Intelligent Scene Text Recognition of Natural Images

*Pooja P. Kohapare[1], Neha Kottawar[2], Ashish Manusmare[3]
[1]*(Electronics and Communication Eng. /Gondwana University, India)*
[2]*(Electronics and Communication Eng. /Gondwana University, India)*
[3]*(Electronics and Communication Eng. /Gondwana University, India)*
Corresponding Author: Pooja P. Kohapare

Abstract : *In the past few years, text in natural scene images has gained potential to be a key feature for content based retrieval. They can be extracted and used in search engines, providing relevant information about the images. Robust and efficient techniques from the document analysis and the vision community were borrowed to solve the challenge of digitizing text in such images in the wild. In this thesis, we address the common challenges towards scene text analysis by proposing novel solutions for the recognition and retrieval settings. We develop end to end pipelines which detect and recognize text, the two core challenges of scene text analysis. For the detection task, we first study and categorize all major publications since 2000 based on their architecture. Broadening the scope of a detection method, we propose a fusion of two complementary styles of detection. The first method evaluates MSER clusters as text or non-text using an ad boost classifier. The method outperforms the other publicly available implementations on standard ICDAR 2011 and MRRC datasets. The second method generates text region proposals using a CNN based text/no text classifier with high recall. We compare the method with other object region proposal algorithms on the ICDAR datasets and analyze our results. Leveraging on the high recall of the proposals, we fuse the two detection methods to obtain a flexible detection scheme.*

Keywords: *Robust reading, character recognition, feature learning, cropped word recognition, part-based tree-structured models (TSMs), posterior probability, word spotting.*

I. Introduction

With rapid growth of camera-based applications readily available on smart phones and portable devices understanding the pictures or a text taken by the devices semantically has gained increasing attention from the computer vision community in recent years. Among all the information contained in the image or text which carries semantic information, could provide valuable cues about the content of the image and thus is very important for human as well as computer to understand the scenes. This paper presents an algorithm for detecting and reading text in city scenes. Text includes stereotypical forms such as street sign, hospital signs, bus numbers, shop signs, house numbers, and billboards. Database of city images were taken in San Francisco partly by normally slighted viewers and partly by blind volunteers who were accompanied by slighted guides using automatic camera setting and little practical knowledge [7], were text was located in image. Any natural scenes located at environment can be used for character recognition. Perceptual content includes colors, shapes, textures, intensities and their temporal changes. While semantic content includes objects, events and their relations [5].

Text contains high level of semantic information as compared to visual information. The importance of digital libraries for information retrieval cannot be denied. The ancient historical books contain invaluable knowledge but it is very time consuming to search the required information in this paper books. Different methods have been devised to facilitate the information search. This includes word spotting, optical character recognition etc. [9]. Although a lot of work is already has done in this field, it still remains an inviting and challenging field of research. . Methods for scene text localization and recognition aim to find all areas in an image that would be considered as text by human mark boundaries of the areas (usually by rectangular bounding boxes) and output for real-world images and video processing (i.e. processing of images / videos taken by standard camera or mobile phone) and reading content of each detected area into a digital text format that can be further processed by a computer. The ICDAR 2003 robust reading challenge published the first data set to highlight the problem of detecting and recognizing scene text. In this benchmark, the organizers identified four subtasks: 1) text localization; 2) robust character recognition; 3) robust word recognition; and 4) robust reading. Since text detection is the premise for recognition, a lot of methods have been proposed to address the problem of detecting and localizing text in scene images and some have reported promising localization performance.

