Volume 2, Issue 4 April 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal

Of

Science, Engineering and Information Technology

IRJIF IMPACT FACTOR: 3.01

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi, Urdu all research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: - Dr. Ashak Hussain Malik House No. 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

TEXT EXTRACTION FROM IMAGE

SHARIF SHAIKH [1], CHAITAN KAMBLE [2], SNEHAL BAVISKAR [3] & ANKIT GANGURDE [4]
Department of Computer Engineering, PVPPCOE.

Under the guidance of- **MS. RUPALI PASHTE** (Assistant Professor, Dept. of Computer Engg. PVPPCOE)

ABSTRACT: - In this paper, we propose a novel text detection approach based on stroke width. Firstly, a unique contrast-enhanced Maximally Stable Extremal Region (MSER) algorithm is designed to extract character candidates. Secondly, simple geometric constrains are applied to remove non-text regions. Then by integrating stroke width generated from skeletons of those candidates, we reject remained false positives. Finally, MSERs are clustered into text regions. Experimental results on the ICDAR competition datasets demonstrate that our algorithm performs favorably against several state-of-the-art methods.

1. INTRODUCTION

Learning to read is a long road for humans the majority of the reading activity is carried out over written words on paper, the texts may also appear written on objects around us: text on street signs, a motto painted on a wall, and even text produced by arranging small stones on the sand. Environment text is an important source of information in our daily life. Due to increase in use of portable devices, equipped with integrated built-in digital cameras Reading text in natural scenes leads to new applications such as automatic translation, license plate readings, and visually impaired persons, which shows that the field has gained increasing attention of the researchers in the last decades.

2. LITERATURE SURVERY

"Pattern Recognition" in which the extraction of this information involves detection, localization, tracking, extraction, enhancement and recognition of the text variations of the text due to differences in orientation, size, style, and alignment as well as low image contrast and complex background make the problem of automatic text extraction.

Technique in which fast identification of existing multimedia documents and mounting demands for information indexing and retrieval has been done on extracting the text from images and videos based on Artificial Neural Network.

Approach is a two stage algorithm that is first stage, apply low pass filter on image in FFT domain to remove noisy element and then we apply Laplacian

operator to the resultant image to highlight high contrast areas in the image, In the second stage the algorithm the extracted text blocks are verified using an SVM classifier.

Methods which are usually based on Edge texture, connected domain feature, or learning are always limited by size, location, language, of artificial text in video.

To solve the problems mentioned above, this paper applied SOM (Self Organizing Map) based on supervised learning to video artificial text detection. To enhance the system performance, it is necessary to consider temporal changes in a frame sequence.

The text tracking stage can serve to verify the text localization results. In addition, if text tracking could be performed in a shorter time than text detection and localization, this would speed up the overall system. In cases where text is occluded in different frames, text tracking can help recover the original image.

3. PROPOSE METHODOLOGY

The proposed methods generally focus scene text detection from different document images. After the detection the more informative scene text image that is free from complex back ground, variation in font, size and orientation, color distortions, and noise be the result. The proposed method consists of five steps: candidate Generation, candidate normalization, non text filtering image partition and character candidate grouping. The figure shows the block diagram of our proposed method.

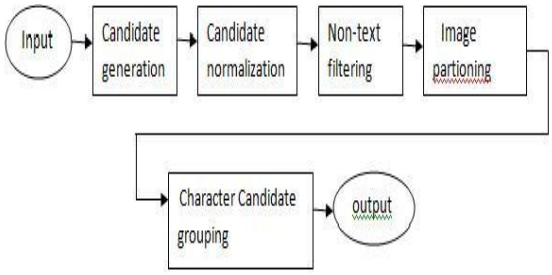

Figure 1: framework of our proposed method

Our candidate generation method is based on popular CC based approaches which consists of a MSER-based CCextraction block and an AdaBoost-based CC clustering block.

The maximally stable extremal region (MSER) algorithm is invariant to scales changes and affine intensity changes, and other blocks in our method are also designed to be invariant to these changes. In our method, both problems and are addressed based on machine learning techniques, so that our method is largely free from heuristics.

We have trained a classifier that determines adjacency relationship between CCs for Problem and we generate candidates by identifying adjacent pairs. In training, we have selected efficient features and trained the classifier with the AdaBoost algorithm. In the proposed method text detection approach

consists of the following modules:

CANDIDATE GENERATION

In the proposed method for the generating a candidates, extract CCs in images and partition the extracted CCs into set of clusters, using clustering algorithm is based on an adjacency relation classifier. In this section first explain the CC extraction method. Then, explain the approaches (i) building training samples, (ii) to train the classifier, and (iii) using that classifier in CC clustering method.

