International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 29, No. 1 (2015) 1555003 (23 pages)

DOI: 10.1142/S0218001415550034

World Scientific
www.worldscientific.com

A Caption Text Detection Method from Images/Videos for Efficient Indexing and Retrieval of Multimedia Data

Samabia Tehsin* and Asif Masood†
MCS, National University of Science & Technology (NUST), Pakistan
*tsamabia@yahoo.com
†amasood@mcs.edu.pk

Sumaira Kausar‡ and Yunous Javed§
College of E & ME
National University of Science & Technology (NUST), Pakistan
‡sum_satti@yahoo.com
§myjaved@ceme.nust.edu.pk

Received 20 September 2013
Accepted 13 August 2014
Published 12 December 2014

Textual information embedded in multimedia can provide a vital tool for indexing and retrieval. Text extraction process has many inherent problems due to the variation in font sizes, color, backgrounds and resolution. Text detection and localization are the most challenging phases of text extraction process whereas text extraction results are highly dependent upon these phases. This paper focuses on the text localization because of its very fundamental importance. Two effective feature vectors are introduced for the classification of the text and nontext objects. First feature vector is represented by the Radon transform of text candidate objects. Second feature vector is derived from the detailed geometrical analysis of text contents. Union of two feature vectors is used for the classification of text and nontext objects using support vector machine (SVM). Text detection and localization results are evaluated on two publicly available datasets namely ICDAR 2013 and IPC-Artificial text. Moreover, results are compared with state-of-the-art techniques and the Comparison demonstrates the superiority of the presented research.

Keywords: Text extraction; image retrieval; caption text; document analysis; ICDAR 2013.

1. Introduction

In recent years, there is a rapid increase in multimedia libraries, which raise the need of efficient retrieval, indexing and browsing multimedia information. Several approaches are introduced in literature to retrieve image and video data. These techniques are based on color, texture, shape and relation between objects, etc. For text-based queries, text embedded in images and videos for retrieval is a very good option.

Visual texts appearing in multimedia data often display information about news headings, title of movie, product brands in commercials, match score and summary, date and time of an event. All these texts are vital for understanding context of multimedia data and hence help in retrieving images and videos.

In literature, text embedded in images and videos is classified into two groups: caption text and scene text. Caption text is laid over the image/video at a later stage e.g. score of match and name of the speaker. It is also known as artificial text or superimposed text. Caption text usually highlights or recapitulates the multimedia's contents. This formulates caption text principally positive for construction of keyword index. Therefore, proposed methodology focuses on the overlay text only. Scene text is actual part of the scene e.g. brands of the products, street signs, name plate and text appearing on t-shirts, etc.

Text extraction and recognition process comprises of five steps namely text detection, text localization, text tracking, segmentation or binarization, and character recognition. Aim of text detection and text localization is to create a bounding box around the text appearing in image or frame of video.

In this paper, segmentation method for images and video frames is presented. This method is used to segregate the image into different regions. This is a two stage process: first step segments image into different regions and second stage merge the potential characters to form a text candidate. This segmentation process can handle a large variation of font types and sizes and can perform very well with complex backgrounds. Two novel feature vectors for text classification are also introduced. The first feature vector uses the Radon transform for the text object classification. Radon transform is introduced for the first time as text classification tool. Second feature vector exploits the basic geometrical structure of characters and words that gives promising results. The proposed methodology also significantly improves the text localization results at par with state-of-the-art methods in literature. The rest of the paper is organized as follows. Section 2 describes gist of related work of the field. Section 3 introduces the proposed method to extract text in images. Section 4 presents the dataset used and results of text location algorithm. Section 5 provides concluding remarks.

2. Related Work

A variety of techniques for text extraction is appeared in recent past.[30–32,37,45,52] Comprehensive surveys can be traced explicitly in Refs. 17, 23 and 41. These techniques can be categorized into two types mainly with reference to the utilized text features i.e. region based and texture based.[24]

Texture-based methods pertain to textural properties of the text, distinguishing it from the background. The techniques mostly use Gabor filters, Wavelet, FFT, spatial variance, etc. These methods further use machine learning techniques such as support vector machine (SVM), Multilayer perceptron (MLP) and adaBoost.[6,12,20,22]

These techniques work in the top down fashion by first extracting the texture features and then finding the text regions.

Region-based approach exploits different region properties to extract text objects. This approach makes use of the fact of sufficient difference existing between the text color and its immediate background. Color features, edge features, and connected component methods are often used in this approach.[10,21,35] These techniques typically work in the bottom up fashion by first segmenting the small regions and then grouping the potential text regions.

Texture-based techniques usually give better results in complex backgrounds than region-based techniques but are computationally very heavy hence not suitable for retrieval systems for hefty databases. Therefore, there is a need to improve the detection results of region-based techniques to be used for retrieval and indexing of large multimedia data.

Anthimopoulos *et al.*[1] proposed two-stage methodology for text detection in video images. In the first stage, text lines are detected based on the Canny edge map of the image. In the second stage, the result is refined using a sliding window and a SVM classifier trained on features obtained by a new Local Binary Pattern-based operator (eLBP) that describes the local edge distribution. The whole algorithm is used in a multi resolution fashion enabling detection of characters for a broad range of font sizes.

Yao *et al.*[50] proposed an edge-based algorithm for text detection. First the canny edge map is generated from the input image then stroke width transform operator is used to group the neighboring pixels to form text objects. Greedy hierarchical agglomerative clustering method, is applied to aggregate the pairs into candidate chains. This method links the characters in arbitrary directions, and text may not necessarily be in the horizontal direction. Random Forest is used as the chain level classifier to get the final results.

Wei *et al.*[46] proposed a pyramidal scheme to detect text in images. First input image is resized into grayscale images of three different sizes. Then, the horizontal gradients, vertical gradients and the maximum gradient difference (MGD) maps of the image pyramid are calculated. K-means clustering is applied on energy uniformity maps of MGD map to segregate text and nontext pixels. Geometrical constraint along with the SVM is used to produce the final results.

Shivakumara *et al.*[36] proposed a method which used edge maps and quad tree to extract text from images. The pixels are grouped together based on their R, G and B values to enhance text information. K-means where $k = 2$ is used to differentiate potential text candidate pixels from nontext pixels. Stroke width-based symmetry is used for further authentication of potential text pixels. These authenticated text objects are then utilized as seed points to reinstate the text information with reference to the Sobel edge map of the original input image. Quad tree is employed to conserve the spatial locations of text pixels. Region growing is applied on Sobel edge map to formulate the text lines.

Fu *et al.*[11] and Liu *et al.*[25] used Gaussian Mixture Modeling (GMM) of three neighboring characters to discriminate between characters and noncharacters. Based on this modeling, the text in an input image is extracted in three steps. First, the image is binarized and the morphological closing operation is used for merging and grouping of regions on the binary image. Then the neighborhood of all the connected components is established by partitioning the image into Voronoi regions of centroids for connected components. Finally, each connected component is labeled as character or not, according to all its neighborhood relationships.

There is a need of in-depth study of text structures and to mathematically model those features to make it workable for machines. Detailed geometrical and statistical study of text objects is also required.

3. Proposed Methodology

The proposed method uses the region-based approach for text detection and localization and comprises of two phases: namely segmentation, and identification.

3.1. *Segmentation*

Segmentation is the procedure of dividing a digital image into multiple fragments called superpixels.[34,44] The objective of segmentation is to reduce the computational complexity of the under-process image and make its representation easier to analyze.

