Computer Systems Science & Engineering
DOI:10.32604/csse.2022.023308
Article

Tech Science Press

CNN and Fuzzy Rules Based Text Detection and Recognition from Natural Scenes

T. Mithila[1,*], R. Arunprakash[2] and A. Ramachandran[3]

[1]Department of Computer Science and Engineering, University College Engineering, BIT Campus, Anna University, Trichy, 620021, India
[2]Department of Computer Science and Engineering, University College of Engineering, Ariyalur, 621704, India
[3]Department of Computer Science and Engineering, University College of Engineering, Panruti, 607106, India
*Corresponding Author: T. Mithila. Email: mithilabit@gmail.com
Received: 02 September 2021; Accepted: 09 October 2021

Abstract: In today's real world, an important research part in image processing is scene text detection and recognition. Scene text can be in different languages, fonts, sizes, colours, orientations and structures. Moreover, the aspect ratios and layouts of a scene text may differ significantly. All these variations appear assignificant challenges for the detection and recognition algorithms that are considered for the text in natural scenes. In this paper, a new intelligent text detection and recognition method for detectingthe text from natural scenes and forrecognizing the text by applying the newly proposed Conditional Random Field-based fuzzy rules incorporated Convolutional Neural Network (CR-CNN) has been proposed. Moreover, we have recommended a new text detection method for detecting the exact text from the input natural scene images. For enhancing the presentation of the edge detection process, image pre-processing activities such as edge detection and color modeling have beenapplied in this work. In addition, we have generated new fuzzy rules for making effective decisions on the processes of text detection and recognition. The experiments have been directedusing the standard benchmark datasets such as the ICDAR 2003, the ICDAR 2011, the ICDAR 2005 and the SVT and have achieved better detection accuracy intext detection and recognition. By using these three datasets, five different experiments have been conducted for evaluating the proposed model. And also, we have compared the proposed system with the other classifiers such as the SVM, the MLP and the CNN. In these comparisons, the proposed model has achieved better classification accuracywhen compared with the other existing works.

Keywords: CRF; rules; text detection; text recognition; natural scene images; CR-CNN

1 Introduction

In the recent days, information is playing an important role in finalizing the economy of knowledge in which perception appears to be a mandatory factor. The text data is being extracted from various

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

environments situations that enable the users in devising effective decisions [1]. The text detection works basically focus on the extraction of the text data from various sources including the natural image scenes. All kinds of people must be able to understand the available text data in the live videos and images. Today, Text processing is necessary for extracting and identifying the text from natural scenes. The text detection process is used for extracting the text objects from the natural scenes and for performing the text recognition process for identifying the available characters for forming a meaningful word [2].

The text data is embedded with darkness in various images along with the different kinds of objects of the images. The various difficulties in the extraction of the text data from the images have been considered together with their objects [3]. The ultimate objective of this work is to extract the text data perfectly. Generally, the text extraction process involves various steps like image pre-processing, objects detection, text recognition and classification. Generally, the text detection and recognition processes from the natural scenes are refined and this appropriates the text detection results. The text recognition is very important for finalizing the text detection process [4]. The text detection process can enhance the text quality as the overall outcome is based on the performance of the text recognition process. Various techniques such as the connected-component based, edge-based, texture-based, and stroke-based methods are available for enhancing the text detection and recognition processes [5].

The text can be acknowledged by gathering the colors and detecting the boundaries of the natural scene images. Moreover, the usual methods cannot detect the text properly when the input images are deformed with the stylized fonts. For this purpose, an optimal method is necessary for automating the text processes. This can be done by applying the machine learning algorithms in which the text detection and text recognition processes would be automated. This in a way enhances the performance by performing the required training procedures on the text detection and recognition models [6].

The Artificial Intelligence (AI) is tremendously contributing in the field of data analysis using the Machine Learning (ML) and the Deep Learning (DL) algorithms. Generally, the ML and the DL algorithms work based on the AI techniques perform by making effective decisions on the various datasets. Effective training processes are therefore incorporated in the process of text detection and recognition. It effectively trains the input dataset and is therefore useful for analyzing the texts and their images automatically. ML considers various applications for recognizing the possible patterns and the learning theories. Moreover, the ML algorithms are capable of learning the datasets in detail and are also capable of constructing new models together with the prediction of the unseen data. In addition, the additional instructions are also to be considered for changing the working flow of the text detection and the recognition processes [7].

Fuzzy Logic is used to make crisp decisions on the various critical situations on different kinds of datasets. Generally, the decisions can be made on the inputs with two options such as “True” and “False”. Moreover, the decision will be either “0” or “1” on the datasets according to the user’s preferences. The fuzzy logic is useful for making crisp decision on various datasets with values between 0 and 1. In this scenario, the percentage of positive and the percentage of negative decisions can be made by using the fuzzy logic. The fuzzy membership functions such as the trapezoidal fuzzy membership function, triangular membership function, Mamdani fuzzy membership functions and Gaussian fuzzy membership functions have been used in the fuzzy logic for making effective decisions on various datasets [8].

The major contributions of this paper are as follows: i) propose a new intelligent text detection and recognition method for detecting the text from natural scenes, ii) propose a new text detection method for detecting the exact text from the input natural scene images, iii) propose a new text recognition method called the CRF based Fuzzy rules Incorporated by the CNN for recognizing the text from natural images, iv) generates new fuzzy rules based on the CRF for making effective decisions on the processes of text detection and recognition. The rest of this paper has been organized as follows: Section 2 describes in

detail the various works done by different researchers in the past in the direction of feature extraction and selection, and image classification. Section 3 explains the system architecture that demonstrates the working flow of the proposed model. Section 4 describes the proposed work with the necessary background details. Section 5 demonstrates the experimental results and discussions. Section 6 concludes the work and discusses about the scope of the future works.

