ISSN 2319 – 7595

Volume 8, No.2, March - April 2019

International Journal of Information Systems and Computer Sciences

Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs01822019.pdf

https://doi.org/10.30534/ijiscs/2019/01822019

Comparative Study on Text Detection and Recognition from Lecture Videos

Ashima Godha[1] Rahul Sharma[2]
[1]M.Tech (CTA)
RKDF School of Engineering, Indore
ashimagodha@maill.com
[2]Assistant Professor (CSE)
RKDF School of Engineering, Indore
Sharma.rahul5656@gmail.com

ABSTRACT

Lecture videos are rich with textual information and to be able to understand the text is quite useful for larger video understanding/analysis applications. Textual information extracted from these sources can be used for automatic image and video indexing, and image structuring. But, due to variations in text style, size, alignment of text, as well as orientation of text and low contrast of the image and complex background make challenging the extraction of text. From the past recent years, many methods for extraction of text are proposed. This paper provides with analysis, comparison of performance of various methods used for extraction of text information from images. It summarizes various methods for text extraction and various factors affecting the performance of these methods.

Key words: Word recognition, Lecture video, Text extraction, Text localization, Text segmentation, Connected component · Edge-based approach.

1. INTRODUCTION

Text in images comprises of valuable statistics and is exploited in many applications that uses image and video applications, such as content-based web image search, video information retrieval, and mobile based text analysis and text recognition [1-5]. Due to composite background, and deviations of font, size, color and orientation, text in natural scene images has to be vigorously detected before being recognized and regained.

With increasing interest in e-learning in the form of OpenCourseWare (OCW) lectures and Massive Open Online Courses (MOOCs), freely available lecture videos are abundant. Understanding lecture videos is critical for educational research, particularly in the context of MOOCs which has become synonymous with distance learning. For example a lecture video can be analyzed to understand a teacher's engagement with the learners, on which frames does the viewers pay more attention [1] etc. The figures, images and text in lecture videos are vital cues for understanding any lecture video. Text is present almost everywhere in a lecture video; particularly in lectures on Science, Mathematics and Engineering. Text alone could be used for a variety of tasks like keyword generation, video indexing and enabling search and extracting class notes [2]–[5].

Text in lecture videos comprise of handwritten text written on a blackboard or a paper, text written using a stylus on a tablet and displayed on a screen or font rendered text appearing in presentation slides (digital text). Lectures are recorded using one or more cameras, and the camera(s) are typically positioned to directly face the blackboard or the presentation slides. Usually text recognition from presentation slides is less challenging as the text is more legible, there is little variation in style and there is more contrast. At the same time text on blackboard is handwritten and not very legible due to poor lighting, smaller size or poor contrast. On blackboard or on paper the lecturer may write over figures and equations, and this makes the scene cluttered, making it harder to detect the text. Figure 1 shows few samples.

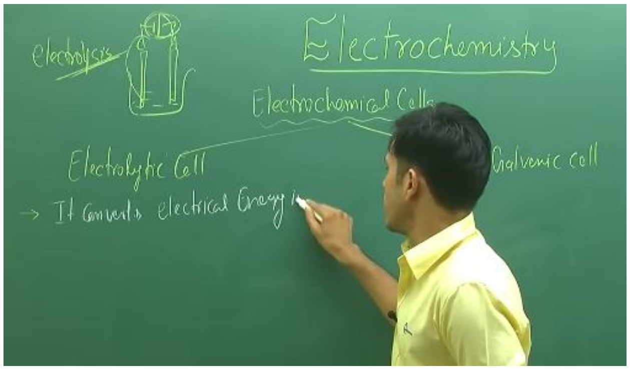

Figure 1: Visualization of text localization and recognition results on frames from the Lecture Videos

2. STEPS OF TEXT EXTRACTION

The text extraction problem is divided into following steps [1]:

i. Text detection

ii. Text localization

iii. Text tracking

iv. Text extraction and recognition

A. Text Detection

As there is no prior information that the images contain text or not so, text detection step determines whether text is present in given image or not. It can be done by making use of pixel intensity. It is assumed that text has higher pixel intensity than background pixels so, pixels with value less than predefined threshold value and having significant color difference from neighboring pixel are considered as text pixel. In videos, text is detected by using scene change information between two consecutive video frames.

