Text Extraction and Categorization of Videos using Image Processing Techniques and Naïve Bayes Classifier

Arunabha Basak[1] Santhi V[2]

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

Email: basakarunabha@gmail.com, vsanthi@vit.ac.in

Abstract

Internet has become a very important part of our lives today. We can search information regarding anything which will be presented to us in various formats like texts, images and videos. There are many different categories of videos present on the internet today such as news, sports and music. But searching for a particular category of video out of all the videos present in a particular dataset becomes a tough and time consuming job. Hence our objective is to categorize the videos into different groups so that it becomes easier for searching or analyzing a certain category of videos. Among the various categories of videos news and cricket videos are searched by almost everyone. So it would be very useful if we are able to categorize them. In this paper, we have proposed a method for categorization of videos by making use of the textual content that is present in the video frames. The dataset involves videos of cricket and news and the average length of the videos is 20 seconds. To categorize the videos according to the text present in them, we need to first extract and store the text from the video frames into a machine readable format. After that we use a machine learning algorithm to categorize the text into their respective groups.

Keywords. Text detection, edge detection, thresholding, dilation, contours, optical character recognition, text classification.

1. INTRODUCTION

After the discovery of internet, we are surrounded by an extensive amount of multimedia content. Information is present on the internet in various forms like texts, images and especially videos. With the rapid growth of internet and a lot of studies performed in compression technology, massive amounts of information in video format are present over the internet today. With every passing day, thousands and thousands of videos are getting uploaded over the internet through various websites. For this reason it has become essential to put videos into their respective categories for easy browsing, indexing, searching and analyzing so that important information can be extracted or computed from those video data [1].

There are two particular categories of videos every Indian loves to watch - one being news and the other cricket. Categorizing news and cricket videos into two classes of videos will

be very useful. To index videos we need information about the content present in them. In the case of news and cricket, this content can be obtained from the textual information appearing in the video frames. We have proposed a model where video classification has been done based on textual content present in the video frames. The dataset used in this model consists of videos of news and cricket. We can gain a lot of knowledge by studying the textual content in the video frames [2]. Usually text present in a video about news is different than the content present in cricket videos. The entire process of video categorization consists of two major steps – first being the text detection and extraction phase and the second being the text categorization phase.

2. LITERATURE SURVEY

Over the past few years, multiple techniques have been designed to detect text in images and videos but they all have certain limitations. In the color-edge combined algorithm for text extraction, first the background is removed which leaves only the text and then the text is extracted using OCR [2]. However this method cannot extract multilingual texts. In another experiment for detecting and localizing text from videos, two edge maps were first created for the edge detection stage [4]. For this Sobel operator was applied to the entire frame. Sobel edge detection can be used to detect the discontinuities in image as well as to apply a smoothing effect to the image. From the results, it was found that a vast majority of the falsely detected text regions were filtered out. Text detection can also be performed using the USTB_TexStar algorithm [5]. Even though this approach was found to be practical and effective in automatic text detection and tracking, the main problem was the motion of the objects or the camera, which makes it hard for the feature extraction process. In the MSER (Maximally Stable Extremal Regions) technique for real time asynchronous text detection, two separate modules were used for real time text detection, one consists of a Multi-Script scene text extraction algorithm related text detection and the other is an MSER based tracking module [6]. The main limitation of this method is the degradation of tracking in the presence of severe motion blur or strong illumination changes. To recognize text regions we need to first localize and isolate the text regions so that OCR can be performed on it. Motion blurs and illuminations can make text tracking difficult as well.

For classification of text many supervised machine learning algorithms exist such as K-NN clustering and Naive Bayes classifier. However the problem with K-NN is that it takes a lot of computational time since we need to iterate the process for multiple values of n to find which gives the best result[1]. Another experiment was performed to recognize song performers based on the lyrics of the song using the Naive Bayes classifier [7]. The accuracy range obtained in the prediction chart for the experiment was in between 0.8 to 0.9.

Naïve Bayes efficiency in prediction and modeling is a major advantage over other classification algorithms and it can handle large datasets and attributes with ease.

3. PROPOSED METHODOLOGY

Most of the videos in news and cricket category contain valuable information in them and they can be used to determine the category in which the videos belong.

