International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July 2018 www.irjet.net p-ISSN: 2395-0072

TEXT EXTRACTION FROM VIDEO CONTAINING ARTIFICIAL IMAGES

Prof. Sharanbasappa A M[1], Dr.Vishwanath C B[2]

[1]Department of Information science and engineering,Appa institute of engineering and technology. Gulbarga. Karnataka, India
[2] Department of Information science and engineering,P D A College of engineering,. Gulbarga. Karnataka, India

---

ABSTRACT— *Videos contain a vast amount of information, which, if channelled properly can provide breakthrough in various research fields. Multimedia content retrieval is an important research field that aims in content based information indexing and retrieval, automatic annotation and structuring of video frames. Large amount of information is embedded in Natural Scene, which often requires automatic extraction and processing; Artificial Text can be termed as one of the important Multimedia content. In this Paper, it is attempted to accurately separate text content from multimedia objects by precisely localizing the text and extracting. Text detection in videos is an important step to achieve Multimedia content retrieval which plays an important part in fully understanding of the video clip. Object retrieval follows the general procedure involving detection, localization, tracking, extraction and enhancement of the text from a given image. The detection step roughly classifies text & non-text regions, the localization step determine accurate boundaries of the text string, the extraction step filters out background pixels in the text string.*

Keywords— Multimedia Object, Text detection, Localization, Tracking, Extraction, Enhancement. etc.

1. Introduction

Images and videos on webs and in databases are increasing. It is a pressing task to develop effective methods to manage and retrieve these multimedia resources by their content. Text, which carries high-level semantic information, is a kind of important object that is useful for this task. The acquisition of a video is generally done using a set of physical captors, Web camera or a Digital Camera, in this Paper. The next step of separating image frames can be accurately modeled as a sampling of the continuous image using a discrete partitioning of the continuous plane. The Image frame thus acquired is the "Digital Image" and the basic procedures of Digital Image Processing are applied.In the case of video-clips; the number of frames containing text is much smaller than the number of frames without text. Binary images considered for analysis generally result from the digitization of frames obtained from Video clip. Similarly, such binary images can be created by thresholding the grey-level at each pixel in grey scale images. This processing represents the most basic operation in the class of segmentation processes. Depending on parameters such as resolution and grey-level threshold, the binary image may be altered by noise; noise refers to either black or white pixels added randomly to the image (i.e. salt and pepper noise) or a group of (black or white) pixels added to the image. The first type of noise typically arises in an acquisition process; whereas the second type of noise typically results from an inaccurate thresholding process. In both cases, this noise is to be removed from the binary digital image for accurate analysis. Noise removal is based on the knowledge of the type of features present in the image. Text Detection follows the Pre-Processing step. The role of text detection is to find the image regions containing only text that can be directly highlighted to the user or fed into an optical character reader module for recognition. It is an essential step for text recognition. In some cases, text detection becomes even meaningful by itself. For example, finding the appearance of a caption in news video can help to locate the beginning of a news item. In majority of the content, either Image or Video, there is no prior information on whether or not the input image contains any text, the existence or non-existence of text in the image must be determined. Several approaches assume that certain types of video frame or image contain text.Text detection and localization are often used interchangeably. The detection of text embedded in images and videos gives rich source of information for content-based indexing and retrieval applications. However, these text characters are difficult to be detected due to their various sizes, grey scale values and complex backgrounds. Text localization is the process of determining the location of text in the image and generating bounding boxes around the text. Text tracking is performed to reduce the processing time for text localization and to maintain the integrity of position across adjacent frames.

A. Objectives:

The aim of the paper, as stated earlier, is to extract text from the video clip. The initial aim is to extract the text from any video clip, but from the vast amount of literature survey it is evident that, it is difficult to perform text extraction on any clip owing to the various reasons like improper illumination, complex background, blurring etc.

2. Related Work

In paper [1] Jie Xi et. al,(2001)have proposed the system, that integrates text detection, tracking and

recognition to extract the text information in news and commercial videos. The image quality of the detected text blocks is enhanced by averaging over multiple frames. Finally, the averaged text blocks are binarized and sent to a conventional OCR engine for recognition. Iterative binarization procedure is used to ensure best binarization results for recognition.

