TEXT REGION EXTRACTION FROM LOW RESOLUTION DISPLAY BOARD IMAGES USING WAVELET FEATURES

M. M. Kodabagi\textsuperscript{1}, S. A. Angadi\textsuperscript{2}, Anuradha. R. Pujari \textsuperscript{3}

\textsuperscript{1} Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot-587102, Karnataka, India
\textsuperscript{2} Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot-587102, Karnataka, India
\textsuperscript{3} Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot-587102, Karnataka, India

ABSTRACT

Automated systems for understanding display boards are finding many applications useful in guiding tourists, assisting visually challenged and also in providing location aware information. Such systems require an automated method to detect and extract text prior to further image analysis. In this paper, a methodology to detect and extract text regions from low resolution display board images is presented. The proposed work is texture based and uses wavelet energy features for text region detection and extraction. The wavelet energy features are obtained at 2 resolution levels on every 50x50 block of the image and potential text blocks are identified using newly defined discriminant functions. Further, the detected text blocks are merged to extract text regions. The proposed method is robust and achieves a detection rate of 97% on a variety of 100 low resolution display board images each of size 240x320.

Keywords: Text Detection, Text Region Extraction, Display Board Images, Wavelet Features
1. INTRODUCTION

As the people move across world for business, field works and/or pleasure, they find it difficult to understand the text written on display boards in foreign environment. In such a scenario, people either look for guides or intelligent devices that can help them in providing translated information to their native language. As most of the individuals carry camera embedded, hand held devices such as mobile phones and PDA’s, there is a possibility to integrate technological solutions into such systems inorder to provide facilities for automatically understanding display boards in foreign environment. These facilities may be provided as an integral solution through web service as necessary computing function, which are not available in hand held systems. Such web based hand held systems must be enabled to capture natural scene images containing display boards and query the web service to retrieve translated localized information of the text written on display boards.

The written matter on display/name boards provides information necessary for the needs and safety of people, and may be written in languages unknown. And the written matter can be street names, restaurant names, building names, company names, traffic directions, warning signs etc. Hence, lot of commercial and academic interest is veered towards development of techniques for web service based hand held systems useful in understanding written text in display boards. There is a spurt of activity in development of web based intelligent hand held tour guide systems, blind assistants to read written text and Location aware computing systems and many more in recent years. A few such works are presented in the following and a more elaborate survey of related works is given in the next section. A point by photograph paradigm where users can specify an object by simply taking picture to retrieve matching images from the web is found in [1]. The comMotion is a location aware hand held system that links personal information to locations. It reminds users about shopping list when he/she nears a shopping mall [2]. At Hewlett Packard (HP), mobile Optical Character Reading (OCR) applications were developed to retrieve information related to the text image captured through a pen-size camera [3]. Mobile phone image matching and retrieval has been used by insurance and trading firms for remote item appraisal and verification with a central database [4].

The image matching and retrieval applications cannot be embedded in hand held devices such as mobile phones due to limited availability of computing resources, hence such services are being developed as web services. The researchers have also worked towards development of web based intelligent hand held tour guide systems. The cyberGuide [5] is an intelligent hand held tour guide system, which provides the information based on user’s location. The cyberGuide continuously monitors the users location using Global Positioning System (GPS) and provides new information at the right time. Museums could provide these tour guides to visitors allowing them to take personalized tours observing any displayed object. As the visitors move across museum floors, the information about the location is pushed to hand held tour guides. The research prototypes used to search information about an object image captured by cameras embedded in mobile phones are described in [6-7].