2 Literature Survey

Various text detection and recognition works have been accomplished by the researchers in the past for performing effective text detection and recognition processes. For that purpose, some image pre-processing methods have been proposed and incorporated along with the text detection and recognition models for enhancing the performance. Darshan et al. [9] has established a detection approach based on the MSER detector and has further explained the text region detection in an input image. Moreover, this process is a general activity over the unstructured scenes of the images while capturing the videos from the unstabilized positions and it also performs the segmentation processes on the text from the cluttered images and recognizes the characters perfectly. Ye et al. [10] have discussed about the technical issues and have performed the text detection and recognition on the color images. Their method has handled the basic issues and has resolved them effectively. They had also considered some problems that had not been addressed earlier and have proved the efficiency of the same by conducting experiments with the benchmark datasets. Yu et al. [11] had developed a new approach that enhanced the width of the image applying two steps such as the edge classification and the recombination processes. They have applied the new idea over the processes of merging the regions and over the segmentations. In their work, the input image's edges were divided into many segments and were merged at their edge segments as well. In their next step, the various boundaries of the images were organized as a text chain and were classified by applying the features of the characters and chains. The grey images were extracted according to the position of the edge after performing the edge recombination process. They have used the standard dataset called the ICDAR dataset for evaluating their approach and proved them to be better when compared with the other existing works.

Zhu et al. [12] developed a new system that identifies the text in the natural scenes by applying the different cues. They have used the CNN along with the k-means as the Convolutional k-means for performing the feature detection process together with the feature mapping of the CNN. Moreover, they have obtained the necessary pixels, characters and the detection of the text by achieving better result in terms of the f-measure value for the ICDAR 2015 dataset. In addition, they have proved their system capability by comparing with the existing systems. Yang Zheng et al. [13] developed a new image operator for extracting the textual data in the natural scenes. They have used the CNN for verifying the text and this improves the text detection process. Moreover, they have used two datasets such as the ICDAR 2011 and the ICDAR 2013 for proving the efficiency of their model. Tang et al. [14] have proposed a new technique for detecting and segmenting the scene text according to the cascaded CNNs. They have developed a CNN based text extraction model using the edges and the text regions. They have used three different datasets for proving the efficiency of their method and showed that their method outperformed the other existing methods. Kanagarathinam et al. [15] have explained the raw image datasets, segmentation and detection of the texts in videos. They have recognized that the text from the segments is carried out by applying the datasets.

Dai et al. [16] proposed a new architectural framework for positioning the regions of the texts that deal with the curved shape scenes. They have exploited the context and the box aware text segmentation for obtaining the scene text positions. They have conducted experiments using the total TEXT, CTW1500 and ICDAR 2015 for demonstrating the superiority of their model. Liu et al. [17] developed a

new approach for detecting the text in the scene images. They have decomposed the scene images by applying the morphological analysis that reduces the text components. In addition, they have improved the image decomposition, discriminative dictionaries and learnt from the training samples. Finally, they have proved the efficiency of their approach in terms of the text detection accuracy. Hou et al. [18] developed a new hidden anchor mechanism for detecting the scene text. Moreover, they have developed a post processing method for detecting the text. The numerous investigations thus accomplished have been the lead for demonstrating the performance of the proposed model and has thus proved its efficiency. Cao et al. [19] developed a new deep learning method for detecting the text that enhances the text. They have considered the RCTW Chinese text and the real scenario dataset. Islam et al. [20] had developed a novel approach for detecting the text and for identifying the locality of the Bangla texts in the scene images. They have incorporated the double filtering process for removing the false positive rates and for increasing the f-measure values. Jose et al. [21] had developed a new solution together with a unary and binary operators for merging the bounding box and for removing the bounding box based on some conditions. Jiang et al. [22] developed a robust deformable structure with a new text region representation for detecting the text instances. The benchmark datasets were used for evaluating their model and its effectiveness was proved in terms of the representation of the text regions and the post-processing procedures.

3 System Architecture

The proposed architecture of the text detection model has been illustrated in Fig. 1, this consists of seven important components such as the natural scene image datasets, user interface module, image pre-processing, text detection and recognition module, decision manager, rule manager and a rule base. The image dataset contains the latest natural scene images with the text including the ICDAR 2003, ICDAR 2005, ICDAR 2011 and the SVT image datasets. The user interface module collects the required images from the dataset and forwards them to the image pre-processing module for performing the effective preprocessing that concentrates on the image segmentation, feature identification and selection processes. Finally, it provides the pre-processed images to the text detection and recognition modules for recognizing the text and detects them perfectly according to the suggestion of the decision manager.

