Text Detection, Extraction and Removal: A Survey

PriyankaDeelipWagh
M.E. Computer Engineering
SES's R C Patel Institute of Technology,
Shirpur.

D. R. Patil
Associate Professor,
SES's R C Patel Institute of Technology,
Shirpur.

ABSTRACT
Text in video and images is an extremely valuable feature to extract brief knowledge of video or image. But sometimes those text components feel unnecessary, and so various methods have been proposed till a day to detect, and remove the text out of video. This paper performs an extensive survey of various techniques available for text detection and removal from video and images automatically to create visually plausible output like videos/images without any of the embedded text.

Keyword
Text Detection, Inpainting, Bandlet, Edge detector, Stroke Width Transformation (SWT).

1. INTRODUCTION
The Embedded text in the form of captions, subtitles, logos or banners etc. in video provides extremely important information. E.g. In news, textual advertisements running at the bottom of TV shows, subtitles in videos etc.

Even if a text plays an important role to provide additional important information, it is also true that in many cases these embedded texts occlude the important portion of video. Then it's best to remove that textual part out of video.

An automatic video text removal scheme has basically following stages: i) an automatic video text detection and tracking, ii) text removal and video completion/ restoration in visually plausible way [1].

Basically, text displayed in the videos can be classified into scene text or graphics text and overlay text or graphics text or artificial text [2]. Scene texts are naturally occurring text in background of scene. While on the other hand, Overlay text is superimposed on the video scene and provides better visual understanding to viewers [3][4].

The separate stages of automatic text detection, tracking and extraction from video and the working of each step is [5]:

1. Text Detection: Presence of text in each frame or image is determined.
2. Text Localization: The exact location of the text is decided and boxed around them are put.
3. Text Tracking: The time of processing for localization of video is reduced and the integrity of position of texts across the adjacent frames of video is maintained.
4. Text Extraction: The segmentation of text component is performed from background.
5. Text removal: Texts are removed from the image and regions are filled using appropriate way of region fill method.

In this paper, section 2 gives the brief information about the related work already performed in the field of text detection and removal from video automatically. Section 3 discusses the detail methodologies of techniques are ready exist for video text detection and removal.

2. RELATED WORK
So many methods are available till today for detecting, localization and extracting/removing text using various methods of inpainting for the artificial texts present in video. Few of the important related works are studied in brief here:

Wonjun Kim et. al. has proposed a novel frame for Overlay text detection and extraction from video. First the transition map is generated. Then candidate regions are extracted and overlay text regions are gets detected on the basis of occurrence of overlay text in each candidate region. At the last localization of overlay text regions is performed by projecting overlay text pixels in transition map and immediatsely a step of extraction is carried out [4].

Yen Lin Chen has proposed an automatic text extraction, removal and inpainting in complex document images, by decomposing the document image into distinct object planes such as textual regions, non-textual objects, background textures etc. The texts with different characteristics from each object plane are detected using knowledge based text extraction and identification. Then an effective adaptive inpainting neighborhood adjustment scheme is applied immediately following text removal [6].

T. Pratheebaet. al. has proposed a novel framework for video text detection and removal by generating a morphological binary map after calculating difference each candidate which are gets generated by connecting candidate

regions using a morphological dilation operation. Then, the text localization takes place by projection of text pixels in morphological binary map and last step is for text extraction [3].

M. R. Lyuet. al. has proposed a comprehensive method for multilingual text detection, localization and extraction. Basic three methods collectively perform text detection: edge detection, local thresholding and hysteresis edge recovery. Then coarse-to-fine localization scheme is applied. And at the last, adaptive thresholding, dam point labeling and inward filling are applied for the text extraction [7].

J. Malobabicet. al. has proposed a methods for detection and localization of superimposed video text using horizontal difference magnitude measure and morphological processing. Then the smoothing and multiple binarisation is used for enhancing the result of

modified version of Wolf-jolion algorithm in the form of character segmentation [8][9][10].

A. Moslehet. al. has proposed a two state framework. In the video text detection stage, unsupervised clustering is performed on connected components produced by stroke width transform (SWT) in order to perform text locations in each frame. Then, localization of video text gets performed by studying motion patterns of text objects. And at the last inpainting is performed in order to remove and restore the video[1].

