Recent techniques of Extraction and Recognition Textual Information from Natural Scene: Review

Ravi Kumar Barwal, Ashok, Neeraj Rohilla, Manisha saini
Assistant Professor(Computer Science), Govt P.G College , Ambala Cantt.

Abstract

With the growing reputation of the digital camera and embedded visual devices, the text extraction commencing natural scene pictures has become the key issue that is considered to alter the real time applications. Text data contain the natural scene pictures for the analysis of the videos, indexing and recovery process. Normally, the textual data in the scene pictures present the high level semantic data regarding the picture and background which are helpful for proper understanding of the pictures. However, an effective text extraction method has been a challenging approach at present years because of different font-style, text magnitude, alignment and complicated background. To overcome this issue, text extraction method for complicated video scene pictures has been proposed in this paper. Text extraction from video analysis is done through various phases; pre-processing, feature extraction and combing of the key-frames. A comprehensive survey has been done on the extraction of the natural scene pictures. In addition, various generalized texts, images has been descried with figures. Moreover, different method of the text extraction has been elaborated along with advantages and disadvantages. The methods includes region, texture, color, and hybrid based and its sub types has been described.

Keywords:*Scene Pictures, Text Extraction Method, Complicated Background.*

1. Introduction

With the advancement of the internet and smart sensors, there have been the vast demands of the applications and the methods of the picture and video analysis and data retrieval. Some of the applications may be beneficial for the text data in the natural scene [1]. Conversely, scene text extraction is the challenging issue because of the cluttered context of the natural scene and diverse design of the scene text itself. Text extraction is distributed into various elements as detection and recognition. Scene text detection is the searching of the areas consisting text from digital camera captured from the picture and videos [2]. The text layout study is dependent on the gradient and analysis of color to extract the users of the strings of text from the cluttered contextual in the normal video scene. After that, the text structural analysis is analyzed to establish an effective text structure characteristic for differentiating text from the non-text outline between the users of the text strings [3]. In contrast, scene text recognition is to convert the picture based text in detecting areas into understandable text codes. Moreover , the text mainly occurred in the natural scene in the form of the data signage such as road traffic and workshop sign. Generally, the text data play an essential role in daily life, because it presents the direct and definite clarifications of the adjacent environment. Presently, with the advancement of the smart sensors, there has been increased demand of the picture/video surveillance like as reading, CBIR(content based picture retrieval), assistive direction finding and so forth. Mostly, complete applications may benefit from automated text data extraction from the natural scenes [4]. In contrast, text is a significant tool for the group effort and connection and also plays an essential role in the modernised society [5]. Moreover, high level semantics of the text may be valuable for understanding the domain of our surroundings environment. For instance, the text data may be utilised in the wide variety of the real time applications like as searching of the image, immediate translation, robot navigation, and engineering automation. The other promising applications that are interrelated to scene text are, understanding of the text and retrieval of text.Hence, automated reading of the text from the natural situations as given in figure 1 , for the text detection and recognition has increased demand and hot topic in the computer vision [6]. Text understanding is to receive the text data from the natural scene to realize the surrounding situation and objects, Whereas, the text retrieval is the validate if the part of the text data take place in the normal scene. In smart sensor devices, the above two applications are mostly used.

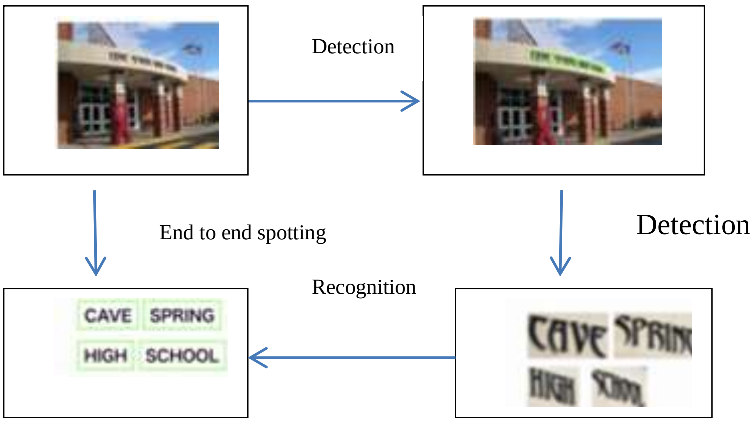


Fig.1 Scene Text Detection and Recognition [6]

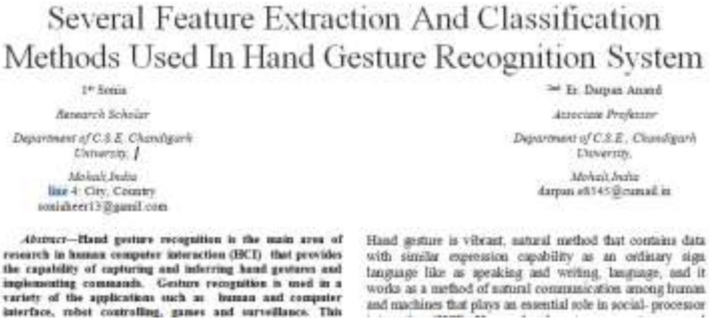
Text identification, division and extraction from the complicated pictures may be applied in different fields where data requires to be examined and investigating. A number of the applications are described as [7],

Generally, the effectiveness of the compiling the digitalized libraries, the video dataset may be enhanced by automatically understanding the pictures. In the content based filtration process, the picture spam may be identified and pornography, fake words may simply filter. A text extraction method may be applied to identify the scene from pictures achieved with mobile phones, robots, intelligent monitoring scheme. In addition, text extraction from traffic condition may be supervised and vehicle license may be identified from traffic accidents that may enhance the efficiency of the transport scheme. In addition, reading the text from the document and translate the pictures in the form which computer may modify. OCR scheme may capable to receive documents and nourishment directly in the form of the electric computer folder, and after that editing the folder through the word processor. Extraction of the text through the video series is done through valued data about the content in pictures and video sequence.