B. Text Localization

It locates the position of text in given image. If the shape of text region is rectangular then text can be efficiently located but text can be aligned in any shape, i.e., rectangular or circular so, it is difficult to locate text. Text is located by using similarity in various text features like color, size, shape, intensity, and distance between two text pixels. Color histogram is used to figure out similarity in color of text pixels.

C. Text Tracking

If the text is not located in text localization step then text tracking is helpful in locating that text. Text tracking maintains the integrity of position across adjacent frame and reduces the processing time for text localization. This step is used to verify text localization results. Text localization, detection, and extraction are often interchangeably used.

D. Text Extraction and Recognition

It describes how to separate text from images. This can be done by separating text pixels from other non-text pixels. By using various text features like similar font, orientation and stroke width, text can be segmented from image. OCR is the one of common method to extract and recognize text. It can easily recognize text in text document but it is unable to recognize text information in images efficiently because of presence of noise and distortion in images. In order to extract the text, different methods are used like region-based- and texture-based method. Region-based methods involve connected component and edge-based methods.

3. APPLICATIONS OF TEXT EXTRACTION

Extraction of text from images can be utilized in many applications. Text extraction proves helpful in indexing of images and videos, structuring of images and in classification of images and videos. Following are some real-time applications of text extraction from images.

A. Applications for Disable People

Text in images or in natural scenes provides significant information that is utilized in text to speech devices. These devices assist visually impaired people in understanding grocery product, signs and pharmaceutical labels, and currency, instructions of ATM and in path navigation.

B. Navigating Systems

Text on sign board and natural scenes are used in navigation systems embed in robots and automatic geolocators to navigate their path.

C. Texts in Web Images

Relevant information can be provided by extraction of text from images on Internet. Indexing of images is performed using this information which will lead to efficient web mining.

D. Industrial Automation

Industrial automation related to large number of applications that used the recognized text on packages, containers, houses, and maps.

4. CHALLENGES IN TEXT EXTRACTION

Text extraction process faces many challenges in segmenting the text from images. Some of these challenges are analyzed as follows:

- Text in images has different font, more than one color, and nonuniform size.
- Natural scene consists of complex background which contains a large number of objects like buildings, signs, bricks, grasses, fences, poles, etc., which resemble text characters causing error and confusion.
- Blurred image can occur with non-focused camera and with moving objects.
- Uneven lightening is very common problem during capturing of image.
- OCR is not used in text recognition in natural scene images as these images contain complex background containing many objects like bricks, fences, signs, etc. This complex background makes OCR to predict wrong word as it considers any object as character if that character resembles any character or alphabet.

5. TECHNIQUES IN TEXT EXTRACTION

The several techniques of text extraction are as follow:

A. Sliding window based method

Region-based technique uses the assets of the color or gray scale in the text region or their alterations to the corresponding properties of the contextual. They are recognized on the condition that there is very small aberration of color restricted by text and this color is sufficiently separate from text's immediate background [30]. Text can be achieved by thresholding the image at intensity level amongst the text color and that of its instant background. This technique is not vigorous to complex background. This method is further sub-divided as connected component (CC) and edge based.

B. CC based approaches

CC-based approaches [23, 25] use a bottom-up approach by grouping small components into successively larger components until all regions are identified in the image. A geometrical analysis is required to merge the text components using the spatial arrangement of those components so as to filter out non-text components and the boundaries of the text regions are marked. This method locate locates text quickly but fails for complex background.

C. Edge based approaches

Edges are a reliable feature of text regardless of color/intensity, layout, orientations, etc. Edge based method [31] is focused on high contrast between the text and the background. The three distinguishing characteristics of text embedded in images that can be used for detecting text are edge strength, density and the orientation variance. This method is not robust for handling large size text.

Texture based Method: This method uses the fact that text in images have discrete textural properties [25] that distinguish them from the background. This method is able to detect the text in the complex background. The only drawback of this method is large computational complexity in texture classification stage.

D. Morphological based Method

Mathematical morphology is a topological and geometrical based method for image analysis.

Morphological feature extraction techniques have been efficiently applied to character recognition and document analysis. It is used to extract important text contrast features from the processed images [25]. These features are invariant against various geometrical image changes like translation, rotation, and scaling. Even after the lightning condition or text color is changed, the feature still can be maintained. This method works robustly under different image alterations.