CCS EXTRACTION

The MSER algorithm [7] is the most efficient in CC extraction because it shows a good performance with a small computation cost. Only the MSER algorithm could provide the stable binary results and also help us find most of the text components our classifier is based on pair wise relations between CCs and CC pairs. Therefore, rather than focusing on this difficult problem, we address a relatively simple problem by adopting an idea of region-based approaches. If we have ci ~cj it will yield a character candidate components consisting of non text CCs and this candidate will be rejected at the non text rejection step. Also, we will perform word segmentation as a post processing step. Based on these observations, we build training sets. Specifically, we first obtain sets of CCs by applying the MSER algorithm to a training set released.

NORMALIZATION

After CC clustering, we have a set of clusters. From that we will normalize corresponding regions for the reliable text/non-text classification it is of two steps as follows

A. Geometric Normalization

For the given set of clusters we first localize its corresponding region. We approximate the shape of text boxes whose left and right sides are parallel to y-axis. This approximation reduce the difficulties in estimating text boxes for this we only have to find a skew and four boundary supporting points.

B. Binarization

By using the geometrically normalized images binary images can be built. We perform the binarization separately by estimating text and background colors.

NON TEXT FILTERING

A text/nontext classifier that rejects nontext blocks among normalized images. The main challenge of the approach is the variable aspect ratio. One possible approach to solve this problem is to split the normalized images into patches covering one of the letters and develop a character/non-character classifier. However, character segmentation is not an easy problem so split a normalized block into overlapping squares and develops a classifier that

assigns a textness value to each square block. Finally, the decision results for all square blocks are integrated so that the original block is classified.

IMAGE PARTITION

To extract text information from complex background, image partition is first performed to group together pixels that belong to the same text character, obtaining a binary map of candidate character components. Based on local gradient features of text characters, we design a gradient-based partition respectively.

CONNECTED COMPONENTS GROUPING

The image partition creates a set of connected components from an input image, including both text characters candidates and unwanted noises. The text information appears as one or more text strings in most natural scene images, we perform heuristic grouping and structural analysis of text strings to distinguish connected components representing text characters from those representing noises. Assuming that a text string has at least three characters in alignment, we develop a method to locate regions containing text strings. A connected component C is described by four metrics: height (), width (.), centroid (.), area (.). In addition, we use D (.) to represent the distance between the centroids of two neighbouring characters.

ADJACENT CHARACTER GROUPING

Text strings in natural scene images usually appear in alignment, namely, each text character in a text string must possess character siblings at adjacent positions. The structure features among sibling characters can be used to determine whether the connected components belong to text characters or unexpected noises. Here, five constraints are defined to decide whether two connected components are siblings of each other.

1) Considering the capital and lowercase characters, the height ratio falls between and 1/T1 and T1.

2) Two adjacent characters should not be too far from each other despite the variations of width, so the distance between two connected components should not be greater than T2 times the width of the wider one.

3) For text strings aligned approximately horizontally, the difference between -coordinates of the connected component centroids should not be greater thanT3 times the height of the higher one.

4) Two adjacent characters usually appear in the same font size, thus their area ratio should be greater than 1/T4 and less than T4.

5) If the connected components are obtained from gradient based partition. The color difference between them should be lower than a predefined threshold because the characters in the same string have similar colors.

4. CONCLUSION:

In this paper, we discuss the way to build a Text extraction from images and Videos and also note down the existing problems that occur in the process. We, therefore, suggest a group of methods for constructing and inputting our project, and show the result's quality and also the run time and scalability. The quality of this approach shows a greater improvement in the quality and run time as compared to the existing text extraction from images and videos.

5. ACKNOWLEDGE

We wish to acknowledge Prof. Rupali Pashte for their throughout support and guidance in every step from conceptualisation to im-plementation of system aiding in successful comple-tion of this survey paper.

6. REFERENCES

1. http://www.cs.ubc.ca/~perfo/papers/forssen_cvpr0
2. http://www.iosrjournals.org/iosrjce/papers/
3. http://www.ijcset.com/docs/
4. http://web.stanford.edu/~hchen2/papers/
5. http://refbase.cvc.uab.es/files/GoK2014

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- Dr. Ashak Hussain Malik House No-221, Gangoo Pulwama - 192301 Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com