ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 10(17), DOI: 10.17485/ijst/2017/v10i17/112354, May 2017* ISSN (Online) : 0974-5645

An Efficient Approach for the Identification and Localization of Texts in Images

D. Selvanayagi* and S. Pannirselvam

Department of Computer Science, Erode Arts and Science College, Erode - 638009, Tamil Nadu, India;
selvasubhika@gmail.com, pannirselvam08@gmail.com

Abstract

Objectives: To reduce false positive rate of text localization and to improve the performance of image segmentation process for better detection of text in images. **Methods:** Text information in images is an important consideration for image based applications such as automatic geo coding, content based image retrieval and understanding scene. But to detect the text from a complex background with different colours is a complex task. There are different techniques has been proposed to detect text in images. One of the techniques is hybrid approach for text localization and text detection in natural scene images. In this approach, the text region detector is used to detect text region and to segment the candidate components by local binarization. Then the non text components are filtered using Conditional Random Field model (CRF). Finally the text components are grouped together as line or words. This approach feels hard to segment some complex images due to lack of colour information. In order to enhance the performance of text detection in images Mahalanobis Distance (MD) metric, cosine based similarity metric and text recognition is introduced. **Findings:** The Mahalanobis Distance (MD) metric, cosine based similarity metric are computed for image segmentation process where colour information of images is considered instead of gray level image information. Based on the color information the input images are segmented. Then CRF is employed to filter out the non text components. After that text components are grouped together to localize the text from images. Moreover in the proposed work, text recognition is done by Kohonen neural network and compares the results of text localization and text recognition to reduce the false positive of text localization. **Improvements:** The experimental results show that the proposed work provides better results in terms of precision, recall, F1 measure and ROC curve.

Keywords: Kohonen Neural Network, Text Detection, Text Localization, Text Recognition

1. Introduction

Nowadays, the content-based image analysis[1] methods are the mostly required due to the utilization of digital image capturing devices. The contents in images are more necessary and the text information is more important because it can be easily identified and understood by different applications like content-based web image search, sign detection and translation, mobile text recognition, license plate reading and etc. Generally, the text information extraction system has four processes such as text detection[2], text localization[3,4], text extraction and enhancement, and recognition[5].

Text extraction from an image or video is mainly concerned for extracting the most relevant text information. Among these different processes, many techniques have been developed for detecting and localizing the image and video text. Since, the output of the text information is mostly depending on these two processes. The text detection is defined as the identification of the presence of text in the given frame. The text localization is referred as the identification of the location of text in the image and generation of bounding boxes around the text. Moreover, the text components are segmented and extracted during the text extraction process. Then, the extracted text components are enhanced since it consists of low-resolution and noise. Finally, the extracted are converted into the plain text by using Optical Character Recognition (OCR) technique.

Normally, the text detection and localization

* *Author for correspondence*

techniques were classified as region-based and Connected Component (CC)-based techniques. The region-based techniques were used for detecting and localizing the text regions based on the texture analysis whereas CC-based techniques were used for directly segmenting the candidate text components based on the edge detection and color clustering. The non-text components were pruned along with the heuristic rules or classifiers. But, these techniques have different issues such as the speed of region-based approaches were slow and CC-based approaches were does not utilized without any prior knowledge about the text location and scale. Therefore, a hybrid approach for detection and localization of texts was developed for avoiding these issues.

In hybrid approach[6], which compute the text existing coefficient and scale information in image pyramid to design a text region detector, which help segment candidate text components by local binarization. The non text components are removed by using CRF model. Finally the text components are grouped together into a line or word. This approach has high false positive rate and poor performance. These problems are resolved in this proposed work by using colour information instead of gray level information during image segmentation and text recognition method.