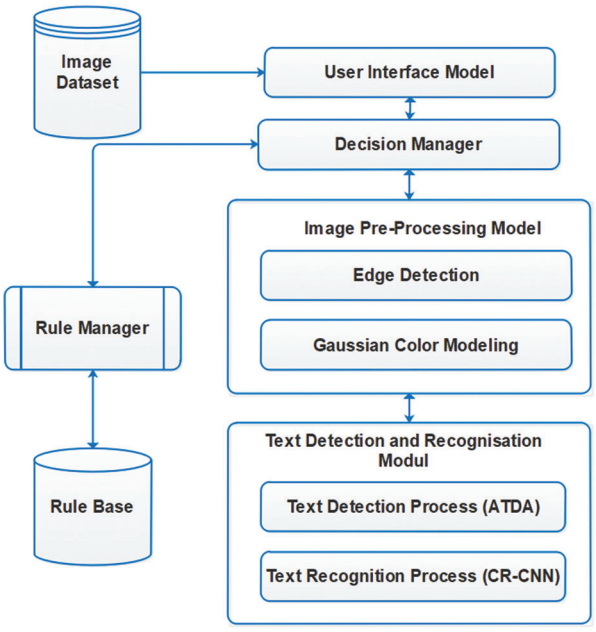


Figure 1: Text detection and recognition model

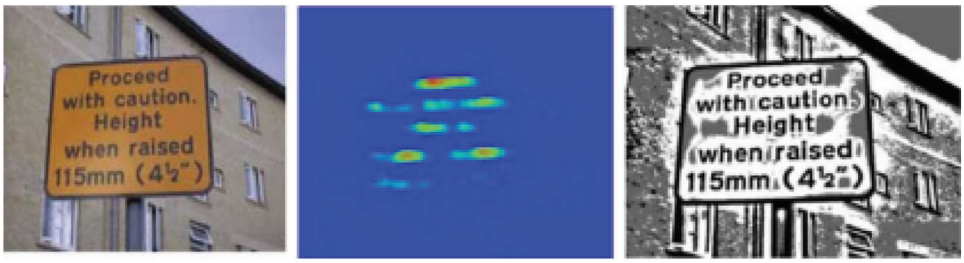
The decision manager would manage the entire architecture and would control the activities. The decision manager can select the essential rules from the rule base with the help of the rule manager. Finally, the decision would be made on the images by applying the fuzzy rules. Here, it applies three algorithms such as the prep-processing algorithm, the text detection and the recognition algorithms.

4 Proposed System

This work proposes a new intelligent text detection and recognition method for detecting and identifying the text from the natural scenes. In this method, two separate algorithms have been introduced for performing the text detection and recognition. Moreover, this work uses a newly proposed pre-processing technique for enhancing the text detection accuracy and also uses the newly generated fuzzy rules for finalizing the decisions on the text detection and recognition processes. This section explains the proposed model with the necessary background information, algorithmic steps and explanation.

4.1 Pre-processing

The image pre-processing is performed in this work by considering the general image pre-processing activities such as edge detection, color modeling and layout relation analysis shown in Fig. 2. In this section, all the three pre-processing activities are described in detail in a separate sub section. Initially, the edge detection process is explained in detail in this section with the necessary background details and the working processes on the input images in this work.



(a) (b) (c)

Figure 2: Image pre-processing a) Original image b) Text region detected image c) Segmented image

4.1.1 Edge Detection

The edge detection process is significant for refining the image segmentation process and the effective segmentation process is used for discovering and recognizing the text perfectly. This work applies an existing edge detection algorithm called the Canny Edge Detection Algorithm (Ref). This section explains in detail about the Canny Edge Detection algorithm and the functionality of the same in this proposed model.

The major objectives of the edge detection algorithm are (i) to reduce the error on the image edges, (ii) text localization for detecting the correct edge and (iii) not to possess one response to an edge of the image. Canny edge detection algorithm is capable of reducing the errors in the edge detection process greatly and it also retains the useful data by applying the filtering method and also maintains the changes from the original input image and removes the multiple responses to the neighborhood edges of the input image. The Canny edge detection algorithm works according to the work (ALK and Sangam 2016) with five important steps such as noise detection by applying the smoothing process, compute image gradient value, apply the Non-Maxima Suppression (NMS), double thresholding and tracking. Among them, the noise detection is very important in this work for performing the effective text detection processes. Here, the Gaussian blur has been applied for performing the smoothening process on the image successfully without noise. The

general Gaussian filter is used in this work for eliminating the noise from the input image and it is defined in the Eq. (1).

$$G(DOH,\ DOV) = \frac{1}{2\pi\gamma^2} EXP\left(-\frac{DOH^2 + DOV^2}{2\pi\gamma^2}\right) \tag{1}$$

where DOV is the distance between the origin and the vertical axis of the image, DOH indicates the distance between the origin and the horizontal axis of the image and γ represents the Gaussian distribution value.

Afterwards, the gradient values (magnitude and angle) are calculated by using the formulas given in the Eqs. (2) and (3).

$$GM = \sqrt{GM_{DOH}^2 + GM_{DOV}^2} \tag{2}$$

$$GA = TAN\left(\frac{GM_{DOH}}{GM_{DOV}}\right) \tag{3}$$

where GM_{DOH} and GM_{DOV} indicate the horizontal gradient and the vertical gradient. Here, four different angles such as 0, 45, 90 and 135 have been considered. It is used for predicting the derivative in both the vertical and the horizontal gradients using the canny edge detection method. Then, the NMS is applied for finding the edges effectively and for reducing the flaws in the edge detection process. Moreover, the double thresholding method is helpful for fixing the two thresholds such as low and high that helps in devising effective decision on the edge detection process according to the pixel intensity values of various text regions. Finally, the edges are to be eliminated from the output image when the edges are not having any connectivity with the other images by applying the hysteresis-based track edge method.

