the robustness of texture based methods depends on texture features extracted from the window/block or the region of interest that are used by the discriminant functions for classification decisions. The probability of misclassification is directly related to the number and quality of details available in texture features. Hence, the extracted texture features must give sufficient information to distinguish text from the background and also suppression/removal of background information is an essential pre-processing step needed before extracting distinguishable texture features to reduce the probability of misclassification. But most of the works cited in the literature directly operate on the image without suppressing the background. Hence, there is a scope to explore such possibilities. The proposed method performs pre-processing on the image for suppressing the uniform background in the DCT domain and further uses texture features for text localization. The detailed description of the proposed methodology is given in the next section.

3. PROPOSED METHOD

The proposed method enhances the low quality video images by using resolution enhancement technique through discrete and stationery wavelet decomposition, texture based method is implemented to detect the text regions. The block diagram of proposed method is shown in Figure 1.

3.1 Resolution Enhancement [2]

The enhancement technique uses DWT to decompose a low resolution image into different subbands. Then the three high frequency subband images have been interpolated using bicubic interpolation. The high frequency subbands obtained by SWT of the input image are being incremented into the interpolated high frequency subbands in order to correct the estimated coefficients. In parallel, the input image is also interpolated separately. Finally, corrected interpolated high frequency subbands and interpolated input image are combined by using inverse DWT (IDWT) to achieve a high resolution output image.

3.2 Algorithm for Texture Based Text Region Extraction [1]

The proposed method comprises of 5 phases; Background removal/suppression in the DCT domain, texture features computation on every 50x50 block and obtaining a feature matrix, Classification of blocks, merging of text blocks to detect text regions, and refinement of text regions.

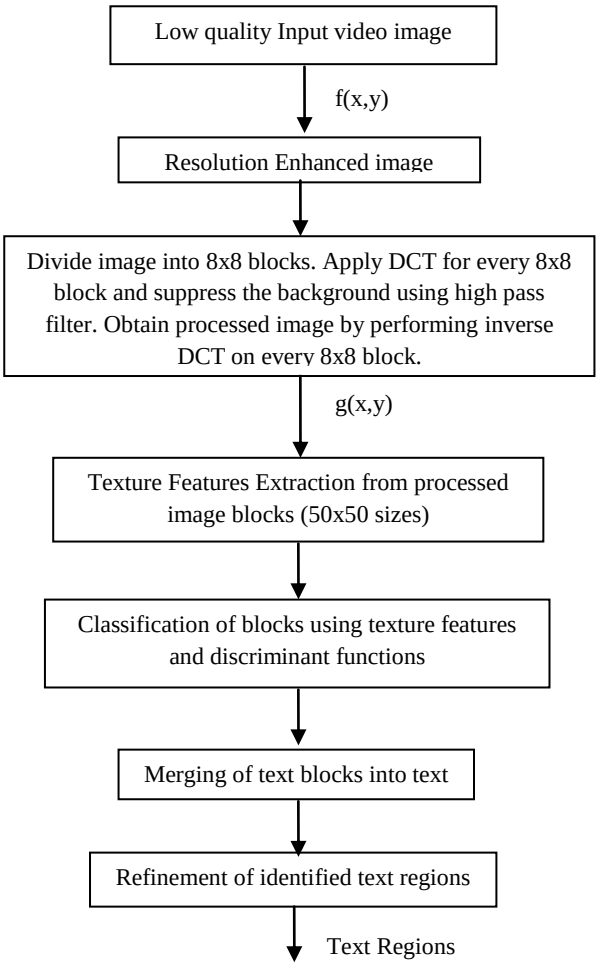

Figure 1: Block Diagram of Proposed Method

The basic steps of the texture based text extraction algorithm are given below.

1. Divide the input image into 8x8 blocks and apply DCT for each block.
2. Suppress the background of image using high pass filter.
3. Perform inverse DCT on each block to obtain processed image.
4. Divide the processed image into 50x50 blocks.
5. Calculate the features homogeneity and contrast at 0^0, 45^0, 90^0, 135^0 orientations for each block using gray level co-occurrence matrix.
6. Filter the non-text blocks using text features and discriminant functions.
7. Merge the obtained text blocks into text regions
8. Refine the size of the detected text regions to cover the missed text present in undetected blocks and unprocessed regions.

The proposed methodology is robust and performs well different sizes of font and image resolution. The block size is an important design parameter whose dimension must be properly chosen to make the method more robust and insensitive to variation in size, font and its alignment.

4. RESULTS AND ANALYSIS

The proposed methodology is implemented with MATLAB for text region detection and localization has been evaluated on a data set containing 30 different images on Intel Celeron (2.4 GHz) computer. It was observed that the processing time lies in the range of 5 to7 seconds due to varying background.

