ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

# Tamil Text detection in videos

A. Thilagavathy, S. A. Mariyam Benazir, S. Monica, V.Sangeetha, A. Chilambuchelvan
Department of CSE, R. M. K Engineering. College, Kavaraipettai, Tamil Nadu, India

***Abstract*-- Detecting text in videos with complex background is challenging. Text appearing in videos contains useful information and it is required for easy understanding of videos. We propose a novel method that improves the performance of detecting Tamil text in videos. In this paper, we propose a different method for detecting Tamil text in videos using Localization method. Initially, the input video is divided into 'n' number of frames and key frame is selected from the frames by avoiding temporal redundancy. Then, edge is detected on the key frame by using Sobel edge detector, then the image is filtered to remove any noise present by bounding box algorithm and text region is separated from non text region by localization process finally the text is detected.***

***Index terms*--Sobel edge, Bounding box, Tamil text, Localization.**

## I. INTRODUCTION

The text contained in videos has wide variety of semantic application. The process to include more semantic knowledge is to use the text included in the images and video sequences. It has wide information but easy to use, e.g., Content-based image analysis. Text extracted from a video sequence provides natural, meaningful content that mirror the video's content. Text detected from the video has wider benefits. It is very useful to describe the contents of video sequence. It is easily detected and compared to other semantic contents. Generally, Text is divided into two types, **scene text** and **caption text**. **Scene text** is the text that appears naturally while capturing any video. It is also called Graphics Text. Example for scene text includes: t-shirt, CD cover, sign board, text on number plate in vehicle, Bottle cover, Text on shop advertising boards, etc. **Caption text** is the texts that are overlaid on the captured video that is text is artificially superimposed on the given video. It is also called Superimposed Text. Example for caption text such as scores in sports videos subtitles in news video, date and time in video, etc. Throughout the past decade, plenty of techniques available to detect the text from a video. Exact behaviour in the process of numerical evaluation. Edge is defined as object border and extracted by features such as gray, colour or texture discontinuities [27]. The text information extraction system (TIE) with four stages [16]: 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). Some text extraction methods are proposed by [6], [15], [16], [28] they are: texture based methods, connected component based methods, edge based methods and gradient based methods. Here, Texture based methods are used to separate text regions from their background or other regions within the image by [18], [23], [37]. In Connected Component based methods, the image is divided into a set of smaller components known as connected components and it repeatedly merge the small components to form larger components [24], [34]. Edge based methods are used to define the boundaries within the regions of the image, that helps to execute image segmentation and object recognition [1], [20], [35]. In this paper, we propose a method for detected text using Localization. Here, first the input video is divided into frames. The key frame is taken from various frames and edge is detected by Sobel edge detector. Then the image is filter using bounding box to remove any noise present and the image is enhanced and text is detected by localization process.

## II. PREVIOUS WORK

In generally, localization text is very expensive task in any 2n subset that corresponds to the text. This problem is resolved by using two techniques sliding window by [17], [5], [19] where search is limited to a subset of image. This technique reduces the count of subset is analyzed for the presence of text to cN where c is the constant that varies values from less than one to greater than one for single scale method to relatively large values. The next technique uses many methods to find individual character pixel into region using connect component analyses [10], [21], [22], [25]. Bounding box algorithm enhances the performances of computers. Bounding box algorithm will possess only objects that intersect. Bounding box is intersecting only if its bounding boxes are intersected. The algorithm that heuristic for rendering [39] that traditional algorithm is used for visible surface determination and [40] says that bounding algorithm also used in animation particularly for path planning in collision detection. Sobel edge detector is more accurate other than canny because Sobel gives more detail for text region and less detail for non-text region [30]. A method which uses soft-threshold wavelet to remove any noise [33], then Sobel edge detection operator is used to do edge detection on the image. The involvement of Image enhancement takes an image and improves its visual effect by taking benefit of the response to visual stimulant [32]. To facilitate the development of the solution for image problems in computers many enhancement techniques are proposed. Most of these techniques involve the use of low illumination or high magnification for the noise persist

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

problems. Because of this reason, noise removal continues to be very important image processing process by [2], [7], [13]. In digital images, spatial domain represents very important enhancement technique that can efficiently use to remove different types of noise.

