——ONLINE FIRST(NOT Peer-Reviewed)——

RECOGNITION OF CHARACTER FROM VIDEO SUBTITLES

Venkatesh H

Department of Digital Communication Engineering, Siddaganga Institute of Technology, India
E-mail: Venkateshdhattani@gmail.com

Abstract

An important task in content based video indexing is to extract text information from videos. The challenges involved in text extraction and recognition are variation of illumination on each video frame with text, the text present on the complex background and different font size of the text. Using various image processing algorithms like morphological operations, blob detection and histogram of oriented gradients the character recognition of video subtitles is implemented. Segmentation, feature extraction and classification are the major steps of character recognition. Several experimental results are shown to demonstrate the performance of the proposed algorithm.

Keywords:

Character Recognition, Classification, Complex Background, Feature Extraction, Segmentation

1. INTRODUCTION

With rapid growth in multimedia data base, content based indexing has gained interest among several image processing experts in present era. As videos carry lot of information, among which semantic information is provided by the text present in it. In experimental results obtained by Judd et al. [1], it was proved that text convey more information. The text present in the videos is of two types: (1) Scene Text and (2) Graphics Text. The text added externally (graphics text) to the video speaks about content of the video. Subtitles are the graphics text present in the videos which help in content based video indexing. The process of analysing the subtitles involves three steps: (1) Segmentation of characters present in the subtitles, (2) Feature extraction from segmented characters and (3) Classification of characters.

The major challenges present in extracting subtitles of the video are complex background, illumination and low resolution. There may be substantial amount of non-text area present in each video frame even after text extraction from it. Due to the presence of non-text area in each video frame, video indexing may be hampered. Therefore, character segmentation is implemented to reduce the non-text area. After segmentation process, the information of the character can be known by extracting important features from it. A feature is a part of information which is useful for solving the computational task associated to certain applications. Feature extraction includes reduction in the amount of resources required to describe a large set of data. As the features of each character are different from one another, it can be used as classification parameter. The classification of characters is same as recognizing the characters. The recognition of characters of subtitles helps in the classification of the videos such as videos with English subtitles and videos with other language subtitles. The Fig.1(a) and Fig.1(b) show the example of video frames with graphics text i.e., subtitles.

The paper is organized as follows. Section 2 briefly reviews the related work. Section 3 describes about the methodology involved in character recognition of video subtitles along with experimental results. Section 4 draws the conclusion.

(b)

Fig.1. Examples of video frames with subtitles

2. RELATED WORK

The previous work can be classified into two categories: traditional OCR (Optical Character Recognition) based method and object recognition based method.

2.1 TRADITIONAL OCR BASED METHOD

In traditional OCR based method, there are three main steps: Binarization, Segmentation and Recognition. In binarization step, the foreground text is segmented from the background. Gllavata et al. [3] used a hybrid approach to localize, segment and binarize the text. In hybrid approach, text is localized using connected component based approach and text is detected using texture based approach. K-means algorithm is used to classify pixels into background and text. Sergey Milyaev et al. [4] proposed a new binarization method which includes three steps: Local binarization producing seed pixels, Seed pixel strength estimation and Global optimization. In [5], Nobuyuki Otsu explains about global thresholding method from gray level histogram. Using Otsu thresholding, an optimum threshold can be calculated separating two classes in binarization.

Another important step in traditional OCR based method is segmentation. As the accuracy of segmentation increases, the character recognition also increases. In [6], the number plate (NP) extraction and each character segmentation of NP (similar to character extraction and segmentation of video subtitles) are discussed using different methods. The author in [7] considers some key features of the character and use split and merge algorithm to achieve segmentation of the characters. Basavaraj A et al. in [8] have explained about video text extraction using image processing operations like morphological edge detection, Sobel filter and dilation. The last step in traditional OCR based method is recognition, for which OCR engine is used. The information loss during binarization step of traditional OCR based method is irreversible. Hence, recognition of the complete text cannot be achieved.

