**TEXT DETECTION AND RECOGNITION IN NATURAL IMAGES**

**Abstract:**

Text detection and recognition has been a well-studied problem in the past. However, when it comes to detection and identification of text in natural images and text from natural scenes, it becomes a much more challenging problem because of the distortion in geometry, variance in the illumination. In this paper, a deep learning technique is used, which is based on the convolutional neural networks(CNNs) to recognise the text in the natural images and a MSER(Maximally Stable Extremal Region) algorithm is used for detection of text in natural images. The MSER algorithm extracts from an image a number of different regions, called MSERs. An MSER is a stable connected component of certain image gray-level sets. A compact architecture is made based on original VGG-net to get the character recognition results under the similar framework. An CHAR74K dataset, which is a benchmark for natural images is used on which a lot of experiments are conducted. The main aim is to convert the images of different types such as printed, natural images into a machine encoded text such that the machine can interpret the data easily.

**Introduction:**

Text plays an important role in vision-based applications. Text detection is basically the identification of text in a given set of images. In most advanced  computer vision applications, such as visual assistance, image retrieval, text reading in natural images is very important. It is because of text in images usually conveys important information. Text detection and identification has therefore received a great deal of attention in recent years.

Over the past few years, Text detection and identification in natural images has evolved as an important and challenging problem. And it also has a lot of practical applications. Unlike recognition of characters in scanned documents, recognising text in natural and unconstrained images is a very complex problem which is complicated by fonts, conditions of lighting, textures and a lot of variations in backgrounds. Unfortunately, text characters in natural images can be of any gray-scale color (It does not always include white), variable length, low resolution and in complex backgrounds.

Many researchers have proposed different kinds of methods to solve these problems. A lot of competitions were held on this topic [2,3,4], to advance the progress on the character text reading. To use an exemplar-based support vector machine (SVM), sheshadri et al. [5] proposed a multi-layered approach for the classification of Indic script language Kannada characters, which is making use of the characteristics of the exemplars in the given dataset. These are some of the proposed methodologies for solving the challenging problems face during detection and recognition of text in natural images.

In this paper, a robust approach is proposed which combines the features of both MSER and a deep neural network. The deep neural networks are of two types, the recurrent neural network (RNN) [7] and the convolutional neural network (CNN) [6]. The deep neural network used in this model is convolutional neural network (CNN). The approach used is used to solve the complex problems of detection and recognition of text in natural images.

This paper proposes a deep learning method for the detection and recognition of  text in natural images. Based on the basic idea of the VGG-Net [8], the original architecture of the given network, and the usage of the data augmentation and initialization made to get a more compact network. The input and output of the compact network are more focused. The recognition of characters and words are done in the same framework.

The contributions of this paper are described as the following: Firstly, a labeled dataset is used to train a convolutionary neural network (CNN). Secondly, a model is built which serves the purpose of detecting the text regions in the image. Here, MSER algorithm with the help of NMS(Non-maximum Suppression) is used for the purpose of detecting the text regions in the given input image. Thirdly, when an image is loaded into the model, then it shall detect the regions in the image with text and also recognize the text characters in the given input image.

**Literature Survey:**

The system which has been designed for the most common writing system is OCR system which includes languages such as Chinese, Bengali, Indic, Devanagari, Arabic, Latin etc., using most common programming languages such as ANN, LABVIEW, TESSERACT, MATLAB. The OCR was first designed in 1965, which is based on the technology proposed primarily by the Jacob Rainbow which was used by the United technology States Postal Services. Later in 1970s, Dr. Sinha of Indian Institute of Technology, Kanpur made efforts to propose pattern analysis system. In 1974, Ray Kurzwell developed Omni-font OCR which is used for recognition of printed text in virtually any form. In 2000, OCR was made available online as a service called WEBOCR.

The recognition of characters is not a new problem but its roots can be traced back to the systems before the invention of the computers. The OCR systems which are earliest were not computers but mechanical devices that were able to recognize characters, but having a very slow speed and low accuracy. In 1951, M. Sheppard, a popular scientist discovered a reading robot GISMO that can be considered as the pioneer work on modern OCR. GISMO can read musical notes and also words on a printed page one by one. However, it can only recognize 23 characters. The machine is also designed that it could copy a typewritten page. In 1954, J. Rainbow, devised a machine that can read uppercase typewritten English characters, it reads one per minute. The early OCR systems were not popularly recognized due to errors and slow speed. Hence, not much efforts were put on this aspect during 1960’s and 1970’s. The only research that were done by large agencies like banks, newspapers and some government organisations. Due to the complexities associated with recognition, it became mandate that there should be a standardized fonts for easing the process of recognition for OCR. Hence in 1970 ANSI and EMCA developed OCRA and OCRB, that provided reasonably acceptable rates of recognition.

For the past thirty years, Extensive research has been done on OCR. This has lead to the emergence of document image analysis (DIA), multi-lingual, handwritten and omni-font OCRs . Although extensive research efforts had been made, the machine’s ability to read text is not trustworthy. Hence, current OCR research is being done on improving accuracy and speed of OCR for variety of style documents printed and written in unconstrained environments. There has not been availability of any commercial or any open source software available for complex languages like Urdu or Sindhi etc.

