AN IMPROVED TEXT EXTRACTION ALGORITHM FOR INTELLIGENT SEARCH IN VIDEOS

1ANJALI T, 2SIJI RANI S

1,2Department of Computer Science, Amrita Vishwa Vidyapeetham University, Amritapuri, Kollam, Kerala, India
E-mail: anjaliraj87@gmail.com, sijiranis@am.amrita.edu

Abstract- With the extensive advancement in multimedia technologies across the globe, usage of internet archival and storage of multimedia files has become a challenge. Accessing specific segments of videos is the competent problem addressed in this paper. A solution to this problem requires designing an efficient video retrieval tool with indexing, content analysis, and intelligent search in order to retrieve the needed video portions by the users. The scope of this solution is to develop a web based application for detecting text and captions from videos. Considering the fact that corners exist more in text and captions, corner based approach methodology is used. The moving caption patterns are detected by the combination of text and motion features. Proposed solution is built in such a way so as to capture the text and motions in any direction across the screen. Text recognition is done by using Optical Character Recognition. Miss rate, recall rate and false rate are the evaluation metrics used to evaluate the efficiency of this work. Experimental results have shown that the proposed approach is very efficient in text detection in comparison with the existing real time ones.

Keywords- Text detection, Harris corner detector, Moving caption, Optical flow, And Video retrieval.

I. INTRODUCTION

Consumers’ video consumption habits are maturing and becoming more sophisticated. As a consequence, detecting text and caption from videos is gaining importance in video retrieval, annotation, indexing, and content analysis. The availability of large multimedia databases demands more efficient and scalable algorithms to retrieve, archive, index and locate desired content in them. During the past decade, lot of techniques have been developed to extract the semantic content from images and videos. Among these techniques, text detection based approaches are of particular interest to many applications due to the rich information content.

Detecting text from videos is a challenging task which often suffers from appearance variations of text, low contrast and complex background. The techniques that are used to extract texture features include support vector machine (SVM), Wavelet, FFT, spatial variance, and neural networks.

Scene text and artificial text are two kinds of text we generally find in any video. Scene text pertains to the environment which generally comes along with the capturing of video.

Artificial text is added at the time of video editing that carry information about semantics of the video content which is considered essential information, for example in news broadcasts we can see the text information on screen carrying critical information about when, where, who of the event. In sports event telecasted on the screens, the details of the players and information on their play are displayed on the bottom of the screen. In summary videos of educational courses, the text shown at the end is important for users to recollect what they have learnt and thus provides important guidelines. Video captions are usually highly compact and structured; however, it is still a challenging task to design a robust algorithm that performs well with different types of videos. Automatic annotation and indexing of videos are some real world applications of this system.

This paper propose an improved and unique search methodology based on “corner detection approach” to make the data retrieval simple, user friendly and to minimize the storage problems. This methodology is tested in various fields like film industry, sports world, news channels, educational videos etc. It helps to detect the text scrolling across the screen in multi-directions. It enables the user to focus on the portion of interest in the video. The paper discusses the design and implementation of an improved algorithm to search the content of interest from the saved videos.

II. RELATEDWORK

There exist a number of algorithms to extract the caption texts from the still images. These algorithms used the following properties of texts to determine the captions.

1. Characters are strictly bounded in fixed size.
2. A text line will be aligned horizontally which contains cluster of characters.
3. Text usually appears in a bright font, relative to the background screen.

In the following section we are trying to specify the methods and problems related to existing video retrieval tools based on the detection of text. The robust detection of the textual blocks in
images/frames is an important step for categorizing video clips based on the textual content. One of the easiest methods for localizing the text in a particular image is the binarization techniques which uses the concepts of global, local, or adaptive thresholding.

There are six main text detection methods namely caption detection using discrete cosine transform, color similarity based, texture based, connected components based, edge based methods and continuous random field based method.

A. Caption Detection Using Discrete Cosine Transform (DCT)
Caption localization method to extract text from images utilizes the unique texture and geometrical arrangements presented in a text line/block. This is captured by DCT coefficients to detect text regions that directly operate in the DCT domain of JPEG images and MPEG video sequences. It mainly utilizes the unique texture characteristics of the text and locates the caption text as regions of high horizontal intensity variations. The texture features are extracted directly from the DCT domain using the quantized DCT coefficients. These coefficients capture the local frequency information. However, Color information is not utilized in this method. Also this method does not efficiently distinguish text and non character regions with spatial orientation similar to text. So result of text detection gives poor result.