1.1 Categories of the General Text Pictures

Images may take place in a variety of the classes. The generalised text, pictures may be analysed in different types which are categorised under [8];

Document pictures. Generally, the document pictures are presented in the form the script and graphics. The kind of the pictures is created by scanners or digital camera that achieve published and handwritten documents , ancient documents so forth. The picture is converted from document based in to the picture format for electrical reading. In the initial phase of the text extraction, the main focus is mainly on the document pictures.



Several Feature Extraction And Classification
Methods Used In Hand Gesture Recognition System

1st Sonia
Research Scholar
Department of C.S.E. Chandigarh University, |
Mohali India
line 4: City, Country
soniaheer13@gmail.com

2nd Er. Darpan Anand
Associate Professor
Department of C.S.E., Chandigarh University,
Mohali India
darpan.e8145@cumail.in

Abstract—**Hand gesture recognition is the main area of research in human computer interaction (HCI) that provides the capability of capturing and inferring hand gestures and implementing commands. Gesture recognition is used in a variety of the applications such as human and computer interface, robot controlling, games and surveillance. This**

Hand gesture is vibrant, natural method that contains data with similar expression capability as an ordinary sign language like as speaking and writing, language, and it works as a method of natural communication among human and machines that plays an essential role in social- processor

Fig .2 Document Picture Example [8]

Scene Pictures. These pictures contain the text, like as banners, advertised boards that are captured through digital cameras, and hence scene text performs with the contextual part of the scene [9]. Such pictures are interesting to detect and identify , because the context is complicated , consisting the text in varied size, style and arrangements. Moreover, the scene text is inclined by the illumination and view alterations. Present OCR s/w are not handling the complicated background interferences and non arranged text lines.

Fig . 3 Example of Scene Picture[8]

Instinctive Digital Pictures. The pictures are created by computer s/w and protected as digitized pictures. When comparison is done with the document and scene pictures, some of the issues are found in instinctive digital pictures like as complicated background, minimum resolution, lossy compression [10]. Hence, text extraction is not easy to differentiate the text from contextual.

Fig .4 Example of Instinctive digital pictures [8]

Heterogeneous Text Pictures. It contains whole types of the pictures as text pictures, document and Instinctive digital pictures.

2. Literature Criteria

In computer visualization, the elimination of the text in picture is the standard issue. Normally, an automated detection and extraction of the text in pictures has been utilised in various aspects. Large number of the papers has been surveyed and some of them are listed below,

Liu, F et al., 2019[11] proposed research on the scene text identification technique that links the (CNN) convolutional neural network and (RNN) recurrent neural network. The pyramid system has been applied in the CNN portion to eliminate the multiple scale structures of the picture for getting the texture at different scales. They used the Bi- LSTM(bi-directional longer term memory) for encoding the features with the output as the sequence characters along with the output a sequence of the texture proposal. The proposed method was associated into the text connection that was adapted to arranged texts. The planned technique was computed on the desired databases, which are ICDAR2013, ICDAR2015, and USTB-SV1K. For ICDAR2013 and USTB-1K. Experimental analysis has determined that they achieved the value of F-measure as 92.5% and 62.6%. The F measure achieved by using the dataset ICDAR2015 was 73%.

2.1 CNN classification and Architecture

The structure of the planned technique is given figure 4. Generally, the CNN part is the method to eliminate the features of the input pictures that contains the feature pyramid. After that, the spatial window required to slide the feature map of the feature pyramid and convolutional of every window are placed in RNN portion of the series input. The RNN portion is realistic to encrypt the background data or the text part. The inner stage of the RNN is represented in required FC layer and creates the estimations. In addition, the ROI pooling consider the last estimations and create a sequence of the text schemes.