Proposed segmentation method consists of two processes: splitting and merging. Splitting is performed by the traditional region-based segmentation techniques, whereas, merging is based on the fuzzy-based method.

3.1.1. *Splitting*

Goal of this process is to repeatedly partition the input image into a small number of regions having consistent color and composition. This process is divided into two sub-processes. One defines the segmentation of image into k regions for specified value of k. Other sub-process describes the mechanism to select the value of k i.e. number of regions. Combination of two subprocesses gives the solution to the image segmentation problem. Color based k-means clustering is chosen for the segmentation of images into k regions.

Determining the optimal number of clusters is very vital task for the clustering process. Performance of the k-means clustering is highly dependent on the choice of K. The right selection of K is usually indefinite. Increasing K may lead to over segmentation of the image, and decreasing the value of K may end up with under segmentation issues. Intuitively, there is a need to determine the optimal choice of K that can give the desirable segmentation results. There are numerous classes of techniques for building this decision.[9,27,40] Three factors are used to determine the number of clusters for the input image. Experiments prove that K with '2' or '3' gives the best results. Number of clusters, are determined using weighed average of three

gray level co-occurrence matrices (GLCM) features applied on L component of input image I. These features are energy ($\mathfrak{E}$), entropy ($\mathfrak{P}$) and contrast ($\mathcal{C}$) and can be defined as

$$\mathfrak{E} = \max_{\theta}\left(\sum_{i}\sum_{j} P_d^2(i,j)\right), \tag{1}$$

$$\mathfrak{P} = \max_{\theta}\left(-\sum_{i}\sum_{j} P_d(i,j)\log P_d(i,j)\right), \tag{2}$$

$$\mathcal{C} = \max_{\theta}\left(\sum_{i}\sum_{j} (i-j)^2 P_d(i,j)\right). \tag{3}$$

A GLCM element $P_{\theta,d}(i,j)$ is the joint probability of the pixel pairs with gray levels i and j in a given direction θ, having d distance between them. Here $d = 1$ and $\theta = \{0, 90, 180, 270\}$.

K-means clustering is the promising and thoroughly researched image segmentation methodology. There exists sharp contrast between the text and its background. Because of this contrasting nature of text, pixels of text object and its background generally are assigned to different clusters.

A color space is *perceptually uniform* if a little disruption to a constituent value creates similar perception for the range of that value. The RGB color space does not demonstrate this perceptual uniformity. This is due to the reason that colors having identical distances in the RGB color space may possibly not be perceived by humans as being evenly disparate.[49] L*a*b* color space is used in the current research for presenting the color information of an image.

Due to the low resolution of web images, contrast enhancement is applied before the clustering process. The objective of contrast stretching is to get sharper contrasts for segmentation.

$$h = 255\frac{[\omega_c(\delta_g) - \omega_c(\delta_{g\min})]}{[\omega_c(\delta_{g\max}) - \omega_c(\delta_{g\min})]}, \tag{4}$$

where δ_g is the sigmoid function for $C \times C$ sliding window that is defined as

$$\omega_c(\delta_g) = \left[1 + \exp\left(\frac{m_c - g}{\sigma_c}\right)\right]. \tag{5}$$

$\delta_{g\min}$ and $\delta_{g\max}$ are the minimum and maximum intensity values for the gray scale input image. m_c and σ_c are the mean and variance of grayscale intensity values of the given window. C is chosen as 20 after the experimentations. Contrast stretching process is applied on the three channels of the colored image independently.

Suppose $I : \mathbf{\Omega} \to \mathbb{R}$ is the input image to be segmented, where $\mathbf{\Omega} \subset \mathbb{R}^3$ is the domain of I. For given image I, the goal of image segmentation is to partition its domain into K distinct regions. Precisely, segmentation defines the set of disjoint

regions $\{\mathbf{\Omega}_i\}_{i=1}^{K}$, such that $\mathbf{\Omega} = \bigcup_{i=1}^{K} \mathbf{\Omega}_i$. Here, $\mathbf{\Omega}_i$ presents the ith region of the image I.

For given set of observations $S = \{\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_n\}$, k-means partition the n observations into k sets $\{\mathbf{\Omega}_i\}_{i=1}^{K}$, for specified value of k. The centers and partition matrices are updated until the solution converges.[18] Different methods to choose the initial cluster centroid positions, sometimes known as *seeds*, are presented in the literature.[13,33] In the projected system, K points are chosen uniformly at random from the domain $\mathbf{\Omega}$.[3] This seeding method capitulate substantial enhancement in the ultimate error of k-means. With this initial selection method, the k-means algorithm converges very swiftly and thus reduces the computational cost of the segmentation process.

The significance of stroke width information has been highlighted in numerous recent researches.[10,38,39] It is experimentally proven that the stroke width of characters has a low variation; components with large standard deviations are excluded from the image. If the ratio of standard deviation to mean of the stroke width exceeds Γ_{sw} or the area of the component is less than Γ_{ar}, that component is eliminated from the component list. Stroke width information is extorted using distance transform[7] and the area is the pixel count of the component.

3.1.2. *Perceptual merging*

For given $\{\mathbf{\Omega}_i\}_{i=1}^{K}$, the set of all the regions of input image I, succeeding section explains the fuzzy merging process.

The problem of merging process can be defined using the graph theory. Let G denote the undirected graph and $V(G) = \mathbf{\Omega}$ represents the vertices of the graph G. Edges of the graph G are $E(G) = \{(\mathbf{\Omega}_i, \mathbf{\Omega}_j) \in \mathbf{\Omega} \times \mathbf{\Omega} \,|\, i \neq j\}$. These edges show the probability of joining of two vertices. Initially $\forall e \in E(G)$ are set to null. This probability $\mathfrak{p}_{i.j}$ is calculated by fuzzy logic and based upon the four factors. Difference of color and height, horizontal and vertical distance are the four input factors that are fed in the fuzzy inference engine. Detail of the factors and fuzzy merging process can be found in Ref. 42.

3.2. *Feature vector*

Feature vector plays an important role in the detection and localization of the text objects. Feature vector is an n-dimensional vector of numerical features that represent some object for classification and recognition. Most of the algorithms in region-based text detection require a numerical representation of objects, since such representations facilitate processing and statistical analysis. This paper provides two feature vectors for the representation and detection of text objects. First is the Radon-Based feature vector and second is based on the geometric examination of text.

Aggregated feature vector, comprising of union of both of these vectors, is used for the classification of text and nontext objects.

3.2.1. *Radon transform-based feature vector*

Methodologies presented in literature used Radon transform to refine the boundaries of bounding box, for skew correction and to segregate the words and sentences in the detected text object.[1,5,47,51] In the proposed technique, this is used as the text detection and identification tool for the very first time. Moreover, Radon-based feature vector for the classification of text and nontext is presented.

The radon transform computes projections of an input image along specific directions. For the efficient processing only two directions are considered for text detection. First Radon transform is applied on Region R, and feature vector is defined.

A projection of a two-dimensional function R forms a set of line integrals. The radon transform calculates the stripe integrals from several sources along parallel shafts, in a certain direction. The rays are having 1 pixel distanced from each other. For image representation, the radon transform takes numerous projections from multiple angles by rotating the source with respect to the center of the image.[4] Figure 1 shows radon projection at a specified angle.

