A Review on Character Segmentation of Video Subtitles

Satish S Hiremath, K V Suresh

Abstract—Subtitle extraction from video sequences finds many useful applications in video classification. The complex background, illumination and font size of text present in the video makes subtitles extraction extremely challenging. Text in the video are classified as scene text and graphics text. In this paper, literature review lists several techniques to extract graphics text from the video. Compared to word segmentation in video subtitles, single character segmentation can achieve higher accuracy because of the simple background. The commonly used method for character segmentation using morphological operation is discussed in the paper.

Index Terms— Character segmentation, Complex background, Font size, Graphics text, Illumination, Morphological operations, Scene text, Subtitle extraction.

1 Introduction

THE topic of image to text processing has attracted many image processing experts. In modern technology, similar to image, videos are also the main source of information media. As multimedia database is increasing, video indexing has become a challenging and gained lot of interest. There are several methods involved in video indexing, among which one of the method is by using text present in the video. The videos with subtitles carries semantic information, which could provide valuable cues about content of the videos and thus is very important to analyze it. Video index can be built by detecting, extracting and recognizing text from the video. In text processing, one of the important step is segmentation and in which character segmentation is formidable, as it involves separation of the touching characters.

As proved by Judd *et al*. [1], when an image containing of text, object, human or any scene is viewed then viewers tends to focus on the text first. Hence, the text present in videos can play a vital role in video indexing. The video text has been classified into two groups [2] namely 'Scene text' (e.g. text on vehicle, building and sign boards on the road etc...) and 'Graphics text' (e.g. subtitle video, news video). In this paper, literature review lists techniques involved in character segmentation of graphics text of the video. To extract graphics text the complex background, illumination and low resolution creates major problem. Even after text extraction from each frame there may be still considerable amount of non-text region in it. The non-text region in the video frame may hamper video indexing. Hence, to minimize the non-text region we aim for word and character segmentation. Word segmentation divides the text region into small portions, which reduces the interaction with background noise. Whereas, character segmentation further divides the words into small portions comprising of single character which in turn further reduces the non-text region in video frames. The video index enables the user to enjoy the videos with specific topics. For example, one can search for the term "Sports News" to get the Sports news of the day. Some examples of video frame with subtitles are shown in Figure 1(a) and 1(b).

2 Literature Review

The extraction of number plate (NP) and segmentation of its character is similar to the segmentation of characters in video subtitles. In [3], the NP extraction and each character segmentation of NP are discussed using different methods. As segmentation is one of the steps in optical character recognition (OCR), the efficiency of OCR increases with increase in accuracy of character segmentation. Hence different methods are proposed to obtain accurate NP character segmentation.
Disadvantage: As the accuracy of OCR depends on the character segment method; the wrongly segmented character can cause misidentification of the character.

The author in [4] considers some key features of the character like Characters are monochrome; Characters are rigid; Characters have size restrictions; the same characters appear in multiple consecutive frames and so on for character segmentation. Text segmentation extracts all pixels out of the video that are part of text characters and discards all pixels which do not belong to characters. The Split-and-Merge algorithm is used to perform segmentation and it is based on a hierarchical decomposition of a frame. Using the proposed method, the accuracy of segmentation obtained is in the range from 96% to 99%.
Disadvantage: Segmentation performance is higher for video samples with moving text and/or moving background than for those where the text and background are stationary. The algorithm cannot profit from multiple instances of the same text in successive frames.

• *Satish S Hiremath is currently pursuing masters degree program in Signal Processing in SIT, Tumakuru, Karnataka, India, E-mail: satishsh36@gmail.com*
• *K V Suresh is currently Professor in Electronics and Communication Department in SIT, Tumakuru, Karnataka, India, E-mail:sureshkvsit@yahoo.com*

(a)

(b)

Figure 1: Examples of video frames with subtitles

Basavaraj A *et.al.* [5] have explained about video text extraction. Authors use morphological edge detection, sobel filter and dilation to extract the text from the gray scale image. The main objective of the text extraction is to reduce the number of false text region to be fed to the OCR.
Disadvantage: The proposed algorithm is insensitive to skew and text orientation.

X Huang *et.al.* [6] presents the method for text extraction in video where the edge map is obtained using the color gradient method and it helps to process the text rows. The adaptive Thresholding and inward filling is used to process the edge map and get contour of the whole text row. Using the proposed method character segmentation is achieved easily.
Disadvantage: The proposed method assumes that the text character has uniform color. Hence, this method is not applicable to the text whose character has no uniform color.

The efficient character segmentation of handwritten devnagari text is proposed in [7]. The authors use the morphological operation dilation and erosion to obtain high accuracy of segmentation of devnagari handwritten characters.
Disadvantage: The accuracy of segmentation of devnagari handwritten characters depends on the preprocessing steps which include removal of noise and distortions of an input image.