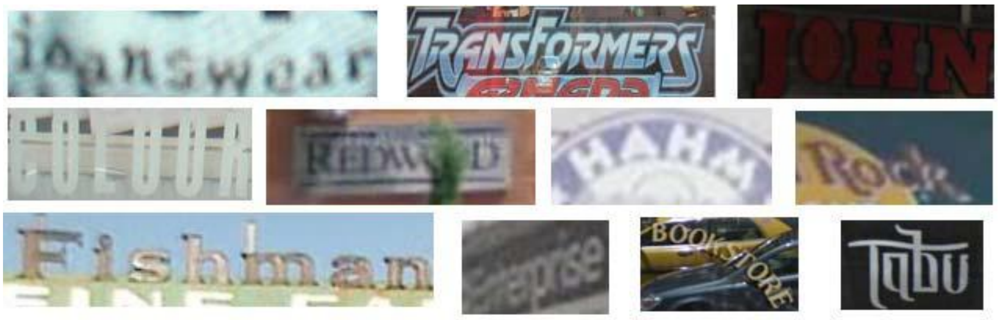


Fig.1 some scene text images from ICDAR 2003 and SVT. The characters in these images have different fonts, shadows, distortions, deformations, low resolutions, and occlusions

Fig.1 shows Detection of text and identification of characters in scene images is a challenging visual recognition problem. As in much of computer vision, the challenges posed by the complexity of these images have been combated with hand designed features and models that incorporate various pieces of high-level prior knowledge. In this paper, we produce results from a system that attempts to learn the necessary features directly from the data as an alternative to using purpose-built, text-specific features or models. Among our results, we achieve performance among the best known on the ICDAR 2003 character recognition dataset. In contrast to more classical OCR problems, where the characters are typically monotone on fixed backgrounds, character recognition in scene images is potentially far more complicated due to the many possible variations in background, lighting, texture and font. As a result, building complete systems for these scenarios requires us to invent representations that account for all of these types of variations. Indeed, significant effort has gone into creating such systems, with top performers integrating dozens of cleverly combined features and processing stages .respectively. But not easily created by hand. Another potential strength of these approaches is that we can easily generate large numbers of features that enable higher performance to be achieved by classification algorithms.

II. Literature Survey

This section describes previously proposed studies that text extracted from any natural scenes around us which depends on their font sizes and thickness of text denoted on any scene, as in [5] paper revealed Y. Song, A. Liu, L. Pang, S. Lin, has mentioned about text images which contain important contents for information indexing and retrieval, automatic annotation and structuring of images. As in [8] L. Neumann and J. Matas, describes regarding end-to-end and real-time scene text localization and recognition method is presented. The real-time performance is achieved by posing the character detection problem as an efficient sequential selection from the set of extremely regions (ERs). The ER detector is robust to blur, illumination, color and texture variation and handles low-contrast text. As paper published [11]. P. Shivakumara, T. Phan, S. Bhowmick, C. Tan, and U. Pal, introduce a novel ring radius transform (RRT) and the concept of medical pixels on character with broken contours in the edge domain for reconstruction. For each pixel, the RRT assign a value which is the distance to the nearest edge pixel. The medical pixels are those which have maximum radius values in their neighborhood.

III. Proposed Methodology

For traditional OCR-based methods as shown in, they focus on the binarization process, which segments text from background, and then the binary image could be segmented into individual characters, which could be recognized by the OCR engine. Various approaches have been proposed to binarize images with low quality or complex background. Gllavata first used a color quantize to determine the color of text and background, and then adopted the modified k-means algorithm to classify pixels into text and background. So it used color-based k-means clustering to segment text from background. . Proposed to integrate local visual information and contextual label information into a conditional random field to segment text from complex background. Proposed to binarize video text images using graph cut algorithm based on the automatic acquired hard constraint seeds. Recently, Shiva kumara introduced a novel ring radius transform and the concept of medial pixels on characters with broken contours in the edge domain for reconstruction to improve the character-recognition rate in video images. Field proposed to use bilateral regression to model smooth color changes across an image region without being corrupted by neighboring image regions and use feedback from a recognition system to choose the best foreground region.