2.2 OBJECT RECOGNITITON BASED METHOD

In object recognition based method, both character recognition and object recognition are assumed to be similar with high intra class variation. In this method, features are extracted directly from the original image. These features are used to train and test for different classifiers to recognize the character. In [9], T. de Campos et al. has used Bag of visual words representation in natural scene for character recognition. Bag of visual words is a common method for representing image content for object category recognition. Kai Wang et al. [10] have implemented Histogram of Oriented Gradients (HOG) for text detection. Due to exceptional performance of HOG in pedestrian detection, it is frequently chosen for feature extraction. The classifiers are trained using the features extracted from real images taken in unconstrained conditions. During testing of the trained classifier, the image feature is given as input. The output of the classifier will belong to the same class as of an input image. Using two extension of HOG [11], object based classification is framed. In first extension, histogram is obtained for various scale of images and whereas in second extension HOG column is used.

3. METHODOLOGY

The text detection and recognition in an image has gained lot of importance for automatic processing of an image. In text detection and recognition, the important steps involved are segmentation, feature extraction and classification. The overall accuracy of text detection and recognition depends on these three steps. The methodology of character recognition is implemented in Microsoft Visual Studio 2010 using Open CV library. The video frames are extracted from the video [12] and one of a frame with text is considered for processing. The original video specifications are: (1) Bits/Pixel: 24, (2) Frame Rate: 23.976, (3) Height: 720, (4) Width: 1280 and (5) Video Format: 'RGB24'.

The video frames must be classified as frames with and without subtitles before segmentation process. The blob detection algorithm is used to achieve classification of video frames. The blob is defined as the group of connected pixel in an image with some similar property. In this paper, the video frame with no subtitles is classified if the number of blobs detected is less than the threshold in the area of region of interest (ROI) and

if the number blobs detected is more than the threshold then the video frame is classified as frame with subtitles. The detected blobs are represented using red circles. Results of blob detector with no subtitle and with subtitle are shown in Fig.2 and Fig.3 respectively.

3.1 SEGMENTATION OF CHARACTERS

The video frame classified as frame with subtitle by blob detection algorithm is considered for character segmentation. The steps involved in character segmentation are: (1) ROI extraction, (2) Resizing of ROI region, (3) RGB to gray conversion, (4) Morphological Gradient, (5) Binarization, (6) Morphological Close, (7) Contour Tracing and (8) Character Segmentation.

The recognition of characters increases with proper segmentation of each character. The video frame with subtitle is shown in Fig.4. Video frame with text is considered for processing. As subtitles are region of interest, they are cropped and resized to increase the character detection rate. The color image is converted to gray scale to ease the image processing operation on resized image. Morphological gradient operation is applied on the resized gray scale image because it is most preferred for segmentation and it helps in finding the outline of the characters.

Let $f(x)$ be the gray scale image [13] and gray scale gradient is defined as

$$\nabla (f) = \max_{x \in g}\{f(x)\} - \min_{x \in g}\{f(x)\} \qquad (1)$$

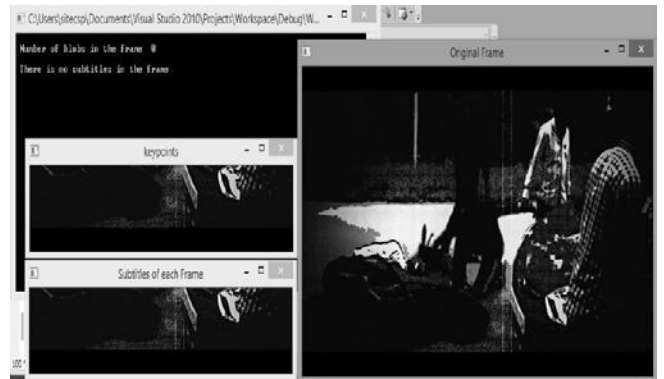where, g is a structuring element.