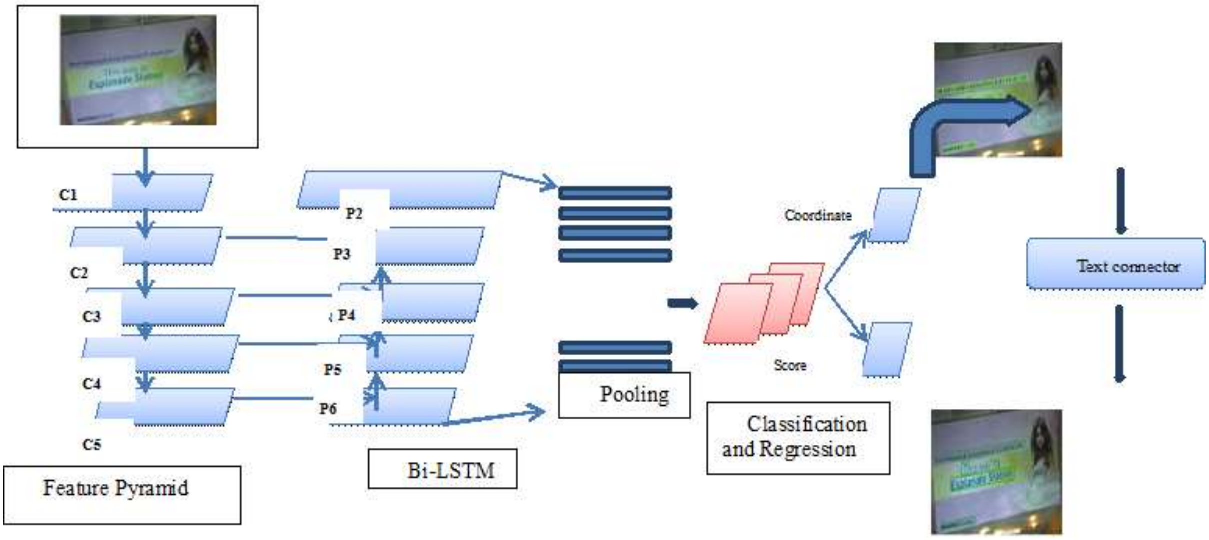


Fig.4 Structure of different layers of the network[11]

In figure 4, the extraction of the features from various layers and ,contains feature pyramid .In subsequent procedure , the spatial-window(SW) moves by the feature maps(FM) of the feature pyramid(FM) and convolution characteristics of every window are stored in Bi-LSTM. Generally, the Bi-LSTM is linked to FC layers where the result of the sequence of the proposals. After that, RNN portion is linked to region of interest, pooling layer followed by external layer that estimates the texture or non texture scores and coordinate value of the texture plan models. In the final approach, the plan model are linked into texture link through texture connector.

2.2 Extraction of Features

Moreover, using features of every layer for not dependent prediction infinitely lead to large calculation resources computation and lead to predictable usedby different non robust-features [12]. In the given method, FPN reserves as the better support network and accordingly resolve the issues. For using great resolution and robust semantic-features, FPN(feature pyramid network) includes the bottom-up , top -down and adjacent networks. The bottom up approach used the output of remained block of the every phase of resnet consisting conv1, con2, con3, conv4 and conv 5 outputs. On other hand , pathway top samples have the superficial map with pyramid level to imagine the high resolution features. After that, the spatial dimensions of the top and bottom pathway are combined with the adjacent networks. In the final process, the group of the feature maps are labelled by P2,P3,P4,P5. In [1],P6 are included into the feature pyramid , then P6 is advances to sub-sampling of the P5. Moreover, the feature pyramid utilised RPN(region proposed network) that is called as P2,P3,P4,P5. Moreover, P6 is uncontrolled due to less resolution for the text identification job. Then , they achieved feature P2,P3,P4,P5.

2.2 Encoding of the RNN

The series feature is the essential transformation among the text and general object identification , that was demonstrated in CRNN experimentation , whereas RNN used the contextual data so as to decrease the fake and lost reviews [13].

2.3 Connecter of Text

The sequence of the regular text proposal and required a text connector to build the absolute result. After stimulated by CTPN(connected text proposed network), the group of P_k is defined for the plan P_j if P_k>P_j if satisfy the needs as;

(i)P_k is closest to P_j and the displacement among themselves is less than $w_k + w_j.$

(ii)P_k and P_j have large number of 0.5 vertical overlapping in which w_k and $w_j$are the breadth of the plan method P_k and P_j.

The two plan methods are combined. The enhancements are made so that the text plan method are interconnected sequentially into a four sided instead of the rectangle. Hence, the text line may be adapted to the arrangement. However, the end to end text identification model is to create in a large number of the pictures which consists the text data [14].

2.4 Multi task Missing

The identification model finals with the text and non-text classification and regression with bouncing box and multi task missing is described as follows,

$$\text{L} (\{ p_j\} ,\{t_j \}) = \frac{1}{N_{cls}}\sum_j L_{cls} (p_j , p_j^*) + \lambda \frac{1}{N_{reg}} \sum_j p_j^* L_{reg} (t_j , t_j^*) \quad \text{..................(i)}$$

Here in equation (i), the value j is the index value in the minimum batch and p_j determines the possibility of the target , p_j^* identifies the ground truth that may be one or zero . In addition, the location of the possible frame and location of the ground box, correspondingly. The missing of the classification is dependent on the regression that is computed as,

$$L_{cls} (p_j ,p_j^*) = - \log [p_j^* p_j + (1- ,p_j^*) (1- p_j)] \quad \text{...........(ii)}$$

$$L_{reg} (t_j , t_j^*) =\sum_{j \in\{z,h\}} smooth\, L_1 (t_j , t_j^* \text{.....................(iii)}$$

$$smooth\, L_1 (\text{y}) = \begin{cases} 0.5x^2 & if\, x<1 \\ x - 0.5 & otherwise \end{cases} \quad \text{............................(iv)}$$

The related possible coordinate is realistic in CTPN Every parameter in the bouncing box regression is computed as;

$$t_d = (d_y - d_y^b) /g^b t_g = \log (\text{g}/g^b) \quad \text{......................................(v)}$$

$$t_d^* = (d_y^* - d_y^b) /g^b t_g^* = \log (\text{g}/g^b) \quad \text{.......................................(vi)}$$

In equation (v) and (vi), t = $\{t_d, t_h \}$ and t^* =$\{ t_d^* t_g^*\}$ are the relative possibility coordinate and coordinate value of the ground truth box, correspondingly. Whereas, the value g , g^b and g^* the possible height , and ground truth box.