Mohammad Khodadadiet. Al has proposed an algorithm for subtitle detection, extraction and inpainting in color image. Using stroke filter, new segmentation and verification algorithm based on image profile a text is detected. Background and text color in candidate block are estimated using color histogram. At the last the inpainting algorithm based on matching algorithm is used to reconstruct the initial image contents in text areas [11]

3. METHODOLOGY

After the text detection, text extraction stage takes place. Several techniques have been proposed for video and image text extraction based on morphological operators, wavelet transform artificial neural network, skeletonization operation, edge detection algorithm, histogram technique etc. Every technique has its pros and cons.

Fig. 1(a) and 1(b) illustrate the detail methodologies proposed by W. Kim et. al. [4]. In first step, using combination of the change of intensity and the modified saturation a transition map is generated. The basic advantage of this method is, even in the complex background a transition maps are generated very well. If the gap of consecutive pixels between two nonzero points in the same row of linked map is shorter than 5% of image width, then 1s are filled in them. Connected components which are smaller than threashold, are removed and all the remaining connected components are reshaped in order to have smooth boundaries. In overlay text detection step, the aspect ratio of overlay text region, density of transition pixels and texture based approach are the basic components to be operated for text region detection. Using, number of different local binary patters and the density of transition pixels in each candidate region, the probability of overlay(POT) text is determined. If, POT is larger than the predefined value, the corresponding region is considered as overlay text regions. For overlay text region refinement, first horizontal projections of transition pixels on the transition map are performed, then a null points are removed. Then vertical projection takes place and again null points are removed. Then, in order to take advantage of continuity of overlay text between consecutive frames for the next frame text detection, an overlay text region updation is used.

Then, in order to have bright text compared to it's surrounding pixels in the every module, the color polarity is checked and the inverse of intensity is performed.

In, the last step, adaptive threasholding and modified dam point labeling are performed and using inward filling of background of text region, overlay text is extracted.

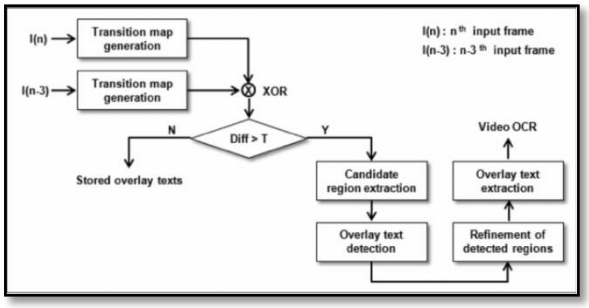

Fig. 1(a).Overall procedure for detection method [4]

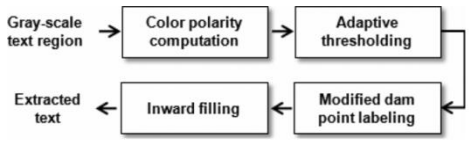

Fig. 1(b).Overall procedure of extraction method [4]

While, Yen-Lin Chen has proposed an automatic approach for text extraction, removal and inpainting of complex document images as shown in fig. 2 [4].by decomposing the document image into distinct object planes such as textual regions, non-textual objects, background textures etc. The texts with different characteristics from each object plane are detected using knowledge based text extraction and identification. Then the adaptive neighborhood adjustment scheme is applied for text removal and inpainting. In order to visually plausible image without text, the inpainting is used. This inpainting is carried out along the isophote lines and arrives at the boundaries of inpainting region. Recursive inpainting and use of surrounding background pictorial information generates the nontext background image [6].

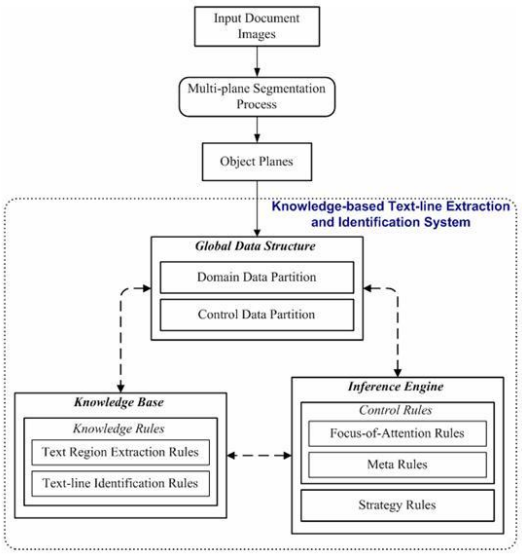

Fig. 2. Overall procedure of Knowledge based method [6]