Fig.2. Video frame with no subtitles
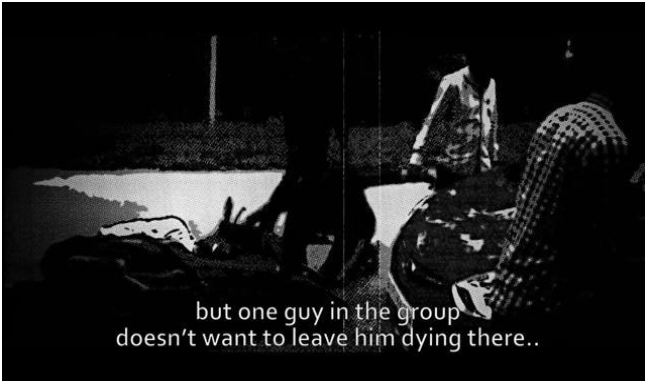

Fig.3. Video frame with subtitles

Fig.4. Video Frame with text

To preserve the edge information of the character, the gradient operation is performed in prior to binarization step. The result obtained from gradient operation is binarized using Otsu thresholding. Otsu thresholding is global thresholding method and the threshold is calculated using the mean and variance of the image pixels. To fill the small holes produced during binarization, morphological close (Dilation followed by erosion) operation is performed on the binarized image. Using the result of morphological close operation, the contour of each character is traced and rectangular box are used to separate each character. The segmentation result for video frame of Fig.4 is shown in Fig.5.
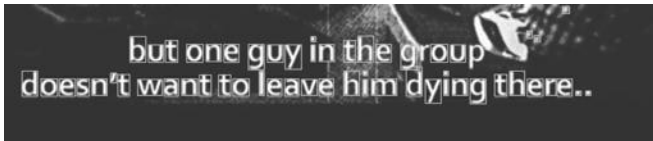
Fig.5. Final result with segmented characters

3.2 FEATURE EXTRACTION

The segmented characters are cropped and from each individual character, features must be extracted. To extract features there are several methods, amongst which Histogram of Oriented Gradients (HOG) [14] is the state-of-the art methods for feature extraction and data representation technique. It is extensively used in the area of computer vision and pattern recognition. Using this technique, an image can be efficiently represented as a feature vector of low dimension. Feature vector provides relevant information for recognition than the raw image. In the following subsection the concept of HOG is briefly explained.

3.2.1 *Histogram of Oriented Gradients:*

Histogram of Oriented Gradients is one of the popular feature descriptor used in computer vision and image processing for the purpose of object detection. In HOG feature extraction, the centered horizontal and vertical gradient is computed with no smoothing. The gradient orientation and magnitude is calculated for the segmented character images.

Centered:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x-h)}{2h} \qquad (2)$$

Filter mask in *x* and *y* directions

Centered:

Gradient:
- Magnitude:

$$s = \sqrt{s_x^2 + s_y^2} \qquad (3)$$

Orientation:

$$\frac{S_y}{s_x} \qquad (4)$$

The HOG features and orientation binning is shown in Fig.6 and Fig.7, respectively. The image of a character is resized to 32×32 which is the window size. The parameters selected for HOG descriptor calculations are: Block Size: 8×8, Block Stride: 8×8, Cell Size: 2×2 and Number of Bins: 9 i.e., orientations from 0–180 degree's.

