Middle-East Journal of Scientific Research 19 (5): 716-722, 2014
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2014.19.5.21019

A Novel Hybrid Method for Text Detection and Extraction from News Videos

[1]*M. Daud Abdullah Asif,* [2]*Umair Ullah Tariq,*
[1]*Mirza Nauman Baig and* [1]*Waqar Ahmad*

[1]University of Engineering and Technology Taxila, Pakistan
[2]COMSATS Institute of Information and Technology Abbottabad, Pakistan

Abstract: Modern era has observed an enormous development in media information in the manifestation of audio, video and image data. Retrieval and Indexing of content-oriented video has evolved as an intriguing research zone with the colossal development in the product of advanced digital mass media. Not with standing varying media information, text showing up in videos can assist in an effective contraption for semantic abstraction, video analysis and recovery of video data. A proficient algorithm and high quality videos of news are required for accomplishing the desired task. This paper recommends a system dependent upon gray-scale edges-features for evenly arranged English ticker text localization and extraction from news videos. The framework exploits edge based localization of text regions to concentrate text based materials from videos. For low quality videos, some contrast enhancement operations are used to enhance the video frames first and then morphological operators are applied to segment out the ticker text regions in news videos. At last, these regions are cropped from the video frames and on satisfying certain geometrical constraints, the results are acknowledged to be text regions. No assumptions about the ticker color, style of text fonts, size of text and the types of ticker is made because no standard format of tickers exist in news videos of different channels and different countries have separate style of ticker texts format and color. The proposed algorithm is evaluated on a data set of CNN and BBC news videos and it displayed promising results.

Key words: Ticker text · Text Detection and Localization · Cropping · Morphological operations · Edge detection · Caption text

INTRODUCTION

The monitoring of news videos has been a paramount job for media experts, business world, political gatherings and brainpower firms. News analyst requires text reports generated by the news videos for their analysis. It requires a lot of time and self-effort to monitor the news round the clock and generate text reports manually. It is likewise plausible that the reports created manually by human minds could be inclined, taking into account their particular observation and might differ from individual to individual. So news videos observed manually from various stations may not be an exceptional alternative. To overcome this issue, an image processing application is required which automatically generate the text from ticker of news videos which can be used by all the analyst agencies.

The text content showing in videos is categorized into two groups i.e. caption or description text content which is also recognized as artificial/graphic content and scene text content. Description text content mentions the content which is rooted in the video like name of the anchors, scorecard and ticker text appearing at the subtitle location in the videos. This caption content is recognized advantageous for video indexing and recovery. This research work mainly focuses on the caption text content.

Customarily, the script detection methodology is isolated in two stages: firstly detection and localization of text and secondly withdrawal of text. In the first step, content areas and the non-content regions are recognized from each other. The regions with useful content areas are localized and upon satisfying the criteria that it only contains text regions is then fed into text extraction stage. As soon as the content sequences are concentrated from

Corresponding Author: Umair Ullah Tariq, COMSATS Institute of Information and Technology Abbottabad, Pakistan.

videos, they can be easily nourished into OCR i.e. an Optical Character Recognition System for character acknowledgment. The main focus in current research is drawing out of text only without its recognition. The research can be helpful in image processing academic and scientific fields.

This research work emphases on developing an efficient algorithm for ticker text detection and extraction based on video frames accessed from various sources like capturing video directly from various sources like the real time using TV capture card, using live news streaming and from prerecorded videos available on internet, CNN and BBC resources. The proposed text detection and extraction system acquires video frames of different news channels which are then enhanced by removing noises in pre-processing stage. These enhanced video frames are further used for identification of areas having ticker text and at last fed into the text extraction stage to get the cropped text regions from news videos.