Yanagi, R et al., 2019[15] implemented new method that depends on the text to picture generic adverbial network(GAN). In this research, they applied the text to pictures scene recovery for demonstrating the presence of the unpleasant created pictures. On other hand, the pictures created were dependent on the GAN that were more appropriate to visualised features to determine the correct scene recovery. The planned technique used an atten GAN as the texture to picture converter. Initially they used the minimum and maximum resolution pictures created by focusing on single word input sentence. In addition , the different resolution were used , and they focused on complete sentence and every word , simultaneously. Using this method, they realised the recovery through sentence which were not suitable for the word groupings. Therefore, the planned technique realised the complex scene recovery which may differentiate a minimum variation among the same scenes.

Huang, Z. et al., 2019 [16] presented a novel mask R-CNN dependent text detection method that may robustly identified the multi organised and curve texture from standard scene picture in the definite way. For the improvement of the feature demonstration capability of mask R-CNN for the detection of the texture , they implemented a PAN(pyramid attention network) as the novel contextual network of mask R-CNN. Experimental analysis was done that PAN may overwhelm the fake alarm rate that may lead to the texture like context in more efficient way. The planned method has acquired the high performance rate on both multi oriented and curve text identification benchmark jobs though single model texting.

The planned R-CNN dependent on the mask R-CNN is based on four models which are;

2.5 PAN(Pyramid Attention Network) backbone network

It is accountable for calculation of the multi scale convolution feature pyramid above complete picture.

2.5.1 RPN (Region proposed network)

It creates the rectangle texture model. Generally , RPN are associated to P2,P3 and P4 accordingly, every slide of the minimum network deeply on the related pyramid level to perform the classification of the text and bounding box regression[17].

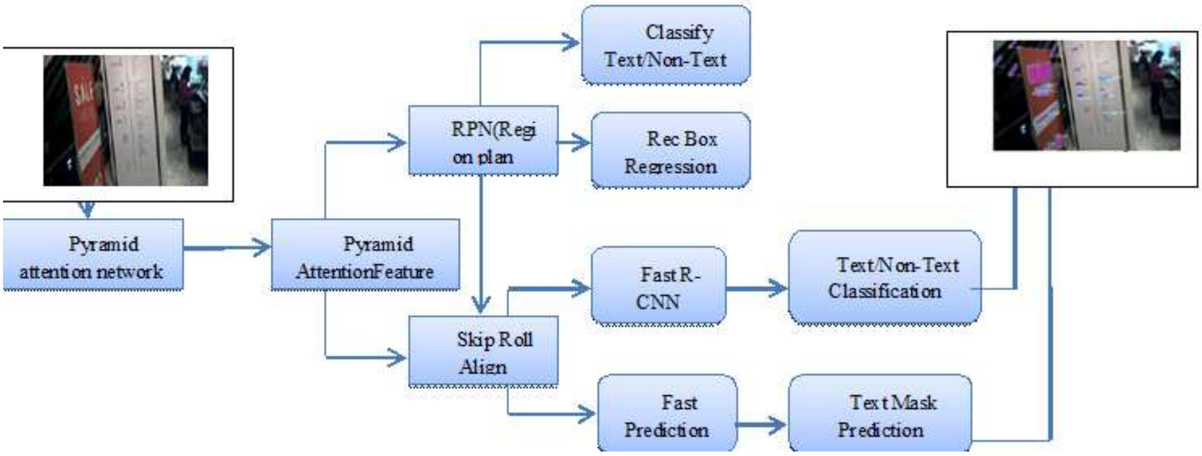


Fig .5 Structural design of mask R CNN based text identifier that contains PAN backbone system, RPN, mask PN[17]

The implemented network is about 3x3 convolution layer considered by double lower convolution layers, that are used for estimating the texture score and rectangle bounding box positions .

2.5.2 Fast R-CNN detector

This classified eliminated proposal and resulted value that is related to quadrilateral bounding box. After the development of the region creation stage , eliminating the efficient features of every proposal is delicate to the performance of succeeding fast R-CNN and mask probable network. In the actual fastest R- CNN , the characteristics of the complete plan methods are eliminated from the final convolution layer on contextual network , that may result in inappropriate features of smaller proposals.

2.5.3 A mask estimation network

It estimated the texture mask for the input plan method. The quadrilateral bounding box is used by the fast R-CNN model of the last identification output. However, the text mask is used for the prediction of the text mask through the prediction network as last detection output.

In fig 5, the pyramid attention network (PAN) that is comprised of pyramid attention feature pyramid with values P2, P3 and P4. In addition the region planned network is segmented as texture and non-texture classification. RPN has planned model as skipping roll alignment having fast R CNN and mask prediction determine the output picture.