In[7] proposed a framework for text string detection in natural scene images. The proposed framework follows two step processes. The image segmentation process is carried out in the first step of framework that determines the text character candidates based on color uniformity of local gradient features and character components. In the second step of framework text strings are detected through character candidate grouping. It is done by joint structural features of text characters in each text string such as character distances, size differences, between neighboring character alignment and characters. There are two string detection algorithms are proposed based on the assumption is that a text string has at least three characters. The two proposed text string detection algorithms are text line grouping method and adjacent grouping method. However, the calculation cost for text line grouping is more expensive.

In[8] presented a robust hybrid method for detection of text in natural scene images. This method detects the text from images by using Partial Differential Equations (PDEs) that integrate both fundamental differential variants with a non linear regressor. Initially learning based PDEs is used to create a text confidence map and then from this map text region candidates are detected by applying connected components clustering and local binarization method. Then character candidates are detected from each text region based on the color similarity of character candidates and by applying simple rules character candidates are grouped into text candidates. At the end of text detection process two level classification scheme is adopted that removes non text candidates. The major disadvantage of this method is it does not achieve a better F-measure.

In[9] presented text detection and text recognition approach based on the edges of images. The segmentation points in the images are detected that distinguish the text edges from background of images. Then analyzed color similarity and stroke width to merge the neighborhood edges. Then extract edge based features to know whether the edges contain text or not. In the text recognition system, the candidate boundaries are determined by using candidate boundary computation method in the text detection method. In order to recognize characters in images random forest and Histogram Of Gradients (HOG) features are used. The major drawback of this method is it has little worse precision values.

In[10] proposed a text detection method from natural scenes. The proposed text detection method follows steps to detect text in an image. The two steps are connected components extraction and text filtering. A multi scale adaptive color clustering is proposed for connected component extraction. It has capability of extracting text from images from different color complexities and it is more robust to contract variations. The Text Covariance Descriptor (TCD) is integrated with Histogram Of Gradients (HOG) for non text filtering that constructs feature vectors and it is used to differentiate the text from background of image. This method is applicable only for English language is the disadvantage of this method.

In[11] proposed a robust two step method to detect text in natural scene images. The robust two step method is based on multi-layer segmentation and higher order Conditional Random Forest (CRF). The text from background images are separated by segmenting images into nine layers through multi layer segmentation method. The candidate text is obtained from the nine layers as connected components. The candidate text CCs are verified by using higher order CRF based analysis. At the end of the process, CC pairs, separate CCs and CC

strings are combined together that forms higher order CRF model which distinguish text from non text in an input image. This method has poor performances.

In[12] a robust text detection approach in natural scene images is presented. This approach is based on color-enhanced contrasting external region (CER) and neural networks. From the gray scale input image six component trees is constructed. Then CREs is extracted from component tree as character candidates. Each candidate patch is labeled by using divide and conquer method. This method labeled by rules as Long, Thin, Fill, Square-large and Square-small, and classified as text or non-text by a corresponding neural network. However this method does not convert text objects which cannot be extracted as extremal regions in current image channels, into extremal regions.

In[13] presented a robust system for text detection in images. The robust system is based on the different concepts are Gradient Vector Symmetry (GVS), Mutual Direction Symmetry (MDS) and Mutual Magnitude Symmetry (MMS) properties to determine text pixel character candidates from natural scene images. It process based on the fact that text patterns present in Canny edge maps and Sobel of input images has a similar behavior. SIFT features are used to refine text pixel candidates. However the Sobel operations lose information for low contrast texts and for complex background this method show poor precision rate.

In[14] presented a new method for text localization in natural scene images. In this candidate texts in the images are detected by using combining of edges and color. Then wavelet coefficient histogram features and Histogram Of Gradients (HOG) features are extracted which represent the text patterns. These features are given as input to Support Vector Machine that classified text and non text components in the images based on the input features. However this method has high false alarm rate.

2. Methodologies

In this section, the proposed image segmentation method and text recognition for text detection in images is discussed. In hybrid approach for detecting and localizing text in natural scene images[5] a binarized image is obtained from gray level images. This method experiences hard to segment text from some images. So this method still needs improvements for text detection in such images. Hence in the proposed work color information is included in the image segmentation process. Moreover in the proposed work text recognition is integrated with text localization which reduces the false positives of text extraction. The proposed work is explained briefly in the following section.