4.1.2 Color Modeling

This research work uses a Gaussian Mixtures based Color Modeling (Ref) which is useful for enhancing the segmentation and the text detection processes. This section explains in detail about the Gaussian Mixtures aware Color Modeling method with necessary steps and explanation. The color modeling is necessary for detecting and extracting the text. For this purpose, this work applies the Gaussian mixtures based adaptive color modeling algorithm for detecting and extracting the text by using the text region and the background. The D-dimensional color vector x is signified as a weighted mixture of the K basis components as given in Eq. (4).

$$p(x) = \sum_{k=1}^{k} w_k.\alpha_k(x) \tag{4}$$

where, w_k represents the mixture weight, $\alpha_k(x)$ indicates the Gaussians form and it also appears as an extended form as given in the Eq. (5).

$$\alpha(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|C|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x - m_x)^T C^{-1}(x - m_x) \right\} \tag{5}$$

m_x and C are the expectation and the covariance matrix x. $|C|$ is the basis of C. In this color modeling, the various basis components were applied for representing the different text regions with a number of color properties. The specific region (k) is used for generating the pixel, the color vector x is to be calculated by using the formula given in Eq. (6).

$$\gamma_k(x) = \frac{w_k.\alpha_k(x)}{p(x)} \tag{6}$$

The steps for finalizing the number of Gaussian mixture (K) are as follows:

Input: Extract the Text Regions

Output: K value

Step 1: Initially, assume 2 as the K value

Step 2: Do {

2.1 Extract the text regions according to the K value.

2.2 Apply Gaussian mixtures color modeling for all the sub regions of the input image.

2.3 If $SR_K(c,\ p,\ s,\ t) \supset SR_{K-1}(c,\ p,\ s,\ t_0)$ then

FLAG = 1;

2.4 If (K < K_{max}) AND (FLAG = 1)

K = K + 1;

Else

K_{max} = K,

Break;

}

Step 3: Return the K value

The above steps were used for determining the K value that plays a major role in the Gaussian Color Model and the text detection process. In this work, the color modeling and the edge detection processes are very important for detecting and recognizing the text in the natural scene images that are available in the dataset.

4.2 Text Detection Process

The text detection process is carried out in this work by applying a newly proposed intelligent adaptive algorithm that incorporates a hierarchical method with various emphases in every layer. This detection algorithm contains the necessary steps of the multi-scale edge detection method, a new searching and coloring method to the text regions in the initial text parts. Finally, the layout is analyzed for detecting the text regions. The steps of the algorithm are as follows:

Algorithm 1: Text Detection

Input: Detected Text cues

Output: Detected Text Regions

Step 1: Read the detected text cues.

Step 2: Let us assume that the text region Text Cue (c, p, s, t_o) is True.

// s: size, t: Text, t_o: The initially detected text.

2.1 Read every region of the detected text cues SR0 (c, p, s, t_o)

2.2 Assign the text cues Text Cue (c, p, s, t_o) into the region of text cues SR0 (c, p, s, t_o)

Step 3: Find the search region SR (c, p, s, t_o) in the neighborhood of SR0 (c, p, s, t_o);

3.1 If for each (c, p, s, t_o), there is no SR (c, p, s, t_o), Then

SR (c, p, s, t_o) $\supset$ SR0 (c, p, s, t_o) and SR (c, p, s, t_o) = SR0 (c, p, s, t_o),

Go to Step 5

Else

Continue;

Step 4: Apply the color modeling process in the SR (c, p, s, t_o) with the consideration of the layout for extracting the text regions in the SR (c, p, s, t_o);

Step 5: Initialize the SR (c, p, s, t_o) value where t is the text in the SR (c, p, s, t_o) and also extracted from SR (c, p, s, t_o) in Step 3.

Step 6: Remove the search regions which does not contain the text on it and also assign SR0 (c, p, s, t_o) = SR (c, p, s, t), go to step 2.

Step 7: Conduct the layout analysis for the SR (c, p, s, t)

Step 8: Provide the detected text regions.

This algorithm would receive the detected text cues as inputs for identifying and detecting the exact text regions. Here, after reading the text cues and the specific parts of the images are to be considered as the text regions in the identified text cues. Moreover, the assumed region from the text cue would be assigned as true. In this text region, the size of the area, text and the area of the initially identified and detected text would be determined. Now, all the possible regions would be read from the detected text cues with the same information. Then, the text cues would be appropriately assigned to the text regions. Afterwards, the search region area from the nearby region of the image would be calculated. Each new search regions would be considered as the search regions for the specific image when the new search region is present in the search regional area of the input image followed by the extraction of a part of the image. Then, the color modeling process would be applied on the search region by considering the text region layouts. If the new search region is not present in the text then the search region would be removed and a new search region would be searched for in the input image. Finally, the layout analysis would be conducted for the specific search regions on the text and the detected text regions would be supplied as the output.

4.3 Text Recognition Using CR-CNN

This section describes in detail about the text recognition process that is essential in the proposed system. In this work, a new CRF rules incorporated CNN has been proposed for recognizing the text. Here, the CRF based rules are generated using the lexicon term scores and the recognition scores. This work considers the window size with 32 pixels and identifies the 62 characters such as the ten numerical digits (1–10), twenty six uppercase alphanumeric (A-Z) and the twenty six lower case alphanumeric (a-z). Initially, the text detection process is also done again on the natural scene images for finding the text line that performs the text segmentation and recognition procedures for achieving better results in the text recognition process. The presence of every character that comes under the list of 62 characters is to be detected and recognized from the input natural images.