In paper[2] Xiaojun Li,et. al, (2002) have proposed a method, where a method to detect text regions by means of the sparse representation is proposed. The proposed method starts with an edge map of the image. Stroke filter is used to obtain four stroke maps which characterize the stroke strengths in horizontal, vertical, left-diagonal, right-diagonal directions. Then the corresponding features are extracted for each sliding window, and a SVM (Support Vector Machine) is employed to classify the sliding windows into text blocks and non-text blocks.

In paper [3] Hrishikesh B, et. al, (2002)have proposed a method, that proposes a novel, non-causal temporal aggregation method that acts as a second pass over the output of an existing text detector over the entire video clip. First, a Multi-scale detection of videotext events is performed followed by spatial segmentation by the use of recursive Paperions and k means clustering is done. Binary search ensures the detection of in-place changes in text content and finally Verification of spatial partitions ensures successful text detection.

In paper [4] Xiaodong Huang,et. al, (2003)have proposed a method, where a new video scene text detection and localization method is proposed. First, a stroke map based on Log-Gabor filter is built, which will suppress some background interference. Second, texture feature on every line of stroke map to locate text lines is calculated. Finally, Harris corner detection is performed on stroke map of detected text lines. Morphological operation is performed to connect these corners into text regions and use CCA (connected component analysis) to remove some non-text regions.

In paper [5] Shi Jianyonget. al, (2004)have proposed a method, where an edge-based video text extraction approach with low computation, is used which can automatically detect and extract text from complex video frames. In the first step, the edge maps of both an intensity image and its binarized image are obtained, and merged with the two into one edge map, containing less edge pixels of background but enriched edge pixels of text. An adaptive thresholding method is applied to identify adjacent pixel rows and columns which contain text. The intersections of these rows and columns are extracted as text regions. Finally, a novel extraction method based on mono-chromatism of text is applied to the regions. The output of the extraction method can be directly fed to OCR (Optical Character Recognition).

In paper [6] Michael R. Lyuet. al, (2005)have proposed a method, that proposes a procedure that performs a detailed analysis of multilingual text characteristics. A comprehensive, efficient video text detection, localization, and extraction method, which emphasizes the multilingual capability is achieved. The text detection is carried out by edge detection, local thresholding, and hysteresis edge recovery. The coarse-to-fine localization scheme is then performed to identify text regions accurately.The text extraction consists of adaptive thresholding, dam point labeling, and inward filling.

In paper [7] Palaiahnakote Shivakumaraet. al,(2009)have proposed, a new method based on wavelet transform, statistical features and central moments for both graphics and scene text detection in video images. The method uses 2D Haar wavelet decomposition for detecting text in the video image. Three high frequency sub-band images LH, HL and HH are used for text detection purpose. The features computed are fed to k means clustering to classify the text pixel from the background of the image.

In paper [8] Yatong Zhou,et. al, (2010)have proposed a method, to video text localization using gradient discrete cosines transform (DCT). Firstly, the video frame is divided into NxN sub-blocks to get the DCT coefficients from every sub-block. After that, the gradient operator value considered as block intensity is calculated. Lastly, the horizontal and vertical Paperion is obtained, the number of text line got and the text region marked with text box. The experimental results show that the proposed method isefficient on the localization of the static and rolling video text.

In paper [9] JayshreeGhorpadeet. al,(2011)have proposed a method, to design algorithms for each phase of extracting text from a video using java libraries and classes. First the input video is framed into stream of images using the Java Media Framework (JMF) with the input being a real time or a video from the database. Then the image is converted to grey scale and removal of the disturbances like superimposed lines over the text, discontinuity removal, and dot removal. Finally the algorithms that train neural network pattern machine are employed for localization, segmentation and recognition

In paper [10] T.Pratheeba,et. al, (2011) have proposed a method, to detect and extract the text from the video scene using a novel framework. A morphological binary map is generated by calculating difference between the closing image and the opening image. Then candidate regions are connected by using aMorphological dilation operation and the text regions are determined based on the occurrence of text in each candidate. The detected text regions are localized accurately using the Paperion of text pixels in the morphological binary map and the text extraction is finally conducted.