Gómez, L et al., 2018[18] addressed the issue of the scene text retrieval , the required text query , the method that required to re-appear whole pictures consisting the query text. The uniqueness of the planned method contained in using the single shot convolution neural network(CNN) structure which predicts at the same time bounding box and compact text demonstration of the words. The texture dependent picture recovery task may be cast as the modest closest neighbour of the query-text demonstration above, the outcome of CNN above the complete picture dataset. The experimental analysis describes the planned structure that performed better than the existing state of art whereas other improved significantly at a fast rate.

Zeng, M et al., 2019 [19] presented an effective picture retrieval technique, namely CATIRI(content and text based picture retrieval through indexing).CATIRI followed a three-stage arrangement system that builds up another ordering structure called MHIM-tree. The MHIM-tree consistently coordinates a few components, including Manhattan Hashing, Modified record, and M-tree. To utilize our MHIM-tree carefully in the inquiry, they presented a lot of significant measurements and uncover their characteristic properties. In light of them, they build up a top-k question algorithm for CTBIR. Result analysis dependingon standard picture datasets exhibit that CATIRI beats the contenders by a request for size.

Chaithanya C.P et al., 2019 [20] identified and classified the text in natural images. The network identified the text and search the interconnected areas , connect to relative location. Generally, the detection and classification of the proposed method used various steps as;

2.6 Text Extraction Phases

2.6.1 Pre-processing

Initially, they considered the text, pictures that are diverse , placed in horizontal and vertical position. Generally, the pre-processing may the mutual operations with the images simultaneously at lower abstraction very input and output measurement intensity images. This method is used to enhance the quality of the picture [21]. Moreover, the pre-processing is done through the median filter that is the non linear filtration technique for the elimination of the noise. This kind of the noise reduction is used to enhance the later processing.

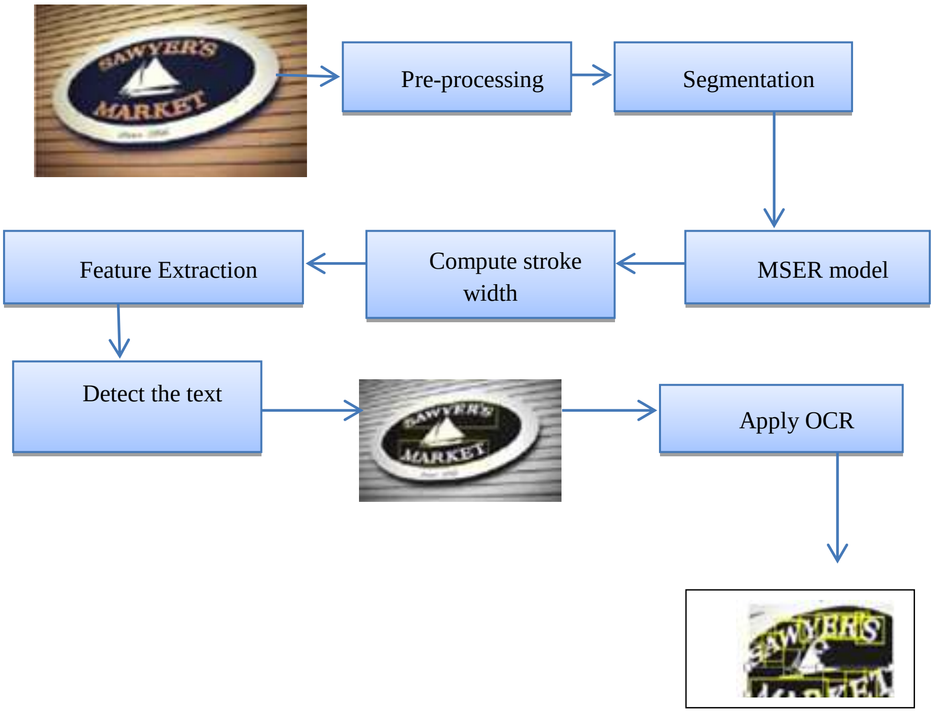


Fig. 6 Structural design of the detection and classification method[20]

2.6.2 Segmentation

It is used for segmenting the digital pictures in multiple areas or the segments. The multiple segments are arranged in the pixel pattern and it results in the group of the segmented pictures.

2.6.3 Searching the maximum stable External Region (MSER)

It is used to achieve the framework of the text in an appropriate way and to prevent the unconnected and non-uniform arrangement of the pixels. Mainly, the co-variance regions are extracted in the picture.

2.6.3 Region and stroke width picture

It is an operator that detects that every pixel in the picture and set the pixels in last candidate that depends on stroke breadth.

2.6.4 Extended bounding box

The detection output may be associated and composed a single text character. The whole character may associate and generate a novel text region. Every bounding box may demonstrate every text.

2.6.5 Distinguished text

The identified text may create an expressive sentence and words are revealed.

2.6.6 Training Dataset

Generally, the classification and training of the dataset are done using CNN. Then automated feature extraction is done. CNN is an essential classification technique for searching the accuracy of the training set.

In table 1 , various papers are surveyed and then methods are given in table with experimental analysis by analyzing different metrics. In table 2 , the advantages and issues are given after surveying different papers.