6. RELATED WORK

Yin et al. [2] proposed a Maximally Stable Extremal Regions (MSERs) based method to extract text form images. Extremal region can be viewed as a connected component in an image whose pixels can have either lower or higher intensity than its outer boundary pixels. In this method, first character candidates, i.e., pixels containing text are extracted based on their difference in variance. Extremal regions are extracted in the form of a rooted tree for the whole image. Second, character candidates are merged into text candidates by the single-link clustering algorithm. Distance metric learning algorithm can automatically learn about distance weights, i.e., distance between two text pixels and clustering threshold and these are used in clustering character candidates. Third, non-text candidates are eliminated by using character classifier. The width, height, smoothness, and aspect ratio of Text region are used by character classifiers.

Le et al. [3], present a learning-based approach which involves three steps. First, in preprocessing the image is binarized using Otsu binarization and the connected component are extracted based on various features like elongation, solidity, height, and width of the connected component, Hue moment which describes the shape of the connected component, stroke width of connected components for discriminating text from non-text. These features provide shape and location information of connected components. Then Adaboosting Decision Trees is used to label these connected components into non-text or text component. Then post processing to correct some connected components which are labeled incorrectly by the classifier.

Vidyarthi et al. [4] purposed a method in which first colored image is converted into gray-level image and histogram is drawn. Otsu method uses this histogram to find global threshold value. This value is further used to find out connected components. The components which are close to each other are merged to form one component. The height histogram is constructed by using the heights of the bounding boxes. Variance is selected for verification of the text and non-text. Finally, filtered components are verified a text on text.

Zhong et al. [8] presented a system which localized text in color images. It roughly localized texts by using horizontal spatial variance and then to find the text, color segmentation was conducted within the localized regions.

Zhong et al. [10] purposed system to localize text in the color images, using DCT. Image patches of high horizontal spatial intensity variation are detected as text components, morphological operations are applied to those patches and thresholding spectrum energy is used to verify the regions.

Kumar et al. [11] discussed about several text extraction techniques grounded on edge detection, connected component analysis, morphological operators, wavelet transform, texture features, neural network etc. and also contributes comparative analysis of different technique which provides efficient performance.

Kim et al. [12] introduce a system to localize the text by using SVMs and texture templates. It classifies the pixels into positives that are connected into text regions by using a mean shift algorithm.

Yu et al. [13] proposed a system to recognize text through candidate edge recombination. First it separates text edges by using canny operator and then divides this edged image into small edge segments. Then the neighbor edge segment having similar stroke width and color are merged to form one candidate boundary. These candidate boundaries are combined into text chains by using character features and chain based features. Character features include stroke width of characters, aspect ratio, orientation difference of text whereas chain based features include histogram of gradient (HOG) which can be used to describe the boundary of object, structure similarity as it is assumed that text have similar structure.

Yi et al. [14] proposed a system in which image first operates on canny operator; then the image is partitioned in order to obtain character candidate components by using two methods gradient-based method and color-based method. In gradient-based method, two pixels in opposite direction and having same local gradient magnitude are coupled together by connecting path. Candidate character component are then extracted by computing the distribution of gradient magnitudes at pixels of the connecting path. In color-based method, color histogram is calculated in order to capture the dominant colors around edge pixels. Then K-mean clustering is used to find character candidate components. These are grouped together to detect text strings based on character features like distances between two neighboring characters, character sizes, and character alignment.

Lu et al. [15] introduced a system for extraction of text from the scene images. The image first operates on canny operator to differentiate text edges from the background and then based on three text-specific features like image contrast, horizontal or vertical direction stroke width text is extracted. These text-specific properties are evaluated at images of various scales because it improves various types of image degradation, because some text edges lost at one image scale, due to the complex image background. The text candidate text boundaries are detected by using Global thresholding. For every detected candidate text boundary, one or more candidate characters are determined by using adaptive threshold that can be calculated based on the neighbor image pixels. Support vector regression model which use bag of word (BOW) is used to identify true characters and words.

Khodadadi et al. [17] introduced a method for extraction of text from images based on stroke filters and color histogram. First, stroke filter is applied to the input image and then local and global threshold values are calculated on stroke filter output. The image is then divided into text blocks by making use of horizontal and vertical projections of binary pixels. The text blocks with overlap or close blocks are merged to form a new text block. It is assumed that the text in every block has the same color, so histogram of color channels are used for extraction of the text characters in the candidate blocks for the text and background areas.

Kumar et al. [18] suggested an algorithm that consists of three stages: first, images are converted into an edge map by using line edge detector which uses vertical and horizontal masks. Second, for identification of the text regions in images, the vertical and the horizontal projection profiles are analyzed and the edged image is then converted into the localized image. After localization, in order to obtain text area segmentation process is done on the localized image by using median filters.