2.1 Pre-processing

The text confidence and the corresponding scale are estimated by a designed text region detector that utilize and extract local text region information. Based on the estimation of text confidence and the corresponding scale, text components are segmented and analyzed more effectively. Initially the color input image is converted into gray level images. Then an image pyramid is built up with nearest interpolation in scale step 1.2 that captures text information of different scale. The widely used feature descriptor named as Histogram Of Gradients (HOG) is combined with WaldBoost a boosted cascade classifier for design of text region detector. It detects the text position and estimates probabilities of the text position and scale information more accurately. Each 16×16 window sliding in a layer of image pyramid are partition horizontally and vertically and then 4-orientation HOG descriptors are extracted and it is given as input to WaldBoost classifier for estimation of text confidence. The output of the WaldBoost classifier is converted into into posterior probabilities (confidence values) based on a boosted classifier calibration method. The posterior probability of a label $l_x, l_x \in \{\text{text}, \text{non-text}\}$, conditioned on its detection state $d_x, d_x \in \{\text{accepted}, \text{rejected}\}$ at the stage s can be computed based on the Bayes formula is described as follows:

$$P_s(l_x|d_x) = \frac{P_s(d_x|l_x)P_s(l_x)}{\sum_{l_x} P_s(d_x|l_x)P_s(l_x)} = \frac{P_s(d_x|l_x)P_{s-1}(l_x|accepted)}{\sum_{l_x} P_s(d_x|l_x)P_{s-1}(l_x|accepted)} \quad (1)$$

In equation 1, all stage likelihoods $P_s(d_x|l_x)$ are calculated on validation dataset during training. The text confidence map was obtained from probability calibration for each layer of image pyramid. All pixel confidence and scale values at different layers of pyramid are projected back to the original image for estimation of final text confidence and scale values. Based on the arithmetic and geometric mean of corresponding pyramid pixel values the final text confidence value is computed. The text confidence value $P(l_x = text|d_x)$ of the pixel x is calculated by arithmetic mean of pyramid pixel value is given as follows:

$$P(l_x = text|d_x) = \frac{1}{F_x^c}\sum_{h\in P_x} w_h . w_h \quad (2)$$

The scale value $scale_x$ of pixel x is computed by geometric mean of pyramid pixel values is given as follows:

$$scale_x = \left[\prod_{h\in P_h} scale_h^{w_h}\right]^{\frac{1}{F_x^s}} \quad (3)$$

In equation 2 P_x represents the set of pixels of all pyramid layers that are projected back to the original image pixel i, w_h represents a weighting coefficient described as the pixel h's coefficient value $P(l_h = text|d_h^m)$ at the pyramid layer m, $F_x^c = \sum_{h\in P_x} w_h$ is the arithmetic normalization factor and $F_x^s = \prod_{h\in P_h} w_h$ is the geometric normalization factor. This text scale map is used in local binarization for candidate connected components (CCs) segmentation and confidence map is used in Connected Component Analysis (CCA) for classification of components.

2.2 Image Segmentation

The candidate CCs in the image are segmented by considering color information of image. The Euclidean distance metric is widely used for this type of segmentation. But it is not efficient for complicated non linear relationship from sets of training images. To resolve this problem, an efficient metric called as Mahalanobis distance (MD) metric is used in the proposed work for image segmentation. One of the main advantages of MD metric is that it is very sensitive to inter variable changes in the training data. The MD metric is calculated as follows:

$$\Delta = \sqrt{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)} \quad (4)$$

In equation 4, Σ represents the covariance matrix of X in each group X represents the vector that contains the measurement made on individual or entity, μ_1 and μ_2 are the mean vectors.