The entire process of video categorization is divided into two phases. First is the text detection and extraction phase where the video will be taken as input and the text present in

the video frames will get extracted into a machine readable format. The second phase consists of the text classification where we will use the Naive Bayes algorithm for classifying the text into their respective categories. We use a series of image processing techniques to detect the text regions in each input video frame. After identifying the regions we isolate them from the rest of the background after which we can apply Optical Character Recognition, or OCR, to recognize and extract the text. The processes to be performed for the text detection and extraction phase are listed in the flow diagram.

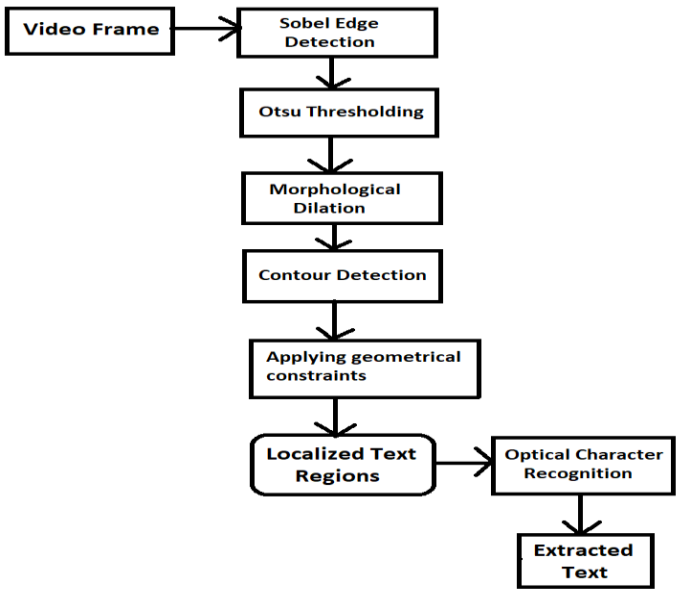

Fig. 3.1. Flow diagram for text detection and extraction.

3.1. Text Detection and Extraction

First we need to detect the locations of text in a frame and extract it. We would first try to identify the regions containing text in each frame and isolate it from the rest of the background. The Videos can be treated as a sequence of video frames. After processing a video frame, we will skip the next hundred frames since most of the consecutive frames will contain the same text content. The text region has sharp differences in intensity between its borders and background. Hence, edges are present in between the boundary of a text region and the background. For this reason edge detection needs to be done. Also a lot of information such as corners and curves can be obtained from the edges which make it easy to identify shapes of objects.

We have used the Sobel edge detector to find the edges of the video frames. Basically it calculates the absolute gradient value at each pixel of the input image.

Fig. 3.2. Sample input video frame

This operator makes use of two kernels. Each has a size of 3x3 and each is used for computing the approximations of the derivatives in horizontal and vertical directions. The final output is a 2 dimensional map which displays the gradient calculated at each pixel of the image.

Fig. 3.3. Result after Sobel edge detection in x direction

After Sobel Edge Detection, the next step is to separate the regions of interest from the regions where no edges were detected. It is important to segment the two parts to have a proper division between regions or segments of the video frame on the basis, that they have textual content in them or not. To solve this problem we use Otsu Thresholding.

In Otsu Thresholding, the threshold value that maximizes the inter-class variance is calculated mathematically. It basically minimizes the sum of intra-class variances present in the background as well as foreground pixels and the result is an optimum threshold value. The final threshold value is the average of the means computed for the two classes.

Fig. 3.4. Result after Otsu thresholding

After the edges have been detected and the text and non text regions have been segmented, connection of the edges needs to be done to create individual blocks. For this morphological dilation is used. Morphological dilation can be defined mathematically with an example. If X = {all Euclidean coordinates of binary image}, E = {coordinates of structuring element}, Kx =translation of K by which origin remains at x then X's dilation by K will be a set of points such that:

$$K_x \cap X \neq \emptyset \tag{3.3}$$

Fig. 3.5. Result after applying Dilation operation

After the Dilation process, the contour extraction process needs to be done. Contours are the boundaries of the text region with same intensity. The contours can be used for detecting and recognizing objects and also for analyzing object shapes. After dilation operation, the detected edges of text characters are combined to form blocks which are basically white in color. Hence the overall procedure involves separating the white color blocks which contain text in them form the background that is black in color. Each contour is a rectangular box which contains text. They give us information about the locations where the textual content is present in a video frame. The second is the retrieval mode of contour and the third is the approximation method. Every contour is basically a set of coordinates that indicate the boundary of the text regions.