Scene Text Recognition

Earlier approaches [19] to scene text understanding generally worked at the character or individual glyph level (bottom-up approaches). Characters from the detected text regions are segmented out and are fed to a classifier which classifies the character patches into the characters in the language. More recent models follows segmentation-free approaches where the words could be recognized without the need for sub-word segmentation. Such models generally use a seq2seq framework built on Recurrent Neural Networks (RNN). Segmentation-free approaches to word transcription along with use of deep features derived from Convolutioanl Neural Networks (CNN) helped to achieve state-of-the art results for scene text recognition.

Handwritten Recognition

Similar to scene text recognition, newer methods in HWR also uses seq2seq formulation [20], [21] which typically uses an underlying CNN-RNN hybrid network for feature extraction and prediction. There are also methods which uses multi-dimensional RNNs (MDRNNs) [22] instead of regular, uni-dimensional RNNs. Given the limited performance of unconstrained prediction, most of the methods in this space, use either a lexicon [21] or language model [22] for arriving at the final output string.

Word Spotting

In the domain of word spotting, the key challenge lies in finding an efficient holistic representation for word images. Most of the recent works use deep neural networks for learning the features. In [23], the author uses the features from the penultimate layer of a deep CNN network, while [24] learns the features by embedding a word image into different attributes spaces such as PHOC, semantic attributes (ngrams, word2vec) etc. These embedding gives a unified representation of both text and its corresponding images and are invariant to different styles and degradations.

Lecture Videos

Text in lecture videos has largely been unexplored, except for few isolated works. One of the early works in this space detect and recognize text in presentation slides to synchronize the slides with the lecture videos [3]. The text detection is based on edge detection and geometry based algorithms and a commercial OCR is used for recognition. Video indexing and keyword search is made possible by text recongition in [4]. Off-the-shelf OCR systems are used for the same. In another work both Automatic Speech Recognition (ASR) and OCR are used to generate keywords for lecture videos [25].

7. CONCLUSION

Detecting and recognizing texts in natural scenes is becoming important in research areas in computer vision as precise and rich information embedded in text that can help in a large number of real-world applications. This paper provides analysis and comparison of performance of various methods used for extraction of text information from images. Text extraction involves text detection, localization of text, tracking of text, extraction of text, enhancement, and text recognition from a given image. Text detection discovers whether text is present in a given image or not. Normally text detection is applied for a sequence of images. Text localization localizes the text within the image and bounding boxes are generated around the text. If the text is not located in text localization step, then text tracking is helpful in locating that text. Text extraction refers to separate of text from images. In order to extract the text, different methods are used like region-based- and texture-based method. Region-based methods involve connected component and edge-based methods. Connected component based method gives poor performance for merged characters or when the characters are not completely separated from the image background. The texture-based methodology has an inability to recognize the characters that reach below the baseline or above other characters, and this ends up in segmenting a character into two components. Edge-based method also makes false prediction when edge of any object in background of image resembles any character. Text extraction from images can be proved useful information for content-based application. In this paper, a study on text detection and recognition on the Lecture Video, using existing state-of-the art methods for scene text and handwritten text is performed.

REFERENCES

[1] Keechul Jung, Kwang In Kim, Anil K. Jain "Text information extraction in images and video: a survey", Elsevier, Pattern Recognition 37 (2004). https://doi.org/10.1016/j.patcog.2003.10.012

[2] Xu-Cheng Yin, Member, IEEE, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao "Robust Text Detection in Natural Scene Images" IEEE transactions on pattern analysis and machine intelligence, VOL. 36, NO. 5, (2014). https://doi.org/10.1109/TPAMI.2013.182

[3] Viet Phuong Le, Nibal Nayef, Muriel Visani, Jean-Marc Ogier and Cao De Trant "Text and Non-text Segmentation based on Connected Component Features" IEEE, 13th International Conference on Document Analysis and Recognition (ICDAR), (2015).

[4] Ankit Vidyarthi, Namita Mittal, Ankita Kansal, "Text and Non-Text Region Identification Using Texture and Connected Components", International Conference on Signal Propagation and Computer Technology (ICSPCT), IEEE (2014). https://doi.org/10.1109/ICSPCT.2014.6884904

[5] Yingying Zhu, Cong Yao, Xiang Bai "Scene text detection and recognition: recent advances and future trends" Front. Comput. Sci., (2016).