In the proposed work one more metric is used for image segmentation named as cosine based similarity. This metric is more reliable and simple method to obtain hue information. The cosine similarity metric is computed as follows:

$$Sim_{\cos(c_x c_y)} = 1 - \left(\frac{c_x c_y}{\|c_x\| . \|c_y\|}\right)\left(1 - \frac{||c_x|| - ||c_y||}{\max(||c_x||, ||c_y||)}\right) \quad (5)$$

In equation 5, c_x and c_y represents the color vectors. The cosine based similarity described in the directional domain computes the changes in color chromaticity whereas magnitude processing like Mahalanobis distance calculates the changes in luminance information. These two information are combined together to process the color image using luminance and chrominance information the performance of image segmentation is improved.

2.3 Connected Component Analysis

After the segmentation of candidate CCs, the CCs are analyzed by CRF to assign the candidate components as either text class or non text class. This method considers both binary contextual component relationships and unary component properties. The total energy function of CRF is defined as follows:

$$E(T, L, C, \Lambda) = \sum_i \left[E_u(t_x, l_x, \lambda_u) + \frac{\lambda_c}{n_x}\sum_{h\in N_i} E_b(t_x, t_h, l_x, l_h, \lambda_b)\right] \quad (6)$$

In equation 6, $E(T, L, C, \Lambda)$ is the real valued energy function with parameters Λ and clique set C, $E_u(., \lambda_u)$ represents two-class unary energy function and $E_b(., \lambda_u)$ represents three class binary energy function, λ_c represents a combining coefficient for balancing unary and binary terms, n_x denotes the number of t_x's neighbors for normalizing different neighborhood sizes. Based on this text and non text components are classified.

2.4 Text Grouping and text Localization

The texts components are grouped together into text regions by using minimum spanning tree algorithm it cluster the neighborhood components into tree and energy minimization model is employed for cutting off between-line edges. Based on Minimum Spanning Tree the text components are clustered into a tree through a distance metric which is defined between two components as a linear combination of two features as

$$D_{metric}(t_x, t_h) = W_D . F_{xh} \quad (7)$$

In equation 7, F_{xh} represents the vector of between-component features and W_D represents the vector of combining weights. Based on MST components belonging to the same text region are spatially close and have similar shapes, two types of binary components features are spatial distance and shape difference are chosen as

distance metric features. Then the MST tree is cut is cut into sub trees that signifies a text unit which is word or line. In order to the edge cutting in the tree as a learning-based energy minimization problem a method is devised. Each edge of the component trees is assigned either cut label or linked label. Through cutting the cut edges each sub tree corresponding to a text line is separated. The main intend is to find the optimal edge labels such that the total energy of the separated sub tree is minimal and it is defined as follows:

$$E(line) = \sum_{x=1}^{num} W_{line} . F_x \quad (8)$$

In equation 8, $E(line)$ denotes the total text line energy and num denotes the number of subtrees and W_{line} denotes the vector of combining weights. The within line similarity and between line dissimilarity is characterized by extracting six different features from text line such as line regression error, cut score, line height, spatial distance, bounding box distance and line number. A greedy strategy method is used to find out the optimal edge labels. Thus the text lines are detected by using an optimal edge labels. The text lines are further partitioned into words which are more similar to line partition. It varies from line partition by word level features are word number, bounding box distance ratio between the cut edge and edges within separated words, component centroids distances of cut edges, the ratio between the component centroids of the cut edge and edges within the separated words, bounding box distances between words separated by cut edges and component bounding box distances of cut edges. The text words corresponding to partitioned subtrees can be extracted and the subtree containing too small components are considered and removed as noises.

2.5 Text Recognition

Text recognition is the process of developing a system which has ability to automatically read the text from images. In our proposed work for text recognition process Kohonen neural network model is used. In addition to the line features and word features, character features are extracted from the text line. These features are fed as input to the Kohonen neural network. The Kohonen neural network is consists of only input and output layers of neurons. The input to a Kohonen neural network is given to the neural network using the input neurons. These input neurons are each given the floating point numbers which make up the input pattern to the network. These inputs are normalized to the range between -1 to 1. Then the normalized input is given to Kohonen neural network will cause a reaction from the output neurons. The output of the Kohonen neural network is either true or false which represents the presence of text and absence of text.

