Scientific Journal Impact Factor (SJIF): 1.711

e-ISSN: 2349-9745
p-ISSN: 2393-8161

International Journal of Modern Trends in Engineering and Research
www.ijmter.com

A SURVEY ON VIDEO EXTRACTION BASED ON CONCEPTUAL FRAMEWORK

Abhishek Kundu

Rajiv Gandhi College Of Engineering, Research And Technology,
Chandrapur, Maharastra, India

Abstract: Since decade, with rapid growth of available multimedia documents and increasing demand for information indexing and retrieval, much effort has been done on text extraction in images and videos. There are many challenges and difficulties for designer and developer of video extraction process. A lot of work has been done in the field of video extraction from multimedia data. But most of the work is application specific and there is still need of work in designing domain independent systems. This is because there are so many challenges when extracting video with variation in fonts, size, color, alignment, orientation, illumination and background. Problem of video extraction get very difficult because of these deviations. This paper presents a conceptual framework for video extraction derived from the research literature and used as a means for surveying the research literature.

Key Words: Text extraction, Video Retrieval, Frame Extraction

I. INTRODUCTION

With the dramatic increase in multimedia data, escalating trend of internet, and amplifying use of image/video capturing devices; content based indexing and text extraction is gaining more and more importance in research community. The amount of digital multimedia data is growing exponentially with time. Thousands of television stations are broadcasting every day. With the vast spread of affordable digital cameras and inexpensive memory devices, multimedia data is increasing every second. Ranging from cameras embedded in mobile phones to professional ones, Surveillance cameras to broadcast videos, every day images to satellite images, all these increasing multimedia data[1].

Video Extraction techniques produce by analyzing the underlying content of a source video stream, condensing this content into abbreviated descriptive forms that represent surrogates of the original content embedded within the video [2]. The multimodal nature of video, which conveys a wide range of semantics in multiple modes, such as sound, music, still images, moving image, and text [3], makes this task much more complex than analyzing text documents. Furthermore, video Extraction research faces the challenge of developing effective techniques for abstracting useful and intuitive semantics from the video stream that are in step with the individual users' comprehension and understanding of video content [4, 5].

The conceptual framework was developed based on a survey and analysis of contemporary video Extraction research literature. The literature included in the survey was identified by searching a range of full text databases for articles which proposed new video Extraction techniques and/or video summaries with a view to collecting a large sample of recent work within the field. To ensure the timeliness and manageability of the included literature, the search was initially limited to articles appearing within the last three years; however, a number of additional articles considered key within the field were also considered if published outside of the specified time frame.

The organization of the paper as first section describes in brief video extraction. In section II reviews the related work. Section III study the conceptual framework for video summarization
and its techniques In Section IV the detailed case study of soccer ball game video clip. results are Finally, in Section VI, conclusions.

II. RELATED WORK

The previous work which has been done on this concept has been basically done on images of documents that have been scanned which are typically binary images or they can be easily converted to binary images using simple binarization techniques i.e. converting the color image into a grayscale image and then thresholding the grayscale image[6].
Thai et.al [7] described an approach for effective text extraction from graphical document images. The algorithm used Morphological Component Analysis (MCA) algorithm, an advancement of sparse representation framework with two appropriately chosen discriminative over complete dictionaries. Two discriminative dictionaries were based on undecimated wavelet transform and curvelet transform. This method overcame the problem of touching between text and graphics and also insensitive to different font styles, sizes, and orientations.
Our work generalizes image-based rendering to the temporal domain. It can thus be thought of as a kind of "video-based rendering." A similar idea has been used in video games, in which hand generated video loops have been created to simulate natural phenomena like fire or water. However, there has been little previous work on automatically generating motion by reusing captured video. Probably the work most closely related to our own is "Video Rewrite" [8], in which video sequences of a person's mouth are extracted from a training sequence of the person speaking and then reordered in order to match the phoneme sequence of a new audio track. Related 3D view interpolation techniques have also been applied to multiple video streams in the Virtualized Reality [9] and Immersive Video [10] projects. Pollard et al. [11] introduced the term "video sprite" for applying such techniques to an alpha-matted region of the video rather than to the whole image. Finkelstein et al. [12] also used alpha-matted video elements in their earlier multi resolution video work, which they called "video clip-art."
The Significant event detection from particular video is essential for video retrieval. However, the existing sports video event detection approaches heavily rely on either video content itself, which face the difficulty of high-level semantic information extraction from video content using computer vision and image processing techniques, or manually generated video ontology, which is domain specific and difficult to be automatically aligned with the video content. Many techniques have been proposed for video event detection and extraction based on supervised and unsupervised learning.
This paper reviews some of the recent work on content-based multimedia information retrieval and discusses their role in current research directions which include browsing and search paradigms, user studies, effective computing, learning, semantic queries, new features and media types, high performance indexing, and evaluation techniques. Based on the current state of the art, we also discuss the major challenges for the future.