The line integral of function R at angle '0' is projected onto the y-axis; the line integral at angle '$\pi/2$' is the projection of R onto the x-axis. Projections can be calculated at any angle θ. Radon transform of function R can be defined as the line integral of R parallel to the y'-axis

$$T_\theta(x') = \int_{-\infty}^{\infty} R(x'\cos\theta - y'\sin\theta, x'\sin\theta + y'\cos\theta)dy', \tag{6}$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

To detect the text in the given image, Radon transform is applied at $\theta = 0$ and $\theta = \pi/2$. It is observed by the experiment that the radon transform of text objects is significantly different from the nontext object. This difference can be exploited to

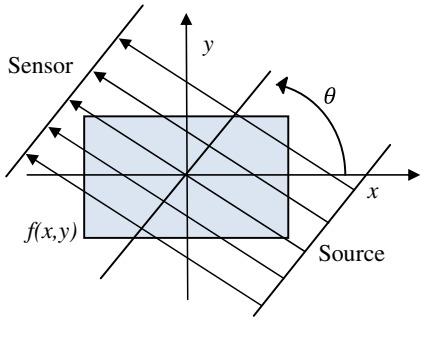
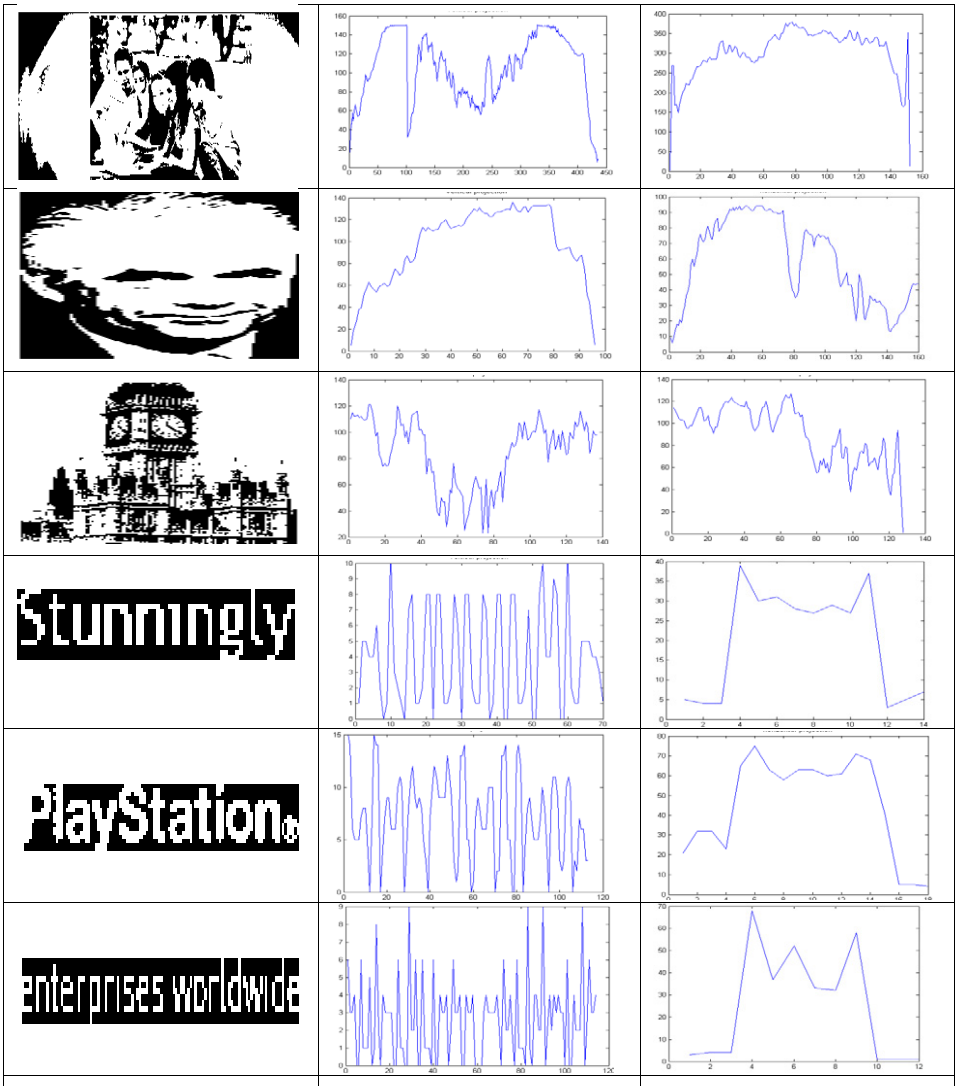


Fig. 1. Parallel-ray projection at angle θ.



(a) Segmentation results (b) Radon Transform at $\theta = 0$ (c) Radon Transform at $\theta = \pi/2$

Fig. 2. Radon transform of segmented objects.

segregate text and nontext objects. Figure 2 shows some results of the radon transform of segmented components of the image.

Radon transform $T_{\theta=0}(x')$, gives the projection of image onto the x-axis and Radon transform $T_{\theta=\pi/2}(x')$, gives the projection onto the y-axis.

All the objects need to be rotated before applying the transform, so that all objects can be put into a common orientation. This is done by aligning the objects horizontally with their major axis. For text detection, two features are introduced in

the proposed method, namely Average Instantaneous rates of change (AIRC) and zero-crossing rate (ZCR). These features are introduced first and their application for text detection is explained later in the subsequent section.

3.2.1.1. Average instantaneous rate of change

The derivative is a measure of how a function changes when its input changes. AIRC (τ_n^θ) is calculated for all the points in Radon transform of the image at θ. Newton's difference quotient is used to calculate the rate of change. AIRC is calculated:

$$\tau_n^\theta = \sum_{w=1}^{\hat{W}} \left| \frac{T_\theta(w'+h) - T_\theta(w')}{h} \right|. \tag{7}$$

h is assumed as 1, $\hat{W}$ is the width of the region R_n and θ can be 0 or $\pi/2$. Due to the different nature of text and nontext regions, τ_n^θ gives the indication for the classification of objects. Attributed to higher stroke density and even distribution of text on the background, $T_{\theta=0}(x')$ Radon signal of text regions have more deviation than nontext regions, whereas, due to similar height of all the characters of the word, $T_{\theta=\pi/2}(x')$ has lesser rate of change.

It is observed that there exist periodic gaps between characters and holes in between the structure of many characters. Therefore, $T_{\theta=0}(x')$ is having higher values for text objects and lower values for non-text objects. Text objects are having almost same height of all the characters in a word and have nearly even distribution in the region. Consequently, its $T_{\theta=\pi/2}(x')$ signal is having very less deviation for text objects.

3.2.1.2. Zero-crossing rate

The ZRC is the rate of sign change of a signal. That change can be from positive to negative or from negative to positive. This acoustic feature is intensively used in literature for speaker identification and music indexing. It is an important feature to categorize percussive sounds. A slight modification of this feature can be used to classify the radon signals of text and nontext objects.

The ZCR is calculated as:

$$\partial_{\mathfrak{y}}^\theta = \frac{1}{\mathfrak{y}} \sum_{w=2}^{\hat{W}} |\operatorname{sign}(T_\theta(w')) - \operatorname{sign}(T_\theta(w'-1)). \tag{8}$$

Here, $\hat{W}$ is the length of the $T_\theta(w')$ signal and $\mathfrak{y}$ is the total numbers of characters in the given text object. $\mathfrak{y}$ is dependent upon the height to width ration $\mathfrak{R}$ of a standard character.

$$\mathfrak{y} = \frac{1}{\mathfrak{R}} X \frac{\mathfrak{H}}{\hat{W}}. \tag{9}$$

The constant $\mathfrak{R}$ is taken as 5/3 here; while $\mathfrak{H}$ and $\hat{W}$ are the height and width of the given object. $\partial_{\mathfrak{y}}^0$ highlights the periodic gap of the text regions and $\partial_{\mathfrak{y}}^{\pi/2}$ portrays

the constant instantaneous height of the text objects. Due to the periodic gaps between characters $\partial_{\mathfrak{y}}^{0}$ has the higher values for the text objects. $\partial_{\mathfrak{y}}^{\pi/2}$ has lower values for the text object, because of nearly constant heights of neighboring characters and even distribution of text on the background.