Kohonen learning algorithm

Input: line features, word features, character features

Output: recognized text in an image

Step 1: Select the random values for initial weight vectors $weight_i(0)$. $weight_i(0)$ different for i=1,2,3,…m where m is the number of neurons in the lattice (images).

Step 2: Draw a simple i from the input space with a certain probability; the vector i signifies the activation pattern which is applied to the lattice. The dimensions of the vector i is equal to n.

Step 3: Find the best matching (winning) neuron h (y) at a time step s by using minimum distance Euclidian criterion:

$$h(y) = avg \min y \|y(o) - weight_i\|, i = 1,2,3,\ldots m$$

Step 4: Adjust the synaptic weight vectors of all neurons by using following equation:

$$weight_i(o+1) = weight_i(o) + \lambda(o) m_{i,h(y)}(o)(y(o) - weight_i(o))$$

//$\lambda(o)$ is the learning rate parameter, $m_{i,h(y)}(o)$ is the Neighborhood function centered around the winning neuron $h(y)$, both $\lambda(o)$ and $m_{i,h(y)}(o)$ varied dynamically during learning for best results.

Step 5: Continue with step 2 until no noticeable changes in the feature map are observed.

Finally the result of text localization and text recognition are compared each other to reduce the false positive rate of text localization. It improves the performance of text detection from images.

3. Results and Discussions

In this section, the performance of the proposed technique is evaluated for identifying its effectiveness compared with the existing text detection and localization approach based on CRF. For the experimental purpose, a public dataset named as ICDAR 2005 is used. The images of the ICDAR 2005 were captured from digital camera in indoor and outdoor conditions. This database has three sets are trail set consists of 258 Trial Train images and 251 Trail Test images, sample set consists of 20 images and competition set consists of 501 images. The text characters

in these images include numeral and English characters. The effectiveness of the proposed work is measured in terms of precision, recall, F1 measure.

3.1 Precision

The precision rate is defined based on rectangle area match ratio between two region rectangles r_i and r_j i.e, ratio between the area of their intersecting rectangle (air) and the region containing rectangle (acr) is defined as

$$m(r_i, r_j) = \frac{\mathbf{air_{ij}}}{\mathbf{acr_{ij}}}$$

The precision rate of each detected region rectangle r_i is defined based on the ground truth text region set GT and the localized text region set LT and the precision rate of each detected region rectangle r_i is given as follows:

$$p_r(r_i) = \mathbf{max}(r \in GT)[m(r_i, r)]$$

The total precision rate is defined as follows:

$$Precision = \frac{1}{LT} \sum_{r_i \in LT} p_r(r_i)$$

In Figure 1 X axis represents the existing text detection method based on CRF and Proposed Mahalanobis distance-Cosine based similarity-CRF-Text Recognition (MD-CS-CRF-TR) and Y axis represents the Precision rate in %. The values of the precision are tabulated in Table 1. From the Figure 1 and Table 1 it is proved that the proposed MD-CS-CRF-TR has high precision rate than the existing text detection method.

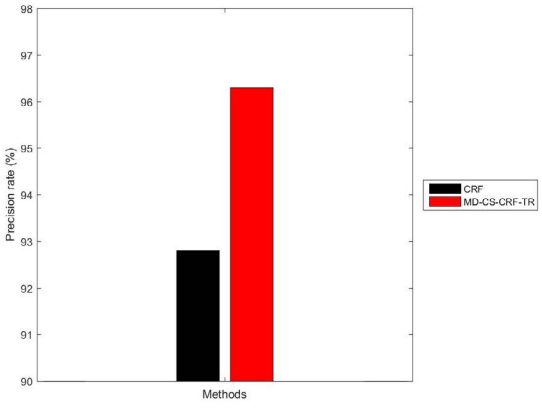

Figure 1. Comparison of Precision.