4.3.1 Training Process for the Text Recognition

In this work, an enhanced deep learning algorithm called the Conditional Random Field based Rules incorporated CNN (CR-CNN) is used for recognizing the text by performing the classification process according to the existing work (Tao Wang et al. 2012). Generally, the CNN has multiple layers including a convolutional layer. The proposed CR-CNN contains two convolutional layers with CL_1 and CL_2 .This multi-neural network detects with the values 96 and 256 for the two convolutional layers while handling

the medium size of the natural images and the values 115 and 720 are to be considered while handling large size of the natural scene images for the text detection and recognition processes. The proposed CR-CNN trains the input images as dataset that are collected from the previous text detection phase with regions identified with the text in the input images according to the work proposed by Coates et al. (2011). Here, the 32×32 pixel grayscale training images are to be extracted as 8×8 and creates the equivalent input vectors as $x^{(i)} \in R^{64}, \in \{1, \ldots, m\}$. Then, it applies the distance variance of the k-means for learning the filters $DS \in R^{64 \times n_1}$. Moreover, it is normalized into 8×8 and also calculates the responses that are received from the first level of the layer with an enhanced activation function which is given in Eq. (7).

$$AF = \max\{0, |DS^{TR}x| - \alpha\} \tag{7}$$

where $\alpha = 0.6$ is a hyper parameter. On the other hand to handle the 32×32 size of images, it calculates the AF value of the individual sub images that are sized with 8×8 for obtaining the 25×25 size of the images in the first convolutional layer. Moreover, the 25×25 size of the image is divided into 5×5 size of the images as a result of the second convolutional layer. In addition, the 2×2 sizes are also to be received as the response for the 8×8 sizes of input images.

4.3.2 Testing Process for the Text Recognition

This module incorporates and considers the output of the image pre-processing using the text region search. This work has assumed the list of the input words that are available in the natural scene images. According to the work [23], the existing knowledge can be applied for finding the text region from various input natural scene images. This text recognition process consists of two phases such as the location/line estimation phase and the text region search phase.

Exact Location Estimation: This section explains the procedure of finding the exact location of the text in the proposed text detection and recognition model. In every scaling area SA, the response time of the detector $RT_{sa}[x, y]$ at every exact location with x coordinate and y coordinate values are calculated. According to the center point of the search area on a specific text/character as a sub-region of the text is $RT_{sa}[x, y]$. Then, it applies the Non-Maximum Suppression on the $RT_{sa}[x, r]$ in every row rw for estimating the location of the text/character on a line of the input image. Moreover, the NMS response can also be defined using Eq. (8).

$$\widetilde{RT}_{sa}[x, rw] = \begin{cases} RT_{sa}[x, rw] & if\ RT_{sa}[x, rw] \geq RT_{sa}[x', rw],\ \forall x'\ s.t\ |x' - x| < \delta \\ 0 & Otherwise \end{cases} \tag{8}$$

where δ indicates the parameter width. This work creates a separate row on the input image with a non-zero $\widetilde{RT}_{sa}[x, rw]$ and creates a row wise bounding box RB_s^r with the same height. The left and right boundaries of the identified text region EL_{sa}^{rw} are defined as the min(x) and the max(x), s.t $\widetilde{RT}_{sa}[x, rw] > 0$. This yields a set of possibly overlapping line-level bounding boxes. It calculates the score for each of the text regional box by considering the average of the non-zero values of $\widetilde{RT}_{sa}[x, rw]$. Moreover, it applies the NMS for removing all the EL's which are the overlapping scores of more than 60% with another text region box and finalizes the exact location of the text region $\widetilde{EL}$.

Text Recognition Process: This section describes in detail about the text recognition process after identifying the exact location of the text region with the line. Then, segments the lines of the text region box into words and also recognizes every text of the word in the exact location. It identifies the text region box of the EL and its exact locations, finds the number of text region boxes by applying a Viterbi-Style method and also identifies the best segmentation method by applying the region search method. To compute the text region box TB, the text recognizer would be divided entirely, this obtains a 62 X N score matrix M(i,j) and recommends a best choice of the text with an index I that is centered over the identified location. Similarly in the detection phase, the NMS is applied on the score matrix of the new

text region window TRM for selecting the specific columns with the text. The other columns of the TRM are set to $-\infty$. Afterwards, the lexicon term lt^* that is more suitable with the score matrix TRM can be identified using Eq. (9).

$$ST_M^{lt} = \max_{l^t \in L^{lt}} \left(\sum_{k}^{|lt|} TRM(t_k,\ l_k^{lt}) \right) \tag{9}$$

where l^t indicates the alignment vector between the characters. S_M^t is calculated efficiently by applying the Viterbi-style alignment method according to the work [23]. It calculates the ST_M^{lt} for all the lexicon terms and the text region box TB with highest term score lt^*. In addition, it also considers $ST_{TB} = ST_M^{lt^*}$ to be the recognition score of the TB. Moreover, the recognition score of the text region box is used to segment the text region using the search region. Here, the Breadth First Search (BFS) is applied for exploring the first ten segments based on the recognition score. In this work, the candidate segmentation is considered for calculating the sum of the recognition scores ST_{TB}'s for all the text region boxes in the specific region/area/line. In addition, a new threshold value is to be fixed by using the text recognition scores for detecting the false alarm rate. In this scenario, the low text recognition score that has secured the text region would be declared as out of the text or "non-text" by applying the CRF based rules.

CRF based Fuzzy Rules: Generally, the CRF is used for identifying the more relevant features in the identification of the text. Here, the CRF based fuzzy rules have been generated by using the lexicon term score and the text recognition score for detecting and recognizing the text. The CRF based fuzzy rules and the standard activation function called the ReLU is used for making the decisions on the natural scene images. Generally, the CRF is derived from the Hidden Markov Model in which various terms and conditions are applied according to the objective of the proposed model. In addition, the standard fuzzy membership function called the Triangular fuzzy membership function is used for generating the fuzzy intervals and also performs the fuzzification and de-fuzzification processes. In this work, the text recognition score is helpful in identifying the text and the non-text.