B. Colour Similarity Based Caption Detection
In this method the text is grouped into clusters. Some of the commonly used clustering methods are K-means clustering Gaussian mixture model (GMM), density based etc. In these methods captions should have a large color contrast from their background, while the intensity contrast should be small. The use of color information can improve the performance of the detection.

C. Texture Based Caption Detection
Texture based approaches are efficient in dealing with complex background with dissimilar textual structure to the text regions, the text regions are treated as a special type of texture with distinct textural properties. But the computational complexity restricts its applications in large databases.

D. Connected Component Based Caption Detection
In this method an image is segmented into connected components and small components are merged successively into larger ones. Finally, merged components are classified as texts. Connected component based approach is efficient when the background mainly contains uniform regions. This kind of approaches encounter difficulties when the text is noisy, multicolored and textured.

E. Edge Based Caption Detection
This method uses the structural and geometrical properties of the text and character. They are effective in detecting text regions if other parts in the image do not have too many strong edges.

F. Continuous Random Field Based Caption Detection
This method is used primarily for document analysis. It takes into account both locational and contextual features of the characters. These features are input into a Multi Layered Perceptron. This method has been found effective when the character density is large as in large documents and when the characters are arranged uniformly.

III. PROPOSED SYSTEM DESIGN

The proposed design for video detection assumes that there exist dense and orderly presence of corner points in characters, especially in text and captions. In the existing methodologies, detailed analysis and description to the shape properties of detected regions are deficient. The paper describes the text regions with the discriminative features, from which the non-text regions formed by the corner points appeared in the background can be filtered out efficiently. The usage of corner points generates more flexible and efficient criteria, under which the margin between text and non text regions in the feature space is discriminative.

In our algorithm, we use Harris Corner Detector to extract the corner points from images. In order to efficiently detect texts, we need to analyse the discriminative properties of text and its basic unit character, as well. Multi-directional text can also be detected by modifying the Harris algorithm. Moving text detection is based on the Lucas-Kanade algorithm. A high level architectural framework of the proposed system is given in Fig 1.

The core functional components of this system are corner detection and feature description. The frames are captured from the video. Then the corner points of each frame are extracted using Harris Corner Detector.

![Fig 1: Frame work of video retrieval System](image)

The non-text region is filtered out with the help of discriminative features like area, saturation etc, there by finding the text region even if it is horizontal.
vertical or inclined. Motion caption is the combination of both text features and motion features. On computing the optical flow we get the motion features of the videos. These caption and text are recognized using an OCR.

A. An Algorithm for Text Extraction from Videos
The algorithm consists of two main steps; detection of text and detection of moving caption. Each of the steps is discussed below.

B. Detection of Text
Corner Point Extraction Corners are one of the essential features, which is generated when the edges intersects each other. Corners are more stable, robust feature than any other low level features like size, colour etc. In text region corners are usually in ordered pattern whereas in non-text region corners are distorted. This method is used very much in the fields like Motion detection, Image matching, tracking image mosaicing and so on. The Harris Corner Detector is a popular interest point detector due to its strong invariance to rotation, scale, illumination variation and image noise. This approach is based on the local auto-correlation function of a signal, which helps to measure the local changes of the signal with the patches shifted by a small amount in different directions. The principle behind corner detection is that when a small window is placed over an image and moved, depending on the orientation of the window with respect to the image, the intensity change occurs in different ways. For example, if the window is over a flat area of the image and moved, there would not be any change in intensity in any direction. If the window is placed over an edge and moved, intensity changes in only two directions. However if the window is placed over a corner and moved, there would be an intensity change in all directions.