Related Work: In this section some of the reputed algorithms are briefly discussed. These algorithms are used for the detection and extraction of text from videos/images. An all-out review of the procedures and algorithms for text information extraction recommended till 2004 presented by K. Jung, K.I. Kim and A.K. Jain can be recognized in [1]. The quality of videos on internet sources is sometimes not good. The image enhancement techniques presented by Shang, Yuan-Yuan, Hou, Xuefeng, Han, Baoyuan [2] produced good results. These techniques are based on median and homomorphic filtering, histogram equalization and average smoothing. The algorithm of blurriness and low quality image enhancement is presented by Bi, Xiao-jun, Wang, Ting [3]. In this framework, three distinctive arrangements are described for the blurriness such as motion blur, defocus obscure and Gaussian obscure. Algorithm utilizes some usual technique for image restoration routines to accomplish the desired enhancement.

The methodologies for text detection are arranged into two prevailing classes: Supervised and unsupervised methodologies. The Supervised method utilizes machine learning strategies for identification of text based substance in video or images. The features extracted from the image are utilized for training of a classifier like Support Vector Machine (SVM) and Artificial Neural Network (ANN) but it requires a lot of time to train a classifier. The unsupervised method however utilizes the image analysis method. This method exploits certain statistical properties of image analysis to recognize the text based regions in images.

Around the renowned regulated systems, SVM has been successfully utilized in [4-6] where qualities based on texture and edge recognition are used to recognize text regions and the non-text areas located in the images. In [7], native binary patterns has been deployed by Jun Ye, Lin-Lin Huang, Xiao Li Hao specifically for extracting features and classification of the text areas and the non-text areas with the help of polynomial neural network (PNN).

Unsupervised methodologies are based on connected components, edge-based and texture-based detection of text content regions. Connected components based strategies [8] commonly utilize gray scale images either fora bottom-up approach including region growing method or a top-down approach including region splitting method to aggregate pixels of texts into clusters. Edge-based methodologies [9-11] usually fragment the text content by discovering the edges residing in the image commonly emulated by certain morphological processing techniques. Texture grounded techniques [12, 13] based on special sort of textures which could recognize text regions and also distinguish these regions from non-texting areas. These aforementioned algorithms comparatively give better performance in complex image backgrounds. Nonetheless, they normally transform additional false positives results when the image background includes textures that produce some comparable properties as the textual content.

Algorithm based on edge detection to locate text in images of videos has been suggested by Akhtar Jamil, Imran Siddiqi *et al*. [14] which utilizes edge detection technique and algorithm of run length smoothing for the location of text in the images but it is made to detect only artificial urdu language text in video based images.

Some of the details regarding edge detection approaches like Sobel edge detector [15] have been implemented in our algorithm and have given good results. We have also examined Hough Transform [15] for the line detection of ticker and localization of subtitles in videos but Hough Transform needs a lot of amendments to detect the ticker boundary accurately.

In this paper a method has been proposed which accurately detects ticker text in the news videos and extracts the text information while preserving its benefits of low computational cost and no prior information. The proposed scheme is aimed to be light weight and inexpensive so that it can be run on real time to detect text from live news videos.

Proposed Methodology: In this section, the methodology is explained in detail which comprises of the following steps. Overview of the proposed methodology has been presented in Figure 1.

These steps are discussed in detail in the below given section. It mainly includes all the details of methodology utilized for extraction and detection of ticker text.

Video Acquisition: The quality of video plays a vital role in the research zone of image processing and computing vision. For this research work the news videos for processing are taken from TV capture card, live news channel streaming and some of the data set is also taken from the pre-recorded news channels videos major from CNN and BBC resources available on the internet. Though in some cases the quality of the videos is not up to the mark, but this algorithm has given good results.

Frame Extraction: There are two standards of videos available around the world namely NTSC and PAL. NTSC is the video framework or standard utilized within North America and the greater part of South America. In NTSC, 30 frames are transmitted every second. PAL is the transcendent video framework or standard basically utilized abroad. In PAL, 25 edges are transmitted every second. So in our case, news videos follow PAL standard, we can check every 25th frame of video for further processing.

The Figure 2 above shows the key frame extraction at the initial stage. These frames are fed to next level for further processing.