The number of features obtained from segmented image is 2304. The original image and HOG result is shown in Fig.8(a) and Fig.8(b) respectively.
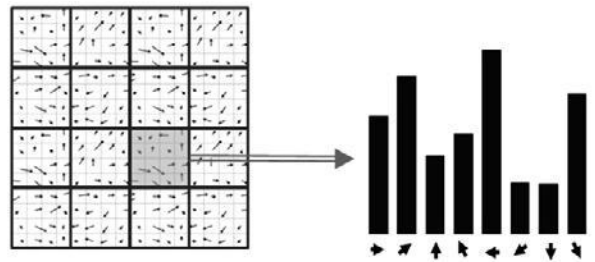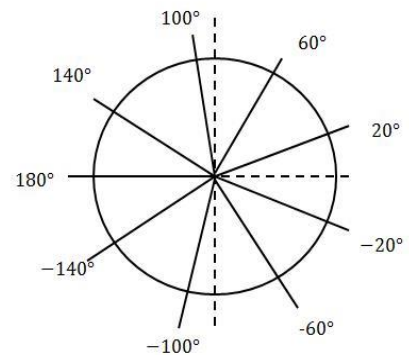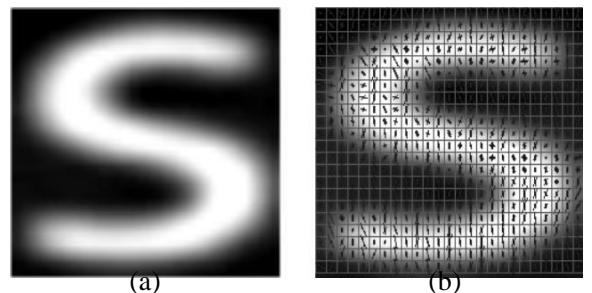

Fig.6. HOG Features

Fig.7. Orientation Binning

Fig.8. (a) Input image, (b) HOG result on scaled input image.

3.3 CLASSIFICATION

Support vector machine (SVM) [15] analyses the input data and classifies the input. In machine learning, SVM is a supervised learning model. The classification is based on the idea of decision hyper-planes that defines decision boundaries in input space. SVM is based on the principle of separation

between the samples of binary classes *i.e.,* positive and negative.

3.3.1 Binary SVM:

The binary SVM resolves the problem of separation of two classes, symbolized by n samples of m attributes each. Consider a problem of separating two classes represented by n samples in a training set,

$$T = \left\{ (x_i, y_i) \mid x_i \in R^p, y_i \in (-1, +1) \right\}_{i=1}^{n} \quad (5)$$

where, x_i are learning samples and y_i are output classes i.e., -1 or +1. The goal of the SVM is to separate two classes by finding a linear function i.e., hyper-plane as shown in Fig.9.

$$y_i = \begin{cases} +1 \; if\, x_i \in class +1 \\ -1 \; if\, x_i \in class -1 \end{cases} \quad (6)$$
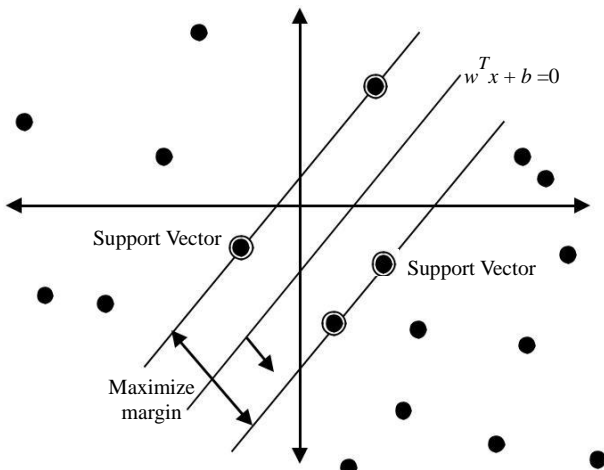

