TEXT DETECTION AND EXTRACTION FROM VIDEOS USING ANN BASED NETWORK

A. Thilagavathy, K. Aarthi, A. Chilambuchelvan

Department of Computer Engineering, R.M.K Engineering College, Kavaraipettai, Tamil Nadu, India.
atv.cse@rmkec.ac.in
aarthi.kme@gmail.com

ABSTRACT

With fast intensification of existing multimedia documents and mounting demand for information indexing and retrieval, much endeavor has been done on extracting the text from images and videos. The prime intention of the projected system is to spot and haul out the scene text from video. Extracting the scene text from video is demanding due to complex background, varying font size, different style, lower resolution and blurring, position, viewing angle and so on. In this paper we put forward a hybrid method where the two most well-liked text extraction techniques i.e. region based method and connected component (CC) based method comes together. Initially the video is split into frames and key frames obtained. Text region indicator (TRI) is being developed to compute the text prevailing confidence and candidate region by performing binarization. Artificial Neural network (ANN) is used as the classifier and Optical Character Recognition (OCR) is used for character verification. Text is grouped by constructing the minimum spanning tree with the use of bounding box distance.

KEYWORDS

Caption text, Pre-processing, Scene text, Classification of text, Text grouping, and Video frame extraction

1. INTRODUCTION

Text data present in images and video contain useful information for automatic annotation, indexing and structuring of images. According to Jung *et al* [13], the text information extraction system (TIE) consists of four stages: text detection (finds the text region in the frame), text localization (groups the text region and generate bounding boxes), text extraction and enhancement (extract the text using some classifier and enhance it) and recognition (verify the extracted text with OCR).

Video consists of two types of text: Scene text and Caption text. *Scene text* is the text that in nature occurs in the area of capturing of video like text on banners, signs, container, CD cover, sign board, text on vehicle. It is also called graphics text. *Caption text or artificial text* on the other hand is the text that is artificially overlaid on the video/image such as the scores in sports videos, subtitles in news video, date and time in the video. It is also called superimposed text. However, the variations of text due to differences in font, size, orientation, style, alignment, complex background, unknown layout makes the text extraction from video a challenging task.

The existing methods to detect and extract the text from video are classified into thresholding based and grouping based methods.

1.1Thresholding based method

In this method, a global threshold is defined for the whole image and a local threshold is defined for a selected portion of the image. There are two types of thresholding based method:

1.1.1 Histogram based thresholding:

It is usually used for monochrome image. It counts the number of each pixel value in the histogram. Threshold is the value between the two peaks. The main disadvantage of this method is it doesn't work for images with complex background.

1.1.2 Adaptive binarization techniques:

It is used for grayscale image. Threshold is defined for the parts of the video. The most common method used is the Niblack's method which considers the mean and standard deviations over the radius of r.

1.1.3 Entropy based method:

This method is used only for the grey scaled image. It makes use of the entropy values of the different grey level distribution.

1.2 Grouping based method

This method groups the text pixel based on some criteria to extract the text. Grouping based method consists of the following types:

1.2.1 Region based method:

This method is based on texture analysis. Region based method can be either top-down and bottom-up. *Top-down* considers the whole image and moves to smaller parts of the image. Top-down approach considers the grayscale value. *Bottom-up* approach starts with the smaller parts of the image and then merges into a single image. The widely used bottom-up approaches are *connected component (CC)* based method and edge based methods.

1.2.2 Learning based approach:

It mainly makes use of the neural networks. Some of classifiers used are Multilayer perceptron (MLP), single layer perceptrons (SLP) etc.

1.2.3 Clustering based approach:

This method groups the text into clusters based on the color similarity. Some of the commonly used methods are K-means clustering, Gaussian mixture model (GMM), density based etc.

Extracting the scene text from video is demanding due to complex background, varying font size, different style, lower resolution and blurring, position, viewing angle and so on. In this paper we put forward a hybrid method where the two most well-liked text extraction techniques i.e. region based method and connected component (CC) based method comes together. Initially the video is split into frames and obtain the key frames. Text region indicator (TRI) is being developed to compute the text prevailing confidence and candidate region by performing binarization. Artificial Neural network (ANN) is used as the classifier and Optical Character Recognition (OCR) is used for character verification. Text is grouped by constructing the minimum spanning tree with the use of bounding box distance.