Fig. 3(a) shows a morphological text detection method proposed by the T. Pratheebaet. al. and it is clear that, the overall procedure of text detection performed by W. Kim et. al (fig. 1(a)) and T. pratheeba (fig. 3(a)) are nearly similar except just a difference that instead of using

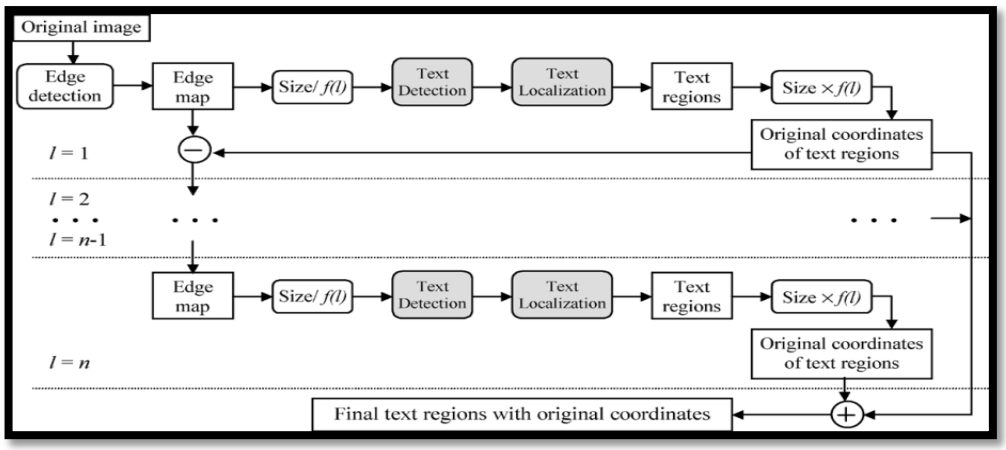

Fig.4 Sequential multiresolution paradigm [7]

transition map, a morphological map is used and same subtasks are performed on that morphological map as mentioned in [4] for text detection.

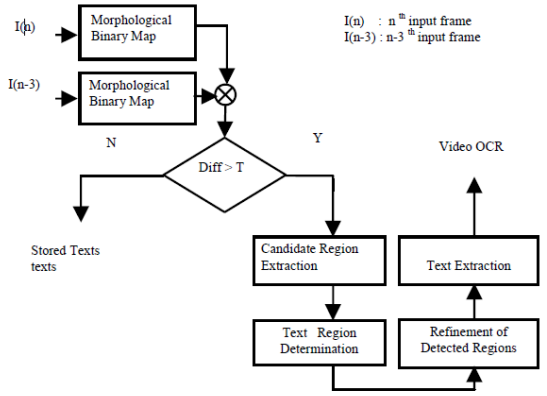

Fig. 3(a).Overall procedure for text detection method [8]

Fig.3(b) shows the steps to generate the morphological binary map from input image, where morphological closing and opening operators are applied on input image and difference is used for binarization. Then the morphological binary image goes through all the same procedure mentioned by W. Kim et. al. for text extraction[4]. But yet the results of T. Pratheeba have stated that average probability of error is reduced by their proposed technique.

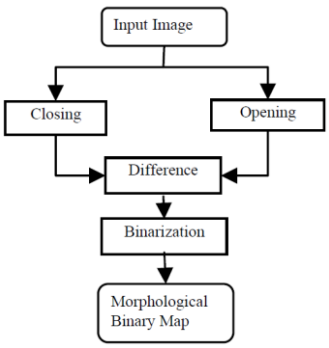

Fig. 3(b). Morphology based technique to extract the contrast features [8]

so many methods have been proposed for the basically English language text detection, localization, extraction or removal from image and video till a day. But if the text of any other language appears in the image or video then due to certain reasons systems fail many times in detecting and removing or extracting those texts from video or image. That's why the comprehensive method for detection, localization and extraction of multilingual text present in video has been proposed by Michael R. Lyuet. al.[7].

