International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
327 – 331

Text Extraction in Video

Dhanashri Holgare, Rutuja Talewar, Prof. Karishma Dhule
Department of Computer Engineering,
BapuraoDeshmukh College of Engineering,Sewagram,Wardha

Abstract:- The detection and extraction of scene and caption text from unconstrained, general purpose video is an important research problem in the context of content-based retrieval and summarization of visual information. The current state of the art for extracting text from video either makes simplistic assumptions as to the nature of the text to be found, or restricts itself to a subclass of the wide variety of text that can occur in broadcast video. Most published methods only work on artificial text (captions) that is composited on the video frame. Also, these methods have been developed for extracting text from images that have been applied to video frames. They do not use the additional temporal information in video to good effect.This thesis presents a reliable system for detecting, localizing, extracting, tracking and binarizing text from unconstrained, general-purpose video. In developing methods for extraction of text from video it was observed that no single algorithm could detect all forms of text. The strategy is to have a multi-pronged approach to the problem, one that involves multiple methods, and algorithms operating in functional parallelism. The system utilizes the temporal information available in video. The system can operate on JPEG images, MPEG-1 bit streams, as well as live video feeds. It is also possible to operate the methods individually and independently.

Keywords *—Detection, Extraction, Frame, Images and Tracking.*

*****

I. INTRODUCTION:

As computer, compress technology, storage media and high speed communication skill are developed dramatically; digital video has become one of the most important elements in many applications such as education, news and games. Multimedia data are also getting bigger than before. In order to extract and search important information from a huge amount of video data, we need to extract text from video. Text is obviously an important element in video. So extracting text appears as a key clue for understanding contents of video and for instance for classifying automatically some videos. Videotext detection and recognition has been identified as one of the key components for the video retrieval and analysis system. Videotext detection and recognition can be used in many applications, such as semantic video indexing, summarization, video surveillance and security, multilingual video information access, etc. Videotext can be classified into two broad categories: Graphic text and scene text. Graphic text or text overlay is the videotext added mechanically by video editors, examples include the news/sports video caption, movie credits etc. Scene texts are the videotexts embedded in the real-world objects or scenes, examples include street name, car license number, and the number/name on the back of a soccer player. This report is to address the problem of accurately detecting and extracting the graph videotexts for videotext recognition. Although the overlay text is manually added into the video, the experiments showed they are even as hard to extract as many video objects, such as face, people etc.

This is due to the following reasons:

1. Many overlay texts present in the cluttered scene background.
2. There is no consistent color distribution for texts in different videos. Consequently, the color-tone based approach widely used in face or people detection application actually cannot be applied in text detection.
3. The size of the text regions may be very small such that when the color segmentation based approach is applied, the small text region may merge into the large non-text regions in its vicinity. Here we used edge detection based method for extracting the text and it is implemented using Mat lab. Here the two critical angles are defines and the text is extracted and recognized using the coincidence of the edges of the image with the threshold defined based on the critical angles.

II. MAIN CONCEPTS:

Text extraction in video consists in three steps. The first one is to find text region in original images. Then the text needs to be separated from background. And finally a binary image has to be produced (for example, text is white and background is black) Difficulties of such a project can be classified in following main categories:

1. Background and text may be ambiguous.
2. Text color may change: text can have arbitrary and non-uniform color.
3. Background and text are sometimes reversed.
4. Text may move.

 ISSN: 2454-4248
327 – 331

5. Unknown text size, position, orientation, and layout: captions lack the structure usually associated with documents.
6. Unconstrained background: the background can have colors similar to the text color. The background may include streaks that appear very similar to character strokes.
7. Color bleeding: lossy video compression may cause colors to run together.
8. Low contrast: low bit-rate video compression can cause loss of contrast.