## III. PROPOSED SYSTEM

We propose a novel method that improves the performance of detecting Tamil text in videos using Localization method. In this paper, we propose a different method for detecting Tamil text in videos using Localization method. The input video is divided into 'n' number of frames and key frame is selected from those frames by avoiding temporal redundancy. Edge is detected by Sobel edge detector then Filtering is done to suppress the noise without the edges of the text frame being blurred using bounding box method. To enhance the text region localization and final text is detected by eliminating the other regions which are not text.

### *A) Splitting of frames*

The main purpose of this process is to extract the frames from the video which contain the texts. Motion pictures/videos are made up of number of consecutive frames. A frame is a **single picture or still image** in a video. Frame is flashed on screen for a short period of time (usually 1/24th, 1/25th or 1/30th of a second known as frame rate) and then immediately replaced by the next one. Frames is a rapid succession of the images blend together gives the illusion of movement. So the video (.avi, .mov, .mpg, .flv) is given as input to the System. To check whether there is a change in the scenario or not the program will compare and calculate the similarity of each and every video frames. If any changes occur, the video will break and finally the video is splitted into frames (shots).Then the frames (i.e., jpg files) are saved in a separate folder. Here, the video which contains Tamil text is uploaded and it is divided into 'n' number of frames by reducing the frame rate of the uploaded video into 1 second or 0.1 second. We consider the 320 x 240 frames and frame rate to be one second i.e., one frame per second, the size of the frame is75Kb. We have revealed that there are large numbers of frames due to reduction of frame rate of the video and this will lead to temporal redundancy.

### *B) Key Frame Selection*

Finding a particular frame in a single continuous stream of frames taken from a single camera source. (i.e., a shot).This process is called **key frame selection**. In this paper the key frame is selected from various frames by removing inter frame redundancy. The temporal redundancy is avoided by making use of the edge comparison where edges of the frames are compared with respect to that of the other frames. The edges of the video frames are obtained by using the Sobel Edge detector. The edges of a single frame are mapped with that of its neighbour ones to check for the frame similarity. When we find the inter frame space difference to be high it indicates that those frames are similar, so we store only one frame and discard all the remaining frames. By performing edge comparison the redundant frames are eliminated. The frame which contains the text is key frame. Fig 1 shows the result of key frame selection.


**Fig 1.Key frame selection**

### *C) Edge Detection*

The digital image which has text can be detected using sobel edge map. Using this edge detector, the image which has text can be detected accurately. In this sobel detector, there are two coordinates namely, x-coordinate and y-coordinate. The x-coordinate will do increasing "right-direction" and the y-coordinate will do increasing "down-direction" for detecting the text in the image. For x-coordinate and y-coordinate namely, $G_x$ and $G_y$ are used. The edge detection which are using sobel edge map is to detect text edge in image better than canny. The sobel edge detector performs 2D convolution operation. The sobel edge detector which give expression for coordinate [13].

$$|G| = \sqrt{Gx^2 + Gy^2} \qquad (1)$$

Then finds approximate gradient magnitude by

$$|G| = |Gx| + |Gy| \qquad (2)$$

This edge detector uses 3*3 convolution masks for detecting text in the image. The mask which slid over the whole image and square of pixels manipulate at a time to detect the text. The sobel which express 3*3 convolution masks which is given in the Fig 2[13].By using this detector, the text is detected and noise is removed slightly. And also which the convolution kernel uses smoothen the text which is detected. The sobel gives fine details for text and less details for non text. Fig 3 shows the sobel edge detected for given input.