The state of art hand held systems available across the world are not automated for understanding written text on display boards in foreign environment. Scope exists for exploring such possibilities through automation of hand held systems. One of the very important processing steps for development of such systems is automatic detection and extraction of text regions from low resolution natural scene images prior to further analysis. The written text provides important information and it is not an easy problem to reliably detect and localize text embedded in natural scene images [8]. The size of the characters can
vary from very small to very big. The font of the text can be different. Text present in the image may have multiple colors. The text may appear in different orientation. Text can occur in a complex background. And also the textual and other information captured is affected by significant degradations such as perspective distortion, blur, shadow and uneven lighting. Hence, the automatic detection and segmentation of text is a difficult and challenging problem. Reported works have identified a number of approaches for text localization from natural scene images. The existing 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 nontext regions. But, the presence of noise, complex background, and significant degradation in the low resolution natural scene 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, a new model for text extraction from low resolution display board images using wavelet features is presented. The proposed work is texture based and uses wavelet features for text extraction. The wavelet energy features are obtained at 2 resolution levels on every 50x50 block of the image and potential text blocks are identified using newly defined discriminant functions. Further, the detected text blocks are merged to extract text regions. The proposed method is robust and achieves a detection rate of 97% on a variety of natural scene images.

The rest of the paper is organized as follows: The related work is discussed in section II. The proposed model is presented in section III. Experimental results are given in section IV. Section V concludes the work and lists future directions.

2. RELATED WORKS

The web based hand held systems useful in understanding display boards requires analysis of natural scene images to extract text regions for further processing. 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 natural scene 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 here; the use of horizontal window of size 1x21 (Mask size) to compute the spatial variance for identification of edges in an image, which are further used to locate the boundaries of a text line is proposed in [9]. However, the approach will only detect horizontal components with a large variation compared to the background and a processing time of 6.6 seconds with 256x256 images on SPARC station 20 is reported. The Vehicle license plate localization method that uses similar criteria is presented in [10]. It uses time delay neural networks (TDNNs) as a texture discriminator in the HSI color space to decide whether the window of an image contains a license plate number. The detected windows are
later merged for extracting license plates. A multi-scale texture segmentation schemes are presented in [11-12]. The methods detect potential text regions based on nine second-order Gaussian derivatives and is evaluated for different images including video frames, newspapers, magazines, envelopes etc. The approach is insensitive to the image resolution and tends to miss very small text and gives a localization rate of 90%.

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]. The techniques for text extraction in complex color images, where a neural network is employed to train a set of texture discrimination masks that minimize the classification error for the two texture classes: text regions and non-text regions are reported in [16-17].

Learning-based methods for localizing text in documents and video are proposed in [15] and [18]. The method [18] is evaluated for various video images and the text localization procedure required about 1 second to process a 352x240 image on Sun workstation. And for text detection a precision rate of 91% and a recall rate of 92.8% is reported. This method was subsequently enhanced for skewed text in [19]. A work similar to the proposed method which uses DCT coefficients to capture textural properties for caption localization is presented in [8]. The authors claim that the method is very fast and gives a detection rate of 99.17%. However, the precise localization results are not reported. The method that use homogeneity and contrast textural features for text region detection and extraction is reported in [20-21]. In recent years, several approaches for sign detection, text detection and segmentation from natural images are also reported [22-34].

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. Hence, there is a scope to explore such possibilities. The detailed description of the proposed methodology is given in the next section.

3. **PROPOSED MODEL**

The proposed method comprises of 4 phases; Preprocessing for gray level conversion and dividing image into 50x50 blocks, Extraction of wavelet energy features on every 50x50 block, Classification of blocks into text and non text categories, and merging of text blocks to detect text regions. The block schematic diagram of the proposed model is given in the fig. 1. The detailed description of each phase is presented in the following subsections.
3.1. Preprocessing and dividing image into 50x50 blocks

Initially, the input image is preprocessed for gray level conversion and resized to a constant size of 300x300 pixels. Further, the resized image is divided into 50x50 blocks. For each block Discrete Wavelet Transform (DWT) is applied to obtain wavelet energy features. There are several benefits of using larger sized image blocks for extracting energy features. One such benefit is, the larger size image blocks cover more details and hence extracted features give sufficient information for correct classification of blocks into text and non-text categories. The other benefits include; robustness and insensitiveness to variation in size, font and alignment.