Table 1. Comparison of Precision

	Methods	
	CRF	MD-CS-CRF-TR
Precision rate (%)	92.8	96.3

3.2 Recall

The recall rate of each ground truth text region is defined as follows:

$$r_r(r_j) = \mathbf{max}(r \in LT)[m(r_j, r)]$$

The total recall rate is defined as follows:

$$Recall = \frac{1}{GT} \sum_{r_i \in LT} r_r(r_i)$$

In Figure 2 X axis represents the existing text detection method based on CRF and Proposed Mahalanobis distance-Cosine based similarity-CRF-Text Recognition (MD-CS-CRF-TR) and Y axis represents the recall rate in %. The values of recall are tabulated in Table 2. From the Figure 2 and Table 2 it is proved that the proposed MD-CS-CRF-TR has high recall rate than the existing text detection method.

3.3 ROC curves

ROC curve is defined as the relationship between Recall rate and the false positive Rate.

In Figure 3 X axis represents the False positive rate ranges for 0 to 0.5 and Y axis represents the recall rate. From the Figure 3 and Table 3 it is proved that the proposed MD-CS-CRF-TR has better ROC curve than the existing text detection method.

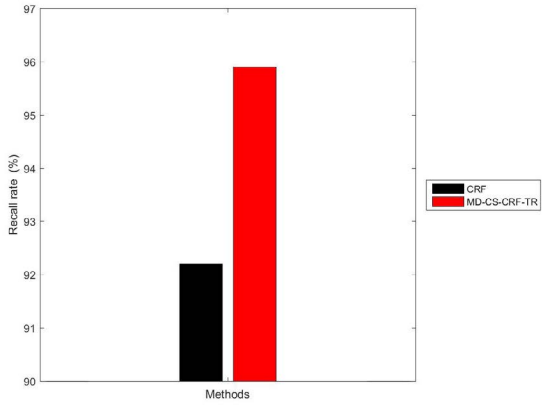

Figure 2. Comparison of Recall.

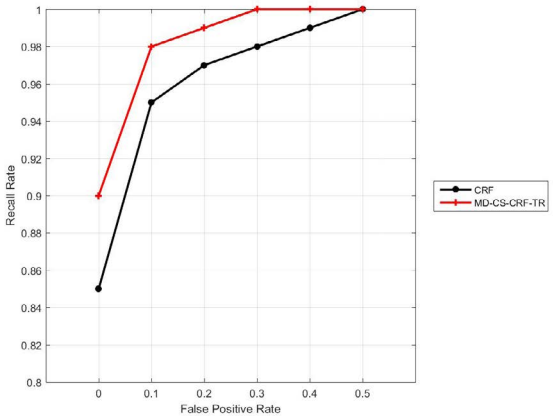

Figure 3. Comparison of ROC curves.

Table 2. Comparison of Recall

	Methods	
	CRF	MD-CS-CRF-TR
Recall rate (%)	92.2	95.9

Table 3. Comparison of ROC curves

False Positive Rate	Recall Rate	
	CRF	MD-CS-CRF-TR
0	0.85	0.90
0.1	0.95	0.98
0.2	0.97	0.99
0.3	0.98	1
0.4	0.99	1
0.5	1	1

3.4 F1 measure

F1 is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. It is defined as follows:

$$F1 = 2.\frac{\text{Precision} \,.\, \text{Recall}}{\text{Precision} + \text{Recall}}$$

In Figure 4 X axis represents the existing text detection method based on CRF and Proposed Mahalanobis distance-Cosine based similarity-CRF-Text Recognition (MD-CS-CRF-TR) and Y axis represents the F1 measure in %. From the Figure 4 and Table 4 it is proved that the proposed MD-CS-CRF-TR has high F1 measure than the existing text detection method.