Fig.9. Binary Classification Using

SVM *3.3.2 Multi Class Classification:*

There are two approaches involved in solving multiclass problem (say *k*-class) using SVM. In one-against-one method, $k/(k-1)/2$. SVMs are trained and each SVM separates a pair of classes as shown in Fig.10. In one-against-all approach, *k* SVMs are trained and each SVM separates a single class from all rest of the classes as shown in Fig.11.
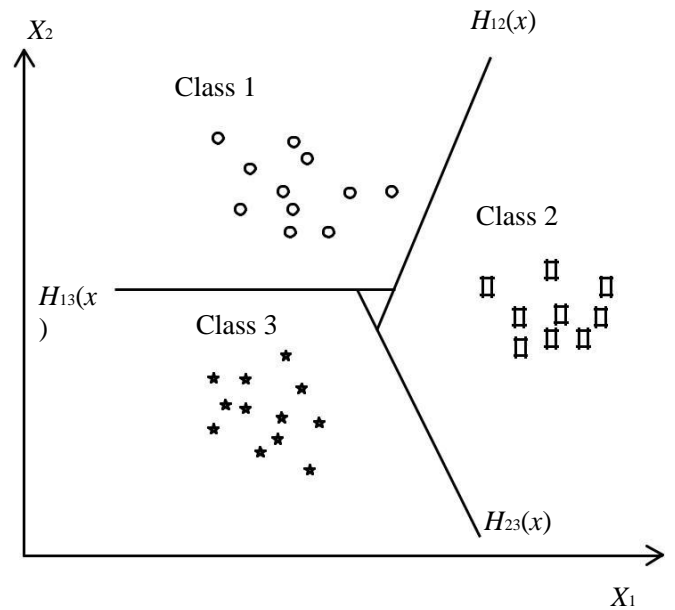

Fig.10. One-against-one approach

In this paper, one-against-one approach is used. In training using one -against-one approach, instead of learning decision functions, it discriminates each class from every other class thus only decision functions are learned. In the database, 13 samples of each uppercase and lowercase letters are considered for training and the segmented characters from video frame with subtitles are used for testing. The database used for training the SVM in this paper is shown in Fig.12.
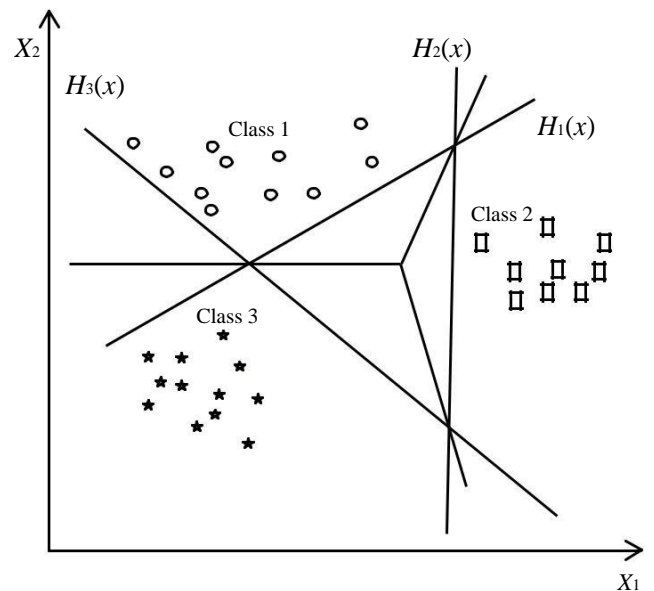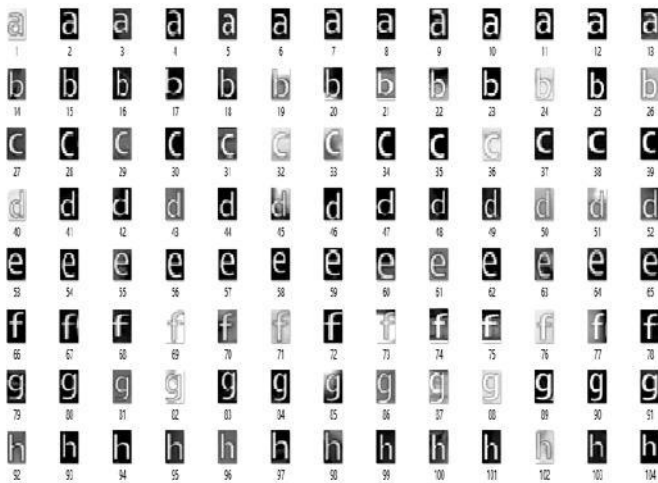