The sample list for the test images and results obtained are shown in the Figure 2. The precision rate and recall rate have been computed to evaluate the efficiency and robustness of text detection in Low Resolution Image (LRI) and Resolution Enhancement Image (REI). The corresponding detailed analysis is presented in Table 1.

Figure 2: Text Detection and Localization results of low resolution image (LRI) and resolution enhancement image (REI)

Table 1: The Performance of the system for test images in Figure 2

Type of the image	Input Image	Correctly Detected Blocks	False Positives	False Negatives	Precision Rate (%)	Recall Rate (%)
Low Resolution Image (LRI)	Figure 1(a)	1	3	3	25	25
	Figure 2(a)	10	8	5	55.5	66.6
	Figure 3(a)	5	4	6	55.5	62.5
	Figure 4(a)	4	10	4	28.5	50
	Figure 5(a)	3	7	5	30	37.5
	Figure 6(a)	1	5	5	50	50
Resolution Enhancement image(REI)	Figure 1(a)	4	2	1	66.6	80
	Figure 2(a)	20	4	1	83.3	95.2
	Figure 3(a)	16	4	5	80	76.19
	Figure 4(a)	4	6	2	40	66.6
	Figure 5(a)	10	5	2	66.6	83.3
	Figure 6(a)	5	2	1	71.4	83.3

The Precision rate is defined as the ratio of correctly detected blocks to the sum of correctly detected blocks plus false positives. False positives are those regions in the image which are actually not characters of a text, but have been detected as text regions.

$$\text{Precision Rate} = \left(\frac{\text{Correctly Detected blocks}}{\text{Correctly Detected blocks} + \text{False Positives}}\right) * 100$$

The Recall rate is defined as the ratio of correctly detected blocks to the sum of correctly detected blocks plus false negatives. False Negatives are those regions in the image which are actually text characters, but have not been detected as text regions.

$$\text{Recall Rate} = \left(\frac{\text{Correctly Detected blocks}}{\text{Correctly Detected blocks} + \text{False negatives}}\right) * 100$$

Figure 3 deals graph obtained for precision rate between six numbers of low resolution images and Resolution Enhancement images and it clearly states that enhanced images has high precision rate than low resolution images.

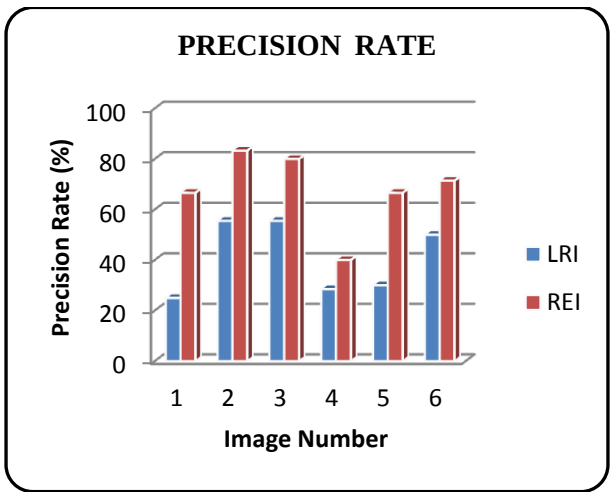

Figure 3: Graph for Precision rate between low resolution images (LRI) and Resolution Enhancement images (REI)

Figure 4 deals graph obtained for recall rate between six numbers of low resolution images and resolution Enhancement images and it clearly states that enhanced images has high precision rate than low resolution images.

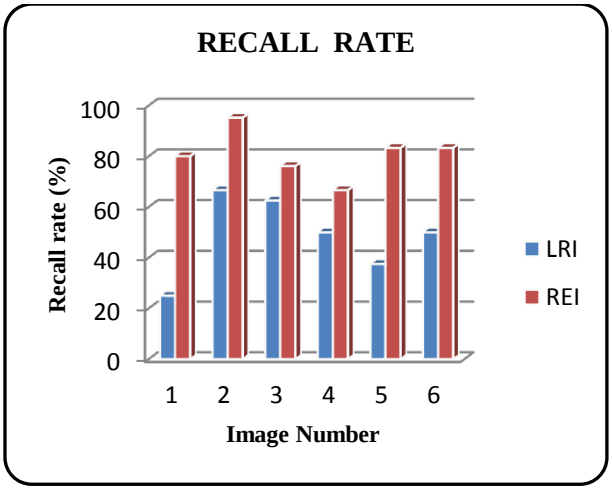

Figure 4: Graph for Recall rate between low resolution images (LRI) and Resolution Enhancement images (REI)

5. CONCLUSIONS AND FUTURE WORK

The effectiveness of the method for text localization from low quality images is presented. The overall precision rate and recall rate of enhanced resolution image is more compared to the low quality images and observed that using image resolution technique extraction of text from the low quality video images is highly efficient. Precision rate and recall rate depends on false positives and false negatives. For improving the performance of the system, false positives and false negatives should be as low as possible.