RGB to Gray Scale Conversion: The key frames are then converted from RGB to Gray scale at this stage. First all these frames are converted to a standard size of 704 x 576.The color component (Chrominance UV) is removed from every frame because every news video and ticker text present in the news video is in different color so by converting to gray scale, it speeds up the whole process with respect to space and time. The 8-bit gray scale value is imposed on each pixel of key frame by using the following formula:

R = Image (i, j, 1)
G = Image (i, j, 2)
B = Image (i, j, 3)

$$Y(i, j) = wr * R(i , j) + wg * G(i , j) + wb * B(i , j)$$

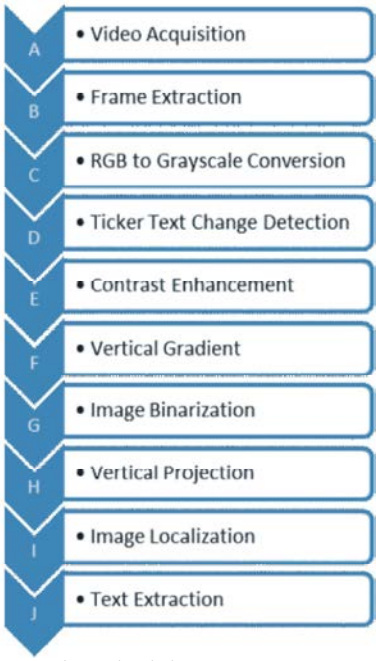

Fig. 1: Proposed Methodology Steps

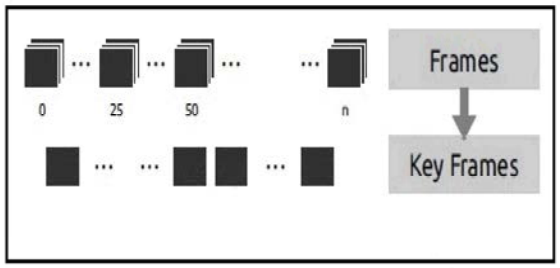

Fig. 2: Key Frames Extraction

where

wr = 0.59
wb = 0.30
wg = 0.11

Now we have all the frames in Gray scale which is then fed to pre-processing stage.

Ticker Text Change Detection: At this pre-processing stage, the frames have clear difference of contrast in Gray scale of ticker text location and the background. As the duration of one news on news ticker is more than one second and it is not changing every second so we can skip most of the frames which contains same news at one time. By selecting only specific frames, this algorithm becomes more reliable and processing becomes faster as rest of the process contains only specific video images which definitely contain different news. For detecting

different news at the ticker location, a module is added in our algorithm which checks the contrast values change at the ticker location it then saves the next frame and starts comparing it with the next coming frames. This is done using Correlation method. By this, most of the same news frames are filtered out. Only selected filtered frames are fed to next level for further processing. The pseudo code for this stage is given below:

If change in Gray scale value is detected by subtraction

- Binarize the frame
- Count change in pixels value in text region using Correlation
- If increase or decrease of pixel value
- Save the frame

End
End

This shows the ticker text change module pseudo code. It works perfectly and saves only frame with different news and these frames are then further processed.

Contrast Enhancement: The pixel values of the video images are histogram equalized for enhancing the quality of every image.256 gray levels (0 ~ 255) are used for contrast stretching where 255 indicates the maximum gray level which is white and minimum gray level value 0 is black. For poor quality videos, this step enhances their quality by contrast enhancement for further processing. If noise is much greater that enhancing its contrast increases more noise than first Median filtering could be applied which is a nonlinear operation which swaps the gray scale value of a picture element by averaging gray values of its neighbors. Median filter is often used in image processing to reduce salt and pepper noise. Then this filtered image is contrast enhanced.

