ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 1, January 2017

Review on Textual Content Detection Techniques

Durgesh Kumar Gupta

Department of Computer Science and Engineering, Galgotias University, Yamuna Expressway Greater Noida, Uttar Pradesh, India

Email Id: durgeshgup@gmail.com

ABSTRACT: Text recognition attempts to interpret the text in the appropriate structure from handwritten hard copy papers. The text recognition technique involves several phases, pre-processing, separation, selection of attributes, classification, post-processing. Preprocessing is required for the critical procedure on the input object, such as binarizing gray image size into a discrete image, which eliminates noise from an image segmentation process for the segment, the provided image into line by line and segment every character from the segmented line. Future extraction calculates the characteristics of a character. The belongings of future extraction are calculated. A text classification includes and compares the details. In today's office, the classroom, etc., plays a crucial role in these phases, and in that paper, various text-related applications for imaging are discussed, as are their corresponding strengths, drawbacks and implementations. This analysis analyzes elementary mechanisms in this area in a quantitative way. Segmentation contains the information and will the comparison. Today, it plays a crucial role within the workplace, university, etc. necessary approaches wont to undergo these stages and their corresponding advantages, disadvantages, and application are presented during this article, numerous text-related applications for imagery also is presented over here. This review performs a comparative analysis of elementary processes during this field.

KEYWORDS*:* Text Detection, Classification, Preprocessing, Segmentation, Object Detection.

INTRODUCTION

In recent years, text detection and acknowledgment have been a very major drawback. Advances in image processing and machine learning have culminated in this trend, similarly, as a rise among the frameworks focused on text detection by means of image analysis techniques. The text recognition of videos is further developed as web content is also being noticed, extensive text detection and the acknowledgment of natural scenes has been carried out. There are also various methods of optical character recognition on the market. However, the question of text recognition is not fully resolved. Texts are still very challenging to section and obtain from natural sceneries. Text identification is used in an official role where significant knowledge is needed to type in real-life implementations, such as postal offices, banks, college establishments and so on anywhere details from text-written pictures should be obtained. People pretend to scan a report and connect directly the document text in.txt or doc format. The document should be available. Different ways of text detection from the image and video are proposed for the last decade. Most of them are divided into the subsequent stages, spatial text extraction, temporal text detection, image binarization image segmentation, and character recognition. The efficiency of the entire system relies on the precise position of text in images, much less research on the timely identification of text in videos has been achieved, while the only static text is generally considered. Several methods for binarizing the text picture have also been suggested, although most of the investigators are using innovative methods from the classical research field for document analysis.

Last but not least, the creation of the correct validation protocols that are required to refine the algorithms in a similar manner to parallels in the literature is an extremely important factor that was not studied adequately. In this portion, this paper describes techniques for different phases of text detection from images and video frames contained in the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 1, January 2017

literature. The text-areas are extracted under the assumption that the image background region and the object-areas are smaller than the text-areas. These methods are however not very successful in extracting large font sized texts[1]–[11].

II.RELATED WORK

The study of a system of text abstraction is not as straightforward as it may seems. This approaches the traditional problem of study of objects, but also has its own issues. Many studies use basic box-based or region-based approaches for testing, while only a few works address the specific empirical issue. The average text extraction test for a container, a box-based method named FDA. The evaluation ways of this type have supported the mapping between ground truth and objects detected. Particularly for the text extraction problem, text lines are thought-about to be the objects wherever a text line is typically defined as aligned series of characters with a small intermediate distance relative to their height. The study of a text emergent phenomenon system is not as simple as it might seem. This addresses, but also has its own problems, the standard problem of object study. Several studies are testing using basic box-or regional strategies, whereas only a few works deal with the precise empirical problem. The median container text removal test, called the FDA.

- Text Detection:

Different levels of text abstraction and identification can be found in this segment. Their position and significance is represented by text identification and location, image classification, image segmentation and text recognition. Many methods are clarified in this segment for these phases. Extraction of text detects the existence in the input picture of the image text, while the place of the texts and types squads of text regions by removing most of the backdrop. Methods for text removal and location are carried out using the analysis for the connected components or regions.

The connected component analysis methodology forms a graph of connected points supported color or edge features from the binarized image. Region-based strategies divide images into small regions using windows and search these regions for the presence of text using texture or morphological operations since text and non-text regions have totally different textual properties. Unsupervised machine learning techniques are employed by, on little 8×8 grayscale patches of image for feature extraction learning. It uses 32×32 pixels of image for feature detection, text detection training, and character classifier training.

- Classification:

After text extraction stage output could contain non-text regions together with text regions as false positives. The text classification stage verifies text regions and eliminates non-text regions using image classification algorithms. This stage may be referred to as verification, classification algorithms are either supervised or unsupervised. Supervised machine learning algorithms are aware of the properties of text detection like color, size, texture, etc. before text classification. Unsupervised machine learning algorithms don't have prior data regarding text features. Supervised classification algorithms would like training before classification. These algorithms undergo training to be able to detect features of the text to be classified and use these features in the classification section. Supervised machine learning algorithm constraints on edge area additionally as area, image height and dimension constraints on block obtained within the text detection stage are employed in for text classification.