We define the mean of all the points of radon transform as the neutral point or the zero-crossing point. Absolute zero is used as the zero-crossing point for edge detection, in image processing. For the proposed method, absolute zero can only work for noiseless images. It will fail in the presence of noise or for joint writing styles. Therefore, we are using the mean value as crossing point (see Fig. 3). It will work even for noisy data where characters are merged due to noise and for the joint writing style.

Character objects usually have some space left on the upper and lower end of word. Some characters like 'y' and 'g' used to hang down, and usually first letter of the word is capital, therefore have the greater height than other characters of the word. Because of these reasons, we will consider the central 60% $T_{\theta=\pi/2}(x')$ signal, for the calculation of ZCR and AIRC.

Figure 4 shows the box plots for the above mentioned four features. Box plots show the strength of the individual feature for text object classification problem. Due to some overlapping areas in box plots (see Fig. 4), these features are not used in isolation but are used as multi-dimensional feature spaces. Feature vector is defined as $f_n^1 = \{\tau_n^\theta, \partial_n^\theta\}$, where $\theta \in \{0, \pi/2\}$.

It is very much evident from the box plots that proposed features vector presents very effective candidates for text object classification.

3.2.2. *Geometric feature vector*

This feature vector is generated by the investigation of the geometric structure of text content. Dissection of textual data shows that geometry of text objects is quite different from the nontext ones. Two very effectual geometric features are introduced in this section. Both are based on the geometric characteristics of text.

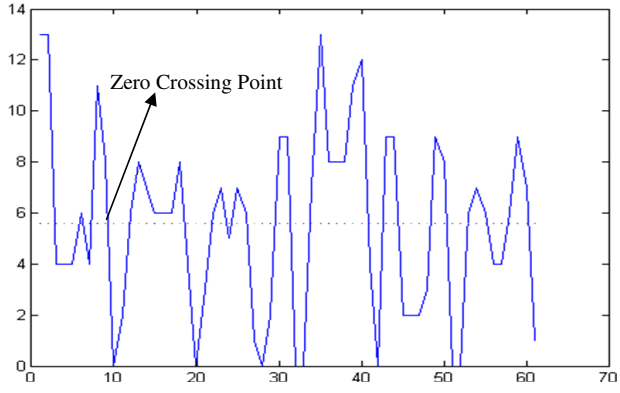


Fig. 3. Zero-crossing in a waveform representing Radon transform of text object.

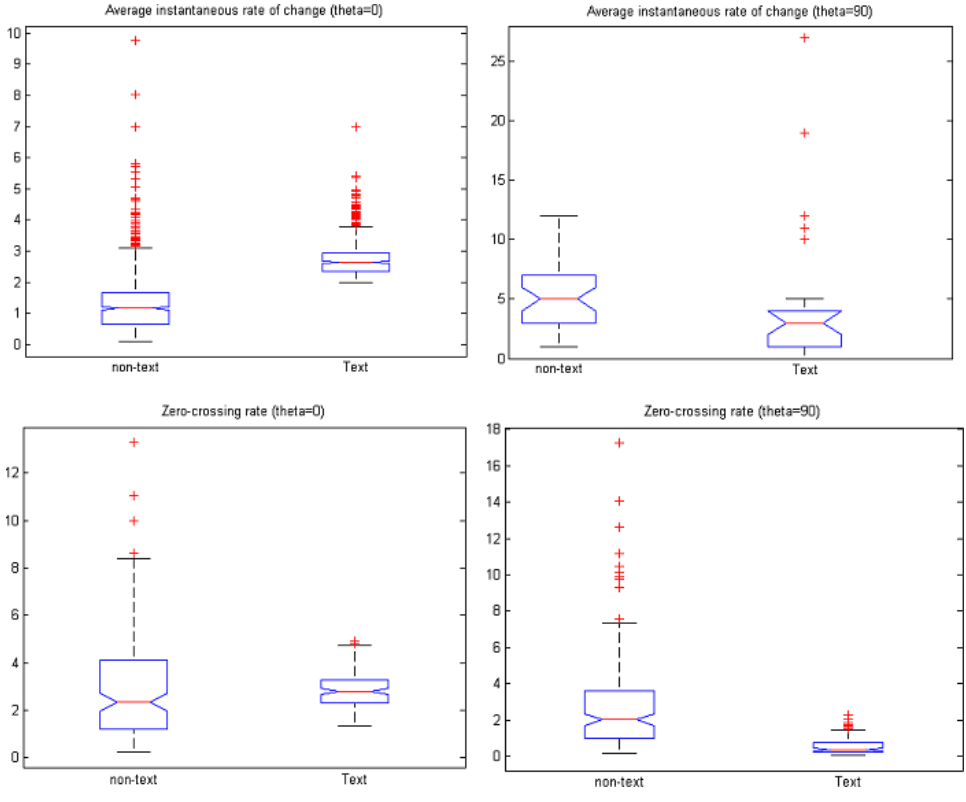


Fig. 4. Box plots for bio-inspired features.

It is observed in the text elements that, these are having transition in their patterns. Transition here is used as the alteration of the foreground pixel to background pixel. In Fig. 5, '*star*' is representing the perpendicular transition $\mathbb{W}$, i.e. transition occurs in particular column of the image. $\mathbb{W}_j$ is the perpendicular transition count of column j. In Fig. 5, 'C' has $\mathbb{W}_3 = 1$ and $\mathbb{W}_{10} = 2$; where 'S' has $\mathbb{W}_3 = 2$ and $\mathbb{W}_{10} = 3$.

Experimentation shows that all the text units have this count between one and four inclusive. Figure 6 shows maximum and minimum perpendicular transition count for all the capital/small English alphabets and numeric digits. It can be observed that most of the digits have this count between one and three except 'g'. Letter 'g' may have four transitions in some fonts but may not have four in other fonts (g). Dark gray color in Fig. 6 represents the possible range of the count for textual data. Therefore, every column of the text object should have the perpendicular transition count between 'one' and 'four'. Mathematically,

$$\mathbb{W}_j = \sum_{i=2}^{\mathrm{H}} \nabla_{i,j}, \tag{10}$$

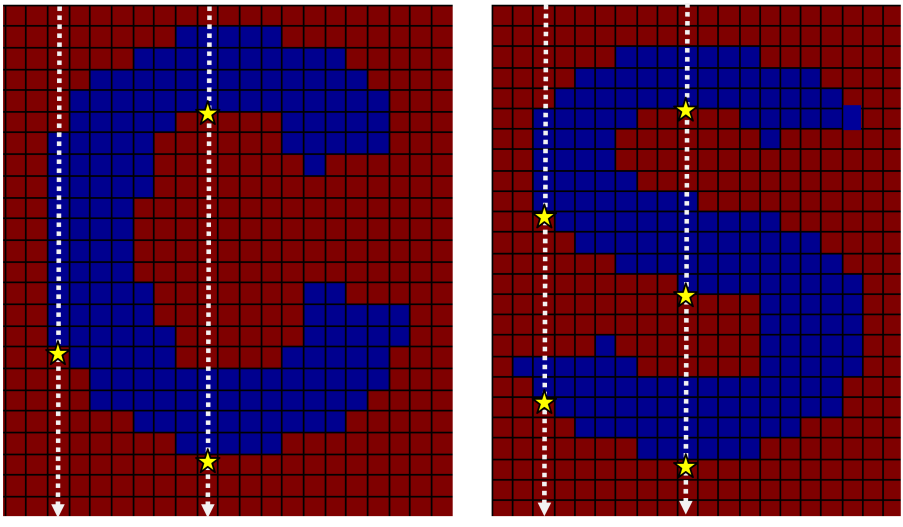


Fig. 5. Perpendicular transition.

$$\nabla_{i,j} = \begin{cases} 1 & \text{if } \boldsymbol{\Omega}(i-1,j)\epsilon \text{ foreground and } \boldsymbol{\Omega}(i,j)\epsilon \text{ background,} \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

Here '$\mathbb{H}$' is the row height of the text candidate $\boldsymbol{\Omega}$.