5 Results and Discussion

This section describes in detail about the results and discussions. The proposed intelligent text detection and recognition model has been implemented using Python programming and MATLAB. This work uses the standard benchmark datasets such as the ICDAR 2003, ICDAR 2005, ICDAR 2011 and the Street View Text (SVT) for evaluating the performance of the proposed model.

Datasets

This section discusses about the four benchmark datasets such as the ICDAR 2003, ICDAR 2005, ICDAR 2011 and the Street View Text (SVT). These four datasets have been used for evaluating the proposed model.

ICDAR 2003: The ICDAR 2003 dataset is a collection of camera captured images. It is a first benchmark dataset for text detection and recognition research. It contains only natural images in which 258 images are used for the training procedures and 251 images are used for the testing procedures.

ICDAR 2005: The ICDAR 2005 dataset is a collection of digital camera captured images in two different conditions such as the indoor and the outdoor. This dataset comprises of three kinds of images such as i) 258 trial train images and 251 trial test images, ii) 20 sample images and, 3) 501 competition images. All the text characters in these images are included for the numerical characters and English.

ICDAR 2011: The ICDAR 2011 dataset is derived/created from the first dataset ICDAR 2003 with small changes. Here, totally 484 natural scene images are available in which 229 are training images and 255 are testing images.

Street View Text (SVT): The street view text dataset contains 350 images that are captured and annotated with alignment from the Google Street View. Moreover, it contains smaller resolution and lower resolution text and not all the text instances.

Various experiments have been performed for proving the efficiency of the proposed text detection and recognition model in terms of image pre-processing, text detection and text recognition processes.

6 Experimental Results

The proposed intelligent text detection and recognition system incorporates the image pre-processing techniques such as the edge detection and the color modeling processes. Initially, the proposed Intelligent Text detection and recognition model performs

The proposed model is also implemented and executed according to the work done by [24]. Specifically, the text region level is measured in this work like the word-level with a lexicon containing all the terms that are taken from the ICDAR 2003 dataset named as the I-WD, and the 50 random "distractor" terms are considered as a test dataset named as the I-WD-50). On the other hand, the SVT dataset provides the lexicons for evaluating the accuracy named as the SVT-WD. Fig. 3 shows the comparative analysis among the proposed model and the existing works that are proposed by [25,26] in terms of recognizing the word on the ICDAR 2003, the ICDAR 2011 and the SVT.

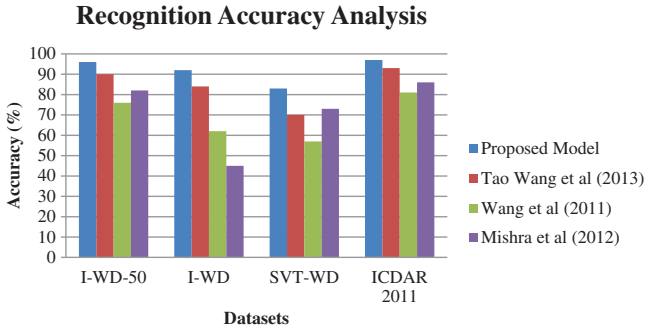


Figure 3: Recognition accuracy analysis

The proposed work has been evaluated using the three standard benchmark datasets such as the ICDAR 2003, ICDAR 2011 and the SVT datasets in terms of recognizing the terms. In this experiment, the proposed model identifies the location of the text in the given natural scene image. This work has considered the three different lexicons such as the I-5, the I-20 and the I-50 words that are provided by [24]. In addition, the full dataset is also considered in the ICDAR 2011 and the SVT. At the end of the accuracy analysis, the proposed model has been found to perform well than the existing works such as the ones in [26–29]. The reason for the performance improvement is the use of effective image pre-processing in this work. The text detection and recognition accuracy has been calculated for the various relative model sizes in this work. The various relative model sizes has been selected according to the various stages of the pipeline and trains the text detection and text recognition processes with different number of filters in the proposed deep learning algorithm called the CR-CNN. The detection modules consisting of 64, 128 and 256 on filter n2 has been used in the proposed model. These detection modules can be called as D-64, D-128 and D-

256 respectively. In addition to this, the text recognition modules have been applied with the consideration of C-180, C-360 and C-720 that corresponds to 180, 360 and 720. Finally, the smaller models consist of 25% and 50% of learnable parameters that are compared with the model that consists of 100% learnable parameters. Fig. 4 shows the text detection and recognition accuracy in two different ways of detection and recognition module.

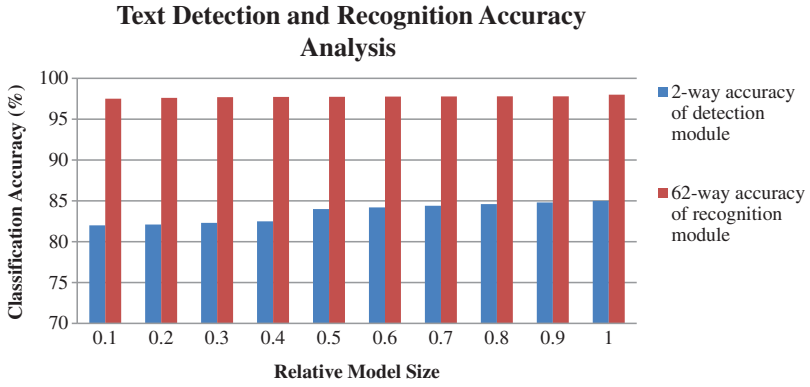


Figure 4: Text detection and recognition accuracy analysis

From Fig. 4, it is proved that the efficiency of the proposed method is ideal with the consideration of 62-way accuracy recognition modules then the module which is considered with the 2-way accuracy detection. This is due to the effective pre-processing activities such as the edge detection and the color modeling together with the newly proposed deep learning algorithm that incorporates the CRF and the fuzzy rules for making decisions on the text detection and recognition process. Fig. 5 shows the classification accuracy of the proposed model and the existing works on the ICDAR 2003 test images. The various recognition modules such as C-180, C-360 and C-270 have been used in this work for performing the classification process. Five different experiments have been conducted for the various models for evaluating the proposed model using the standard benchmark dataset. The various sizes of the images were considered for performing the different experiments.