The paper is organized into the following sections. We discuss the related works in section 2, System overview in section 3 and Pre-processing in section 4, Classification of text in section 5, Text grouping in section 6, Conclusion in section 7 and future contribution in section 8.

2. RELATED WORK

Liu *et al* [1] proposed two text extracting methods: region based and connected component (CC) based method. The region based method is used for segmentation and CC for filtering the text and non-text components. In [2] Hu *et al* used a corner based approach to detect and extract the caption text from videos. It is based on the assumption that the text has a dense corner points. In [3] Weinman *et al* proposed the unified processing. The method involves the following information: appearance, language, similarity to other characters, and a lexicon.

Tsai *et al [4]* proposed the method for detection of signs from videos. It uses connected component analysis to detect the candidate region. A single detector is used for all the color instead of a separate detector for each color. The radial basis function network is used as the classifier. A rectification method was proposed to rectify a skewed road sign to its correct shape.

In [5] Nicolas *et al* implemented the Conditional Random Field (CRF) in the document analysis. It takes into account both the local and contextual feature. These features are extracted and feed as input to the Multilayer Perceptron (MLP). In [6] Chen and Yuille used Adaboost classifier where the weak classifiers are applied to train for a strong classifier in order to construct the fast text detector. This region identified by the classifier is given as input to the binarization and followed by CC analysis.

In [7] Kim *et al* presented an approach where they used SVM as a classifier and then perform the CAMSHIFT to identify the text regions. In [8] Lienhart *et al,* Complex-valued multilayer feed forward network is trained to detect text at an unchanging scale and position. In [9] Li et al makes use of the neural network as classifier and the extracted text is compared with the successive frames. It will identify the presence of the text in 16 X 16 windows only and SSD (Sum of Squared Difference) for frame similarity.

In [10] Zhong *et al* extract text from compressed video using the DCT. It applies horizontal thresholding to obtain the noise. In [11] Zhu *et al* employs Non-linear Niblacks method to perform the grey scale conversion and then fed into the classifier which is trained by Adaboost algorithm for filtering the text and non-text regions. In [12] Liu *et al* used edge detection algorithm to obtain the text color pixels. Connected component analysis is done to obtain the text confidence.

In [15] Bouman *et a*l proposed a low complexity method for detecting the sign and localization of text in natural scene images. The image is divided into block and search for the homogenous block. There is the need to find the hole in the hull of homogenous block in order to detect the text region. In [16] Shivakumara *et al* proposed a method to identify the candidate region in the video frame using Fourier- Laplacian filtering which is followed by the k means clustering. The segmentation of candidate region is performed by the skeleton of connected component. In order to have a better result the false positives are discarded by applying straightness and edge density.

In [17] Hua *et al* make use of the corners and edges to detect the text region in the frames but it would not account for the description of the shape properties of the detected text region. In [18] Vasileios *et al* proposed a method to segment the video into shots and extract the key frames for each shot using the fast global k-means clustering algorithm. The shot clustering is performed using the visual similarity and they are labelled in accord to the cluster they are assigned.

3. SYSTEM OVERVIEW

The main objective of the proposed system is to extract the text from videos. The extracting text from videos comprises many stages namely text detection, text localization and text extraction. The text detection is used to identify the presence of text in the video frame whereas text

localization is used to determine the location of the text in the video frame and generate the bounding box in order to indicate the candidate region. The candidate region is a portion of the frame which contains the text. In text extraction stage the text are extracted from the frame and passed on to the OCR for character verification. The proposed system consists of the following stages: preprocessing, classification of text and text grouping. The preprocessing stage performs the function of the text detection and text localization whereas the classification of text performs the text extraction function. Video is splitted into frames based on the shots. Redundant frames are discarded by performing frame similarity which results in selection of key frames. The key frames are the one that contains the scene text.

In pre-processing stage, the text prevailing confidence is identified and its scale in the key frames. This stage identifies the region where the text is present i.e. candidate region. The adaptive thresholding (binarization) is applied to identify the presence of text in the key frame. After the detection of text region, the connected component analysis is performed where both horizontal and vertical projection in the key frame is used to detect the text.

In Connected Component Analysis (CCA) the CRF model is used to classify the candidate region into two classes: {text, non-text}. The Artificial neural network is trained to be a classifier to filter out the text and non-text components. The extracted text is passed to the OCR (Optical character recognition) for character confirmation. Then the texts are grouped into words and in turn into lines by using the horizontal and vertical bounding box distances by building minimum spanning tree.