On the other hand, nontext instances have very low probability of having such distribution. Here is the mathematical description of vertical spread for nth object

$$\underline{\mathfrak{w}}_n = \frac{\text{length}(\text{find}(0 < \ddot{\mathbb{W}} \leq 4))}{(\hat{W} - \underline{\mathfrak{Z}})}. \tag{12}$$

Here, $\ddot{\mathbb{W}} = \{\mathbb{W}_1, \mathbb{W}_2, \cdots, \mathbb{W}_{\hat{W}}\}$, $\hat{W}$ is the column width of the object and $\underline{\mathfrak{Z}}$ are the empty columns i.e. the columns with no foreground pixel. Numerator shows the length of the vector having perpendicular transition counts between the specified range.

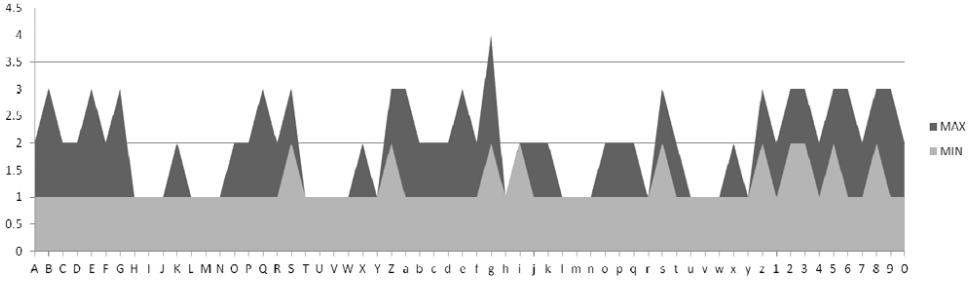


Fig. 6. Range of perpendicular transition count for textual data.

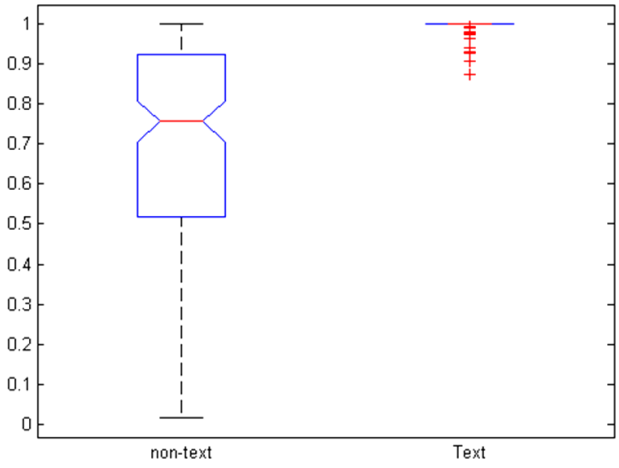


Fig. 7. Box plot for vertical spread.

For ideal text object, the feature $\underline{\mathfrak{w}}$ is exactly 'one'. But in practice this may not achieve the 'ideal' value due to noise, blur and low resolution images. Experimentation has proved this feature as a good option for classification. Figure 7 present the box plot for the vertical spread and it is evident from the plot that majority of the text object has its value ≈ 1.

Horizontal transitions are also studied for classification of text objects. Interesting findings are achieved through the analysis of text content. It is observed through the experimentation that at least one horizontal transition per character is visible in each and every row of the text objects. Horizontal spread $\overline{\mathfrak{M}}$ of nth object can be defined as

$$\overline{\mathfrak{M}}_n = \frac{\text{length}(\text{find}(\ddot{\mathbb{M}} \geq \mathfrak{y}))}{(\mathtt{H} - \overline{\mathfrak{Z}})}. \tag{13}$$

Here $\mathfrak{y}$ is the number of characters, $\mathtt{H}$ is the height and $\overline{\mathfrak{Z}}$ are the empty rows of the given text object. $\ddot{\mathbb{M}} = \{\mathbb{M}_1, \mathbb{M}_2, \ldots, \mathbb{M}_{\mathtt{H}}\}$ is the vector of the horizontal transition counts. Mathematically,

$$\mathbb{M}_j = \sum_{j=2}^{\hat{W}} \Delta_{i,j}, \tag{14}$$

$$\Delta_{i,j} = \begin{cases} 1 & \text{if } \boldsymbol{\Omega}(i, j-1)\epsilon \text{ foreground and } \boldsymbol{\Omega}(i,j)\epsilon \text{ background,} \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Figure 8 shows the box plot for the horizontal spread. This plot presents the potential strength of this feature as text object identifier.

The proposed geometric feature vector has two elements namely, vertical spread and horizontal spread and is defined as $f_n^2 = \{\underline{\mathfrak{w}}_n, \overline{\mathfrak{M}}_n\}$. Combined strength of

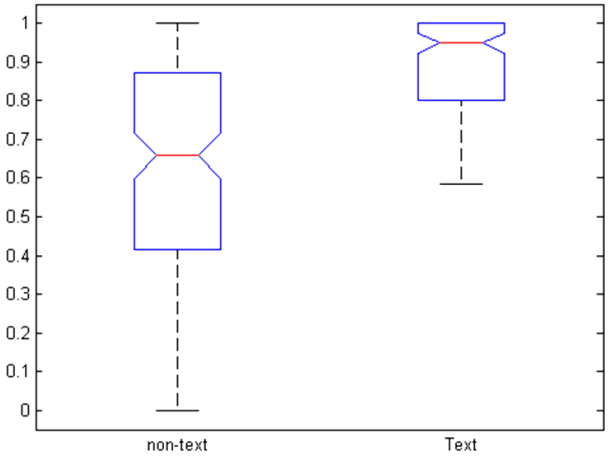


Fig. 8. Box plot for horizontal spread.

horizontal and vertical spread can be visualized in Fig. 9. The scatter plot explicitly defines the collective potency of geometric feature vector. Reduced length of feature vector ensures fast computation in the classification stage.