Fig.11. One-against-all approach

Fig.12. Database used for training the SVM

The one-against-one approach is supervised learning method, where features of each alphabet are labelled individually. The database used for testing the classifier is shown in Fig.13 and the output of the classifier is shown in Fig.14.
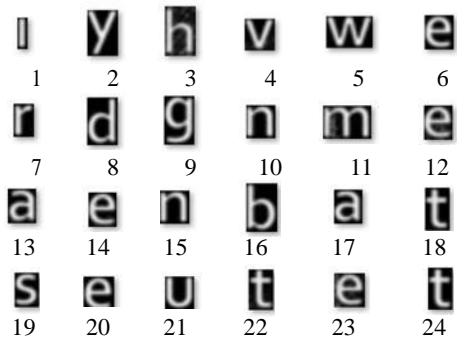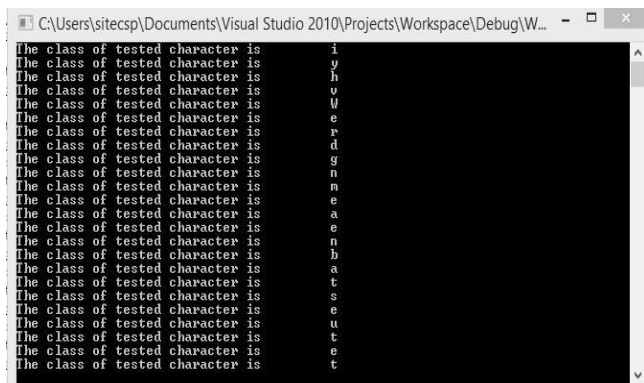

Fig.13. Database used for testing the SVM

Fig.14. Output of the SVM

Fig.15. Kannada Segmented Characters used for testing the SVM

Using the classification of English characters, videos of different subtitles can be classified as English subtitle videos and other language subtitle videos. The segmented Kannada characters and Classification of video (Video Frame) based on subtitles is shown in Fig.15 and Fig.16, respectively.

Unlike Artificial Neural Network (ANN), the computational complexity of SVM does not depend on the dimensionality of the input space. Hence using of SVM is computationally effective. The accuracy of classifier (SVM) [16] can be defined as:

Accuracy = (Number of correctly recoginized characters by SVM / Total number of characters in the test dataset) × 100

The data set used to calculate the accuracy of SVM is shown in Fig.17 and result of SVM is shown in Fig.18. The number of correctly recognized characters by SVM is 21 out of 23 i.e., 91%.
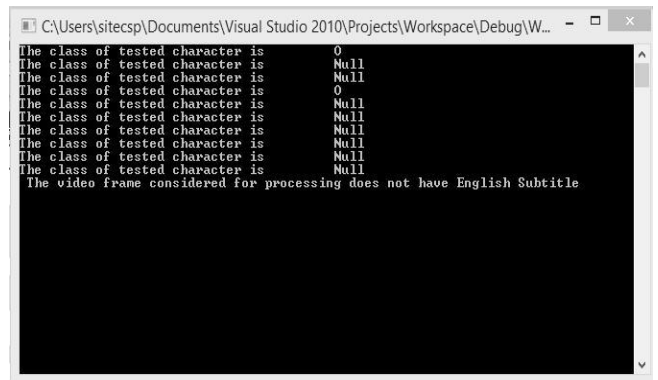

Fig.16. Output of the SVM for Kannada Characters

Due to similarity in the structure of the characters and presence of texture in a segmented image, results ambiguity in recognizing the characters. The Fig.19 shows the examples of characters of misclassified class and the result of the classifier is shown in Fig.20. In Fig.19, first character is '0' (zero), second character is 'o' (small letter o), third character is 'd', fourth character is 'n' and last character is 'l' (small letter l) but the classifier classify the first character as 'o' (small letter o), second character as 'O' (capital letter O), third character as 'a', fourth character as 'null' and last character as 'i' (small letter i).