| | | |
|---|---|---|
| -1 | 0 | +1 |
| -2 | 0 | +2 |
| -1 | 0 | +1 |

Gx

| | | |
|---|---|---|
| +1 | +2 | +1 |
| 0 | 0 | 0 |
| -1 | -2 | -1 |

Gy

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

**Fig 2. Sobel convolution kernel**

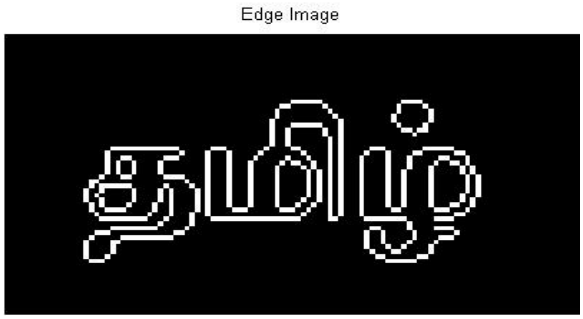


**Fig 3.Sobel edge for input image**

***D) Bounding Box***

Noise reduction is the process of removing unwanted distortions and it involves smoothening of the image. Images are often falsified by random variations in brightness, radiance, or having poor contrast and cannot be used. Image noise is an undesirable by-product of image that adds spurious and extraneous information. Filtering transforms pixel intensity values to reveal certain image characteristics. Foreground/background text segmentation of an image is an ambiguous problem. Interactions with colour distributions and contrast edges observed in the image on which the practical systems rely. Several types of interaction paradigms implemented the most natural and most economical in terms of user interaction is the bounding box interaction. Initially, it limits the segmentation process's attention to the interior of the bounding box. It is an easy property to blend into any algorithm, because all the exterior pixels are simply assigned to the 'Background classes. Another property is very sedulous to incorporate or even difficult to formalize. This can be expressed informally as users provide bounding boxes which are sufficiently tight but not too loose, or by other words, the segmentation desired must have parts which are close enough to each sides of the bounding box. The goal of this module is to develop a new partition of framework which is capable of imposing tightness. For some extent, this property of tightness may be taken into criteria by process which is based on local curve evolution. This type of method can assume the"foreground" region to acquire the complete bounding box then the energy-driven shrinking is performed. The local minimum found by the process might likely be tight in the manner stated above, it is because the shrinking process does not go too far, which the only hope is. Our framework is a different approach which exploits the drawbacks of local curve evolution. Alternatively, it imposes global optimization techniques, convex continuous optimization and graph cuts namely. In brief the important aspects and the contributions are as follows [12]. 1) It proposes a representation of the idea of tightness. Hence the result is the task segmentation of the image by the integer program (IP). This program, applies the low-level cues namely consistency with distribution of colour segments and edge image, respect to the bounding box.

2) This module investigates the approximate solution for IP. Considering the integer program relaxation solved by a linear programming solver. Then a new graph cut-based algorithm called pinpointing which is approximate is presented used for rounding procedure as relaxed solution for original IP.

3) The module evaluates on the proposed algorithms with ground truth for available dataset for segmentation problems.

Fig 4(a) shows the image with noise and Fig 4(b) shows the image without any noise.


**Fig 4(a) input image with noise**.

**Fig 4(b) Filtered image**

***E) Localization process***

The process of **localization** consists of enhancing the text regions by eliminating the other regions that are not text. In a text usually all characters appears closer to each other. Thus morphological dilation operation can be performed in order to eliminate the pixels that are far away from the region. It is basically an operation which expands or enhances the region of interest using the structuring element. The outcome of this process consists of image with some non-text regions are noise which will be eliminated by the upcoming phase. Morphological techniques typically probe an image with a small shape or template known as a **structuring element.** The structuring element which is in morphological dilation is

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

to be positioned at all locations in the image and is to be compared with the comparable neighbourhood of pixels. The shapes and size which are adapted to property of geometry objects are represented in Fig 5 (a)

**Fig 5(a). Geometry property for object**

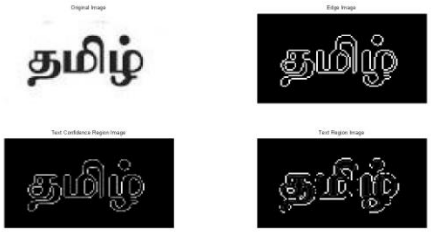

**Fig 5 (b). Localization for input image**

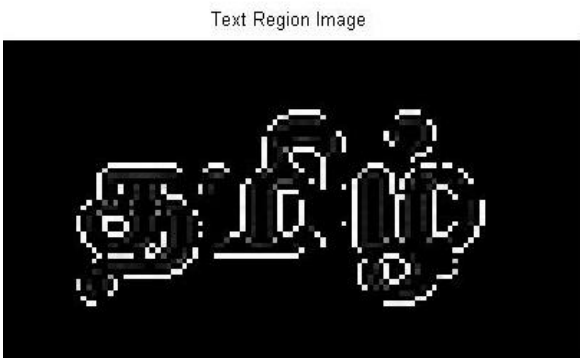


**Fig 5 (c). Text detected**

The localization process which is next to filtering which gives enhancing text regions and eliminating non-text regions, removes noise. In this morphological dilation technique is used. The dilation operator takes input as two pieces of data. For a binary image processing, **white pixels which are normally taken as to represent foreground regions**, whereas black pixels which are to denote background. Fig 5 (b) shows the localization for the input image and text is recognized and shown in Fig 5(c).