3.3. *Classification*

Feature vector used to classify the text objects comprised of union of f_n^1 and f_n^2.

$$\bar{f}_n = \bigcup_{1}^{m} f_n^m. \tag{16}$$

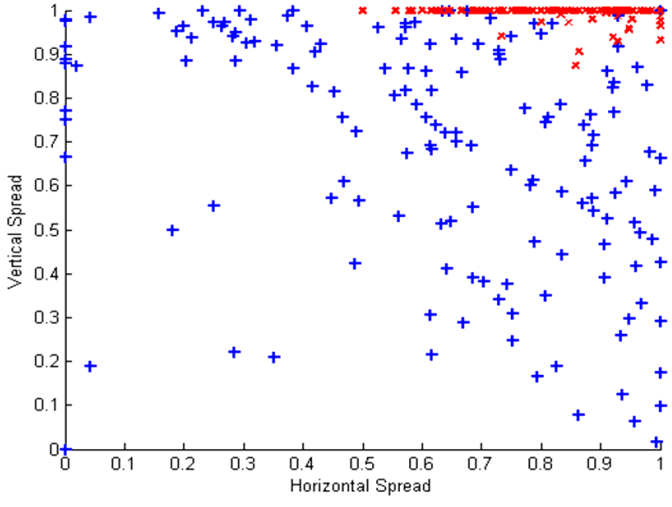


Fig. 9. Scatter plot for geometric feature vector.

Table 1. Comparison of different kernel functions.

	RBS	Polynomial	MLP	Linear
Accuracy	100%	94%	88%	91%

Here $m = 2$ and length of final feature vector is 6.

$$\overline{\boldsymbol{f}}_{\boldsymbol{n}} = \{\tau_n^0, \partial_n^0, \tau_n^{\frac{\pi}{2}}, \partial_n^{\frac{\pi}{2}}, \underline{\mathfrak{w}}_n, \overline{\mathfrak{M}}_n\}. \tag{17}$$

Given the feature vector space of candidate text regions, vectors are trained and classified using SVM.[14] Four kernel functions, naming linear, polynomial, Radial Basis Function (RBS) and MLP, are tested. This experiment is performed on the ICDAR dataset. RBS performed better than other kernel functions because it maps the input space to the feature space without complicated inner product (see Table 1). Four kernel functions used for comparison are defined as:

$$\text{RBS: } K(\bar{f}_i, \bar{f}_j) = \exp\left(-\frac{\|\bar{f}_i - \bar{f}_j\|}{2\sigma^2}\right), \tag{18}$$

$$\text{Polynomial: } K(\bar{f}_i, \bar{f}_j) = (1 + \bar{f}_i^T \bar{f}_j)^d, \tag{19}$$

$$\text{MLP: } k(\bar{f}_i, \bar{f}_j) = \tanh(\gamma_1 \bar{f}_i^T \bar{f}_j + \gamma_2), \tag{20}$$

$$\text{Linear: } k(\bar{f}_i, \bar{f}_j) = \bar{f}_i^T \bar{f}_j. \tag{21}$$

Accuracy rate of different kernel functions is computed by

$$\text{Acc}_i = \frac{\aleph_i}{\max_{j=1,\ldots,n} \aleph}. \tag{22}$$

Here, $\aleph$ is the harmonic mean obtained by particular kernel function and n is the total number of kernel functions used for comparison. Based on these experimental results, RBS is chosen in the presented research. Bootstrap learning is used to further improve the performance of the classifier. False positives and false negatives are used for retraining, which enhances the overall performance of the classification process.

4. Results and Evaluation

Several evaluation metrics are introduced in the literature.[2,15,28,29] Wolf and Jolion's harmonic mean is used as the ranking metric for the evaluation of presented research. Harmonic mean is a combination of two measures: precision and recall. Cumulative precision and recall are calculated for all detections for the complete test set. The reason for opting this metric is its goal oriented evaluation and its capability to deal with one-to-one matches, one-to-many matches (splits), many-to-one matches (merges) and many-to-many matches (splits and merges). This evaluation scheme is computationally complicated, but offers affective assessments. It retains balance between the quantity and quality of the results to be evaluated.

$$\text{recall} = \frac{\sum_{i=1}^{|G|} \text{match}_G(G_i)}{|G|}, \tag{23}$$

$$\text{precision} = \frac{\sum_{j=1}^{|D|} \text{match}_D(D_j)}{|D|}, \tag{24}$$

$$\text{harmonic mean} = 2\frac{\text{recall.precision}}{\text{recall} + \text{precision}}. \tag{25}$$

Here G is the set of ground truth bounding boxes and D is the set of detected ones. The online version of this evaluation scheme is also available in Ref. 8.

Fig. 10. Text localization results on ICDAR 2013 dataset.

Table 2. Comparison of results on ICDAR 2013 dataset.

Method	Recall (%)	Precision (%)	Harmonic Mean (%)
Proposed	**89.40**	**88.83**	**89.11**
USTB_TexStar	82.38	93.83	87.74
Blindsight2012	73.81	90.11	81.15
TH-TextLoc	75.85	86.82	80.96
I2R_NUS_FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27

The proposed detection and localization methodology is evaluated using two datasets namely ICDAR 2013 and IPC-Artificial text Datasets.

Figure 10 shows the localization results on the ICDAR dataset. Text bounding boxes are shown with the black rectangles. It is visible from the results that proposed

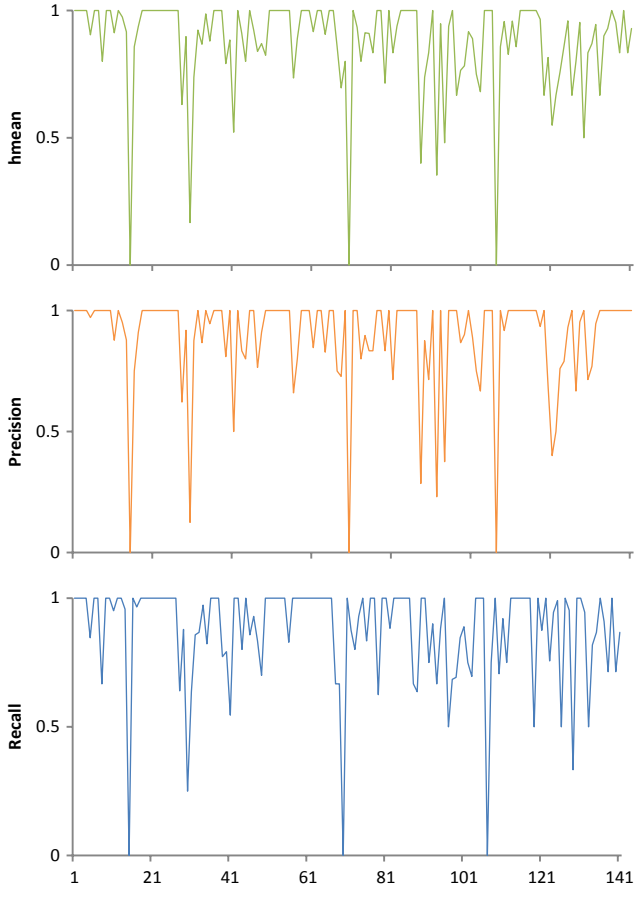


Fig. 11. Image level results of the proposed method on ICDAR 2013 dataset.

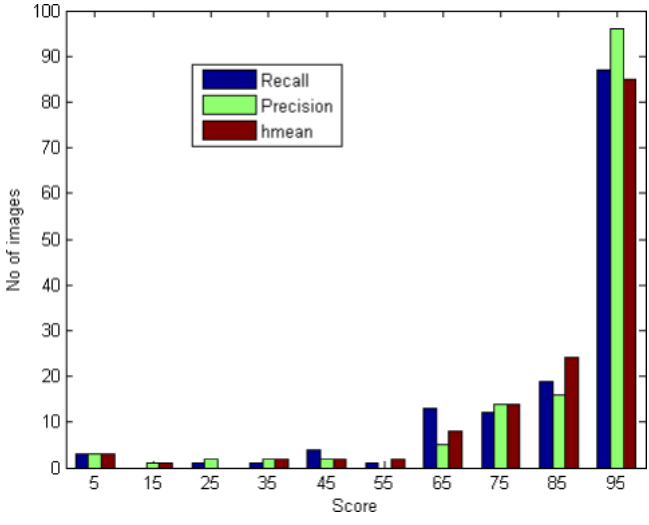


Fig. 12. Cumulative scores of the proposed method on ICDAR 2013 dataset.

methodology is effective with variation in color, size, font style and background. Results of the proposed methodology are compared with USTB_TexStar, Blindsight2012, TH-TextLoc and I2R_NUS_FAR. These are the submitted methods of ICDAR 2013 Robust reading competition,[19] except Blindsight2012 that is received in the continuous mode of competition. Table 2 shows the comparative results of the dataset.