Table 1. Comparative analysis of the techniques and parameters of different surveyed papers

Author	Year	Technique	Parameter
Liu, F et al., [11]	2019	CNN	F measure =92.5%
Liao, M., Shi, [12]	2017	CRNN	Accuracy =0.93
Yanagi, R., [15]	2019	GAN	Recall =0.51
Gómez, L., [18]	2018	CNN	Precision=93.5
Zeng, M., [19]	2019	CATIRI	Accuracy dimension =128
C.P. Chaithanya, [20]	2019	MSERand CNN method	Accuracy=92.3%

Table 2. Comparison table of methods used, benefits and drawbacks

Citation	Technique Used	Benefits	Drawbacks
[11]	CNN	Multi-oriented	Text classification issue
[14]	Deep CNN	Simplification ,Accessible	Degraded gradients
[13]	Segmentation based method	Industrial automation	Curve text detection
[17]	PAN	Extract deep features and pixels	Multi scale classification

3. Scene Text Extraction Process

In the scene text extraction process, the layout of the text is differentiated from the non-texture contextual outliers that is mainly the scene of the text region recognition. The main question arises about the feature of the scene texture characters because the group of the character may be easily predicted which is scene text character estimation. Generally, the scene texture extraction is segmented into a binary function element in the standard output [22] [23]. These standard outputs are the scene text detection and recognition. The two models, creates the text identifier, and detector correspondingly. In addition, the image areas consisting the texture and strings are positioned and localized that is filtered out by the major contextual interference in the scene text detection. The detection techniques may segment the texture string in identified text areas into liberated characters for the recognition. In scene text detection, the transformation of the picture based text strings in identifying text areas into readable ASCII codes.In fig 7 , demonstrate the flow process of the scene text extraction method. The digital camera based natural scenes captured the layout of the text and structure. The main goal of the proposed method is the extraction of the text from the videos. The planned method takes place in various stages, pre-processing, Extraction of the text and grouping or alignment. Initially the videos are fragmented into the different frame that depends on the shots. Redundant frames are rejected by performing the frame resemblance that leads to the key -frame selection. Generally, the key –frames consisting of the scene text.

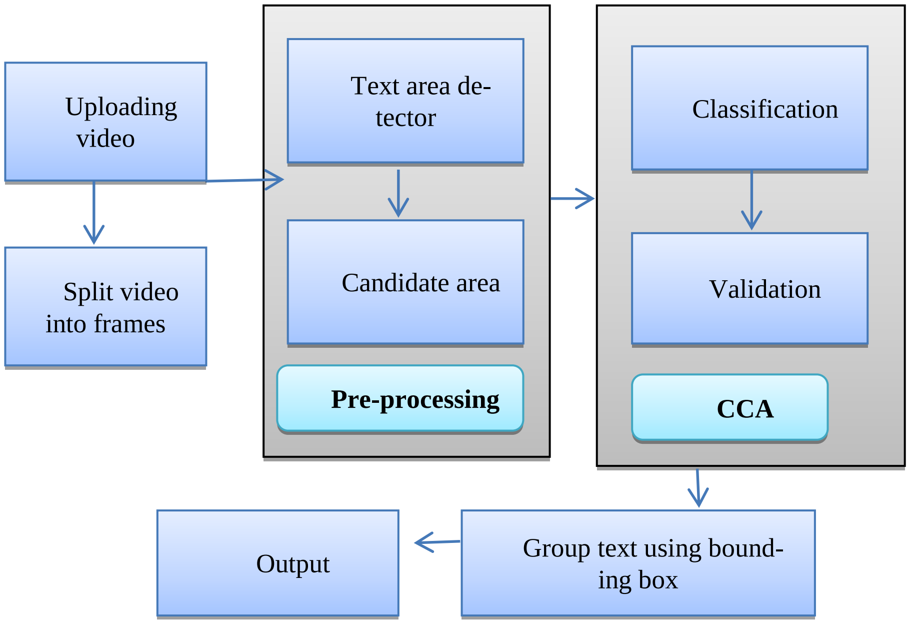


Fig. 7 Process Of Detection And Extraction Of The Scene Text

During the Pre-processing phase, the text prevalent assurance is detected with the scaling in the key-frames. In this stage, the regions are detected in the presence of the text that is also called as the candidate section. After that, the adaptive thresholding is realistic to detect the existence of the text within the key-frame. In addition ,afterwards the identification of the text area , the associated component analysis is analysed where the vertical and horizontal key frame is utilised to identify the text. In connect component analysis(CCA), the CRF model is utilised for the classification of the candidate area on two programs as; text and non-text. During the training stage, the filtration of the text as well as non text elements/components is determined. In the next stage, the extraction of the text delivered to the (optical character recognition)OCR for the validation of the characters. Besides, texts are combined to form the words and then two lines through vertical and horizontal bounding box by constructing the minimum spanning tree. The final scene text is received in the output.