Algorithm for text extraction from videos

Input: video which contains text

Output: text extracted from video frame

Begin

1. Video frame splitting followed by frame similarity check and key frame selection.
2. Detects the text region i.e. candidate region in the frame.
3. Determine the text region location and generate the bounding box.
4. ANN is trained to classify the text and non-text components.
5. Extracted text classes are grouped based on bounding box distance (horizontal and vertical)

End

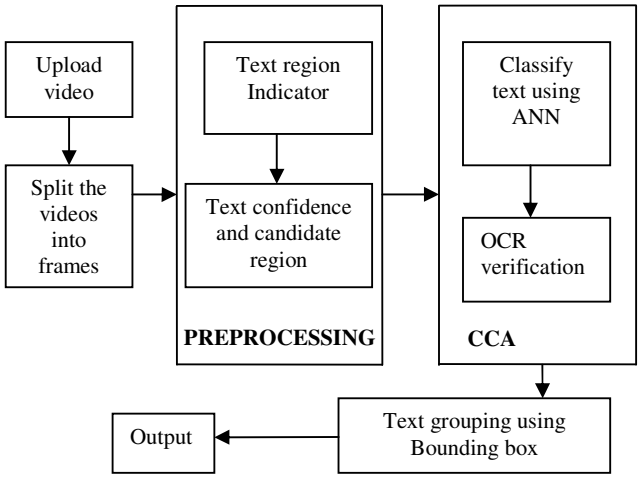

Figure 1 Architecture diagram for text extraction process

4. PRE PROCESSING

The first stage in pre-processing is video frame extraction. In this stage, the video containing the text is splitted into frames after reducing the rate of the video to 1 or 0.1 second. For the one frames per second, the size for the 320 x 240 frame is 75Kb and the size per seconds' video: 4 MB size per minute video: 263 MB. Experimental results showed that when we consider this frame rate it will lead to the sizeable number of frames and so it will lead to redundant frames.

To overcome the temporal redundancy, edge comparisons between frames are done. Canny Edge detection is used for this purpose. The edges of a single frame are mapped with that of its neighbour ones to check for the frame similarity. When the inter frame space difference is high, it indicates that the frames are similar and we store only one frame and discard all the remaining frames. Thus by using this comparison we would be able to eliminate the redundant frames and will result in the distinct and unique frames (non- redundant video frames set). There is the need to choose the key frame from these non-redundant frames set. The key frames are those frames which contain the text in it. We make use of MODI (Microsoft Office Document Imaging) which identifies the frames containing text by discarding the frames containing special characters. The video frames that have only the non-special characters are stored. Thus we can filter out the non key frames easily and use only the key frames for further processing of the proposed method.

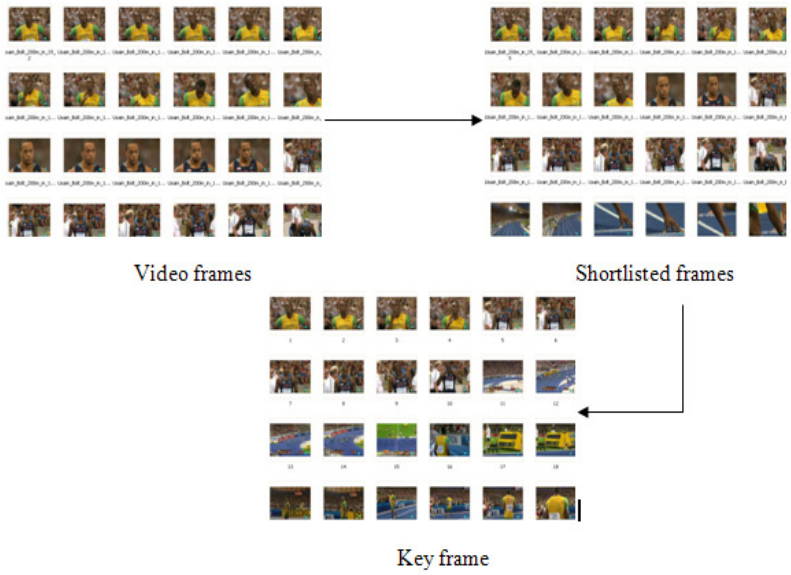

Figure 2 Video splitting and key frame selection

The next stage in pre-processing is to design the TRI. It is used to identify the candidate region. The transition map generation is used to differentiate the text from background. The text will have some kind of properties like vertical alignment, horizontal alignment, inter-character space, static motion, 2D motion, 3D motion to differentiate itself form the background.