Text classification is done using the features detection in the previous step, which corresponds to each character recognize. These text features are analyzed using the set of rules and labeled as belonging to different classes. This text classification is generalized such that it works for a single font type (Figure 1).

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 1, January 2017

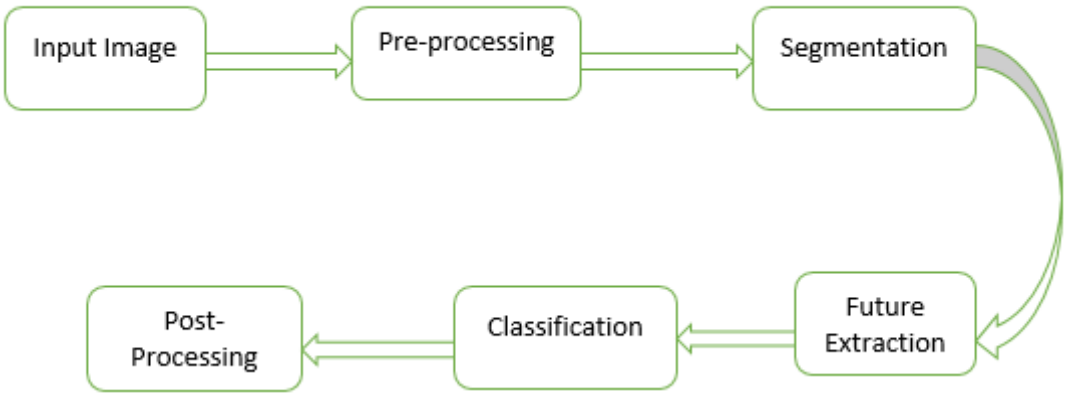

Figure 1: Text Detection and Recognition

After text classification, algorithm will check the classified data with the database where we have already saves the classification of text, text numbers. Then algorithm will display the result according to the comparison. Here this algorithm will check the pattern of inputted character with database. Unsupervised text classification algorithm doesn't undergo the training phase. The text extract features throughout the image classification phase-only not like supervised classification and that they use features detection within the previous text classification for the next one. This can be similar to an adaptive learning algorithm. Wavelet transform which provides sequential approximation through low-pass filter and details of edges and different features from the high-pass filter is used in.

- Segmentation:

The image segmentation method is employed to separate text detection from the background and to detect bounded text from an image. Integrated ways that focus on word matching/recognition usually mix or replace advanced segmentation stage with recognition stage but stepwise strategies undergo segmentation to get exactly extracted characters that are fed to the recognition stage. Binarization, text segmentation are few of the segmentation algorithms studied during this article. Binarization converts color or gray-scale images into black and white images. To realize good segmentation result no matter dark or bright text or background, adaptive thresholding for binarization uses the k-means clustering algorithm for binarization. It uses k=4 and k=5 as cluster parameters and text classifies the binarized image into texts using probabilistic models.

The connected component algorithm is employed in computer vision to detect connected regions in binary digital images, though color images and data with higher-dimensionality may also be processed. When integrated into an image recognition system or human-computer interface, connected component labeling will operate on a spread of data technique[12]–[16].

III.TEXT RECOGNITION ALGORITHMS

The text has played a significant role in human life, far from prehistoric times, one of the most successful developments of civilization. The rich, precise textual knowledge is very valuable in many specific vision applications and thus text identification, recognition and active area of research topics in machine perception and record processing are becoming essential in realistic scenes. Particularly during the last few years, research efforts have improved and considerable progress has been made in these regions, but various challenges also exist (e.g. vibration, haze, distortion, occlusion and variance. This study has three purposes: 1) presenting up-to-date science, 2) defining state-of-the-art technologies, and 3) predicting possible future research paths. In fact, this paper includes comprehensive connections to available public tools, including databases for comparisons, source codes and online tutorials. In short, this review can represent as more of a reference in the fields of text identification and appreciation for research scientists. The algorithm is described below:

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 1, January 2017

Algorithm 1: TextProposals

Input: RGB image I

Output: Set of bounding box proposals B and their scores

Initialize $B = \emptyset$;

Initialize F = list of similarity features;

Extract input channels and scales $C = \{c_1, \ldots, c_n\}$ from I;

foreach *channel* $c \in C$ **do**

  Obtain MSER [31] regions $R_c = \{r_1, \ldots, r_m\}$ from c;

  Obtain coordinates of regions' centers $\{x_1, \ldots, x_m\}$, $\{y_1, \ldots, y_m\}$;

  foreach *feature* $f \in F$ **do**

    Calculate feature set $S_{fc} = \{s_1, \ldots, s_m\}$ with $s_i = (f(r_i), x_i, y_i)$;

    Build dendrogram D_{fc} applying SLC clustering over S_{fc};

    Extract bounding boxes B_{fc} and scores for each *node* $\in D_{fc}$;

    $B = B \cup B_{fc}$;

  end

end

Localizing scene text is an extremely important step in the acknowledgment of scene text. The major problems, including different sizes, textures, irregular orientations, a wide range of colors, patterns, occlusion and local and global environment differences, vary from the common identification of artifacts. The public benchmarking quantitative test results ICDAR2015, ICDAR2017-MLT, ICDAR2019-MLT and Total-TXT prove how efficient and robust our methodology has been compared to recent strategies presented (Figure 2).