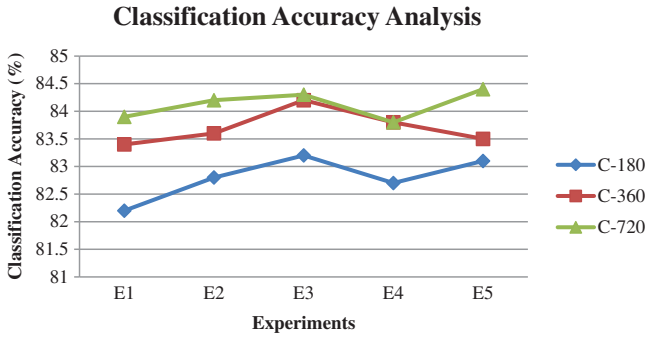


Figure 5: Classification accuracy analysis with respect to the recognition modules

From Fig. 5, it can be understood that the classification accuracy of the planned model in the recognition module C-720 is better than the other recognition modules C-180 and C-320. This is due to the incorporation and use of the image pre-processing activities such as the edge detection and the color modeling, text detection algorithm and the text recognition method. Fig. 6 shows the text recognition accuracy analysis of the proposed model which is the combination of the pre-processing activities such as the edge

detection and the color modeling, text detection method and CR-CNN. Here, three datasets such as the ICDAR 2003, ICDAR 2011 and the SVT have been used for conducting the five different experiments. The equal numbers of natural scene images were used for conducting all the five experiment images. Fig. 6 establishes the performance of the projected model in terms of text recognition accuracy on the three standard benchmark datasets. From this figure, it can be seen that the presentation of the proposed ideal is better on the ICDAR 2011 dataset images than the other two dataset images.

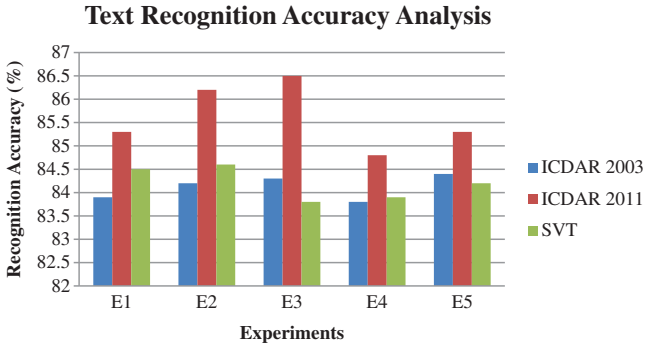


Figure 6: Text recognition accuracy analysis

Fig. 7 shows the comparative analysis in terms of the classification accuracy between the proposed CR-CNN and the other classifiers that are available and applied in this direction earlier. Five different experiments have been conducted for evaluating the various classifiers using the datasets such as the ICDAR 2003, ICDAR 2011 and the SVT.

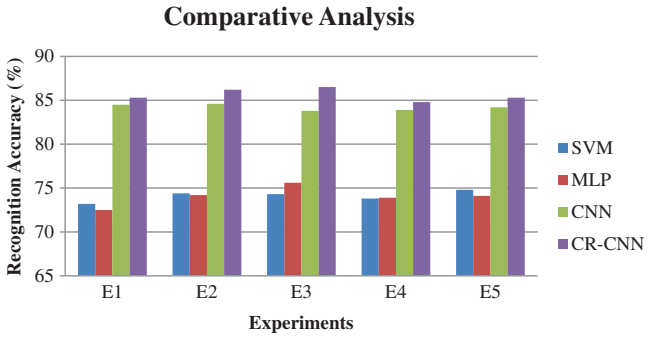


Figure 7: Comparative analysis in terms of text recognition accuracy

From Fig. 7, it is seen that the text recognition accuracy of the proposed CR-CNN is executed well than the other classifiers such as the SVM, the MLP and the CNN. The Cr-CNN has been found to perform well in all the five trials. The reason for the improvement is the application of the CRF based rules along with the standard activation function ReLU on the standard CNN.

7 Conclusion and Future Work

This paper proposes a new intelligent text detection and recognition method for detecting the text from natural scenes and also recognizing the text by applying the newly generated fuzzy rules and the CRF incorporated CNN. This model incorporates the pre-processing activities such as the edge detection and the color modeling. The edge detection process has been accomplished using the existing Canny method

and also applies the Gaussian matrix based color modeling. Moreover, a new text detection method has been proposed and implemented in this work for detecting the exact text from the input natural scene images. In addition, the newly generated fuzzy rules and the CRF incorporated CNN have been proposed and implemented for making effective decisions on the processes of text detection and recognition. The proposed intelligent text detection and recognition model has been evaluated by conducting various experiments using the standard datasets such as the ICDAR 2005, ICDAR 2003, ICDAR 2011 and the SVT datasets and has been proved to be better than the other existing works. The proposed model can be enhanced by the introduction of temporal constraints for handling the text detection and recognition processes on running videos and camera captured image frames.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. C. Yin, Z. Y. Zuo, S. Tian and C. L. Liu, "Text detection, tracking and recognition in video: A comprehensive survey," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2752–2773, 2016.