Vertical Gradient: Like numerous languages, vertical strokes and scripting texts are very prominent in English language. The vertical edges of the text have maximum edges count and row vice variance due to its characters as they appear in groups rather than isolation and share some height. Nowhere else in the video image this pattern appears other than the ticker location. Therefore Sobel operator is used for computing the vertical gradient.

$$Sobel_{Mask} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

In Sobel mask given above, a 3x3 Sobel edge mask is used to find the gradient edges of the contrast enhanced video image in the x direction. The vertical edge at point (I, j) in contrast enhanced video image is found using the following equation.

$$GradVx(j,i) = \sqrt{\left(\sum_{a=-1}^{+1} \sum_{b=-1}^{+1} Sobel_{Mask} X \operatorname{Im} g_{Ench}(j+b,+a)/4 \right)^2}$$

This equation is used in algorithm to find the vertical gradient of video frames.

Image Binarization: At this point, the video image is Binarize using a threshold formula given below.

$$ThreshVx = Mvx + 3.5 * MVx$$

where MVxis the mean value of Gradient Vx video image and is calculated without including the non-zero elements. Mvx is calculated as given by the following formula.

$$MVx = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} ThreshVx(j,i)}{NxM}$$
$$where,\ GradVx(j,i) \neq 0$$

This equation shows the calculation of Mvx which is used to calculate the threshold value to binarize the video frame.

Vertical Projection: After image Binarization, the vertical projection of the video images is calculated. The vertical projection includes the row wise sum and variance of edges in each video frame. The graph of sum and variance is calculated which leads to image segmentation and it help in localization of text regions present in the video image. The variance graph value is scaled to 0 and image width to make a mask based on a threshold calculated as following.

$$ThreshVarMax = \frac{max(\operatorname{var} ianceROW)}{3}$$

The mask is applied on the video image to segment out the initial regions of interest which further leads to localize the video image.

Image Localization: From the aforementioned steps of our algorithm, we can viably recognize the text regions oppressing the greater part of the non-text areas by seeing the edges variance graph. For suppressing the non-text areas in the video frames, image built on the number of edge counts are localized in the variance graphs of the video images. To find the maximum number of edges on the ticker location of the news videos, the whole image under process is localized to find the location of the ticker. We can localize the image either by connected component analysis or gray scale edges count method. In this algorithm we have used edges count method which can suppress all the non-text regions when a specific number of edges count is taken to be the non-text region. By this way all the text regions are recognized. The region with maximum number of edges in the video frame is taken as localized ticker text location area.

Text Extraction: Text Extraction from news videos is the ultimate goal of the current research. As a closing step, certain geometrical checks are engaged to extract the localized ticker text areas from localized text region with maximum number of edges count. With the application of this algorithm, most of our dataset videos produced ideal results which are true positive and text is extracted from news videos perfectly.

RESULTS

The proposed algorithm is developed using Intel Core i5 processor with 4 Giga Byte RAM and this algorithm is optimized to be directly used with live news video streams and TV capture card videos.

The proposed algorithm is applied on numerous videos dataset. One of the news video frames is shown in Figure 3(a). Then this image is converted from RGB to Gray scale and contrast enhanced as we described earlier. Image with contrast enhancement is shown in Figure 3(b). In figure 3(c) and 3(d) the graphs of variance and sum of edges is shown respectively. At the end, in Figure 3(e) the successful text extraction figure is shown.