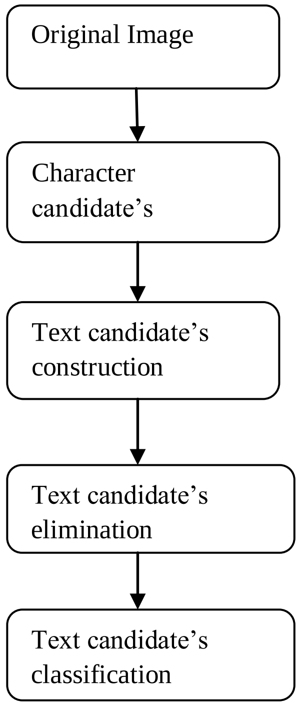

Original Image
Character candidate's
Text candidate's construction
Text candidate's elimination
Text candidate's classification

Fig.2 shows the flowchart of the system with character candidates as MSERs

A contrasting approach for detection also came into the picture in this time period. Focusing on the role of recall in the detection part of an end to end pipeline, the need for an explicit localization of text could be bypassed using a region proposal method. They were presented as an alternative to sliding window strategy for the end to end framework with lesser number of windows generated, relying on a recognition engine to improve the precision of the detection result. In the recent years, several generic object region proposal algorithms were proposed in the field of object recognition which can be considered for text region proposal. Super-pixel based methods like were proposed, where the authors group the super-pixels hierarchically and use several diversification strategies comprising of complementary similarity measures to get the proposals. Region proposal was also presented in a Bayesian framework in which combined several cues like location, size, color contrast, edge density, etc. to extract set of super pixels which appear different from their surroundings and form a closed boundary. Multi-scale hierarchical segmentation with grouping strategies to combine multi-scale regions into object candidates by exploring combinatorial space was suggested. An edge based method was proposed in where box abjectness scores were computed based on the number of edges inside the box minus the number of contour members overlapping the box boundary.

IV. Result

Algorithm used in this system is **stroke width transform** algorithm. In this section I will describe the Stroke Width Transform algorithm as it is presented in [1], with several additions and enhancements. These additions will be discussed in further extent in the next section 'The Application'. The algorithm receives an RGB image and returns an image of the same size, where the regions of suspected text are marked. It has 3 major steps: the stroke width transform, grouping the pixels into letter candidates based on their stroke width, and finally, grouping letter candidates into regions of text. A method is presented for adaptive document image binarization, where the page is considered as a collection of subcomponents such as text, background and picture. Using characteristics analysis, two new algorithms are applied to determine a local threshold for each pixel. An algorithm based on soft decision control is used for thresholding background and picture regions. An approach utilizing local mean and variance of gray values is applied to textual regions' resizable square rim in the viewfinder of the mobile camera, referred to here as a 'focus', is the interface used to help the user indicate the target text. The binarization is targeted at creating layers of binary images for processing by OCR engines.