## IV. EXPERIMENTAL RESULTS

Our proposed system is evaluated and effectiveness of our method is checked on camera based images by

**Table 1.Comparison of precision rate and recall rate**

| | Precision rate | Recall rate |
|---|---|---|
| Our method | 0.81 | 0.88 |
| Chen | 0.60 | 0.60 |
| Ashida | 0.55 | 0.46 |
| HW David | 0.44 | 0.46 |
| Q.Zhu | 0.33 | 0.40 |
| J.Kim | 0.22 | 0.28 |
| Todoran | 0.19 | 0.18 |
| N.Ezaki | 0.18 | 0.36 |

Our new database with standard dataset as Hua's data of 35 video frames [10]. Our dataset includes 100 arbitrarily-oriented video frames (almost all scene text) 300 horizontal text frames (300 English graphics and 500 Tamil text frames). Our result on precision and recall rate is compared with other existing methods that are, [34], [3], [7], [38], [18], [41], [11] is illustrated in Table 1. The main reason to compare these existing methods with our method is because the existing system has few constraints for the complex background. The experimentation of the proposed algorithm was carried out on the various types of videos. These videos will vary on their background, orientation and size of the text. The performance of the proposed system is evaluated on the basis of the precision, recall. The performance measures are defined as follows [26].

$$Recall\,(R) = TDB\,/\,ATB \quad (3)$$

$$Precision\,(P) = TDB\,/\,(TDB + FDB) \quad (4)$$

Where,

Truly Detected Block ($TDB$), a detected block that contains at least one true character.

Falsely Detected Block ($FDB$), a detected block that does not contain text.

Actual Number of Text Blocks ($ATB$), in the images.



ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014


**(a) Original image with text** **(b) Detected text**

## IV. CONCLUSION

In this paper, we have proposed efficient method for Tamil text detection in video by splitting the video into n number of frames and key is selected from the frames and edge is detected by using with the help of the Sobel edge detector. This Filtering process is done by bounding boxes to remove noise in complex background of video frames. Localisation process enhances the text and separates the text region from non-text region. Thus, finally text is detected. Even though the proposed system

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

will extract the text from videos effectively. It has some limitations like scrolling text cannot be detected properly.

## V. FUTURE WORK

In future we try to detect scrolling Tamil text from videos and extraction of the scene text from complex background.

## REFERENCES

[1] Abdullah A. Alshennawy and Ayman A. Aly."Edge Detection in Digital Images Using Fuzzy Logic Technique," World Academy of Science, Engineering and Technology51 2009.DOI - 10.1.1.193.3230.

[2] N. Afford, Digital Image Processing a Practical Introduction using Java TM, Essex; Pearson Education Limited, 2000.

[3] Ashida, H., Kuriki, I., Murakami, I., Hisakata, R. and Kitaoka, A. (2012). Direction-specific fMRI adaptation reveals the visual cortical network underlying the "Rotating Snakes" illusion.

[4] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In ICCV, pages 694–699, 1995

[5] X. Chen and A. L. Yuille. Detecting and reading text in nat-ural scenes. CVPR, 2:366–373, 2004.

[6] D. Chen, J.M. Odobez and J.P. Thiran, "A Localization/Verification Scheme for Finding Text in Images and Video Frames based on Contrast Independent Features and Machine Learning", Signal

[7] Processing: Image Communication, 2004, pp. 205-217.

[8] V. David, Machine Vision-Automated Visual Inspection and Robot Vision, Prentice Hall, 1991

[9] David Doermann,Jian Liang, Huiping Li, "Progress in Camera-Based Document Image Analysis", In Proceedings of seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003, pp. 606-616.

[10] B. Epshtein, E ofek and Y. Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform", CVPR, 2010, pp. 2963-2970.

[11] N. Ezaki, M. Bulacu and L. Schomaker,"Text detection from natural scene images: Towards a system for visually impaired persons", in Proceeding of the International Conference on Document Analysis and Recognition (ICDAR'04), 2004, pp, 683-686.

[12] Grabcut dataset: http://tinyurl.com/grabcut

[13] R. Gonzalez, R. Woods, Digital Image Processing, Second Edition, Prentice-Hall, 2002

[14] X. S Hua, L. Wenyin and H.J. Zhang, "An Automatic Performance Evaluation Protocol for Video Text Detection Algorithms", IEEE Trans. on CSVT, 2004, 498-507.