Let I be a grayscale 2-dimensional image. Consider a image patch over an area \((u, v)\) and shifting it by \((x, y)\). The weighted sum of squared differences (SSD) between these two image patches can be denoted by \(S\), as given below

\[
S(x, y) = \sum_\alpha \sum_\beta w(\alpha, \beta)(I(u + x, v + y) - I(u, v))^2
\]

This equation \(I(u + x, v + y)\) can be approximated using Taylor expansion. For a given \(I\) let \(I_x\) and \(I_y\) be partial derivatives of \(I\) such that

\[
I(u + x, v + y) \approx I(u, v) + I_x(u, v)x + I_y(u, v)y
\]

This produces the following approximation

\[
S(x, y) = \sum_\alpha \sum_\beta w(\alpha, \beta)(I_x(u, v)x + I_y(u, v)y)^2
\]

This can be written in matrix form as below:

\[
S(x, y) = (x, y)A\begin{bmatrix} x \\ y \end{bmatrix}
\]

Here \(A\) is referred as the structure tensor,

\[
A = \sum_\alpha \sum_\beta w(\alpha, \beta) \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} = \begin{bmatrix} \langle I_x^2 \rangle & \langle I_xI_y \rangle \\ \langle I_xI_y \rangle & \langle I_y^2 \rangle \end{bmatrix}
\]

The above matrix is a Harris matrix and the angle brackets denote average. If a circular window is used, then the response will be isotropic [16][1]. A corner will be characterized by a large variation of \(S\) in any directions of the vector \((\alpha, \beta)\). By analysing the eigenvalues of \(A\), the characterization can be expressed in the following ways: The matrix \(A\) should strictly have two large eigen values \(\lambda_1\) and \(\lambda_2\) for a corner point on the account of magnitudes of the Eigen values, the following properties can be made based on these arguments [1].

1. If \(\lambda_1 \approx 0\) and \(\lambda_2 \approx 0\) then this pixel \((\alpha, \beta)\) has no corner point features.
2. If \(\lambda_1 \approx 0\) and \(\lambda_2\) has large positive value, then an edge is present.
3. If \(\lambda_1\) and \(\lambda_2\) gives large positive values, then a corner is present.

The exact computation of the eigenvalues is quite expensive, since it requires the computation of a square root. Hence they suggested the following function \(M_k\), where \(k\) is a tunable sensitivity parameter [1].

\[
M_k = \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2 = \det(A) - k \text{trace}^2(A)
\]

Here \(M_k\) is positive in the corner region, negative in the edge region, and small in the flat region.

Fig 2 and Fig 3 shows the corner detection using Harris corner detector.

![Fig 2: (a) original video frame, (b) corners detected frame, (c) background removed frame.](image)

![Fig 3: (a) original frame (b) corners detected frame.](image)
In Fig 4 and Fig 5 text region is extracted using the following discriminative features:
1. Area: The area of a region is defined as the number of foreground pixels in the region enclosed by a rectangle bounding box. Area is the basic feature for text detection.
2. Saturation: According to our context, the saturation is the proportion of the foreground pixels in a bounding box that belongs to a region.
3. Aspect Ratio: It can be defined as the ratio of its width to its height. In videos, text and captions are usually placed regularly along the horizontal axis.
4. Position: We describe the position of a region with its centroid. The position information will help to locate the text regions with its specific type and style.

C. Inclined Text Detection
Usually, characters are written in a straight line. However, there are cases where there is an inclination to the character. To detect inclined text, the inclination angle with the x-axis also should be considered while detecting corners. In case of a considerable change in intensity while moving the plane, and a corner not being detected, the plane needs to be inclined to check if a valid corner can be detected, and if found, analysis is to be done in that angle so as to get a character string. In Fig 6 we took an image frame having inclined text inside a video. Based on a threshold value we detected the corners of the inclined text. A threshold value is required to distinguish between corners and interior points. In practice, the threshold must be set high enough to avoid the detection of false corners which may have a relatively large corner value due to noise.

IV. EXPERIMENTAL RESULT.
A. Dataset
Sample dataset selected includes film, spiritual video, news, and video lectures. The motivation behind selection of different videos is that each video has different features like horizontal text, moving caption, vertical text etc. Harris algorithm was implemented in a serial way using Visual Studio and C# .Net 4.0. All tests were done using an image with dimensions 3872x2592. The format of the extracted images is JPEG. In the experiments, our text detection approach is tested in both static and dynamic scenario.

B. Evaluation metric
The performance of the method has been tested with a number of different video clips that contain various embedded captions and texts. Recall rate, Miss rate, and false rate are the three parameters used for the evaluation of our algorithm. Recall rate defines the percentage of frames that successfully detected the text as text. Miss rate is calculated by taking the ratio of missed detection frames to the total no of frames which contain text. False rate is the percentage of frames that detected non-text as texts.