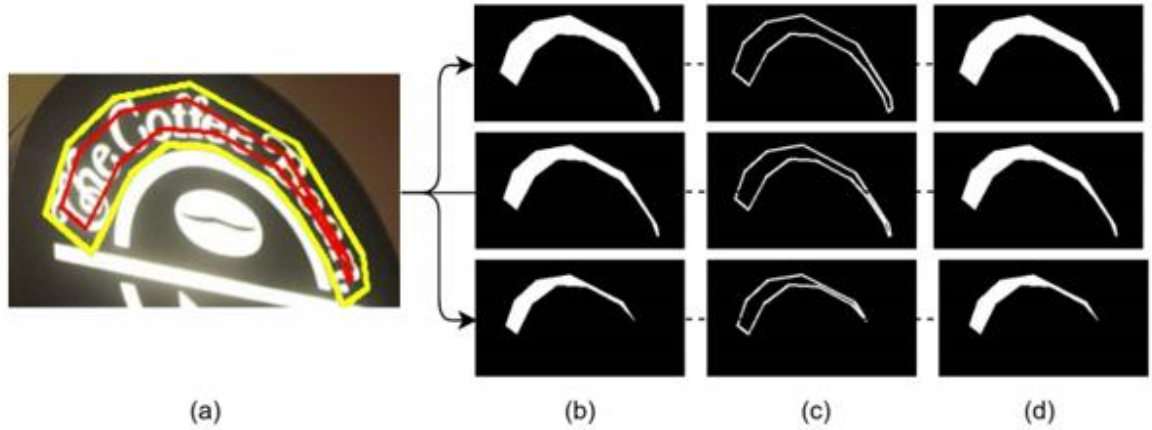

Figure 2: Proposed Text Representation

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 1, January 2017

In this work, they also implemented a mosaic data raise technology in which several picture patches are randomly combined with a new picture. It can improve the percentage of training data and the reliability of the algorithm. The input and output image samples are presented in the Figure 3.

Figure 3: Result of a Mosaic Data Augmentation Example from Four Input Images (Best Viewed in Color)

IV.CONCLUSION

Text is raised as a high-level semantic zed clear operator. This distinguishes text from many other common visual markers such as highlighter, color and texture by this unique property. In this paper, introduce a text detection using image processing that supported a scalable feature learning algorithm and applied it to the pictures of text in natural scenes Consequently, texts in natural action sequences have now become essential and vibrant fields of machine learning detect and recognize. The study seeks to follow up on the progress made in the identification and recognition of text on the site and to provide future researchers with complete access to useful resources in these fields. A systematic literature review on text identification and recognition was provided. Its innovative thoughts and activities have been implemented and public data sets and test guidelines have been identified in these fields. Interest in research areas of fast development and persistent growth can be seen from the preceding paragraph of text scene identification and recognition.

REFERENCES

[1]Q. Ye and D. Doermann, "Text Detection and Recognition in Imagery: A Survey," IEEE Trans. Pattern Anal. Mach. Intell., 2015.
[2]A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic Data for Text Localisation in Natural Images," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016.
[3]H. Cho, M. Sung, and B. Jun, "Canny Text Detector: Fast and Robust Scene Text Localization Algorithm," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016.
[4]Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented Text Detection with Fully Convolutional Networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016.
[5]B. Kessler, G. Numberg, and H. Schütze, "Automatic detection of text genre," 1997.
[6]N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," Decis. Support Syst., 2010.
[7]C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, 2010.
[8]X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast Oriented Text Spotting with a Unified Network," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.
[9]G. D. Miner, J. Elder, and R. A. Nisbet, Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. 2012.
[10]K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in Proceedings of the IEEE International Conference on Computer Vision, 2011.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijareeie.com

Vol. 6, Issue 1, January 2017

[11]P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.
[12]P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018.
[13]J. Ma et al., "Arbitrary-oriented scene text detection via rotation proposals," IEEE Trans. Multimed., 2018.
[14]K. Kukich, "Techniques for Automatically Correcting Words in Text," ACM Comput. Surv., 1992.
[15]X. Zhou et al., "EAST: An efficient and accurate scene text detector," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017.
[16]M. Liao, Z. Zhu, B. Shi, G. S. Xia, and X. Bai, "Rotation-Sensitive Regression for Oriented Scene Text Detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.
•V.M.Prabhakaran, Prof.S.Balamurugan , S.Charanyaa, "Data Flow Modelling for Effective Protection of Electronic Health Records (EHRs) in Cloud", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2015
•R. Santhya, S. Latha, S. Balamurugan and S. Charanyaa, "Further investigations on strategies developed for efficient discovery of matching dependencies" International Journal of Innovative Research in Science, Engineering and Technology Vol. 4, Issue 1, January 2015