4.Categories of the Text Extraction Method

The researchers have presented the different techniques for identifying the texts in natural pictures of the objects and videos. Different methods for the text detection from the scene pictures have been implemented over the last few years. In this section, a review of the different methods for the text detection, segmentation, character detection along with the merits and demerits are given below,

4.1 Region Based Method

This method used the features of the color and grayscale in text area or the variation to the related features of the context or background [24]. The texture is achieved by threading the picture at the intensity level among the texture, color and which is instantaneous background. This technique is not robust to compound contextual. The region based method is also called as sliding window. It is used in sliding window for finding the probability text in the picture

4.1.1 Advantages: Identification of the texts. It is vigorous to the complex-background.
4.1.2 Disadvantages: It has slow processing time because it is used in multiple scales.

4.2 Connected Components (CC) Based Method

This technique used a bottom up method associating minor elements into a major element unless all area is detected in the picture. Consequently, the geometrical method combined the text elements through spatial orientation so to filter out the non textual elements and edges of the text areas are labelled. This technique position the text at a fast rate, but unsuccessfully, for the complicated background. The ***Advantage is*** Receive the stating of artistic output. The complex nature may not depend on the features of text arrang**e**ment and font style. The ***Disadvantages*** *is* Unsuccessful in few scene pictures that have the reduced contrast test and suitable illumination.

This above method is classified into various classes as:

4.2.1 Edge: The boundaries have the consistent feature of the text instead of the color and intensity. The edge based technique is mainly focused on the high contrast between the texture and background. The three differentiating features of the text are rooted in pictures which may be used for identifying the text are the strength of boundary, density and variance arrangement [25]. The extraction method is the general purpose technique that may easily and efficiently position and eliminates the texture from inside/outside pictures and documents.This technique is not vigorous for controlling the large dimension texture.

4.2.2 Color: In the color based method, the pixels are classified through the color clustering with the similar colors and generating the candidate area. After that, the candidate area is examined and connected component is expected. The major issue of this technique is the grade of clustering. If the information is above cluttered , the context and the text area are combined. And , if the information is not clustered , then the amount of the clustering is increased and the performance level gets degraded. And, if the information is below the clustered value, then the quality of the clustering get incremented.

4.2.3 Color and Edge Both: By combining the edge and color text, the detection process is done. An appropriate output is received by linking the features as compared to differentiating of the features.

4.2.4 Texture: This technique is related to the text area as the specific texture. Area or the region is detected as the text area or may not be in accordance to the extracted related texture of the applicant regions. To overwhelm this issue, a hybrid method is proposed that is based on the benefits of the texture and CC based technique, which vigorously detect and position the text in the natural scene pictures.Text region indicator is established that depends on the texture. It is utilised to expect the possibilities of the location and scale of the text and after that examining the existence of the text region . This technique used the notion that the texture in pictures having different text features that differentiates it from contextual. The method is dependent on the Gabor filters(GF), Fast Fourier Transform (FFT), and SVM(support vector machine) classifier. It is utilised to identify the text features of the text area in the picture. This technique is capable to identify the text in complicated contextual [26]

4.2.4.1 Advantages: May efficiently determine the words in scene pictures with high amount of the class markers.

4.2.4.2 Disadvantages: Depends on lexicon and difficult handle large words with distortion.

4.3 Morphological Process

It is the numerical morphology that depends on the analysis of the pictures. The feature extraction is mainly determined to character detection and analysis of the documents. It is used to eliminate the essential text contrast characteristics from the administrated pictures. The features are not variant in contradiction of geometrical pictures, modifications such as transformation , rotation and scaling. Afterwards, the brilliance situation or the texture color is modified , the features are managed. The technique works in various picture modifications.

4.3.1 Advantages: Localization of objects and textures.

4.3.2 Disadvantages: Maximum processing time.

4.4 Hybrid Method

It is the method of linking both the methods which are region and texture based method. In the hybrid approach, the region based method is utilised to identify the texture or character. After that, the text based technique extracted complete features from the texture region [27]. The major benefit of the

method is that unique approach is inappropriate for complete scene pictures because of different characteristics like as color, dimensions and font difference.
4.4.1 Advantages: Region based data are valuable for the text component division and inquiry. The condition random field (CRF) model differentiates the text elements from the non-text element that have better performance as compared to local classifiers.
4.4.2 Disadvantages: Difficult to divide the text.
In table 3, after citation of various papers, the benefits and problems of the various methods of the text detection and extraction have been examined.

Table 3. Comparison table of various methods of scene text pictures.

Reference	Method	Merits	Demerits
[13]	Region based method	Robust to different backgrounds. Handling the manual written text picture .	Not capable to handle slope and bend texture.
[11]	Learning based	Cope up with background picture effectively. Provide accurate images	When texture and non-texture regions has same features , the error rate increases.
[128]	Edge based	Complex background pictures and overlapped texture are handled	May not perform on similar kind of the pictures
[28]	Region based	Robust to non linear texture.	Small texture regions are not identified easily
[36]	Cluster based	Different classifiers are detected	Not easy to identify the texture of similar background.

