Subtitles Extraction from Video by Stroke Width Method and Clustering

Ms. Rima A. Pitroda
Department of Computer Engineering
Assistant Professor, Aditya Silver Oak Institute of Technology, India

Abstract- *With the increment and wide use of practical vision systems and smart phones, text detection in videos has become a critical yet challenging task. In today's world and advancement in easily available and different featured cameras, text recognition has become a tough work and changes are occurring in this field. Texted content present in the video provides useful information likewise scene understanding, for narration, it adds details to the music playing or video playing. With the expansion of digital libraries therefore there is intense need of a method to retrieving the image using the embedded content in them. The proposed system is to detect text and extract them from the videos containing subtitles. The algorithm proposed here aims to get better subtitle extraction form these texted frames of videos. Our approach is to extract the subtitles from the videos by using edge detection and clustering and stroke width transform.*

Keywords- Text Information Extraction, Caption text, subtitles, Edge Detection, Stroke Width Transform, Editable text

I. INTRODUCTION

Today's world in which we are living is a brain child of latest video and capturing devices like mobile phones, digital cameras with the high resolutions and updated technology. So the progress of these hand held video/images (HID) with and /or without text.

With the advancement of multimedia collection and digital libraries expanding more and more has led the need for some tool that does indexing retrieving the information content in the images or videos. There urgent need for development of algorithms for retrieving the image using the embedded content in them. Hence, there arise a need for text extraction from videos and images. Text extraction from video has many applications like automatic license plate reading, sign detection and translation, mobile text recognition, content based web image search, navigation, text and logo detection in CCTV video feeds. It is also very useful for visually impaired person in daily life to give them access to text and, coupled with the text to speech algorithm, make them read book covers, banknotes, labels on doors, medicine labels and so on.

The video document processing community has classified the text in video frames based on its origin. The text information which is artificially overlaid on the image is often known as 'Subtitles' [3] or 'Caption text' or 'graphic text' (e.g. subtitles in news video, sports scores, etc.). Some researchers also called these texts as superimposed text or artificial text. Whereas, the text which naturally exists in the image is known as 'Scene text' (e.g. text on vehicles, commodities, buildings, and sign boards on roads, etc.). Scene text has additional complexities such as multi-orientation and multi-lingual issues, whereas Caption text is usually horizontal or vertical. Figure 1 (a) is an example of Caption text; figure (b) is an example of Scene text [3]. A large number of techniques have been proposed by researchers towards text detection.

Fig(a) Caption Text

Fig(b) Scene Text

Figure 1. Examples of Video Text

II. RELATED WORK

What is Text Information Extraction?

Text Information Extraction [1][2][3][4] system receives an input in form of a still image or a sequence

of images can be in gray scale or color, compressed or uncompressed, and text in the images may or may not move. Text extraction is generally sub divided into detection, localization, extraction and recognition.

Basic Stages of Video Text Information Extraction

Video text Extraction consist of five stages[1][2][3]. That is shown in figure 2.

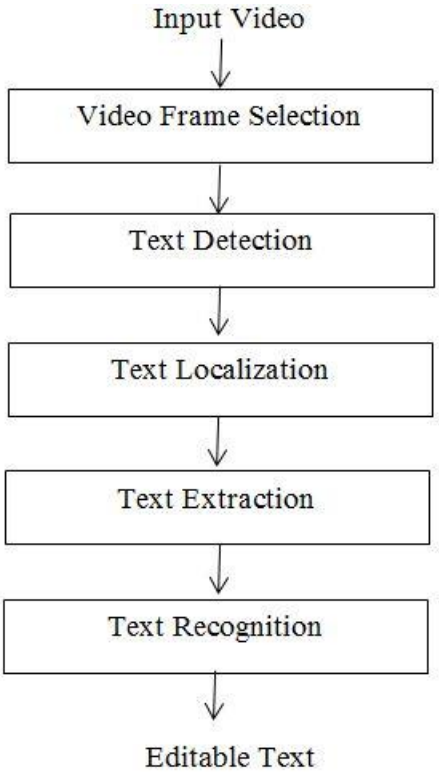

Figure 2: Video Processing Sytem

Text frame selection determines whether a frame contains text information. Text detection and localization finds and defines the actual location of the text present in the frame by formaing bounding boxes around the text. After the bounding boxes are found, the actual charcters are extracted and binariazed for OCR.