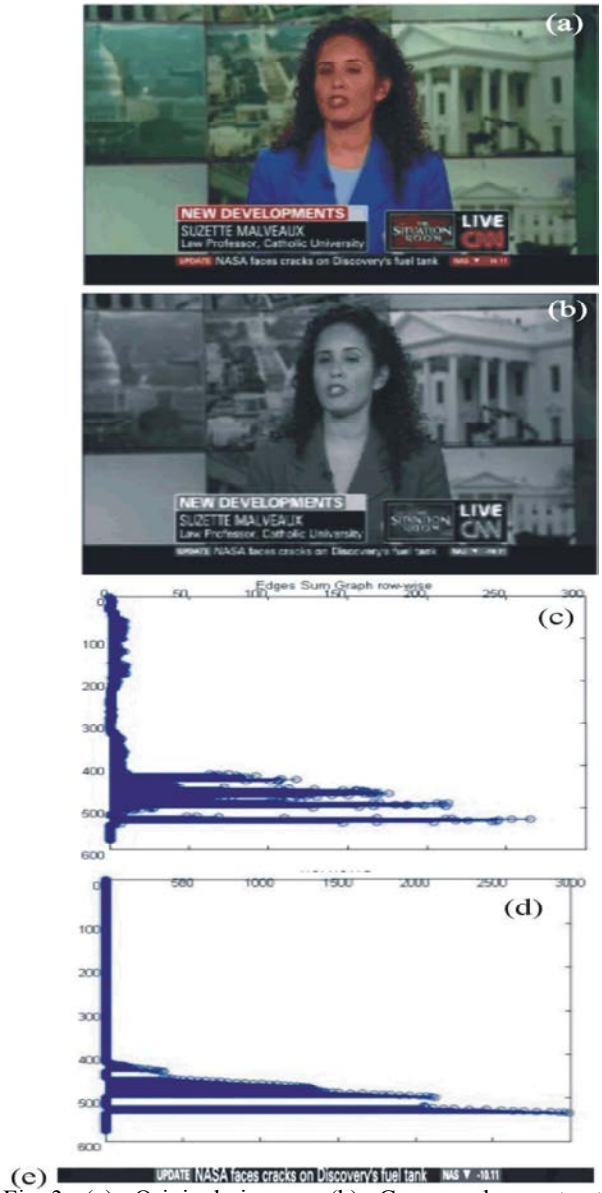

Fig. 3: (a) Original image, (b) Gray scale contrast enhanced image,(c) Variance graph, (d) Edges sum graph, (e)Ticker Text extracted from video

Another result of different video is shown in figure 4 below which also gives perfect result. Only one key frame is shown in Figure 4 from the whole video frames result to illustrate the precision of this algorithm.

The results in Figure 5 show that if the quality of video is very poor, this algorithm gives perfect results. The precision of proposed algorithm is very high and it gives flawless results even when the video background is very complex as shown below. The symmetry of the figure 5 is same as described in previous results.