The proposed method has highest recall rate. This is achieved with the help of very powerful feature vector. Some other methods in comparison (USTB_TexStar and Blindsight2012) has higher precision rate but have low recall, which results in lower harmonic mean than the proposed one.

Figure 11 shows the image level results of the proposed scheme for recall, precision and harmonic mean. Results are consistent for most of the images and gives good balance of trade-off between precision and recall.

A different visualization of the proposed method can be seen in Fig. 12. This figure shows the cumulative scores in different bins of the histograms.

IPC-Artificial text dataset[16] is also used to assess the performance of the presented localization scheme. Some resultant images from the dataset are shown in Fig. 13. Proposed technique give promising results on this dataset as well. Table 3 shows the comparison of proposed methodology with the winner of ICDAR 2011[43] and Liu *et al.*'s method.[26]

IPC artificial text dataset can be divided into two categories naming news data and non-news data. Non-news data comprises of cartoon and comic images. Table 4 shows results of different methods for two categories of IPC dataset. Proposed scheme provides slightly better results for non-news data.

Fig. 13. Text localization results on IPC-Artificial text dataset.

Table 3. Comparison of proposed work on IPC dataset.

Method	Recall	Precision	Harmonic Mean
Proposed	**81.35**	**86.85**	**84.01**
Textorter[43]	71.07	83.85	76.93
Liu *et al.*[26]	61.23	69.95	65.3

Table 4. Comparison of proposed work for different categories of IPC dataset.

	News Data			Non-News Data		
	R	P	$H.M$	R	P	$H.M$
Proposed	**79.54**	**85.22**	**82.28**	**82.25**	**87.85**	**84.95**
Textorter	70.2	81.6	**75.47**	72.07	84.85	**77.93**
Liu *et al.*[26]	60.6	68.6	**64.35**	62.23	70.95	**66.30**

Notes: R = Recall, P = Precision and $H.M$ = Harmonic Mean.

5. Conclusions and Future Work

Content-based image and video retrieval attracts immense interest for researchers due to its applicability and utility. Text objects appearing in multimedia content is a vital tool for content-based image/video retrieval and indexing. This text presents a much useful semantic knowledge about the contents of the multimedia document. However, extraction of text poses many difficulties due to variation in size, color, font style and complexities of background. The proposed work can deal with the varying font style, sizes and backgrounds.

In this paper, a novel supervised methodology is presented for text detection and localization task. Two novel feature sets are also introduced in the research. There are some open issues in the presented work that can be addressed in future research.

(a) This work focuses on the caption text in the images, which can be extended to scene text extraction.
(b) The majority of presented work is not language dependent, yet some tuning may be required for other languages like Persian-Arabic script.
(c) The presented research usually works in group of characters.
(d) Radon transform is presented as the text detection tool in the presented research. Projection of two angles is considered for detection process. Other projections can also be explored for auxiliary improvements. Furthermore, Radon transform can be explored as a shape descriptor for object detection other than text.
(e) The proposed methodology mainly focuses on the horizontal alignment of text. Other alignments may also be explored.
(f) Text detection and localization mechanism can be embedded in different applications like goggles for blinds, automated driving, etc.

References

1. M. Anthimopoulos, B. Gatos and I. Pratikakis, A two-stage scheme for text detection in video images, *Image Vis. Comput.* **28**(9) (2010) 1413–1426.
2. M. Anthimopoulos, N. Vlissidis and B. Gatos, A pixel-based evaluation method for text detection in color images, in *Proc. Int. Conf. Pattern Recognition* (IEEE, Washington, DC, 2010), pp. 3264–3267.
3. D. Arthur and S. Vassilvitskii, k-means++: The advantages of careful seeding, in *Proc. Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, January 2007), pp. 1027–1035.
4. G. Beylkin, Discrete radon transform, *IEEE Trans. Acoust., Speech Signal Process.* **35** (2) (1987) 162–172.
5. M. Cai, J. Song and M. R. Lyu, A new approach for video text detection, in *2002 Int. Conf. Image Processing 2002. Proc.*, Vol. 1 (IEEE, Washington, DC, 2002), pp. 1–117.
6. D. Chen, J. M. Odobez and H. Bourlard, Text detection and recognition in images and video frames, *Pattern Recogn.* **37**(3) (2004) 595–608.

7. H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk and B. Girod, Robust text detection in natural images with edge-enhanced maximally stable extremal regions, *2011 18th IEEE Int. Conf. Image Processing (ICIP)* (IEEE, Washington, DC, September 2011), pp. 2609–2612.
8. DetEval-Evaluation Software for Object Detection Algorithms, Available at http://liris.cnrs.fr/christian.wolf/software/deteval/index.html (accessed on August 2013).
9. S. Dudoit and J. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biol.* **3**(7) (2002) research0036.
10. B. Epshtein, E. Ofek and Y. Wexler, Detecting text in natural scenes with stroke width transform, in *2010 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (IEEE, Washington, DC, June 2010), pp. 2963–2970.
11. H. Fu, X. Liu, Y. Jia and H. Deng, Gaussian mixture modeling of neighbor characters for multilingual text extraction in images, in *Image Processing, 2006 IEEE Int. Conf.* (IEEE, Washington, DC, October 2006), pp. 3321–3324.
12. J. Gllavata, E. Qeli and B. Freisleben, Detecting text in videos using fuzzy clustering ensembles, in *Eighth IEEE Int. Symp. Multimedia, 2006. ISM'06* (IEEE, Washington, DC, December 2006), pp. 283–290.
13. G. Hamerly and C. Elkan, Alternatives to the k-means algorithm that find better clusterings, in *Proc. Eleventh Int. Conf. Information and Knowledge Management* (ACM, New York, NY, USA, November 2002), pp. 600–607.
14. M. A. Hearst, S. T. Dumais, E. Osman, J. Platt and B. Scholkopf, Support vector machines, *IEEE Intell. Syst. Appl.* **13**(4) (1998) 18–28.
15. X. S. Hua, W. Liu and H. J. Zhang, An automatic performance evaluation protocol for video text detection algorithms, *IEEE Trans. Circuits and Syst. Video Technol.* **14**(4) (2004) 498–507.
16. IPC-Artificial Text Dataset, Available at https://sites.google.com/site/artificialtextdataset/ (accessed on September 2013).
17. K. Jung, K. I. Kim and A. K. Jain, Text information extraction in images and video: A survey, *Pattern Recogn.* **37**(5) (2004) 977–997.
18. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7) (2002) 881–892.
19. D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez, S. Robles, J. Mas, D. Fernandez, J. Almazan and L. P. de las Heras, ICDAR 2013 robust reading competition, in *Proc. 12th Int. Conf. Document Analysis and Recognition* (IEEE CPS, Washington, DC, 2013), pp. 1115–1124.
20. K. I. Kim, K. Jung and J. H. Kim, Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm, *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(12) (2003) 1631–1639.
21. M. León Cristóbal, V. Vilaplana Besler, A. Gasull Llampallas and F. Marqués Acosta, Region-based caption text extraction, *11th Int. Workshop on Image Analysis for Multimedia Interactive Services (Wiamis)* (IEEE, Washington, DC, 2010), pp. 1–4.
22. C. Li, X. G. Ding and Y. S. Wu, An algorithm for text location in images based on histogram features and ada-boost, *J. Image Graphics* **3**(3) (2006) 325–331.
23. J. Liang, D. Doermann and H. Li, Camera-based analysis of text and documents: A survey, *Int. J. Doc. Anal. Recogn.* **7**(2–3) (2005) 84–104.
24. R. Lienhart, Video OCR: A survey and practitioner's guide, in *Video Mining* (Springer, US, 2003), pp. 155–183.
25. X. Liu, H. Fu and Y. Jia, Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images, *Pattern Recogn.* **41**(2) (2008) 484–493.