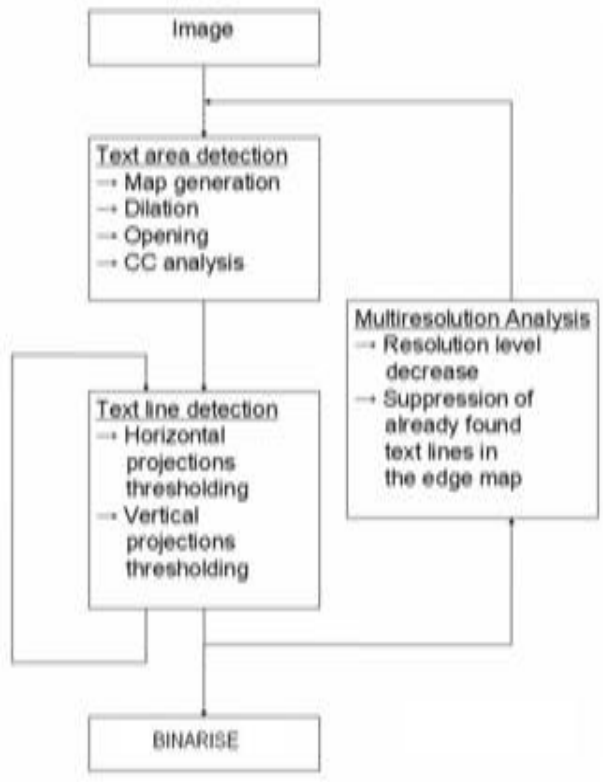


Figure 1: Flow Chart for Text Extraction.

Figure 2: Text Extraction Results

This can be extended by including Text Recognition to it. Also extending it to video, real time operation, the program can work surprisingly well and useful. Adding all these features we can use this project for class automatically some images, for the retrieval and analysis system and in many applications, such as semantic video indexing, summarization, video surveillance and security, multilingual video information accessed. During the course of this research effort several contributions were made to enable reliable extraction of text in video. A multi-threaded multi-architecture prototype system was developed in C++ which operates on JPEG images, uncompressed video, MPEG compressed video, and live video. Various methods which were proposed for extraction of text from images and video in the literature were studied and a critical analysis of their success in detecting text from unconstrained general purpose video was presented. Some of the methods which seemed promising were implemented and some enhanced to enable a better performance. A detailed evaluation of these methods was conducted on a challenging dataset. For this evaluation a ground truth was developed which allowed pixel strict objective evaluation. Potential problems with using such a ground truth were discovered and presented. It was noticed that, cumulatively, different methods were able to localize almost all instances of the text appearing in the video. To address this observation, algorithm fusion methods were studied and solutions presented. The solution utilized human supervision on a small portion of the dataset to enable improved results. In addition, a multifeature text extraction algorithm was developed that applied promising feature extraction algorithms in cascaded fashion. It is seen from the results that further research in this area is needed to make indexing of video a reality. It is shown that algorithm fusion is a promising direction to achieve good localizations.

III. TEXT DETECTION

Since we have no a priori information on whether a given video frame contains text, the first step is to detect whether text of any sort exists in the frame. This serves as a pre filtering operation to avoid more expensive processing on all frames and seeks to give a binary answer to the question of whether text exists in the frame. This stage needs to be fast and should prefer false alarm errors to missed detection errors since the former can be rejected at later stages.

The term text detection here means the distinguishing the letters or the characters from the image part. This is the process of determining whether a given part or part of the image is a text or some other figures. Text detection generally can be classified into two categories:

 ISSN: 2454-4248
327 – 331

3.1 *BOTTOM-UP METHODS*: They segment images into regions and group ―character‖ regions into words..The input image is segmented based on the monochromatic nature of the text components using a split-and-merge algorithm. Segments that are too small and too large are filtered out. After dilation, motion information and contrast analysis are used to enhance the extracted results. The methods, to some degree, can avoid performing text detection. Due to the difficulty of developing efficient segmentation algorithms for text in complex background, the methods are not robust for detecting text in many camera-based images and videos.