Text Information Extraction

Classification of text Extraction techniques [1-8]

The existing methods for detection and extraction of text from video are broadly classified as

- **Thresholding based**

It would define a global and local threshold for the whole image and for a selected portion of the image respectively. The Thresholding based method is further classified into,

Histogram-based one of the most widely used techniques for monochrome image segmentation. These methods work well with low computational resources but are applied mostly on gray-scale images or color channels

Adaptive Techniques define several thresholds for different image parts depending upon the local image characteristics. Adaptive may handle more degradation (uneven lighting, varying colors) than global ones but suffer to be too parametric which is not versatile.

- **Group Based**

Region Based It is further classified into two groups: top-down and bottom-up. Top-down will consider the entire image and then will move towards the smaller parts by considering the grayscale value. Bottom-up approach on the other hand will start with the smaller parts and then will merge into a single image. It is expensive and too slow.

Learning Based mainly makes use of neural networks. Some of the classifiers it makes use of are Multilayer perceptron (MLP), single layer perceptron (SLP). A training database is needed and with the wide range, this task is difficult to realize. Moreover it implies storage problems and labeling of the whole training database before being effective.

Clustering Based approach group color pixels into several classes assuming that colors tend to form clusters in the chosen color space. They belong to unsupervised segmentation while learning-based approaches belong to supervised segmentation. The most popular method is k-means but its generalization, Gaussian Mixture Modeling (GMM), is more and more exploited. But all of the clustering methods explored till now are computational slow. Hence we have tried to introduce a new technique called BVQ which is explained in following section

According to the above analysis, we find that most of previous works cannot completely extract the text when the text regions have complicated background. Thus, we present a novel text extraction method based on strokes[2][7] and BVQ clustering.[8][10]

A. Clustering Technique

Clustering techniques is generally categorized majorly as Simple and Hierarchical. Again Simple Clustering is subdivided as K nearest (Kn), Mixture of Gaussian, K means

Clustering. Hierarchical is subdivided as Partitive and Agglomerative. One of the methods of Partitive Method is as follows.

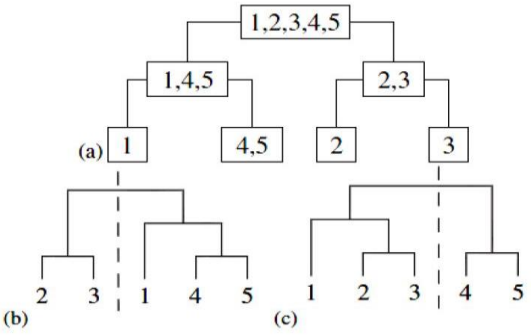

Figure 3. Partitive clustering tree and dendogram

Combines tree-structured vector quantization and partitive k-means clustering. Less sensitive to data preprocessing and data normalization. Strong similarity to those of self organizing maps(SOMs). Significant reduction in the computational load is achieved. Time required for implementation of the algorithm is much less than K-means and GMM which are extensively used for Text Information Extraction.

In this method firstly we will read the image and then compute its covariance and Eigen values and Eigen vectors and initialize the cluster. The cluster with highest Eigen value is selected and split into two parts. And display the newly formed cluster.

B. Stroke Width Transform

Another class for text Extraction is Stroke based method [2][7].First, we extract the gradient image of the text rows and get its edge map. Then we use Laplacian method and edge map to retrieve the character stroke image of text row. Finally, based on Laplacian image, we get the candidate text extraction image. Then, based on the character stroke image, we perform the character filtering on candidate text extraction image to get the final text extraction results.

III. IMPLEMENTATION AND RESULTS

The algorithm presented for subtitles extraction from video. After the detection and localization of text the next part to recognize the text from the videos or images OCR method is used. Texts are recognized by template matching algorithm where after creating bounding box and getting individual characters. These characters are then resized to match with the templates of alphabets and numbers. The characters with highest match are found and are displayed to the screen.

Algorithm for Subtitles Extraction from the video.