Table 4. Comparison of F1 measure

	Methods	
	CRF	MD-CS-CRF-TR
F1 (%)	92.5	96.1

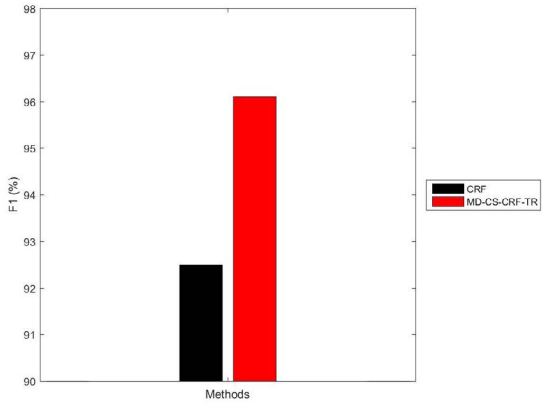

Figure 4. Comparison of F1 measure.

4. Conclusion

In this paper, efficient techniques for text detection from images are proposed by considering color information for image segmentation and text recognition process. Initially the images are preprocessed by text region detector where the text confidence and scale maps are computed for local binarization. To segment CCs in the text region color information of the images are considered by computing Mahalanobis distance (MD) metric ad cosine based semantic similarity. The non text components are filtered out by CRF and then the text components are grouped by MST and energy minimization model which detects the text in images. Then Text recognition process is carried out by KNN model that recognized the text in images and compares the results of both text localization and text recognition which reduce the false positive rate of text localization process. Thus the performance of the text detection process is enhanced. The experimental results are conducted in terms of precision, recall, F1 measure and ROC curve to distinguish the effectiveness of the proposed method.

5. References

1. Singh J, Saxena G. Development Of Content Based Image Retrieval System Using Neural Network & Multi-Resolution Analysis. International Journal of Engineering Sciences and Research Technology. 2016; 5(7):6–9.
2. Naik S, Nayak S. Text detection and character extraction in natural scene images. International Journal of Emerging Technology and Advanced Engineering. 2015; 5(2).
3. Chavre PB, Ghotkar A. A Survey on Text Localization Method in Natural Scene Image. International Journal of Computer Applications. 2015; 112(13).
4. Prabakaran S, Luthra M. Text Localization in the Image of Complex Background using Discrete Wavelet Transform. Indian Journal of Science and Technology. 2016; 9(37). Crossref
5. Wadhawan K, Gajendran E. Automated Recognition of Text in Images: A Survey. International Journal of Computer Applications. 2015; 127(15):15–9. Crossref
6. Pan YF, Hou X, Liu CL. A hybrid approach to detect and localize texts in natural scene images. IEEE Transactions on Image Processing. 2011; 20(3):800–13. Crossref
7. Yi C, Tian Y. Text string detection from natural scenes by structure-based partition and grouping. IEEE Transactions on Image Processing. 2011; 20(9):2594–605. Crossref
8. Zhao Z, Fang C, Lin Z, Wu Y. A robust hybrid method for text detection in natural scenes by learning-based partial differential equations. Neurocomputing. 2015; 168:23–34. Crossref
9. Yu C, Song Y, Meng Q, Zhang Y, Liu Y. Text detection and recognition in natural scene with edge analysis. IET Computer Vision. 2015; 9(4):603–13. Crossref
10. Wu H, Zou B, Zhao Y Q, Chen Z, Zhu C, Guo J. Natural scene text detection by multi-scale adaptive color clustering and non-text filtering. Neurocomputing. 2016; 214:1011–25. Crossref
11. Wang X, Song Y, Zhang Y, Xin J. Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis. Pattern Recognition Letters. 2015; 60:41–7. Crossref
12. Sun L, Huo Q, Jia W, Chen K. A robust approach for text detection from natural scene images. Pattern Recognition. 2015; 48(9):2906–20. Crossref
13. Risnumawan A, Shivakumara P, Chan CS, Tan CL. A robust arbitrary text detection system for natural scene images. Expert Systems with Applications. 2014; 41(18):8027–48. Crossref
14. Darab M, Rahmati M. A hybrid approach to localize farsi text in natural scene images. Procedia Computer Science. 2012; 13:171–84. Crossref