3.2 *TOP-DOWN METHODS:*They first detect text regions in images using filters and then perform bottom-uptechniques inside the text regions. These methods are able to process more complex images than bottom-up approaches. Top-down methods are also divided into two categories

1. Heuristic methods: they use heuristic filters
2. Machine learning methods: they use trained filters

Here we are using heuristic method of text extraction. This method of text extraction can be performed in two different approaches. Each of both uses the characteristics of artificial text.

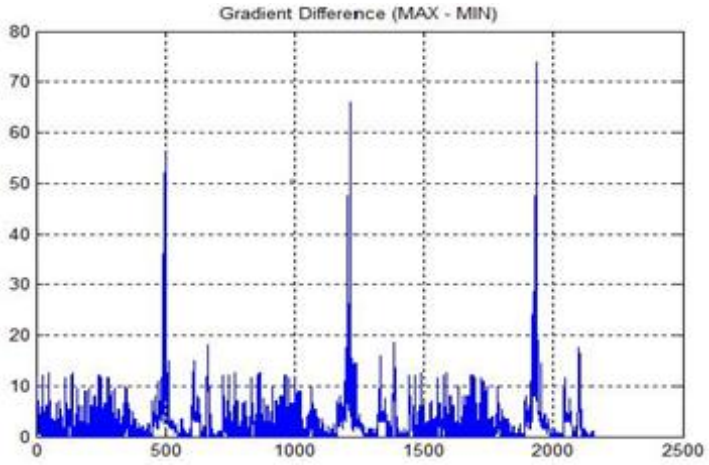
1. Connected regions approach: The main idea of this approach is that a letter can be considered as ahomogeneous region (using our restrictions), and thus it could be very useful to divide the frame into homogeneous regions. To compute such a division, a split-and-merge algorithm seems to be very adequate. Its concept is: while there is a non homogeneous region, then split it into four regions. And if two adjacent regions are homogeneous, then they can be merged. Then, using some size characterizations of the text (not too big and not too small), the inadequate regions will be deleted. The same process will be executed for the different frames, and the results will be temporally integrated in order to keep only the elements which are present in all the frames.

2. Edge detection approach: The main idea of this approach is that text contrasts a lot with background. The text willhave a well defined edge that makes it possible to identify the text clearly. Thus, using edge detection concepts seems to be a good idea. The edge detection is based on certain predefined critical angles called the threshold angles. The lines that coincide with these thresholds are identified as the text edges. After that the edge lengths will be computed, the length number of edges in x and y direction will be calculated and if it is higher than a certain threshold then it will be considered as a text area. Then each text area will be binarized using the luminance. For binarizing also we are using thresholds. The intermediate luminescence is taken as the threshold value and the areas which are brighter than the threshold is white and the other areas are black. So in a final result, the text will be in white and the background in black (or the inverse). Finally, the same process will be executed for the different frames, and the results will be temporally integrated in order to keep only the elements which are present in all the frames.

3. Extracting Text from Video.Thus, methods depending on page models and top-down page segmentation methodsare not appropriate—we need to use more locally adaptive bottom-up methods with less dependence on regularity, uniform font size, availability of multiple lines etc. Also, most document analysis methods for text detection depend on an initial high quality binarization step from the original gray-scale image. In this case the background: Text recognition in document images has been an active research area for some time. However text recognition in broadcast quality digital video is a problem requiring different approaches. Unlike document images, video frames tend to have text not in orderly columns but in widely scattered areas, and fewer, separated lines. Also, video frames are typically noisy, low-resolution, and full-color with interlace artifacts. The text in a video frame can be multi-colored, multi-font, and/or be transparent, with the background showing through.

Some of the text pixels can share the colors of the background. Also the text region itself may be translucent. Thus, a global binarization step is unlikely to isolate the text. Video does, however, offer the advantage of temporal redundancy that can be exploited. This is that successive frames containing text are usually similar in the text region.