C. Result
Table 1 shows the missed detection rate of our tested videos. Here we tested four types of videos having different number of frames. Out of the input videos
we need to detect the frames which contain the text and which does not contain the text. Non-text frames are falsely detected when videos have large no of corners. In such cases non-texts are identified as texts while extracting the text region using discriminative features.

Frames contain blurred images increases the missed detection rate since the corner detection is not possible. With our proposed approach, more than 77% of text regions have been detected correctly with a false detection rate lower than 21% in the case of spiritual video. Table 2 compares our results with that of the existing one. In Harris algorithm only horizontal texts were detected. Most of the videos contain text in all inclinations like vertical, skewed etc. So by applying discriminative features in different ways we could extract all types of text. Here we tested the same in spiritual video containing inclined text. Table 3 depicts the details of recall and false rate. The reduction of false rate in our approach compared with that of Harris’s proves the validity of the combining of text and motion features. Execution time varies with each input video and it depends on the no of frames. intensity difference of two consecutive frames is also presented. After detection of text, the recognition part is done by optical character recognition. Our next focus will be developing a digital library application that can retrieve specific video parts, based on these results. Future work would include exploring the use of this algorithm for languages other than English.

### TABLE 1: Missed Detection Rate

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total number of frames in the video</th>
<th>Number of frames which contain text</th>
<th>Correctly detected text frame</th>
<th>False detected text frame</th>
<th>Missed detection</th>
<th>Miss rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film</td>
<td>350</td>
<td>259</td>
<td>221 (85.3%)</td>
<td>23 (8.8%)</td>
<td>15</td>
<td>5.7</td>
</tr>
<tr>
<td>Spiritual video</td>
<td>280</td>
<td>150</td>
<td>129 (86%)</td>
<td>14 (9.3%)</td>
<td>7</td>
<td>4.6</td>
</tr>
<tr>
<td>News</td>
<td>300</td>
<td>280</td>
<td>238 (85%)</td>
<td>25 (8.9%)</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Lecture video</td>
<td>320</td>
<td>290</td>
<td>245 (84.4%)</td>
<td>24 (8.2%)</td>
<td>21</td>
<td>7</td>
</tr>
</tbody>
</table>

### TABLE 2: Comparison with Harris approach

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total number of frames in the video</th>
<th>Number of frames which contain text at all inclination</th>
<th>Correctly detected text frame</th>
<th>False detected text frame</th>
<th>Missed detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using Harris</td>
<td>280</td>
<td>150</td>
<td>116 (77%)</td>
<td>21 (14%)</td>
<td>13 (8.6%)</td>
</tr>
<tr>
<td>Extension of Harris</td>
<td>280</td>
<td>150</td>
<td>129 (86%)</td>
<td>14 (9.3%)</td>
<td>7 (4.6%)</td>
</tr>
</tbody>
</table>

### TABLE 3: Recall and false rate of proposed method

<table>
<thead>
<tr>
<th>Video Type</th>
<th>Recall rate (%)</th>
<th>False rate (%)</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film</td>
<td>85</td>
<td>8.8</td>
<td>1.87</td>
</tr>
<tr>
<td>Spiritual Video</td>
<td>86</td>
<td>9</td>
<td>1.2</td>
</tr>
<tr>
<td>News</td>
<td>85</td>
<td>8.9</td>
<td>1.01</td>
</tr>
<tr>
<td>Lecture Video</td>
<td>84.5</td>
<td>8</td>
<td>1.2</td>
</tr>
</tbody>
</table>

### CONCLUSIONS

In this paper, an automated video text and caption detection and video retrieval system developed based on corner points present in the video frame is discussed. It shows that the performance of text detection can be improved by incorporating the discriminative features of corner points. An algorithm for finding the moving caption based on the
ACKNOWLEDGMENT

We sincerely express our gratitude for the motivation, research solutions and facilities provided by Satguru Sri Mata Amritanandamayi Devi, the Chancellor, Amrita University. We would like to thank Dr. M. R Kaimal, Mr Swaminathan, and other faculty members in Computer Science Department for providing valuable comments and support.

REFERENCES

[1] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Member, IEEE, Yuxiao Hu, Member, IEEE, Yuncai Liu, Member, IEEE, “Text From Corners: A Novel Approach to Detect Text and Caption in Videos”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 20, NO. 3, MARCH 2011