3.2. Feature Extraction

In this phase, the wavelet energy features are obtained from every 50x50 block of the processed image at 2 resolution levels. Totally 6 features are extracted from every block and are stored into a feature vector $X_i$ (Subscript “i” corresponds to $i^{th}$ block). The feature vector $X_i$ also records block coordinates which corresponds to minimum and maximum row and column numbers of the block. Feature vectors of all $N$ blocks are combined to form a feature matrix $D$ as depicted in equation (1). The feature vector is described in equation (2).

$$ D = [X_1, X_2, X_3, \ldots, X_N]^T $$  \hspace{1cm} (1)

$$ X_i = [r_{\text{min}}, r_{\text{max}}, c_{\text{min}}, c_{\text{max}}, f_{j1}, f_{j2}, f_{j3}]; \quad 1 \leq j \leq 2 $$  \hspace{1cm} (2)
\[ f_{j1} = \frac{\sum H_j (x, y)}{(M \times N)} \]  

\[ f_{j2} = \frac{\sum V_j (x, y)}{(M \times N)} \]  

\[ f_{j3} = \frac{\sum D_j (x, y)}{(M \times N)} \]

Where,
- \( r_{min}, r_{max}, c_{min}, c_{max} \) corresponds to coordinates of \( i^{th} \) block in terms of minimum and maximum row and column numbers.
- \( f_{j1}, f_{j2}, f_{j3} \) are wavelet energy features of corresponding detail band/coefficients at \( j^{th} \) resolution level.
- \( H_j, V_j \) and \( D_j \) correspond to detail coefficients/bands of wavelet transform.
- \( M \times N \) represents size of detail band at the corresponding resolution level.

The features of all image blocks are further used to identify potential text blocks as detailed in the classification phase.

### 3.3. Classification

The classification phase uses discriminant function to categories every block into two classes’ \( w1 \) and \( w2 \). Where, \( w1 \) corresponds to text blocks and \( w2 \) corresponds to non-text blocks category. The discriminant functions thresholds wavelet energy features at both decomposition levels for categorization of image blocks. Then, the coordinates \( r_{min}, r_{max}, c_{min}, c_{max} \) which corresponds to minimum and maximum row and column numbers of every classified text block \( C_i \) will be stored into a new vector \( B \) as shown in equation (6), which are later used during merging process to obtain text regions.

\[ B = [C_i, i = 1,NI] \]  

\[ C_i = [r_{min}, r_{max}, c_{min}, c_{max}] \]

Where,
- \( C_i \) corresponds to \( i^{th} \) text block.
- \( r_{min}, r_{max}, c_{min}, c_{max} \) corresponds to the coordinates of \( i^{th} \) block in terms of minimum and maximum row and column numbers.
- \( NI \) is Number of text blocks.

After this phase, the classified text blocks are subjected to merging process as given in the next section.

### 3.4. Merging of text blocks to detect text regions

The merging process combines the potential text blocks \( C_i \) connected in rows and columns, to obtain new text regions \( r_s \) whose coordinates are recorded into vector \( R \) as depicted in equation (8).

\[ R = [r_s, i = 1, W] \]
\[ r_i = [r_{\text{min}}, r_{\text{max}}, c_{\text{min}}, c_{\text{max}}] \] (9)

where,

- \( r_i \) corresponds to the \( i^{th} \) text region.
- \( r_{\text{min}}, r_{\text{max}}, c_{\text{min}}, c_{\text{max}} \) corresponds to the coordinates of the \( i^{th} \) text region.
- \( W \) is the number of text regions.