There are so many text characteristics are frequently used in video text detection, localization, extraction. Among those, Contrast, color, orientation, stationary location are the language independent characteristics while, stroke density, font size, aspect ratio, stroke statistics are the language dependent characteristics [7].

In order to overcome the drawback of parallel multiresolution paradigm of redundancy of text region, a sequential multiresolution paradigm has been proposed by M. R. Lyuet. al[7]. Initially, a video segment is sampled at two frames per second, and the each frame is converted into 256-level gray scale image. And these images are considered as Original image for further operations to detect, localize and extract the text as shown in Fig. 5. The working of sequential multiresolution paradigm is divided in to levels from 1 to n. In each level from 1 to n, the current edge map is first scaled down by factor f(l). and then text detection and localization operations are performed. At the last, again image is scaled up by same factor f(l) and the original image is generated.

then text detection is carried out by using local threasholding which separates text string from background and hysteresis edge recovery. The text localization is performed using a coarse-to-fine localization method[12] which keeps multiple passes of horizontal and vertical projection in order to get rid of region growing for dealing with multilingual texts as well as complex layouts.

Then, using multiframe verification, the already detected texts from previous frame or which are transitory are eliminated. The signature comparison is performed to decide if two text regions on consecutive frames have same locations and if those are same texts. For dealing with the possible location offset and edge density changes of same text in different frames, the signature distance metric is used.

In the last step of text extraction in this methods consists of three basic subtasks:

1) Adaptive threasholding: for producing a better binary image for different background intensities.

2) Dam point labeling: the pixels of text strokes are called as dam points, which are used to prevent the text pixels to be filled by flooding.

3) Inward filling: the every pixel from Extended-Region are scanned and if a pixel is "white", all the connected white pixels are find using flood fill and set to 'black'. And all the none black pixels are set to 'white' after inward filling.

Fig. 5 shows a system block diagram of the method invented by J. Malobabic [8], where only certain number of frames ie. I frames are considered at a time where same text appears on those all consecutive frames.

Then in the first step of detection, edge map is generated and smoothed using binomial filter followed by blurring horizontal. And then morphological operations like dilation and erosion are carried out as preprocessing steps.

In next step of binarization as per proposed method, to separate text regions from the rest of frame using Otsu's global thresholding method [9][10]. To fit a single bounding box to enclose all the text areas in frame are used in final step of localization. To separate the text pixels from background and form the image with highlighted text pixel with black color and background with white color, a segmentation is carried out. At the last recognition of the text takes place as the basic goal of OCR, by using freely available OCR software known as clara OCR.

All the previous methods had been not really perfect for automatic detection and removal of text from video. And so a new approach of automatic inpainting for video text detection and removal proposed as a great milestone in the area of automation by A. Mosleh [1].

In this methods, basically the operations are performed by splitting the video stream into frames and considering each frame as a separate image as an input for the stream of operations mentioned into fig. 7. The text are detected in each frame over here by generating edge map, and connected components are generated by using stroke width transform[13] and unsupervised clustering for detecting only text regions and rejecting other.

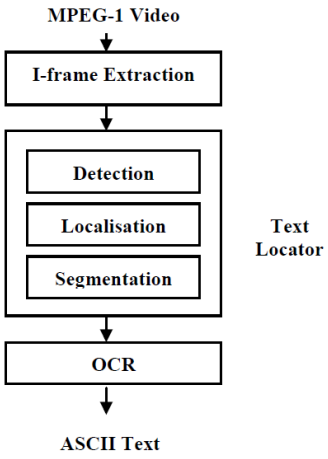

Fig. 5. System block diagram [8]

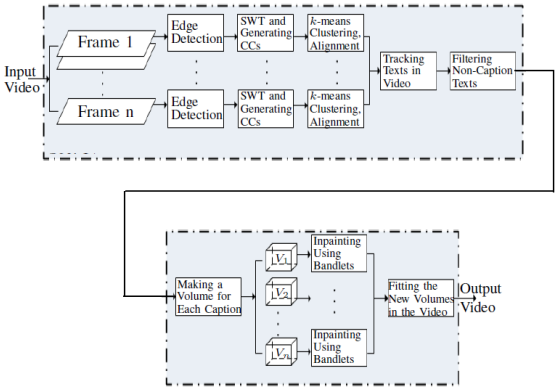

Fig. 6 Main stages of video text detection and removal by A. Mosleh [1]