Fig. 3.6. Contour Extraction

After the completion of the dilation process and detection of contours, the bounding boxes need to be extracted. Among the contours detected, some contours will contain regions that are wrongly detected as text. Also for our project we assume that all the text that appear the video frames are in horizontal direction hence we only focus on the horizontal bounding boxes and ignore the vertical ones. To make the results more accurate, we make use of certain constraints, or conditions, which specify the properties of the extracted rectangles. These are called geometrical constraints. Extraction of the text is done by applying Optical Character Recognition (OCR) on it. OCR helps to convert the text which is form of a video frame segment and an image format to a machine readable format such as at text file in txt format

3.2. Text Categorization

After Text extraction we collect a dataset of text files corresponding to the Videos containing the text that was present in videos. Before the categorization process takes place we need to divide our text dataset into two divisions – Train and Test. The train dataset will be used to train our model and the test dataset will be used to check how well our model performs.

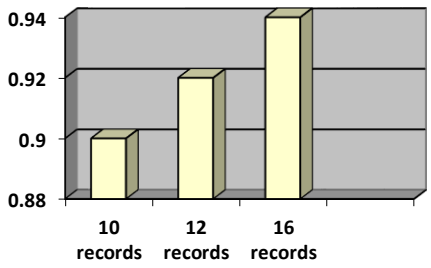

Fig. 3.7. Accuracy of classification for 10, 12 and 16 records in training dataset

4. CONCLUSION

A model for video categorization based on textual content present in video frames has been presented on this project. Using various image processing techniques we were able to extract a lot of text from the frames which helped us gather information on the videos and categorize

them. The results obtained after preprocessing had a better outcome over the efficiency of classification. We were able to achieve an accuracy of 0.94 using this classification process.

This project can be expanded to include various other categories of videos. A lot of videos are available over the internet and if the video frames contain text, then video categorization can be performed for each and every one of those videos.

5. REFERENCES

[1] Zulfany Erlisa Rasjida , Reina Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques", 2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13- 14 October 2017

[2] Xin Zhang, Fuchun Sun, Lei Gu ,"A Combined Algorithm for Video Text Extraction" , Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)

[3] Sebastian Romy Gomes, Sk Golam Saroar, Md Mosfaiul Alam Telot, Behroz Newaz Khan, Amitabha Chakrabarty, Mr. Moin Mostakim, "A Comparative Approach to Email Classification Using Naive Bayes Classifier and Hidden Markov Model", Proceedings of the 2017 4th International Conference on Advances in Electrical Engineering (ICAEE),2017.

[4] Tsung-Han Tsai, Yung-Chien Chen, Chih-Lun Fang, "A Comprehensive Motion Videotext Detection Localization and Extraction Method", 2006.

[5] Ze-Yu Zuo, Shu Tian, Wei-yi Pei, XU-Cheng Yin, "Multi-Strategy Tracking Based Text Detection in Scene Videos", 2015 13th International Conference on Document Analysis and Recognition,2015.

[6] Llu´ıs Gomez and Dimosthenis Karatzas, "MSER-based Real-Time Text Detection and Tracking", 2014 22nd International Conference on Pattern Recognition,2014.

[7] Dalibor Bužić , Jasminka Dobša, "Lyrics Classification using Naive Bayes", MIPRO 2018

[8] Shoji Morita, HitoshiTabuchi, Hiroki Masumoto, TomofusaYamauchi & Naotake Kamiura, "Real-Time Extraction of Important Surgical Phases in Cataract Surgery Videos", The Author(s) 2019

6. BIOGRAPHIES

Arunabha Basak received the bachelor's degree in computer engineering from Jalpaiguri Government Engineering College in 2019, the master's degree in computer engineering from Vellore Institute of Technology in 2022 in the department of Computer Science and Engineering. His fields of interests include data structures, programming, machine learning and image processing.