**Learning Architechture:**

This section explains the architecture used to learn the feature repesentations and train the classifiers that are used for the recognition process .The basic configuration is similar to a convolutional neural network.

The system runs in several stages:

1. First, an enhanced-MSER, which is an extension of the well-known MSER algorithm by incorporating saliency detection methods

2. The Second step is compression of the bounding boxes that are determined in previous step by using NMS

3. The Third step is a text filtering pipeline with a deep CNN. In the classification stage , a powerful convolutional neural network is trained which incorporates pixel-level and character-level information.

**A.MSER Algorithm**

This section describes the text detection algorithm that is MSER (Maximally Stable Extremal Region) algorithm. The MSER algorithm extracts number of intensity regions from image. MSER is based on the idea of considering regions that stay nearly the same through a whole wide range of thresholds. MSER uses two important properties to remove non text regions from image first is Stroke Width Variation Properties and another is Geometric Properties. The summarized common attributes of text as that are required to use MSER algorithm are:

a) Text in image should contain lots of edges in it.

b) The height of text is smaller than its width.

c) Text should be bounded in size.

d) Text has texture but this texture can be irregular.

**B. Non Maxima Supression(NMS):**

Non-maximum suppression (NMS) was an integral part of many proposed detection approaches, although it could be the detection of corners, edges or objects. It has been widely used in computer vision in several key aspects. Its necessity stems from the lack of detection algorithms capable to locating the concept of interest, resulting in the detection of several groups of localized areas near the actual location. In object detection, sliding window based approaches typically produce multiple windows with high scores close to proper object position. Generally speaking, this relatively dense output is not satisfying to understand the context of an image. In addition, the number of window hypotheses is simply uncorrelated to the actual number of objects in the image at this point. Therefore the purpose of NMS is to result in only one window per group, corresponding to the exact local response function maximum, ideally retaining only one detection per object. As a result, NMS also has a significant impact on performance measures that penalize double detections .

**Pseudo code for NMS used:**

For each class:

While predictions are not empty:

Pick the bounding box with biggest probability

Add it along with the class to the result set.

Remove all the other bounding boxes having IOU greater than the threshold from predictions.

IOU(intersection over union) = Area of Overlap / Area of Union

**C.Feature extraction and Detection:**

The system uses a fixed-sized input image of 32\*32 pixels for character classification,which is applied to the character images in a set of labeled train and test dataset that is obtained by dividing the present dataset into 70% train and 30% testdata.

The Dataset is composed of the 7738 images available in the CHAR74K English character dataset. We are not dealing with a binary classification anymore as in this case the number of classes is 62:

■ integers [0-9] : 10 classes

■ lowercase letters of English alphabet [a-z] : 26 classes

■ Uppercase Letters of English alphabet [A-Z]:26 classes

The dataset contains approximately 7740 labeled character images .with large numbers of features it is useful to have even more data.To satisfy the needs we have also experimented with synthetic augmentation of these datasets.and then trained the sysytem on the derived dataset which got an accuracy of 85%.The compare our proposed method with other competing methods.

|  |  |  |
| --- | --- | --- |
| **Author(s)** | **OCR Application** | **Accuracy(%)** |
| Shah, Parul, et al [21] | Chassis – Number  Recognition | 95.49 |
| Zhai, Xiaojun, et al [22] | Number Plate  Recognition | 97.30 |
| Shamsher, Inam, et al [23] | OCR for printed  Urdu Script | 98.30 |
| Yetirajam, Manas, et al | Classification and  Recognition of  Broken charcters | 68.33 |
| Proposed model | Text detection  using MSER and  CNN | 84.75 |

**Experimental Results:**

**Step1:** Choosing the appropriate Dataset

**Data Sets:**

This system can be trained using various datasets such as ICDAR Dataset, Stanford OCR, Text Detection, Scene Text, Natural Environment OCR, Street View House Numbers, Street View Text, On-line Handwriting, Document database, Arabic Printed Text, Chinese Characters, Mathematics Expressions, Devanagri Characters, MNIST Database, NIST Database, ICDAR Database etc. So the CNN in this model is trained with CHAR74K Character Dataset which consists of 7738 labelled images which indeed consists of Digits, Upper and Lowercase letters.

**Step2**: Training the CNN



As seen in above Figure , the model is trained for 70 epochs and achieved low loss with limited overfitting. With additional training data the model could achieve higher accuracy as well. The model showed 85.75% accuracy on the *testing set*.

**Step3:** Text Detection Techniques

**Different Text Detection Techniques:**

**EAST ( Efficient accurate scene text detector):**

It is a simple text detector which outputs the “quadrangle” which is merely a rotated bounding box with coordinates of all vertices.



**CRNN(Convolutional-recurrent neural network):**

Feature columns are intended to represent a certain section in the text.