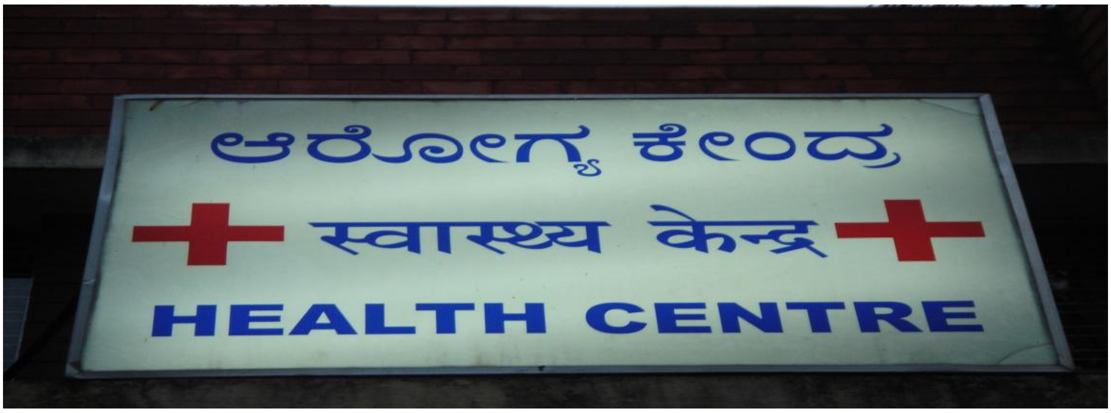
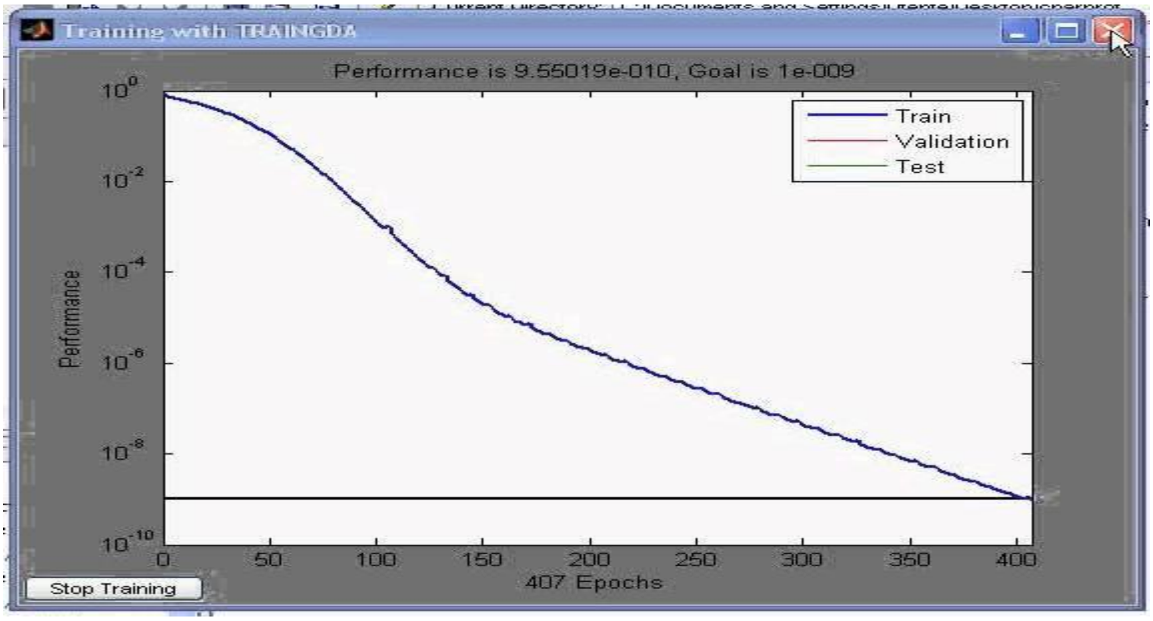


Fig.3 Recoginition scene text image

A method is presented for adaptive document image binarization, where the page is considered as a collection of subcomponents such as text, background and picture. Using characteristics analysis, two new algorithms are applied to determine a local threshold for each pixel. An algorithm based on soft decision control is used for thresholding background and picture regions. An approach utilizing local mean and variance of gray values is applied to textual regions' resizable square rim in the viewfinder of the mobile camera, referred to here as a 'focus', is the interface used to help the user indicate the target text. The binarization is targeted at creating layers of binary images for processing by OCR engines.