The main of aim of the paper in [8] is to provide an appreciation for the range of techniques that have been developed for sementation of text. In this paper, segmentation methods are listed under four main headings. First is "classical" approach consists of methods that partition the input image into sub images, which are then classified and the whole process of portioning is called as "Dissection". The second class of methods avoids dissection, and segments the image either explicitly, by classification of prespecified windows, or implicitly by classification of subsets of spatial features collected from the image as a whole. The third strategy is a hybrid of the first two, employing dissection together with recombination rules to define potential segments. Finally, holistic approaches that avoid segmentation by recognizing entire character strings as units are described.
Disadvantage: The survey paper has not attempted to compare the effectiveness of algorithms, or to discuss the crucial topic of evaluation.

Santosh *et al.* have proposed a method in [9] where the text is segmented in three form i.e., line segment, word segment and character segment. The highest down phaseation approach is used to segment a document image into text lines and using edge detection & connected component based technique the word and characters are segmented.
Disadvantage: The accuracy of character segmentation depends on the binarization of input image.

The paper in [10] presents a two-stage method for multioriented video character segmentation. Words segmented from video text lines are considered for character segmentation. In the first stage, the isolated (non-touching) characters are segmented, and in the second stage the touching characters are segmented. Piecewise Linear Segmentation Lines (PLSL) is used for character segmentation, which allows both vertical and partially curved segmentation paths based on the orientation and slant of the word.
Disadvantages: The proposed method fails to segment the characters if the input image has low resolution, blur and noise in the background.

The authors of [11] propose a method where the character segmentation regions are determined by using projection profiles and topographic features extracted from the gray-scale images. The projection profiles in gray-scale images are defined as:
Let $g(x,y)$ be the intensity of a pixel (x,y) in gray-scale images. Then $g(x,y)$ has the value of range as follows:

$$0 \le g(x,y) \le L-1 \qquad (1)$$

Where L is the level of intensity. Let $H_x(g)$ and $H_y(g)$ be the histograms of column x and row y with intensity g of , respectively. The vertical projection profile, $P(x)$ can be defined as follows:

$$P(x) = \sum_{g=0}^{L-1} H_x(g).c(g) \qquad 0 \le c(g) \le 1 \qquad (2)$$

Where

$$c(g) = \sum_{y=0}^{h} \frac{g(x,y)}{L}$$

(3)

$c(g)$ is a ratio contributing to the projection with intensity of g and h is the height of the image. In similar way, the horizontal projection profile $P(y)$ can be defined as.

$$P(y) = \sum_{g=0}^{L-1} H_y(g).c(g) \quad 0 \le c(g) \le 1 \tag{4}$$

Disadvantages: The proposed method is effective for only touched or overlapped characters.

3 Methodology

The video frames are extracted from the video and the video frame with subtitle is considered for character segmentation. In the methodology, the commonly used procedure for character segmentation is discussed. The flow chart of the method is shown in Figure 2.

Step 1: The video frames are extracted from the video and one of a frame with text is considered for processing.

Step 2: In videos, subtitles appear at the bottom of the video frame. As subtitles are region of interest (ROI), crop the bottom portion of the video frame so subtitles can be used for the character segmentation.

Step 3: The cropped image is resized to twice its resolution, so the detection of character increases.

.

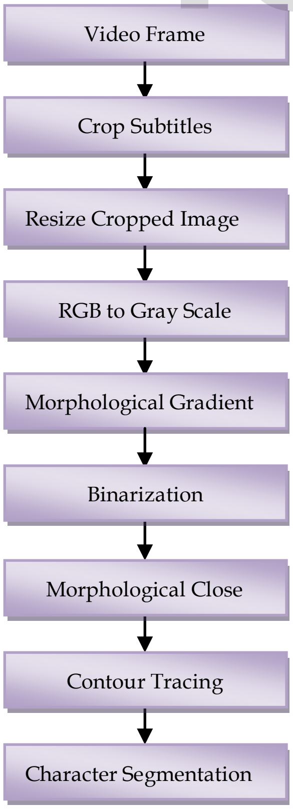

Figure 2: Flow Chart for Character Segmentation.

Step 4: Converting of video frames from RGB to gray scale to make the operations simple.

Step 5: Morphological gradient [12] is the difference between the dilation and the erosion of a given image. It is an image where each pixel value indicates the contrast intensity in the close neighborhood of that pixel. It is most preferred for the segmentation application.

Step 6: After applying morphological gradient to the cropped subtitle image, Otsu Thresholding is applied to convert gray scale image into binary image. The Otsu Thresholding [13] algorithm assumes that the image contains two classes of pixels. It calculates the optimum threshold separating the two classes so that their combined spread is minimal and which in turn maximizes inter-class variance.