The performance of the method needs further exploration by combining the edge, connected component based and texture based algorithms, each of the algorithms is by itself quite robust in extracting text regions from low resolution images. A hybridization of two or more techniques may produce more efficient outputs.

6. REFERENCES

[1] Angadi, S.A. and Kodabagi, M.M, Text region extraction from low resolution natural scene images using texture features, Advance Computing Conference (IACC), 2010 IEEE 2ndInternational.

[2] H. Demirel and G. Anbarjafari, "Satellite image resolution enhancement using complex wavelet transform" IEEE Geosci. Remote Sens. Lett., vol. 7, no. 1, pp. 123–126, Jan. 2010.

[3] Eve Bertucci, Maurizio Pilu, Majid Mirmehdi, 2003 "Text Selection by Structured Light Marking for Hand-held Cameras," Seventh International Conference on Document Analysis and Recognition(ICDAR'03), pp.555-559, August 2003.

[4] Tom yeh, Kristen Grauman, K. Tollmar et al., 2005, A picture is worth a thousand keywords: image-based object search on a mobile platform, In Proceedings of conference on Human Factors inComputing Systems**,** pp.2025-2028, 2005.

[5] Abowd Gregory D., Christopher G. Atkeson, Jason Hong, Sue Long, Rob Kooper, Mike Pinkerton, 1997, "CyberGuide: Amobile contextaware tour guide", Wireless Networks, Vol.3, No.5, pp. 421- 433,1997.

[6] Fan X., Xie X., Li Z., Li M., and Ma, 2005, "Photo-to-search: using multimodal queries to search web from mobile phones", Proc., of 7th ACM SIGMM international workshop on multimedia informationretrieval, 2005.

[7] Lim Joo Hwee, Jean Pierre Chevallet, Sihem Nouarah Merah, 2005, "SnapToTell: Ubiquitous information access from camera", Mobile human computer interaction with mobile devices and services, Glasgow, Scotland, 2005.

[8] Yu Zhong, Hongjiang Zhang, Anil.K. Jain, 2000, "Automatic caption localization in compressed video", IEEE transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.4, pp.385-389, April 2000.

[9] Yu Zhong, Kalle Karu, and Anil K. Jain, 1995, "Locating Text In Complex Color Images", Pattern Recognition, Vol.28, No.10, pp.1523-1535, 1995.

[10] S. H. Park, K. I. Kim, K. Jung, and H. J. Kim, 1999, "Locating Car License Plates using Neural Networks", IEE Electronics Letters, Vol.35, No.17, pp.1475-1477, 1999.

[11] V. Wu, R. Manmatha, and E. M. Riseman, 1999, "Text Finder: An Automatic System to Detect and Recognize Text in Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No.11, pp.1224-1229, 1999.

[12] V. Wu, R. Manmatha, and E. R. Riseman, 1997, "Finding Text in Images", Proc. of ACM International Conference on Digital Libraries, pp.1-10, 1997.

[13] B. Sin, S. Kim, and B. Cho, 2002, "Locating Characters in Scene Images using Frequency Features", Proc. of International Conference on Pattern Recognition, Vol.3, pp.489-492, 2002.

[14] W. Mao, F. Chung, K. Lanm, and W. Siu, 2002, "Hybrid Chinese/English Text Detection in Images and Video Frames", Proc. of International Conference on Pattern Recognition, Vol.3, pp.1015-1018, 2002.

[15] A. K. Jain, and K. Karu, 1996, "Learning Texture Discrimination Masks", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.18, No.2, pp.195-205, 1996.

[16] K. Jung, 2001, "Neural network-based Text Location in Color Images", Pattern Recognition Letters, Vol.22, No.14, pp.1503-1515,December 2001.

[17] K. Y. Jeong, K. Jung, E. Y. Kim, and H. J. Kim, 1999, "Neural Network-based Text Location for News Video Indexing", Proc. Of IEEE International Conference on Image Processing, Vol.3, pp. 319- 323, 1999.

[18] H. Li, D. Doerman, and O. Kia, 2000, "Automatic Text Detection and Tracking in Digital Video", IEEE Transactions on ImageProcessing, Vol.9, No.1, pp.147-156, January 2000.

[19] H. Li and D. Doermann, 2000, "A Video Text Detection System based on Automated Training", Proc. of IEEE InternationalConference on Pattern Recognition, pp.223-226, 2000.

[20] W. Y. Liu, and D. Dori, 1998, "A Proposed Scheme for Performance Evaluation of Graphics/Text Separation Algorithm", K. Tombre and A. Chhabra (eds.), Lecture Notes in Computer Science, Vol.1389, pp. 359-371, 1998.

[21] Y. Watanabe, Y. Okada, Y. B. Kim, and T. Takeda, 1998,"Translation Camera", Proc. of International Conference on Pattern Recognition, Vol.1, pp. 613-617, 1998.