[15] A. Jamil, I. Siddiqi, F. Arif and A. Raza, "Edge-based Features for Localization of Artificial Urdu Text in Video Images", ICDAR, 2011, pp. 1120-112.

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

[17] L. Jung-Jin, P.-H. Lee, S.-W. Lee, A. Yuille, and C. Koch. Adaboost for text detection in natural scene. In ICDAR 2011, pages 429–434, 2011.

[18] K. Kim, K. Jung, and J. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 12, pp. 1631–1639, Dec. 2003.

[19] R. Lienhart and A. Wernicke. Localizing and segmenting text in images and videos. Circuits and Systems for Video Technology, 12(4):256 –268, 2002.

[20] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 2, pp. 243–255,Feb2005.

[21] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In ACCV 2010, volume IV of LNCS 6495, pages 2067–2078, November 2010.

[22] L. Neumann and J. Matas. Text localization in real-world images using efficiently pruned exhaustive search. In ICDAR2011, pages 687–691, 2011.

[23] W. Mao, F. Chung, K. K. M. Lam, and W. Sun, "Hybrid Chinese/English text detection in images and video frames," in Proc. 16th Int. Conf. Pattern Recognit., 2002, vol. 3, pp. 1015–1018.

[24] V.Y. Mariano and R. Kasturi, "Locating Uniform-Colored Text in Video Frames," Proc. Int'l Conf. Pattern Recognition, pp. 539-542, 2000.

[25] Y.-F. Pan, X. Hou, and C.-L. Liu. Text localization in natural scene images based on conditional random field. In ICDAR2009, pages 6–10. IEEE Computer Society, 2009.

[26] Palaiahnakote Shivakumara, Trung Quy Phan, Shijian Lu and Chew Lim Tan, Senior Member, "Gradient Vector Flow and Grouping based Method for Arbitrarily-Oriented Scene text Detection in Video Images" ,IEEE tans on 2013, R252-000-402-305.

[27] Russo F: Edge detection in noisy images using fuzzy reasoning, Proceedings of Instrumentation and Measurement Technology Conference, 1998,vol 1, pp.369-372

[28] X. Tang, X. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," IEEE Trans. Neural Netw., vol. 13, no. 4, pp. 961–971, Jul. 2002.

[29] P. Shivakumara, T, Q. Phan and C. L. Tan, "A Laplacian Approach to Multi-Oriented Text Detection in Video", IEEE Trans. on PAMI, 2011, pp.412-419.

[30] X. Tang, X. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," IEEE Trans. Neural Netw., vol. 13, no. 4, pp. 961–971, Jul. 2002.

[31] E. Umbaugh, Computer Vision and Image Processing. A Practical Approach Using CVIP Tools, Prentice Hall, 1998

[32] WenshuoGao; Xiaoguang Zhang ; Lei Yang ; Huizhong Liu ," An improved Sobel edge detection" Computer

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 9, March 2014

Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Volume:5 ), 2010

[33] E.K. Wong and M. Chen, "A New Robust Algorithm for Video Text Extraction," Pattern Recognition, vol. 36, pp. 1397-1406, 2003.

[34] C. Xu and J. L. Prince, "Snakes, shapes, and Gradient Vector Flow", IEEE Transactions on Image Processing, 1998, 359–369.

[35] J. Zhang and R. Kasturi. Character energy and link energy-based text extraction in scene images. In ACCV 2010, volume II of LNCS 6495, pages 832–844, November 2010.

[36] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 4, pp. 385–392, Apr. 2000.

[37] Zhu, Q., Yeh, M.C., Cheng, K.T.: Multimodal fusion using learned text concepts for image categorization. In: Proc. of ACM Int'l. Conf. on Multimedia, pp. 211–220. ACM Press, New York (2006).

[38] FOLEY, J. D., VAN DAM, A., FEINER, S. K., AND HUGHES, J. F. 1996. Computer Graphics (in C): Principles and Practice. 2nd ed. Addison-Wesley systems programming series. Addison-Wesley Longman Publ. Co., Inc., Reading, MA.

[39] LATOMBE, J.-C. 1991. Robot Motion Planning. Kluwer Academic, Dordrecht, Netherlands.

[40] Leon Todoran, Marco Aiello, Marcel Worring, and Christof Monz. Document understanding for a broad class of documents. Technical Report 2001-15, Intelligent Sensory Information Systems Group, University of Amsterdam, October 2001