Then the video global motion patter is calculated using Lucas-Kanade optical flow computation algorithm and text objects are tracked using CAMSHIFT algorithm[1] and filtering of noncaption text takes place. In the last step of inpainting after extraction of text from video, bandlet transform is used as bandlet transform effectively represent image geometry, and exploits spatio temporal regularities to perform regularities to perform a regularization to address the restoration. In this video inpainting method, each caption is considered as volume and so smooth video without text is generated without further any more operation for maintaining visual consistency.

Mohammad Khodadadi's proposed method states to get stroke filter image for all the RGB channels of image. And then performing text localization and extraction using histogram and of text region and background. At the last, inpainting process using texture synthesis and matching is performed with a little bit modified algorithm by considering the high intensity variation for detecting border of text region properly.[11]

4. CONCLUSION

There so many video text detection and extraction and region filling techniques are available. All have different performance as per different type of data or background present in video. The approaches presented for video text detection and extraction considered the different attributes of text such as size, font, style, orientation, alignment, contrast, color, intensity, connected-components, edges etc. Using these attributes the text regions is easily differentiated from their background or other regions within the image and easily detected. This paper does an extensive survey and comparison of the many previously proposed techniques for video text detection, extraction and region filling. A great challenge for this work is to perform the performance survey for all the previous methods.

5. REFERENCES

[1] A. Mosleh, N. Bouguila, and A. B. Hamza, "An Automatic Inpainting Scheme for Video Text Detection and Removal," in IEEE Transactions on Image processing, vol. 22, no. 11, pp.4460–4472, Nov. 2013.

[2] J. Gllavata, R. Ewerth, and B. Freisleben, "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients,"

in Proc. of International Conference on Pattern Recognition (ICPR), vol. 1, Aug. 2004, pp. 425–428.

[3] T. Pratheeba, Dr. V. Kavitha, S. Raja Rajeswari, "Morphology Based Text Detection and Extraction from Complex Video Scene," in Internationl Journal of Engineering and Technology, Vol. 2(3), pp. 200-206, 2010.

[4] W. Kim and C. Kim, "A new approach for overlay text detection and extraction from complex video scene," IEEE Trans. Image Process., vol. 18, no. 2, pp. 401 –411, feb. 2009.

[5] Keechul Jung, KwangIn Kim, Anil K. Jain, "Text information extraction in images and video: a survey", Pattern Recognition, Vol. 37, No. 5. (May 2004), pp. 977-997

[6] Yen-Lin Chen, "Automatic Text Extraction, Removal and Inpaiting of Complex Document Images," in International Journal of Innovative Computing, Information and Control, Vol. 8, No.1(A), pp. 303-327, January 2012.

[7] 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, feb. 2005.

[8] J. Malobabic, N. O'Connor, N. Murphy, and S. Marlow, "Automatic detection and extraction of artificial text in video," in WIAMIS 2004 - 5th International Workshop on Image Analysis for Multimedia Interactive Services, April 2004.

[9] C. Wolf , J.M. Jolion and F. Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Documents", Proceedings of the Int. Conference on Pattern Recognition (ICPR) 2002, vol.4, IEEE Computer Society,Quebec City, Canada, pp.1037-1040, August 2002.

[10] C. Wolf and J.M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents", Technical Report RVF-RR-2002.01, Available: http://rvf.insa-lyon.fr/~wolf/papers/tr-rfv-2002, February 2002.

[11] Mohammad Khodadadi and AlirezaBehrad, "Text Localization, Extraction and Inpainting in color images", ICEE2012, Vol.12, IEEE, May 2012, pp. 1035-1040.

[12] M. Cai, J. Song, and M. Lyu, "A new approach for video text detection," in Proc. of IEEE International Conference on Image Processing (ICIP), 2002, pp. I–117–I–120.

[13] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," IEEE conference on Computer Vision and Pattern Recognition(CVPR), 2010, pp. 2963–2970, June 2010.