The 24 bit colored frame is converted into 8 bit grey scale by binarization where the adaptive thresholding is applied. The threshold value is based on the minimum size of text region. The binarization technique [14] is used which converts the grey scale image into binarized image. In binarization algorithm the threshold value is selected according to the mean and standard deviation by sliding the small window over the key frame. Threshold value is calculated by using the maximum standard deviation of all the calculated windows.

Figure 3 Edge detection using Canny edge detector

The grey scaled frame is subjected to canny edge detector which makes use of the sobel masks for the detection of edge magnitude of the frame. It is followed by a two stage post processing where the non-maxima suppression and hysteresis thresholding is applied to get rid of the non-maxima pixel. The resulting binarized image with the edge is subjected to the morphological operations: dilation and opening. The dilation is applied to the resulting frame to connect the characters whereas the opening operation is performed to remove the noise and to smooth the candidate region shape. The resulting frame after the morphological operations will contains the blobs where the component with height less than the 11 is removed from the further processing. The blobs are the indication of the text region which is to be further processed for localization of the candidate region.

The connected component analysis will compute the candidate region along with the generation of the bounding box. To perform Connected Component, the successive pixels are examined by projecting the binarized image into two dimensions and a rectangular bounding box is generated by linking the four edges of the each character in the candidate region.

In order to decrease the false alarm rate the horizontal and vertical projections are used in each of the bounding boxes. The horizontal edge projection is performed on the bounding boxes and the text lines with the projection value much below the specified threshold should be eliminated. The threshold accounts for the height less than 12 and width less than 50. Box with more than one text lines are further divided and the box which contain no text are discarded from further processing.

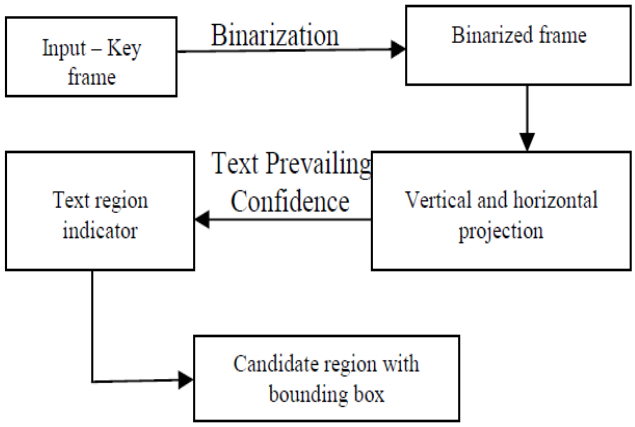

Figure 4 Block diagram for text detection and localization

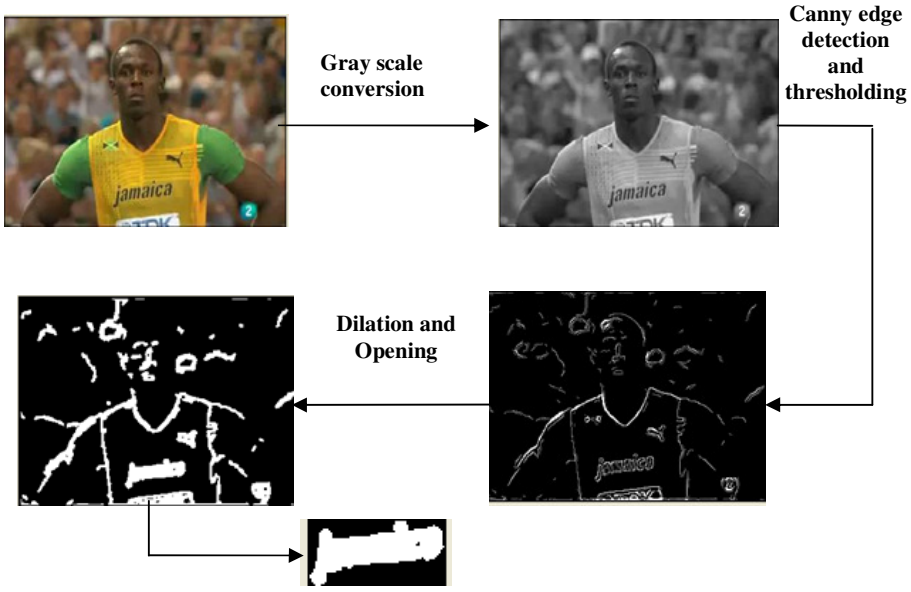

Figure 5 Text detection process and blob extraction

Figure 6 A Key frames with the bounding box (Candidate region)