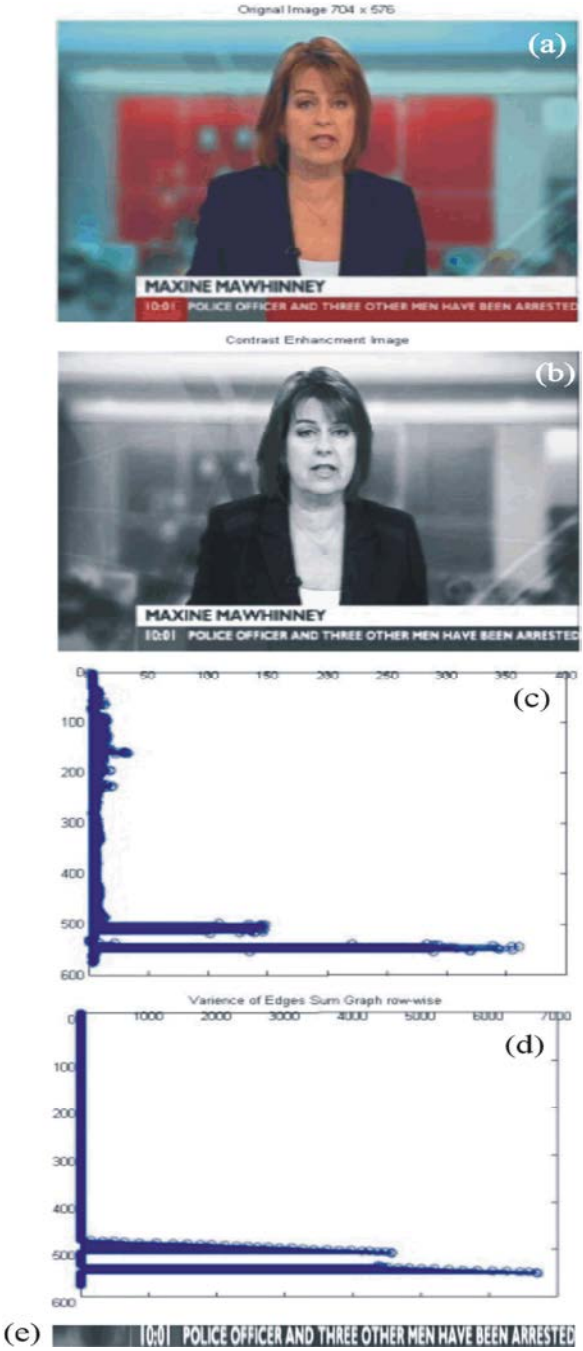

Fig. 4: (a) Original image, (b) Grayscale contrast enhanced image (c) Variance graph, (d) Edges sum graph, (e) Ticker Text extracted from video

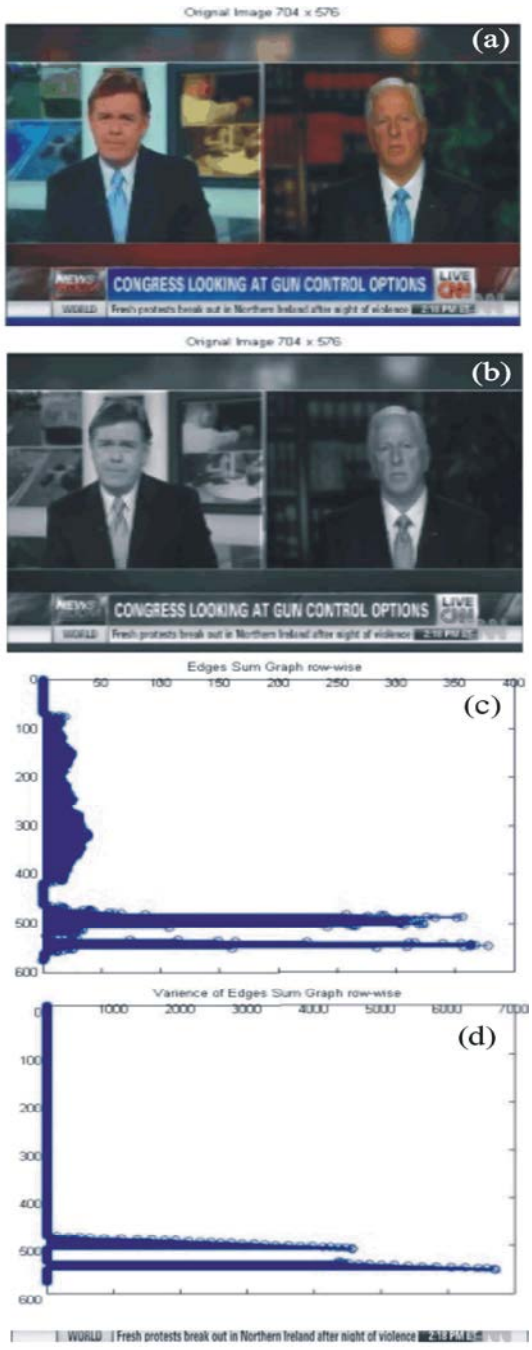

Fig. 5: (a) Original image, (b) Gray scale contrast enhanced image (c) Variance graph, (d) Edges sum graph, (e) Ticker Text extracted from video

Among the both forest site, species richness value was maximum in pine forest at HB (10.5) and minimum in oak forest at HB (7.4). Beta diversity showed pronounced effect at both sites. The value for oak forest varied marginally from 4.5 (HB) to 4.6 (HS), respectively.