[6] Qixiang Ye, and David Doermann, "Text Detection and Recognition in Imagery: A Survey" IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 7, (2015). https://doi.org/10.1109/TPAMI.2014.2366765

[7] N. Senthilkumaran and R. Rajesh, "Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, (2009). https://doi.org/10.1109/ARTCom.2009.219

[8] Zhong Y, Karu K, Jain A K. "Locating text in complex color images." in Proceedings of the 3rd IEEE Conference on Document Analysis and Recognition, pp-146–149, IEEE (1995).

[9] Kim K I, Jung K, Kim J H. "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm." IEEE Transactions on Pattern Analysis and Machine Intelligence, pp-1631–1639, IEEE(2003). https://doi.org/10.1109/TPAMI.2003.1251157

[10] Y. Zhong, H. J. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 4, pp. 385–392, IEEE (2000). https://doi.org/10.1109/34.845381

[11] Kumar, A., & Kumar, S. (2016). Comparative Study on Text Detection and Recognition from Traffic Image. INTERNATIONAL JOURNAL ONLINE OF SCIENCE, 2(9). Retrieved from http://ijoscience.com/ojsscience/index.php/ojsscience/article/view/114. https://doi.org/10.24113/ojsscience.v2i9.109

[12] K. I. Kim, K. Jung, and H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1631–1639, (2003). https://doi.org/10.1109/TPAMI.2003.1251157

[13] Chong Yu, Yonghong Song, Quan Meng, Yuanlin Zhang, Yang Liu, "Text detection and recognition in natural scene with edge analysis", IET Comput. Vis., Vol. 9, Iss. 4, pp. 603– 613, (2015). https://doi.org/10.1049/iet-cvi.2013.0307

[14] Chucai Yi, Ying Li Tian, "Text String Detection From Natural Scenes by Structure-Based Partition and Grouping", Vol. 20, No. 9, IEEE Transactions on Image Processing (2011). https://doi.org/10.1109/TIP.2011.2126586

[15] Shijian Lu, Tao Chen, Shangxuan Tian, Joo-Hwee Lim, Chew-Lim Tan, "Scene text extraction based on edges and support vector regression", 18:125–135, IJDAR (2015). https://doi.org/10.1007/s10032-015-0237-z

[16] K. C. Kim, H. R. Byun, Y. J. Song, Y. W. Choi, S. Y. Chi, K. K. Kim, Y. K. Chung, "Scene Text Extraction in Natural Scene Images using Hierarchical Feature Combining and Verification", 17th International Conference on Pattern Recognition (ICPR'04), IEEE (2004). https://doi.org/10.1109/ICPR.2004.1334350

[17] Mohammad Khodadadi, and Alireza Behrad, "Text Localization, Extraction and Inpainting in Color Images", IEEE, 20th Iranian Conference on Electrical Engineering, (ICEE2012), (2012). https://doi.org/10.1109/IranianCEE.2012.6292505

[18] Anubhav Kumar "An Efficient Text Extraction Algorithm in Complex Images", IEEE, (2013). https://doi.org/10.1109/IC3.2013.6612171

[19] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in ICCV, 2011. https://doi.org/10.1109/ICCV.2011.6126402

[20] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in BMVC, 2012. https://doi.org/10.5244/C.26.127

[21] P. Voigtlaender, P. Doetsch, and H. Ney, "Handwriting recognition with large multidimensional long short-term memory recurrent neural networks," in ICFHR, 2016. https://doi.org/10.1109/ICFHR.2016.0052

[22] P. Krishnan and C. V. Jawahar, "Matching handwritten document images," in ECCV, 2016. https://doi.org/10.1007/978-3-319-46448-0_46

[23] S. Sudholt and G. A. Fink, "PHOCNet: A deep convolutional neural network for word spotting in handwritten documents," in ICFHR, 2016. https://doi.org/10.1109/ICFHR.2016.0060

[24] P. Krishnan, K. Dutta, and C. V. Jawahar, "Deep feature embedding for accurate recognition and retrieval of handwritten text," in ICFHR, 2016. https://doi.org/10.1109/ICFHR.2016.0062

[25] Kartik Dutta, Minesh Mathew, Praveen Krishnan and C.V. Jawahar, "Localizing and Recognizing Text in Lecture Videos", International Conference on Frontiers in Handwriting Recognition, 2018. https://doi.org/10.1109/ICFHR-2018.2018.00049