The merging procedure is described in algorithm 1;

**Algorithm 1**

**Input:** Vector \( B \) which contains coordinates of identified text blocks  
**Output:** Vector \( R \) which records text regions

**Begin**  
1. Choose the first block \( C_s \) from vector \( B \).
2. Initialize coordinates of a new text region \( r_i \) to coordinates of block \( C_s \).
3. Select next block \( C_p \) from the vector \( B \).
4. **if** (the block \( C_p \) is connected to \( r_i \) in row or column) **then**
   begin  
   Merge and update coordinates \( r_{\text{min}}, r_{\text{max}}, c_{\text{min}}, c_{\text{max}} \) of block \( r_i \).
   \[ r_{\text{min}} = \min\{ r_i[r_{\text{min}}], C_p[r_{\text{min}}] \} \]
   \[ r_{\text{max}} = \max\{ r_i[r_{\text{max}}], C_p[r_{\text{max}}] \} \]
   \[ c_{\text{min}} = \min\{ r_i[c_{\text{min}}], C_p[c_{\text{min}}] \} \]
   \[ c_{\text{max}} = \max\{ r_i[c_{\text{max}}], C_p[c_{\text{max}}] \} \]
   end
5. **else**
   Store text region \( r_i \) into vector \( R \).
   Initialize coordinates of a new text region \( r_i \) to coordinates of current block \( C_p \).
   end
5. Repeat steps 2-5 until \( p = N \).

4. **EXPERIMENTAL RESULTS AND ANALYSIS**

The proposed methodology for text region detection and extraction has been evaluated for 100 indoor and outdoor low resolution natural scene images (having 3600, 50x50 size blocks) with complex backgrounds. The experimental tests were conducted for most of the images containing Kannada text and few containing English text and results were highly encouraging. The experimental results of processing several images dealing with various issues and the overall performance of the system are given below;

![Fig.2(a) Input Image](image1)
![Fig.2(b) Extracted Text Region](image2)
The proposed methodology has produced good results for natural scene images containing text of different size, font, and alignment with varying background. The approach also detects nonlinear text regions. Hence, the proposed method is robust and achieves an overall detection rate of 97%, and a false reject rate of 3% is obtained for 100 low resolution display board natural scene images. The method is advantageous as it uses only wavelet texture features for text extraction. The reason for false reject rate is the low contrast energy of blocks containing minute part of the text, which is too weak for acceptance to classify blocks as text blocks. However, the method has detected larger region than the actual size of the text region, when display board images with more complex backgrounds containing trees, buildings and vehicles are tested.
As the texture features used in the method does not capture language dependent information, the method can be extended for text localization from the images of other languages with little modifications. To explore such possibilities the performance of the method has been tested for localizing English text without any modifications, but the thorough experimentation is not carried out for various images containing English and other language text. The use of different block sizes can also be experimented to improve the detection accuracy and reduce false acceptance rate. The overall performance of the system of testing 100 low resolution natural scene display board images dealing with various issues is given in Table 1.

**TABLE 1: Overall System Performance**

<table>
<thead>
<tr>
<th>Total no of blocks tested(# of text blocks)</th>
<th>Correctly detected text blocks</th>
<th>Falsy detected text blocks (FAR)</th>
<th>Missed text blocks (FRR)</th>
<th>Detection rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>3600 (1211)</td>
<td>1175</td>
<td>19 (1.5%)</td>
<td>36 (3%)</td>
<td>97%</td>
</tr>
</tbody>
</table>

5. **CONCLUSION**

The effectiveness of the method that performs texture analysis for text localization from low resolution natural scene images is presented. The wavelet texture features have performed well in detection and segmentation of text region and are the ideal choice for degraded noisy natural scene images, where the connected component analysis techniques are found to be inefficient. The proposed method is robust and has achieved a detection rate of 97% on a variety of 100 low resolution natural scene images each of size 240x320.

The proposed methodology has produced good results for natural scene images containing text of different size, font, and alignment with varying background. The approach also detects nonlinear text regions. However, it detects larger text regions than the actual size when the background in the image is more complex containing trees, vehicles, and other details from sources of outdoor scenes for some images.

As the texture features does not capture language dependent information, the method can be extended for text localization from the images of other languages with little modifications. The performance of the method has been tested for localizing English text, but needs further exploration.

**REFERENCES**