26. C. Liu, C. Wang and R. Dai, Text detection in images based on unsupervised classification of edge-based features, in *2005 Proc. Eighth Int. Conf. Document Analysis and Recognition* (IEEE, Washington, DC, August 2005), pp. 610–614.
27. R. Lletı, M. C. Ortiz, L. A. Sarabia and M. S. Sánchez, Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes, *Anal. Chim. Acta* **515**(1) (2004) 87–100.
28. S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young, ICDAR 2003 robust reading competitions, in *Proc. Int. Conf. Document Analysis and Recognition* (2003), pp. 682–687.
29. Y. Ma, C. Wang, B. Xiao and R. Dai, Usage-oriented performance evaluation for text localization algorithms, in *Proc. Int. Conf. Document Analysis and Recognition* (IEEE, Washington, DC, 2007), pp. 1033–1037.
30. R. Minetto, N. Thome, M. Cord, N. J. Leite and J. Stolfi, T-HOG: An effective gradient-based descriptor for single line text regions, *Pattern Recogn.* **46**(3) (2012) 1078–1090.
31. L. Neumann and J. Matas, A method for text localization and recognition in real-world images, in *Computer Vision–ACCV 2010* (Springer, Berlin, Heidelberg, 2011), pp. 770–783.
32. L. Neumann and J. Matas, On combining multiple segmentations in scene text recognition, in *12th Int. Conf. Document Analysis and Recognition (ICDAR)* (Washington, DC, 2013), pp. 523–527.
33. R. Ostrovsky, Y. Rabani, L. J. Schulman and C. Swamy, The effectiveness of Lloyd-type methods for the k-means problem, in *Proc. 47th Annual IEEE Symp. Foundations of Computer Science (FOCS'06)* (IEEE, Washington, DC, 2006), pp. 165–174.
34. N. Senthilkumaran and R. Rajesh, Edge detection techniques for image segmentation–A survey of soft computing approaches, *Int. J. Recent Trends Eng.* **1**(2) (2009) 250–254.
35. C. Shi, C. Wang, B. Xiao, Y. Zhang and S. Gao, Scene text detection using graph model built upon maximally stable extremal regions, *Pattern Recogn. Lett.* **34**(2) (2012) 107–116.
36. P. Shivakumara, H. T. Basavaraju, D. S. Guru and C. L. Tan, Detection of curved text in video: Quad tree based method, in *2013 12th Int. Conf. Document Analysis and Recognition (ICDAR)* (IEEE, Washington, DC, August 2013), pp. 594–598.
37. P. Shivakumara, T. Q. Phan and C. L. Tan, A Laplacian approach to multi-oriented text detection in video, *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(2) (2011) 412–419.
38. A. Srivastav and J. Kumar, Text detection in scene images using stroke width and nearest-neighbor constraints, in *TENCON 2008–2008 IEEE Region 10 Conf.* (IEEE, Washington, DC, November 2008), pp. 1–5.
39. K. Subramanian, P. Natarajan, M. Decerbo and D. Castañòn, Character-stroke detection for text-localization and extraction, in *Ninth Int. Conf. Document Analysis and Recognition, 2007. ICDAR 2007.*, Vol. 1 (IEEE, Washington, DC, September 2007), pp. 33–37.
40. C. A. Sugar and G. M. James, Finding the number of clusters in a dataset, *J. Am. Stat. Assoc.* **98**(463) (2003) 750–763.
41. C. P. Sumathi, T. Santhanam and G. Gayathri, A Survey on various approaches of text extraction in images, *Int. J. Comput. Sci. Eng. Survey* **3**(4) (2012) 27–42.
42. S. Tehsin, A. Masood, S. Kausar and F. Arif, Fuzzy based segmentation for variable font sized text extraction from images/videos, in *Mathematical Problems in Engineering* (Hindawi Publisher, 2014).
43. S. Tehsin, A. Masood, S. Kausar and Y. Javed, Text localization and detection method for born-digital images, *IETE J. Res.* **59** (2013) 343–349.
44. O. J. Tobias and R. Seara, Image segmentation by histogram thresholding using fuzzy sets, *IEEE Trans. Image Process.* **11**(12) (2002) 1457–1465.

45. K. Wang and S. Belongie, Word spotting in the wild, *Computer Vision–ECCV 2010* (Springer, Berlin, Heidelberg, 2010), pp. 591–604.
46. Y. C. Wei and C. H. Lin, A robust video text detection approach using SVM, *Exp. Syst. Appl.* **39**(12) (2012) 10832–10840.
47. A. Wernicke and R. Lienhart, On the segmentation of text in videos, in *2000 IEEE Int. Conf. Multimedia and Expo, 2000. ICME 2000*, Vol. 3 (IEEE, Washington, DC, 2000), pp. 1511–1514.
48. C. Wolf and J. M. Jolion, Object count/area graphs for the evaluation of object detection and segmentation algorithms, *Int. J. Doc. Anal. Recogn.* **8**(4) (2006) 280–296.
49. G. Wyszecki and W. S. Stiles, in *Color Science: Concepts and Methods, Quantitative Data and Formulae* (John Wiley & Sons, New York, 1982).
50. C. Yao, X. Bai, W. Liu, Y. Ma and Z. Tu, Detecting texts of arbitrary orientations in natural images. in *2012 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (IEEE, Washington, DC, June 2012), pp. 1083–1090.
51. Q. Ye, Q. Huang, W. Gao and D. Zhao, Fast and robust text detection in images and video frames, *Image Vis. Comput.* **23**(6) (2005) 565–576.
52. M. Zhao, S. Li and J. Kwok, Text detection in images using sparse representation with discriminative dictionaries, *Image Vis. Comput.* **28**(12) (2010) 1590–1599.

Samabia Tehsin is a Ph.D. Scholar at MCS, NUST. She completed her M.S. degree Software Engineering from NUST in 2007. Her areas of research are digital image processing, computer vision and document analysis.

Asif Masood completed his B.E. degree in Software Engineering from Military College of Signals (MCS), NUST in 1999. He completed his M.S. degree and Ph.D. in Computer Science at University of Engineering and Technology Lahore in 2007. Currently, he is working in MCS, National University of Science and Technology.

Sumaira Kausar is a Ph.D. Scholar at CEME NUST. Her research interests are digital image processing, computer vision and machine learning.

Yunous Javed is the Dean of Computer Engineering at CEME, NUST. His research areas are algorithms, adaptive and predictive modeling, digital image processing, operating systems and encoding/decoding systems and parallel processing.