Step 7: The morphological close operation [14] is a dilation followed by erosion, using the same structuring element for both the operations. Closing of set A by structuring element B, denoted $A \bullet B$. The morphological close operation is applied on the binarized subtitle image.

Step 8: Contour tracing is a technique that is applied to digital images in order to extract their boundary. After applying the morphological operation, the boundary of the characters is traced.

Step 9: The contour tracing helps to separate each characters. The accuracy of segmentation can be defined as

$$Accuracy = \frac{No.\text{ of Characters Segmented}}{\text{Total No. of Characters}} \tag{5}$$

Another way of measuring the accuracy is using region based [15]. Most semantic segmentation measures evaluate pixel-level classification accuracy. Consequently, these measures use the pixel-level confusion matrix C, which aggregates predictions for the whole dataset D:

$$C_{ij} = \sum_{l \in D} |\{z \in I \text{ such that } S_{gt}^{I}(z) = i \text{ and } S_{ps}^{I}(z) = j\}| \tag{6}$$

Where $S_{gt}^{I}(z)$ is the ground-truth label of pixel z in image I, $S_{ps}^{I}(z)$ is the predicted label and $|A|$ is the cardinality of the set A . In other words, C_{ij} is the number of pixels having ground-truth label i and whose prediction is .j

Using morphological operation characters of video subtitle can be segmented effectively.

5 CONCLUSIONS

The character segmentation using different techniques is discussed in detail in literature survey. The most commonly used method is explained in the methodology section. The accuracy describes the effectiveness of the method in character segmentation of video subtitle text (Graphics Text).

REFERENCES

[1] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look", *IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 2106-2113, October 2009.

[2] Jiamin Xu, Palaiahnakote Shivakumara, Tong Lu, Trung Quy Phan and Chew Lim Tan, “Graphics and Scene Text Classification in Video”, *22nd International Conference on Pattern Recognition (ICPR),* pp. 4714 - 4719, August 2014.

[3] Chirag Patel, Atul Patel and Dipti Shah,”A Review of Character Segmentation Methods”, *Internationl Journal of Current Engineering and Technolgy*, Vol. 3, No. 5, pp. 2075 - 2078, December 2013.

[4] Rainer Lienhart, “Indexing and retrieval of digital video sequences based on automatic text recognition", *Proc. 4th Association for Computing Machinery (ACM) International Multimedia Conference*, pp. 11-20, November 1996.

[5] Basavaraj Amarapur and Nagaraj Patil, “Video Text Extraction from Images for Character Recognition”*, Canadian Conference on Electrical and Computer Engineering*, pp. 198-201, May 2006.

[6] Xiaodong Huang, Huadong Ma and He Zhang, “A New Video Text Extraction Approach”, *IEEE International Conference on Multimedia and Expo*, pp. 650-653, July 2009.

[7] Saniya M.Ansari and Dr. Udaysingh Sutar, “An efficient method of segmentation for handwritten devnagari word recognition”, *International Journal of Scientific & Engineering Research*, Vol. 6, No. 5, pp. 1126-1131, May 2015.

[8] Richard G. Casey and Eric Lecolinet, “A Survey of methods and Strategies in Character Segmentation “, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 18, No. 7, pp. 690 – 706, July 1996.

[9] Santosh, Dr. Jenila Livingston L.M, “Text Detection From Documented Image Using Image Segmentation”, *International Journal of Technology Enhancements and Emerging Engineering Research*,Vol. 1, No. 4, pp. 144-148, 2013.

[10] Nabin Sharma, Palaiahnakote Shivakumara, Umapada Pal, Michael Blumenstein and Chew Lim Tan, “A New Method for Character Segmentation from Multi-Oriented Video Words”, *12th International Conference on Document Analysis and Recognition (ICDAR),* pp. 413 – 417, August 2013 .

[11] Seong-Whan Lee, Dong-June Lee andHee-Seon Park, “A New Methodology for Gray-Scale Character Segmentation and Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 10, October 1996.

[12] Evans.A.N, “Morphological gradient operators for colour images”, *International Conference on Image Processing (ICIP),* Vol. 5, pp. 3089 - 3092, October 2004.

[13] "OpenCV: Image Thresholding (Otsu’s Binarization)", http://docs.opencv.org/master/d7/d4d/tutorial_py_thresholding.html#gsc.tab=0.

[14] Rafel C. Gonzalez and Richard E. Woods, *Digital Image Processing*, Pearson Prentice Hall, 2013.

[15] Gabriela Csurka, Diane Larlus and Florent Perronnin, “What is a good evaluation measure for semantic segmentation?”, *24th British Machine Vision Conference (BMVC),* pp. 1-11, September 2013.