**SSD(Single Shot MultiBox Detector):**



**Proposed Technique:**

**MSER(Maximally stable extermal regions):**

As the main objective of the model is to detect the character text individually, so the model cannot use the above specified techniques. To solve this problem the system uses MSER technique to identify the intensity regions and contours and hulls to detect the text along with the bounding boxes.

**MSER to remove Non-Text Regions Based On the Variation in Width of the Stroke**



**Contour Approximation Retriving Bounding Areas using Hulls**  

**RESULTS:**



Input : Natural Image with Text Output : Natural Image with detected text

**Future Work:**

To overcome the limitation of punctuation marks

* We update the dataset with more labels which consists of images with punctuation marks and train the model.

To overcome the limitation of falsely identifying the non text image

* We propose to add another classifier to the model which differentiates images of text with images with no text.
* We can use SVM classifier and take a dataset consists of two classes i.e positive and negative class.
* Positive class consists of images with text in it.



Figure 24: Image with text

* Negative class consists of images with no text in it.



Figure23:Image with no text

* Then we train the model with this dataset, If any image is given and if the image contains any text in it , the model would return ‘1’ and the image is passed into the CNN. If the image doesn’t contain any text , it returns ‘0’ and is discarded.
* Thus we only feed the model with images with text in it ,so accuracy is increased.

**Conclusion:**

This project presents an enhanced scene text detection method based on MSER which is capable of differentiating the text regions from the natural scene image and can recognize the text from the selected text part using Convulutional Neural Networks(CNN). The MSER detects the intensity differences in the image which is then passed to the next step where contours outline the text area and the convex hull produces numerous bounding boxes around it which are then passed to the NMS function where it supresses the unwanted bounding boxes. Finally, The model is applied on to the selected text region to recognize text. The proposed system exhibits good performance over CHAR74K CHARACTER dataset. Some work is proposed to overcome the limitations of this model.

**References:**

**[1]** S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “Icdar 2003 robust reading competitions,” in IAPR International Conference on Document Analysis and Recognition, 2003.

**[2]** S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J. Jolion, L. Todoran, M. Worring, and X. Lin, “Icdar 2003 robust reading competitions: Entries, results, and future directions,” International Journal on Document Analysis and Recognition, 2005.

**[3]** A. Shahab, F. Shafait, and A. Dengel, “Icdar 2011 robust reading competition challenge 2: Reading text in scene images,” in IAPR International Confernce on Document Analysis and Recognition, 2011.

**[4]** C. Yi, X. Yang, and Y. Tian, “Feature representations for scene text character recognition: A comparative study,” in International Conference on Document Analysis and Recognition, 2013.

**[5]** K. Sheshadri and S. K. Divvala, “Exemplar driven character recognition in the wild,” in British Machine Vision Conference, 2012.

**[6]** Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, November 1998.

**[7]** T. Mikolov, M. Karafiat, L. Burget, J. Cernock ´ y, and S. Khudanpur, ` “Recurrent neural network based language model.” in Interspeech, vol. 2, 2010.

**[8]** T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, “End-to-end text recognition with convolutional neural networks,” in Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012, pp. 3304– 3308.

**[9]** A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” in International Conference on Computer Vision and Pattern Recognition, 2016.

**[10]** K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015.

**[11]** Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, “Multioriented text detection with fully convolutional networks,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

**[12]** M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “Textboxes: A fast text detector with a single deep neural network,” in AAAI, 2017.

**[13]** B. Shi, X. Bai, and S. Belongie, “Detecting oriented text in natural images by linking segments,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

**[14]** 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.

**[15]** H. Cho, M. Sung, and B. Jun, “Canny text detector: Fast and robust scene text localization algorithm,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3566–3573.

**[16]** X. Bai, C. Yao, and W. Liu, “Strokelets: A learned multi-scale mid-level representation for scene text recognition,” IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2789–2802, 2016. [28] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, “Detecting oriented text in natural images by linking segments,” in AAAI Conference on Artificial Intelligence (AAAI), 2016.

**[17]** B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

**[18]** A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connection- ´ ist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 369–376.

**[19]** A. Poznanski and L. Wolf, “Cnn-n-gram for handwriting word recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2305–2314.

**[20]** T. E. de Campos and B. R. Babu, “Character recognition in natural images,” in International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, February 2009

**[21]** Shah P, Karamchandani S, Nadkar T, Gulechha N, Koli K,Lad K. OCR-based chassis-number recognition using artificial neural networks. InVehicular Electronics and Safety (ICVES), 2009 IEEE International Conference on 2009 Nov 11 (pp. 31-34). IEEE.

**[22]** Zhai X, Bensaali F, Sotudeh R. OCR-based neural network for ANPR. In2012 IEEE International Conference onImaging Systems and Techniques Proceedings 2012 Jul 16 (pp. 393-397). IEEE.

**[23]** Shamsher I, Ahmad Z, Orakzai JK, Adnan A. OCR for printed urdu script using feed forward neural network.InProceedings of World Academy of Science, Engineering and Technology 2007 Aug (Vol. 23, pp. 172-175).