[2] C. Yao, X. Bai and W. A. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.

[3] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in *Proc. Fourth Int. Conf. on Advanced Computing & Communication Technologies*, Rohtak, India, vol. 12, pp. 5–12, 2014.

[4] M. S. Das, B. H. Bindhu and A. Govardhan, "Evaluation of text detection and localization methods in natural images," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 6, pp. 277–282, 2012.

[5] B. Soundes, G. Larbi and Z. Samir, "Pseudo zernike moments-based approach for text detection and localisation from lecture videos," *International Journal of Computational Science and Engineering*, vol. 19, no. 2, pp. 274–283, 2019.

[6] C. Yao, X. Bai and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.

[7] M. Woschank, E. Rauch and H. sifkovits, "A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics," *Sustainability*, vol. 12, no. 9, pp. 3760–3781, 2020.

[8] V. Tadic, T. L. Turukalo, A. Odry, Z. Trpovski, A. Toth *et al.*, "A note on advantages of the fuzzy gabor filter in object and text detection," *Symmetry*, vol. 13, no. 4, pp. 678–685, 2021.

[9] H. Y. Darshan, K. Gopalkrishna and H. raju, "Text detection and recognition using camera based images," in *Proc. 3rd Int. Conf. on Frontiers of Intellegent Computing (FICTA)*, Cham, Switzerland, vol. 14, pp. 573–581, 2014.

[10] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1480–1500, 2015.

[11] C. Yu, Y. Song, Q. Meng, Y. Zhang and Y. Liu, "Text detection and recognition in natural scene with edge analysis," *IET Computer Vision*, vol. 9, no. 4, pp. 603–613, 2015.

[12] S. Zhu and R. Zanibbi, "A text detection system for natural scenes with convolutional feature learning and cascaded classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, USA, pp. 625–632, 2016.

[13] Y. Zheng, J. Liu, H. Liu, Q. Li and G. Li, "Integrated method for text detection in natural scene images," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 11, pp. 5583–5605, 2016.

[14] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1509–1520, 2017.

[15] K. Kanagarathinam and K. Sekar, "Text detection and recognition in raw image dataset of seven segment digital energy meter display," *Energy Reports*, vol. 5, no. 8, pp. 842–852, 2019.

[16] P. Dai, H. Zhang and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 1969–1984, 2020.

[17] S. Liu, Y. Xian, H. Li and Z. Yu, "Text detection in natural scene images using morphological component analysis and laplacian dictionary," *IEEE/CAA Journal of AutomaticaSinica*, vol. 7, no. 1, pp. 214–222, 2020.

[18] J. Hou, X. Zhu, C. Liu, K. Sheng and L. Wu, "HAM: Hidden anchor mechanism for scene text detection," *IEEE Transactions on Image Processing*, vol. 29, no. 21, pp. 7904–7916, 2020.

[19] L. Cao, H. Li, R. Xie and J. Zhu, "A text detection algorithm for image of student exercises based on CTPN and enahnced YOLOv3," *IEEE Access*, vol. 8, pp. 176924–176934, 2020.

[20] R. Islam, R. Islam and K. Talukder, "An enhanced MSER pruning algorithm for detection and localization of bangla texts from scene images," *the International Arab Journal of Information Technology*, vol. 17, no. 3, pp. 375–384, 2020.

[21] L. Jose, F. Campana, A. Pinto, M. Albeto, C. Neira *et al.*, "On the fusion of text detection results: A genetic programming approach," *IEEE Access*, vol. 8, no. 2, pp. 81257–81270, 2020.

[22] X. Jiang, S. Xu, S. Zhang and S. Cao, "Arbitrary-shaped text detection with adaptive text region representation," *IEEE Access*, vol. 8, no. 12, pp. 102106–102118, 2020.

[23] S. Sarawagi and W. W. Cohen, "Semi-markov conditional random fields for information extraction," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1185–1192, 2004.

[24] K. Wang, B. Babenko and S. Belongie, "End-to-end scene text recognition," in *Proc. International Conference on Computer Vision*, Barcelona, vol. 11, pp. 1457–1464, 2011.

[25] A. Mishra, K. Alahari and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, vol. 3, pp. 2687–2694, 2012.

[26] T. Wang, D. J. Wu, A. Coates and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. 21st Int. Conf. on Pattern Recognition (ICPR2012)*, Tsukuba, vol. 3, pp. 3304–3308, 2012.

[27] N. Krishnaraj, K. Vinothkumar, T. Jayasankar and V. Eswaramoorthy, "Effective hybrid technique in security based wireless sensor network," *Journal of Computational and Theoretical Nanoscience*, vol. 18, no. 4, pp. 1300–1305, 2021.

[28] N. Krishnaraj, M. Elhoseny, E. Laxmi Lydia, K. Shankar and O. ALDabbas. "An efficient radix trie-based semantic visual indexing model for large–scale image retrieval in cloud environment," *Software: Practice and Experience*, vol. 51, no. 3, pp. 489–502, 2021.

[29] N. Krishnaraj, M. Elhoseny, M. Thenmozhi, M. M. Selim and K. Shankar, "Deep learning model for real-time image compression in internet of underwater things (IoUT)," *Journal of Real-Time Image Processing*, vol. 17, no. 6, pp. 2097–2111, 2020.