3. Proposed Work

The aim of the Paper, as stated earlier, is to extract text from the video clip. The initial aim is to extract the text from any video clip, but from the vast amount of literature survey it is evident that, it is difficult to perform text extraction on any clip owing to the various reasons like improper illumination, complex background, blurring etc. For a clear video clip with ample amount of illumination the procedure followed to achieve the result is depicted in the block diagram. The video clip is first taken in .avi format that is supported by the tool used in our Paper.

The mmread function is called that performs the initial step of capturing video-frame from the video clip. The frame which has the best possible chance for the clear extraction of the text is selected. Another real time feature would be to have a web-camera installed that captures the video, save it and then link that particular video to MATLAB. Image acquisition tool of the MATLAB can be used for this. The selected frame is then subjected to various pre-processing steps that include de-noising. Then the pre-processed image is subjected to filtering and enhancement that detects the region of the text, if present. Then the steps of Localization and Extraction are performed with the use of Morphological operations and the use of region properties. The aim is to extract text from any video clip, but owing to the availability of vast type of text and different font types and sizes, the basic motivation is to develop a code that works for at-least one video clip each with and without text.

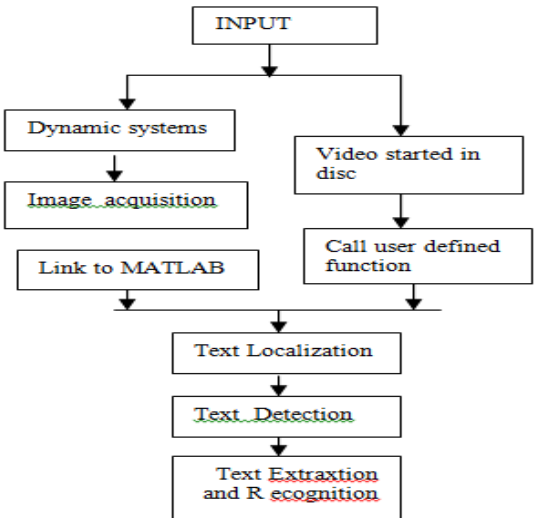

Figure 1 Outline of the proposed solution.

4. Methodology

Humans are primarily visual creatures. Not all animals depend on their eyes, as we do, for most of the information received about their surroundings. There are some important differences between human vision, the kind of information it extracts from images, and the ways in which it seems to do so, as compared to the use of imaging devices based on computers for scientific, technical, or forensic purposes. The most common and affordable way of acquiring images for computer processing was with a video camera. Humans are especially poor at judging color or brightness of objects and features within images unless they can be exactly compared by making them adjacent. Human vision is inherently comparative rather than quantitative, responding to the relative size, angle, or position of several objects but unable to supply numeric measures unless one of the reference object is a measuring scale. The mapping from a continuous to a discrete image forms the first step in any digital image processing application. Discrete data resulting from this digitization process is then stored in a form which is suitable for further processing. Computer-based image processing and analysis use algorithms based on human vision methods in some cases, but also employ other methods that seem not to have direct counterparts in human vision. In particular, some image processing methods are based on the physics of the image formation and detection process.

A. Algorithm

The algorithm uses Morphological operation for text localization along with some preprocessing and post processing steps. This algorithm is tested with natural Images as well as Images taken from ICDAR 2003 Robust Reading Competition dataset.

First, the input image is filtered by the Median filter to remove any noises. Then edges are detected using Laplacian of Gaussian (LOG) edge detector. Then the morphological dilation &erosion operations are applied for object localization. All the connected Components are then extracted and all non-text character components are discarded by a two-step process. Features are then extracted from the extracted Components.

B. Pre-Processing

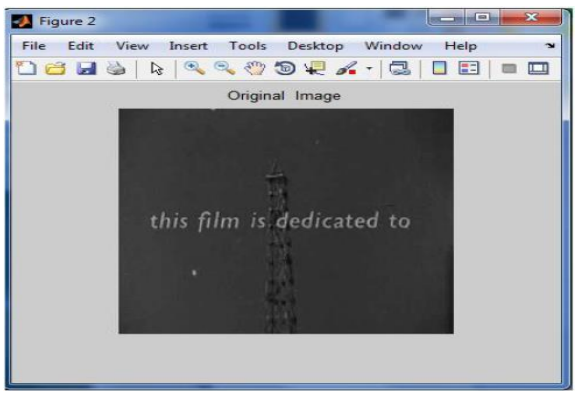

Figure 2 The Image Captured from sample Video rendered to MATLAB

In the preprocessing step, first the input RGB image is converted to grey-scale image. This conversion is done in order to reduce the processing overload. Median filtering is then applied to the Grey-scale image to remove any noises present in that. Next edges are extracted from the resultant Image using LOG edge detection algorithm. The choice of using LOG edge detector is for the reason that it finds the correct places of edges and testing wider area around the pixel.