IV. WAVEFORM:



Consider a frame image F having I rows and J columns. Then the set of text blocks T can be determined using the dynamic range.

$$M_{k,l} = \max(F (i, j)) \quad (4.1)$$

$$M_{k,l} = \min(F (i, j)) \quad (4.2)$$

Where,

$$\gamma k \leq i < \gamma (k + 1)$$

 ISSN: 2454-4248
327 – 331

$\gamma k \leq j < \gamma(l + 1)$

The bounds of k and l are given in Equation 4.3.

$k = 0, 1. . . (I/\gamma) - 1$

$l = 0, 1. . . (J/\gamma) - 1$ (4.3)

The dynamic range, d, is then defined as shown in Equation 4.4

$d_{k,l} = | M_{k,l}$

$- m_{k,l}$(4.4)

The set of blocks in the frame image F classified as text blocks are then given by 4.5

$T=\{(K,l)\} : =\{d\ K,l >=t$

$d\ K,l=0\}$(4.5)

In this case, γ is set to 4. The above equations determine if the dynamic range of the block is either greater than or equal to a distinct threshold τ or is nearly 0, the block is classified as a text block. In addition, the detected blocks are validated by thresholding the number of classified text blocks to a preset threshold set to 0.1. In practical application, the method excludes certain areas of the frame image boundary to reduce false alarms. Empirically determined values of τ, where the method performs fairly well, range from 45 to 60.

V. RESULTS

Figure 5.1 shows sample resulting image frames after applying the intensity stage of the Method-A. It is seen that the application of such a cascaded set of constraints enables removal of most false alarms, while detecting almost all instances of the text appearing in the video including Arabic language credit titles. In Figure 5.1 almost all instances of text are detected save for the text in the backdrop behind the news anchor which are of very low contrast.

Fig: 5.1 Text localization reults

VI. CONCLUSION:

In many ways the result of these experiments are both surprisingly good and surprisingly bad. For images without definite edges the program may not work properly. But it will work perfectly for image texts which have prominent edge. This can be extended by including Text Recognition to it. Also extending it to video, real time operation, the program can work surprisingly well and useful.

 ISSN: 2454-4248
327 – 331

Adding all these features we can use this project for classifying automatically some images, for the retrieval and analysis system and in many applications, such as semantic video indexing, summarization, video surveillance and security, multilingual video information access etc. During the course of this research effort several contributions were made to enable reliable extraction of text in video. A multi-threaded multi-architecture prototype system was developed in C++ which operates on JPEG images, uncompressed video, MPEG compressed video, and live video. Various methods which were proposed for extraction of text from images and video in the literature were studied and a critical analysis of their success in detecting text from unconstrained general purpose video was presented. Some of the methods which seemed promising were implemented and some enhanced to enable a better performance. A detailed evaluation of these methods was conducted on a challenging dataset. For this evaluation a ground truth was developed which allowed pixel strict objective evaluation. Potential problems with using such a ground truth were discovered and presented. It was noticed that, cumulatively, different methods were able to localize almost all instances of the text appearing in the video. To address this observation, algorithm fusion methods were studied and solutions presented. The solution utilized human supervision on a small portion of the dataset to enable improved results. In addition, a multifeature text extraction algorithm was developed that applied promising feature extraction algorithms in cascaded fashion. It is seen from the results that further research in this area is needed to make indexing of video a reality. It is shown that algorithm fusion is a promising direction to achieve good localizations. Yet, much work is needed on this topic to enable detection of information from the video source to provide unsupervised algorithm fusion.

REFERENCES:

[1] J.B. Bosch and E.M. Ehlers. Remote Sensing of Characters on 3D Objects. Computers&ndustrial Engineering, 33(1 -2):429–432, 1997.

[2] N. Chaddha, R. Sharma, A. Agrawal, and A. Gupta. Text Segmentation in Mixed– Mode Images. In 28th Asilomar Conference on Signals, Systems and Computers, Pages 1356–1361, October 1995.

[3] N.-S. Chang and K.-S. Fu. Query-by-Pictorial-Example. IEEE Transactions on Software Engineering, 6(6):519–524, 1980.