5. CLASSIFICATION OF TEXT

CRF model is used for classification of the candidate region into text and non-text. The CRF is a graphical model and depends on the Markovian property. The component locality graph is constructed by projecting the binarized image into two dimensions. The construction of graph assumes that the neighbouring text has the same height and width. Euclidean distance is calculated between centroid of the two components. It also takes into consideration the height and width of the bounding box.

The Artificial Neural Network (ANN): Multi-layer Perceptron (MLP) is used as the classifier to categorize the text and non-text components. The Linear separately neural network is used to classify the extracted text from the videos into text and non-text classes. Training is done based on back propagation. Training is through supervised learning where the input and the matching prototype are provided based on the learning rule. The Linear separately network consists of two hidden layer H1 & H2 and an output layer. Layer H1 is for upper case alphabets and layer H2 is for lower case alphabets. The hidden layer of the network contains the matrix value of the alphabets (both upper and lowercase) in the data array which is used for comparing the text from video frame. The input layer is completely connected to H1 and H2 is fully connected to the output layer. Layer H1 and H2 consists of 26 units (one for each English alphabets) whereas the input and output layer has only one units each. The interconnections (link) between the inputs and layer H1 and between layers H1 and H2 are developed. The gradient descent technique is used to obtain the weights for the connection between the processing units. The link

weights are calculated and it is followed by processing of the inputs by the network and it compares the resulting outputs against the desired outputs. Propagation of errors back through system causes the system to fine-tune the link weights which power the network. Throughout the training of a network the unchanged set of data is processed several times as the interconnection weights are ever refined. The set of data which facilitates the training is called the "training set." Once the training is finished now the network is ready to perform as the classifier where it will distinguish between the text (1) and non-text (0) classes.

Algorithm for ANN training

Input: Extracted text from video
Output: text (1) class or non-text (0) class

Begin

1. Network has one input layer (1 unit), two hidden layer (26 units) and one output layer (1 unit).
2. Develop interconnection between the input layer and layer H1 and between layers H1 and H2.
3. Propagate the error function to hidden layers to calculate error derivative.
4. Calculate the link weight.
5. Copy the training set to the input layer.
6. Process the input and perform the comparison between the resulting output and desired output.
7. Propagate the calculated errors back through system to refine the weight and update it.
8. On the completion of training classify the input into text or non-text components.

End

6. TEXT GROUPING

After text extraction, texts are grouped into words and then the words are grouped to lines by constructing the Minimum Spanning Tree (MST) using Kruskal algorithm. The tree is built on the basis of the bounding box distance between the texts: horizontally for the word and vertically for line partition. The spatial distance which is the distance between the bounding boxes is used for this purpose. The edge cuts are used as the partition in the tree construction which will lead to the text localization.

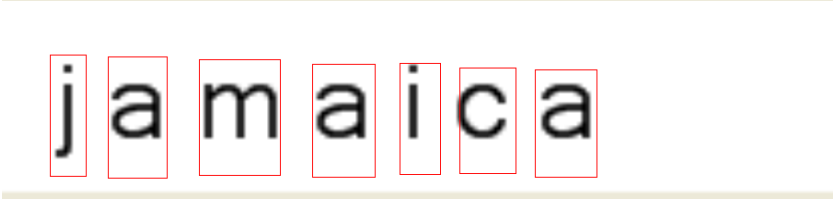

Figure 7 Bounding box generation

7. CONCLUSION

In this paper we proposed a hybrid method where the two most well-liked text extraction techniques i.e. region based method and connected component (CC) based method comes together. The video is split into frames and key frames obtained. Text region indicator (TRI) developed to compute the text prevailing confidence and candidate region by performing

binarization. Artificial Neural network (ANN) is used as the classifier and Optical Character Recognition (OCR) is used for character verification. Text is grouped by constructing the minimum spanning tree with the use of bounding box distance.

Figure 8 A example of text extraction process

8. FUTURE CONTRIBUTION

In this paper the main contribution is only on the English text extraction from the videos. In future, the work can be explored for multilingual languages.