C. Text localization and Extraction

In this phase, the edge image obtained from the previous step is binarized and then the Morphological dilation operation is performed on this edge map. Since texts are normally aligned in the horizontal direction we have used a 2 X 4 rectangular structuring element. All Connected Components are then extracted. Labeling the connected components is introduced apart from the proposed algorithm for the simplicity of successful extraction.

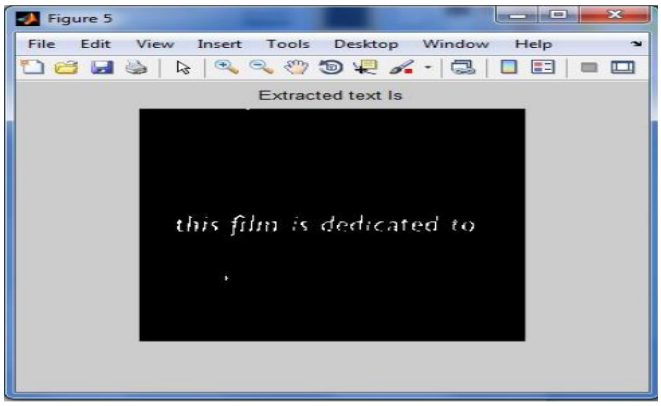

Figure 3.The extracted text

D. Text Detection: Text/non-text classification

All these extracted components may contain both text components and Non-text components. They are separated and eliminated by a two way process. First, the initial Bounding Boxes are drawn for all objects. Then the Connected Components statistical metrics are used to remove non-text components. For this, the following two rules [a] & [b] are applied to remove very big components and very small components that may result from the background objects.

[a] Ai < Ta..

[b] BBw> 2 * BBh.

[c] Pixel distance (Ci ,Cj) < 1.

Where A is the area of the component, BBh and BBw are the Bounding Boxes' height and width. Here [a] discards all very small components and [b] discards all big and wider objects. Occurrences of false positives are the major drawback in using Connected Components based approach. When thick characters are extracted, the inner and the outer circles are separately extracted and thereby increasing false positive counts. Rule [c] is used to discard such components.

E. Algorithm for Text Extraction Using Morphology

1. Input video clip

Start pre-processing:

2. Obtain digitized image.

a(x ,y) → [m ,n]

3. Convert the RGB image to grayscale image.

4. Resize the frame.

5. Filter the Grayscale image using Laplacian of Gaussian.

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x^2+y^2)}{2\sigma^2}}$$

6. Contrast enhancement by thresholding.

Start localization:

7. Edge detection of the filtered image using Laplacian of Gaussian

8. Convert the edge image to binary image

9.Perform morphological Dilation or Erosion

$$I \square H \equiv p \in Z^2 \mid (p + q) \in I, \text{ for every } q \in H$$

10. Find and extract all Connected Components of the Image

11. Label all the Connected Components

Start Detection:

12. Measure all the properties of the image

13. Draw the Bounding Boxes for All the Objects

14. Remove very big (BBw> 2 * BBh) and very small components (Ai < Ta)

15. Extracted Text

16. End

5. Results

Aim of the Paper is accepted to be achieved if the applied strategy gives the expected output at each stage. For simplicity of understanding the output of the acquired images are obtained at each major process. The Paper has been programmed to properly extract the text present in the frame that is acquired from the video-clip. In case there is no text present in the clip, or rather the frame acquired, a notification stating the absence of text is expected.

The Proposed solution, based on the standard algorithm gives the extracted text as its final solution. With the Paper executing above satisfactory levels the future extension of the code can be easily thought about. In this section a brief review of all the available accurate results is presented. The implemented method is executed on various video clips with & without text content in it; and it has been found that for the video clips which do not contain any text content, the method gives a result stating the absence of the text with 100% accuracy. For the video clips containing text, the method performs extraction for 8 out of the 10 input video clips, thus reaching an efficiency of 80% as shown in figure 4.