Table 4: Forest wise ratio of species, genera and family (F, Family; G, Genus; S, Species)

Forest	F:G	F: S	G: S
Oak	1.2	1.3	1.0
Pine	1.4	1.5	1.1

While for pine forest, it remained approximately same at all sub-sites. Between the forests, the value was higher in oak forest than pine forest. The lowest value of beta-diversity in oak forest was observed at HB (4.5) and for pine forest at HS (2.8). Equitability/evenness value ranged from 17.0 (HT) to 31.7 (HB) in the oak forest. A reverse pattern was observed in the pine forest (31.4 at HT and 27.3 at HB).

DISCUSSIONS

An efficient algorithm is recommended for detecting, localizing and extracting text from news videos. The algorithm comprises of ten major steps in order to get the ticker of news videos extracted. The efficiency of the developed algorithm in this particular research work can be proved from the experimental results. Algorithm provides desired results for both the good and the poor quality videos. Extracted region of text could be given to the OCR engine module for extraction of characters from the ticker in ASCII.

REFERENCES

1. Jung, K., K.I. Kim and A.K. Jain, 2004. Text information extraction in images and video: a survey. Pattern Recognition, 37, 2004. Smith MD, Wilcox JC, Kelly T, Knapp AK. Dominance not richness determines invasibility of tallgrass prairie. Oikos, 106(2): 253-62.
2. Shang, Yuan-Yuan, Hou, Xuefeng, Han, Baoyuan, 15-17 Oct 2011. Research on image enhancement algorithms based on Matlab, Image and Signal Processing CISP, 2: 733-736.
3. Bi, Xiao-jun, Wang, Ting, 27-30 May 2008. Adaptive Blind Image Restoration Algorithm of Degraded Image, Image andSignal Processing, 2008. CISP '08. Congress, pp: 536-540.
4. Marios Anthimopoulos, Basilis Gatos and Ioannis Pratikakis, 2008. A Hybrid System for Text Detection in Video Frames, The Eighth IAPR Workshop on Document Analysis Systems.
5. Rongrong Wang, Wanjun Jin and Lide Wu, 2004. A Novel Video Caption Detection Approach Using Multi Frame Integration. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), pp: 1051-4651.
6. Guangyi Miao, Qingming Huang and Shuqiang Jiang, 2008. Wen Gao. Coarseto-fine video text detection.
7. Jun Ye, Lin-Lin Huang and XiaoLiHao, 2009. Neural network based text detection in videos using local binary Patterns, pattern recognition.
8. Fan, W., J. Sun, Y. Katsuyama, Y. Hotta and S. Naoi, 2009. Text Detection in Images Based on Grayscale Decomposition and Stroke Extraction Proc. Chinese Conf. Pattern Recognition CCPR 2009.
9. Palaiahnakote Shivakumara, TrungQuyPhan and Chew Lim Tan, 2009. A laplacian approach to multi-oriented text detection in video, IEEE transactions on pattern analysis and machine intelligence, pp: 33.
10. Teo Boon Chen D. and Ghosh S. Ranganath, 2004. Video-text extraction and recognition.
11. Wolf, C., J.M. Jolion and F. Chasseing, 2002. Text Localization, Enhancement and Binarization in Multimedia Documents. In Proceedings of the 16th International Conference on Pattern Recognition ICPR'02, Quebec, Canada, pp: 1037-1040.
12. Zhu, C., W. Wang and Q. Ning, 2006. Text Detection in Images Using Texture Feature from Strokes. LNCS - Advances in Multimedia Information Processing, pp: 295-30.
13. Wu, V., R. Manamatha and E. Riseman, 1999. Text finder: an automatic system to detect and recognized text in images, IEEE Trans, On PAMI, pp: 20.
14. Akhtar Jamil, Imran Siddiqi, FahimArif and AhsenRaza, 2011. Edge-based Features for Localization of Artificial Urdu Text in Video Images, International Conference on Document Analysis and Recognition.
15. Rafael Gonzalez and Richard Woods, 2002. Image Enhancement, Image Segmentation Digital Image Processing. 2nd Edition, Prentice Hall, Newyork.