Training with TRAINGDA

Performance is 9.55019e-010, Goal is 1e-009

Fig.4 The graph of iterations based on Neutral network trainning

Image which get recognized are get converted into audio and this images get extracted in canny edge detector. we give detailed evaluation of the proposed character detection and word-recognition method. Since character detection and recognition are the basis for word recognition, we first evaluate the detection-based character recognition method. We also compare the proposed character detection method with state-of-the-art sliding window-based detection methods. In addition, we compare the proposed TSM with deformable part-based model (DPM), which treats part locations as latent variables. Since we choose training samples for all the structures from Chars74k data set, all the remaining images except those chosen as training samples comprise the test set for chars74k data set. While for ICDAR03-CH data set, since we do not use the training set to learn the model parameters, we evaluate the performance on both the training and test sets. The results show that the proposed TSM outperforms HOG+KNN by more than 10% on ICDAR03-CH data set,which reduces the error rate by more than 30%, when only considering the first candidate. When we consider the second best result, we find that the performance of TSM improves more quickly with an increase of about 8% whereas the recognition rate of HOG+KNN only increases 3%–5%. The high possibility of achieving better wordrecognition result if we incorporate linguistic knowledge to deal with ambiguities between similar structures. When we consider the recognition rates of the first five candidates, the result is quite encouraging, reaching 86.38%, 91.88%, and 89.74% on Chars74k, ICDAR03-CH-Train, and ICDAR03- CH-Test, respectively.

By this mechanism, we label each pixel with a score according to whether that pixel is part of a block of text. These scores are then threshold to yield binary decisions at each pixel. By varying the threshold and using the ICDAR bounding boxes as per-pixel labels, we sweep out a precision-recall curve for the detector and report the area under this curve (AUC) as our final performance measure. Specifically, for detection and character recognition, we trained our classifiers with increasing numbers of learned features and in each case evaluated the results on the ICDAR 2003 test sets for text detection and character recognition.

VI. Conclusion

For the text detection challenge, we pursue detection via segmentation approach using hierarchical clustering. We find MSERs on the scene image which is clustered using the single linkage clustering method, thus generating several overlapping text candidates. The candidates are classified using a text/non-text adaboost classifier and the positively classified ones are used to create a binarized image consisting of text pixels as foreground. We also implement a text region proposal method which generates several possible word level regions on the scene image with high recall. The method relies on a patch level text/non-text CNN classifier which generates a score map with per pixel probability of text occurrence. The score map is threshold and components are combined using RLSA to generate the proposals.

Acknowledgements

This research was supported by publisher of this paper. We thank our colleagues from Ballarpur institute of technology who provided expertise that greatly assisted the research. Also, for sharing their pearls of wisdom with us during the course of this paper.

References

[1] K. Wang and S. Belongie, “Word spotting in the wild,” in Proc. 11th ECCV, Sep. 2010, pp. 591–604.

[2] A. Mishra, K. Alahari, and C. V. Jawahar, “Scene text recognition usinghigher order language priors,” in Proc.23rd BMVC, 2012, pp. 1–11.

[3]J. Gllavata, R. Ewerth, T. Stefi, and B. Freisleben, “Unsupervised textsegmentation using color and wavelet features,” in Image and VideoRetrieval. New York, NY, USA: Springer-Verlag, 2004..

[4] Y. Song, A. Liu, L. Pang, S. Lin, Y. Zhang, and S. Tang, “A novelimage text extraction method based on K-means clustering,” in Proc.7th IEEE/ACIS ICIS, May 2008, pp. 185–190.

[5] B. Epshtein, E. Ofek, and Y. Wexler, “Detecting text in naturalscenes with stroke width transform,” in Proc.IEEE CVPR, Jun. 2010,pp. 2963–2970.

[6] X. Chen and A. Yuille, “Detecting and reading text in natural scenes,”in Proc. IEEE CVPR, vol. 2. Jul. 2004, pp. 366–373.

[7] L. Neumann and J. Matas, “Real-time scene text localization andrecognition,” in Proc. IEEE CVPR, Dec. 2012, pp. 3538–3545.

[8] W. Niblack, An Introduction to Digital Image Processing. Mundelein,IL, USA: Strandberg, 1985

[9] A. Newell and L. Griffin, “Multistage histogram of oriented gradientdescriptors for robust character recognition,” in Proc. IEEE ICDAR,Sep. 2011, pp. 1085–1089.

[10] P. Shivakumara, T. Phan, S. Bhowmick, C. Tan, and U. Pal, “A novel ringradius transform for video character reconstruction,” Pattern Recognition.,vol. 46, no. 1, pp. 131–140, 2012.