The following bar graph represents in percentage, the amount of text extracted from the video clip against the samples used.

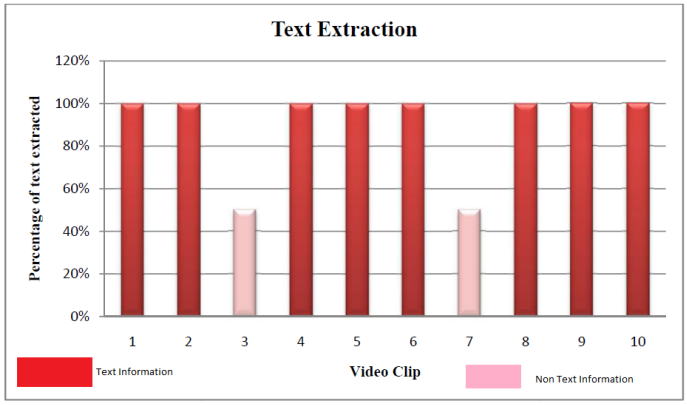

Figure 4: Represents percentage of text extracted from a video clip

6. Conclusion and Future work

The discussed Methodology and Results give an overview about the Paper. At this juncture of the thesis, it is worthwhile to give a brief overview about the work performed. The video clip is linked to the tool used; the clip is present in the format compatible with Matlab. The clip can be related to anything with the relaxation of presence or absence of text. The linked clip is subjected to various Image processing steps that are modeled along the referred algorithm, and as expected, the results are obtained. There are instances where a relaxation is made in following the actual algorithm, but the sheer fact that the Paper provides the expected results justifies the risk of adopting the change.

Change & Development in any field is the rule of nature. Efficiency of the Paper is also a matter of concern, There are several algorithms that boast of high efficiency as compared to the one used. The implementation of the Paper on the lines of different algorithms can also be considered. Given the mode of execution of the code, the efficiency is limited to very small video clips. The huge data base of vivid and complex text content as well as the availability of

3-Dimentional text, the extraction of text under conditions like varying illumination, complex movement and complex text pattern can be considered. Videos with varying text in consecutive frame also pose a challenge for future minds.

REFERENCES

[1] StephaneMarchand-Maillet, Yazid M. Sharaiha, Binary, *"Digital Image Processing: A Discrete Approach"*, Academic Press

[2] John C. Russ, *"The Image Processing Handbook"*- 6 E, CRC Press.

[3] Bernd Jahne, *"Digital Image Processing"*, Springer.

[4] TinkuAcharya and Ajoy K. Ray, *"Image Processing Principles and Applications"*.

[5] BalázsEnyedi, LajosKonyha, KálmánFazekas, TuránJán, *"Character Localization In Video Sequences".*

[6] Jie Xi, Xian-Sheng Hua, Xiang-RongChen ,Liuwenyin , Hong-Jiang Zhang, *"A Video Text Detection And Recognition System"*, 2001 Ieee International Conference On Multimedia And Expo.

[7] Rainer Lienhart and Axel Wernicke, *"Localizing and Segmenting Text in Images and Videos"*, IEEE Transactions On Circuits And Systems For Video Technology, Vol. 12, No. 4, April 2002.

[8] Minoru Mori, *"Video text recognition using feature compensation as category-dependent feature extraction"*, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003 IEEE

[9] Qixiang Ye, Qingming Huang, Wen Gao, Debin Zhao, *"Fast and robust text detection in images and video frames"*, January 2005

[10] PalaiahnakoteShivakumara, Weihua Huang and Chew Lim Tan, *"Efficient Video Text Detection using Edge Features"*, 2008 IEEE

[11] Yatong Zhou, Dan Li, Kewen Xia, *"An Approach To Video Text Localization Based On Gradient Discrete Cosines Transform"* 2010 Ieee

[12] R. Chandrasekaran, RM. Chandrasekaran, *"Morphology based Text Extraction in Images"*, IJCST Vol. 2, Issue 4, Oct. - Dec. 2011