Fig.17. Dataset to calculate accuracy of SVM

Fig.18. SVM result of the dataset
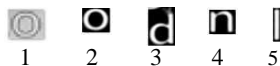
1 2 3 4 5

Fig.19. Examples of characters of misclassified class

Fig.20. Result of the classifier for characters of misclassified class

4. CONCLUSION

Character recognition of video subtitles plays crucial role in analysing the video. In this paper, morphological operations are used for the character segmentation. The morphological operations involved in character segmentation are gradient and close. HOG is used as the feature descriptor for individual characters and the multi class SVM i.e., one -against-one approach classify the characters effectively. With the help of classification of characters, videos of English subtitles and other language subtitles can be classified.

REFERENCES

[1] Tike Judd, Krista Ehinger, Fredo Durand and Antonio Torralba, "Learning to predict where humans look", *Proceedings of IEEE 12th International Conference on Computer Vision*, pp. 2106-2113, 2009.

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

[3] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben, "A Text Detection, Localization and Segmentation System for OCR in Images", *Proceedings of IEEE 6th International Symposium on Multimedia Software Engineering*, pp. 310-317, 2004.

[4] Sergey Milyaev, Olga Barinova, Tatiana Novikova, Pushmeet Kohli and Victor Lempitsky, "Image Binarization for End-to-End Text understanding in Natural Images", *Proceedings of IEEE 12th International Conference on Document Analysis and Recognition*, pp. 128-132, 2013.

[5] Nobuyuki Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, No. 1, pp. 62-66, 1979.

[6] Chirag Patel, Atul Patel and Dipti Shah, "A Review of Character Segmentation Methods", *International Journal of Current Engineering and Technology*, Vol. 3, No. 5, pp. 2075-2078, 2013.

[7] Rainer Lienhart, "Indexing and Retrieval of Digital Video Sequences based on Automatic Text Recognition", *Proceedings of 4th Association for Computing Machinery International Multimedia Conference*, pp. 11-20, 1996.

[8] Basavaraj Amarapur and Nagaraj Patil, "Video Text Extraction from Images for Character Recognition", *Proceedings of Canadian Conference on Electrical and Computer Engineering*, pp. 198-201, 2006.

[9] Toefilo E. De Campos, Bodla Rakesh Babu and Manik Varma, "Character Recognition in Natural Images", *Proceedings of IEEE International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 1-4, 2009.

[10] Kai Wang and Serge Belongie, "Word Spotting in the Wild", *Proceedings of 11th European Conference on Computer Vision Part I*, pp. 591-604, 2010.

[11] Andrew J. Newell and Lewis D. Griffin, "Multiscale Histogram of Oriented Gradient Descriptors for Robust Character Recognition", *Proceedings of IEEE International Conference on Document Analysis and Recognition*, pp. 1085-1089, 2011.

[12] Goa alla Gokarna New Short Film Kannada, Available at: https://www.youtube.com/watch?v=bvEKJHqVkXY, 2015.

[13] A.N. Evans, "Morphological Gradient Operators for Colour Images", *Proceedings of International Conference on Image Processing*, pp. 3089-3092, 2004.

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 886-893, 2005.

[15] Djeffal Abdelhamid, Babahenini Mohamed Chaouki and Taleb-Ahmed Abdelmalik, "A Fast Multi-Class SVM Learning Method for Huge Databases", *International Journal of Computer Science Issues*, Vol. 8, No. 3, pp. 544-550, 2011.

[16] S.Vijayarani and A.Sakila, "Performance Comparison of OCR Tools", *International Journal of UbiComp*, Vol. 6, No. 3, pp. 19-30, 2015.