5. Conclusion

Scene Text Extraction methods are mainly dependent on the classification of the single areas or the patches through the usage of the previous script. Generally, the texts that originate in the pictures are the valuable data. Generally, the extraction and detection of the text areas in the picture are the major problem in the computer visualization field. A new approach for the scene text extraction has been provided, that is motivated by the human observation of the text components. An overview of the scene text extraction , applications and types has been explained in this paper. Then, the comparative analysis of the complete surveyed papers has been done by analysing the techniques and performance metrics. In addition, a hybrid process is analysed for the extraction of the text from the scene. In general, the region based and connect component is the hybrid approach that has been utilised to receive the extracted text. In the initial process, the videos are fragmented into the acquired frames. Then, the classification is through the training model and validated using OCR (optical character recognition). After that, the final output picture is received by grouping the texts through the bounding box. Afterwards, different methods of the text extraction process have been described that includes region based, connect component based, text and color based with a hybrid approach. Additionally, the sub-categories of the methods are also explained along with advantages and disadvantages. Then , the survey of the methods of different papers has been done that clarified the merits and demerits of different methods.

References

1. Kulkarni, C. R., Barbadekar, A. B. :Text detection and recognition: a review. Int Res J Eng Technol (IRJET), 4(6), 179-185. (2017).
2. Ye, Q., Doermann, D. :Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence, 37(7), 1480-1500(2014)..
3. C. K., Chan, C. S.; Total-text: A comprehensive dataset for scene text detection and recognition. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 935-942). IEEE(2017).
4. Xi, J., Hua, X. S., Chen, X. R., Wenyin, L., Zhang, H. J. :A video text detection and recognition system. In IEEE International Conference on Multimedia and Expo, 2001. ICME 2001. (pp. 873-876). IEEE, (2001).
5. Yang, H., Quehl, B., Sack, H.: A framework for improved video text detection and recognition. Multimedia tools and applications, 69(1), 217-245 (2014).
6. Long, S., He, X.,Yao, C. Scene text detection and recognition: The deep learning era. arXiv preprint arXiv:1811.04256 (2018).
7. Zhang, J., Cheng, R., Wang, K., Zhao, H. :Research on the text detection and extraction from complex images. In 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies (pp. 708-713). IEEE,(2013).
8. Karanje, U. B., Dagade, R. :Survey on text detection, segmentation and recognition from a natural scene images. International Journal of Computer Applications, 108(13) ,(2014)..
9. Sumathi, C. P., Santhanam, T., & Devi, G. G. (2012). A survey on various approaches of text extraction in images. International Journal of Computer Science and Engineering Survey, 3(4), 27.
10. Sun, L., Huo, Q., Jia, W., & Chen, K. (2015). A robust approach for text detection from natural scene images. Pattern Recognition, 48(9), 2906-2920.
11. Liu, F., Chen, C., Gu, D., & Zheng, J. (2019). Ftpn: Scene text detection with feature pyramid based text proposal network. IEEE Access, 7, 44219-44228.
12. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W. Textboxes: A fast text detector with a single deep neural network. In Thirty-First AAAI Conference on Artificial Intelligence(2017).
13. Tian, Z., Huang, W., He, T., He, P., Qiao, Y. Detecting text in natural image with connectionist text proposal network. In European conference on computer vision (pp. 56-72). Springer, Cham. (2016).
14. He, K., Zhang, X., Ren, S., Sun, J :Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). (2016).
15. Yanagi, R., Togo, R., Ogawa, T., & Haseyama, M. :Scene Retrieval from Multiple Resolution Generated Images Based on Text-to-Image GAN. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5). IEEE, (2019).
16. Huang, Z., Zhong, Z., Sun, L.,Huo, Q: Mask R-CNN with pyramid attention network for scene text detection. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 764-772). IEEE, . (2019).
17. Li, H., Xiong, P., An, J., Wang, L. :Pyramid attention network for semantic segm entation. arXiv preprint arXiv:1805.10180, (2018).
18. Gómez, L., Mafla, A., Rusinol, M.,Karatzas, D:Single shot scene text retrieval. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 700-715), (2018)..
19. Zeng, M., Yao, B., Wang, Z. J., Shen, Y., Li, F., Zhang, J., .Guo, M ;:CATIRI: An Efficient Method for Content-and-Text Based Image Retrieval. Journal of Computer Science and Technology, 34(2), 287-304, (2019).
20. C.P. Chaithanya, N. Manohar, Ajay Bazil Issac :Automatic Text Detection and Classification in Natural Images. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5S3, (2019).
21. Jacob, J., Thomas, A. :Detection of multioriented texts in natural scene images. In 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) (pp. 625-628). IEEE, (2015).

22. Gera, D., Jain, N. (2015). Comparison of Text Extraction Techniques—A Review. International Journal of Innovative Research in Computer and Communication Engineering, 3, 621-626.
23. Chen, D., Odobez, J. M., Bourlard, H .:Text detection and recognition in images and video frames. Pattern recognition, 37(3), 595-608. (2004).
24. Zhang, X. W., Zheng, X. B., Weng, Z. J. :Text extraction algorithm under background image using wavelet transforms. In 2008 International Conference on Wavelet Analysis and Pattern Recognition (Vol. 1, pp. 200-204). IEEE, (2008).
25. Audithan, S., Chandrasekaran, R. M :Document text extraction from document images using haar discrete wavelet transform. European journal of scientific research, 36(4), 502-512(2009).
26. Song, Y., Liu, A., Pang, L., Lin, S., Zhang, Y., Tang, S. :A novel image text extraction method based on k-means clustering. In Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008) (pp. 185-190). IEEE(2008).
27. Gupta, Y., Sharma, S.,Bedwal, T. Text extraction techniques. International Journal of Computer Applications, 975, 8887(2015).