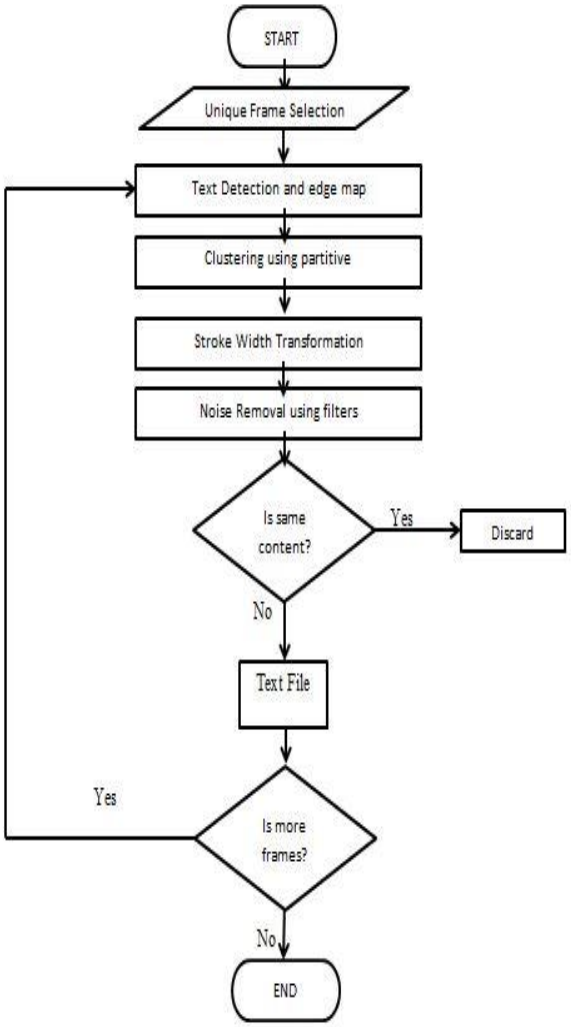

Figure 4. Flow Chart

The first part to work with videos is to get the frames from the videos and saving them to some f-\\older. These frames are saved as a sequence of images. The following screen shots show the frames that are saved to a folder.

Figure 5. Frames Extracted to a folder

Figure. 6 SWT Map

Figure 7. Bounding Box

Performance Parameters

Three metrics are commonly used to report results: recall, precision, and false alarm rate. Recall, shown in equation is simply the number of correct detects over the total number of targets in the ground truth. A perfect score for Recall is 100%. Precision, shown in equation is the number of correct detects over the total number of detects reported by the algorithm. Ideally this would also be 100%. Finally false alarm rate is the number of false alarms over the number of correct detections, shown in equation. Ideally this would be zero.

$$Recall = \frac{correct\ detects}{correct\ detects + missed\ detects}$$

$$Precision = \frac{correct\ detects}{correct\ detects + false\ alarms}$$

$$False\ Alarm\ Rate = \frac{False\ Alarms}{Correct\ Detections}$$

TABLE I. Computed values of Precision Rate and Recall Rate

Test Data	Precision Rate	Recall Rate
Commercials	80.5	86.5
Sports	85.6	92.4
Average	86.05	89.45

IV. CONCLUSION

From the above study of various papers regarding text extraction about subtitles and its increasing importance in today's advanced world were most of information are in multimedia form. Hence there is acute need so some mechanism that could able to extract data from this multimedia which is also known as content based retrieval. To carry out our research objectives we examine first the extraction of still subtitles and then on scrolling. Extracting those texts from subtitles and saving them to some text document. To fulfill the objective of this research, first we have done preprocessing steps of texted image where we could get text characters extracted successfully and converting to text documents. In, next step frames are extracted and key frames are also extracted from the videos and on single frame first by partitive clustering method and then by stroke width method and segmentation is being done and converted to text documents.

REFERENCES

[1] K. C. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: A survey," Pattern Recognition., October 2003, Elsevier.

[2] Xiaodong Huang, Qin Wang, Lishang Zhu, Kehaua Lia "A New Video Text Extraction Method Based on Stroke", 6th International Congress on Image and Signal Processing, IEEE 2013

[3] Nabin Sharma, Umapada Pal, and Michael Blumenstein "Recent Advances in video based document Processing: A Review" 10th IAPR International Workshop on Document Analysis System, IEEE 2012

[4] Mukesh Kumar, Sonam, "Implementation of MD Algorithm for Text Extraction from Video" , Nirma University International Conference on Engineering, IEEE 2013

[5] Xie Guang yi, " Automatic Caption Extraction of News Video and its Implementation" ,IEEE 2012

[6] Teo Boon Chen, D. Ghosh, S. Ranganath, " Video Text Extarction and Regocnition",IEEE 2004

[7] Boris Epshtein, Eyal Ofek, Yonatan Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform", Microsoft Corporation

[8] M. Sultan, D.A.Wigle, "Binary tree structured Quantization approach to clustering and visualizing microarray data".

Web References

1. http://webaim.org/techniques/captions/
http://en.wikipedia.org/wiki/Cluster_analysis