REFERENCES

[1] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, Senior Member, IEEE, "A Hybrid Approach to Detect and Localize Texts in Natural Scene Images", IEEE transactions on image processing, vol. 20, no. 3, March 2011.

[2] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Member, IEEE, Yuxiao Hu, Member, IEEE, Yuncai Liu, Member, IEEE, and Thomas S. Huang, Life Fellow, 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.

[3] J. Weinman, E. Learned-Miller, and A. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, 2009.

[4] L. W. Tsai J. W. Hsieh C. H. Chuang Y. J. Tseng K.-C. Fan, C. C.Lee1,"Road Sign Detection Using Eigen Colour".

[5] S. Nicolas, J. Dardenne, T. Paquet, and L. Heutte, "Document image segmentation using a 2-D conditional random field model," in *Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR'07)*, Curitiba, Brazil, 2007, pp. 407–411.

[6] X.R.Chen and A.L.Yuille, "Detecting and Reading Text in Natural Scenes", in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'04), Washington, DC, pp. 366–373, 2004.

[7] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm," IEEE transaction on. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1631–1639, 2003.

[8] Rainer Lienhart, Member, IEEE, and Axel Wernicke, "Localizing and Segmenting Text in Images and Videos", IEEE transactions on circuits and systems for video technology, vol. 12, no. 4, April 2002.

[9] H. P. Li, D. Doermann, and O.Kia, "Automatic Text Detection and Tracking in Digital Video," IEEE transaction on Image Processing, vol. 9, pp. 147–156, Jan. 2000.

[10] Yu Zhong, Hongjiang Zhang, and Anil K .Jain, Fellow,"Automatic Caption Localization in Compressed Video", IEEE transactions on pattern analysis and machine intelligence, vol. 22, no. 4, April 2000.

[11] K. H. Zhu, F. H. Qi, R. J. Jiang, L. Xu, M. Kimachi, Y. Wu, and T. Aizawa, "Using Adaboost to Detect and Segment Characters From Natural Scenes," in Proc. 1st Conf. Caramera Based Document Analysis and Recognition (CBDAR'05), Seoul, South Korea, pp. 52–59., 2005.

[12] Y. X. Liu, S. Goto, and T. Ikenaga, "A Contour-based Robust Algorithm for Text Detection in Color Images," IEICE transaction. Inf. Syst., vol. E89-D, no. 3, pp. 1221–1230, 2006.

[13] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: A survey," *Pattern Recogn.*, vol. 37, no. 5, pp. 977–997, 2004.

[14] W. Niblack, "An Introduction to Digital Image Processing," Birkeroed, Denmark: Strandberg Publishing, 1985.

[15] Katherine L. Bouman, Student Member, IEEE, Golnaz Abdollahian, Member, IEEE, Mireille Boutin, Member, IEEE, and Edward J. Delp, Fellow, IEEE, " A Low Complexity Sign Detection and Text Localization Method for Mobile Application", IEEE transactions on Multimedia, vol. 13, no. 5,pp 922-934, October 2011.

[16] P. Shivakumara, T. Q. Phan and C. L. Tan, " A Laplacian approach to multi-oriented text detection in videos", IEEE transactions on Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 412–419, February 2011.

[17] X. S. Hua, X. R. Chen, L. Wenyin, and H. J. Zhang," Automatic location of text in video frames," in Proc. ACM Workshops Multimedia: Multimedia Inf. Retriev.,2001,pp. 24-27.

[18] Vasileios T. Chasanis, Aristidis C. Likas, and Nikolaos P. Galatsanos, "Scene Detection in videos Using Shot Clustering and Sequence Alignment", IEEE transactions on Multimedia, vol. 11, no. 1, January 2009.

Authors

Ms. A. Thilagavathy received her B.E degree in Computer Science and Engineering from Madras University. She has received her Post Graduation degree (M.E) in Computer Science and Engineering at SSN College of Engineering affiliated to Anna University. She is currently working as an Associate Professor in R.M.K Engineering College. Her are of interest include Image processing, Pattern Recognition and Artificial Intelligence. The author has published a book in System Software.

Ms. K. Aarthi received her B.E degree in Computer Science and Engineering from Easwari Engineering College affiliated to Anna University, Chennai. She has completed her post graduation in Computer Science and Engineering at RMK Engineering College affiliated to Anna University. Her are of interest include Image processing, Pattern Recognition. She has published a paper titled "A hybrid approach to extract scene text from videos" in IEEE Xplore.