III. CONCEPTUAL FRAMEWORK FOR VIDEO EXTRACTION

In order to identify and extract the various audiovisual cues to be included within video summaries, the underlying content of a video must first be analyzed [13]. The framework contains four major parts: web-casting text analysis, broadcast video analysis, text/video alignment, and semantic annotation and indexing for personalized retrieval[14].

A. Video Summarization

Enormous popularity of the Internet video repository sites like YouTube, Yahoo Video, lecture videos, and social networking sites like face book, Google+ etc. have caused increasing amount of the video content available over the Internet. In such a scenario, it is necessary to have automatic mechanisms of generating concise representation of the video content as a sequence of still or moving pictures i.e. video summary.

The major task in video summarization is to segment the original video into shots and extract those video frames from the original video that would be the most informative and concise representation of the whole video. Such frames are referred as *key frames* [1]. Fig. 1 represents the structural hierarchy of a video. Key frames can be extracted locally or globally using various visual or audio features [1, 2].

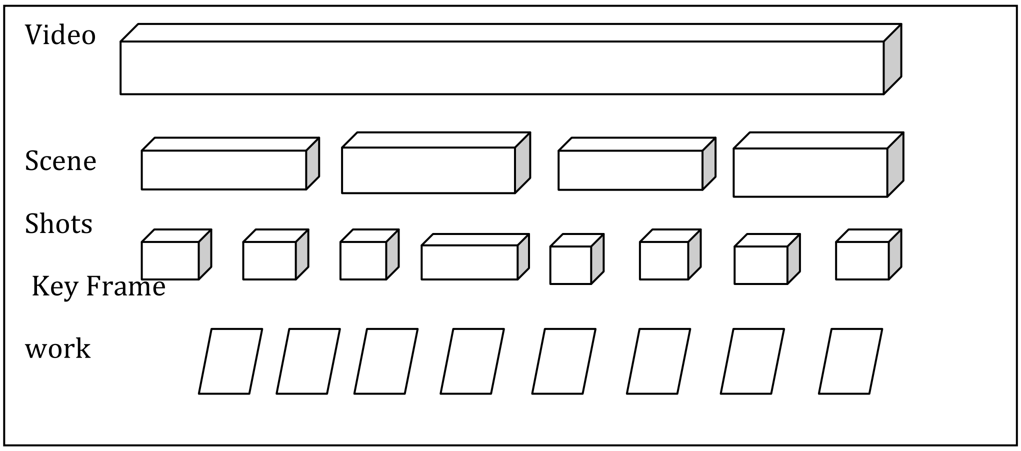

Figure 1: Structural Hierarchy Of A Video

Input Video The video can be in the format of AVI (Audio Video Interleave). To process this video, frames have to be extracted. The AVI format was developed by Microsoft. The AVI format is supported by all computers running Windows, and by the entire most popular web browser [1].

Frame Extraction As video consist of number of frames depend upon size of video. These frames occupy large space in memory. Frame rate is about 20 to 30 frames per second. The video taken as input is divided into frames in this section.

Feature Extraction The feature extraction process can be based on visual or audio features.

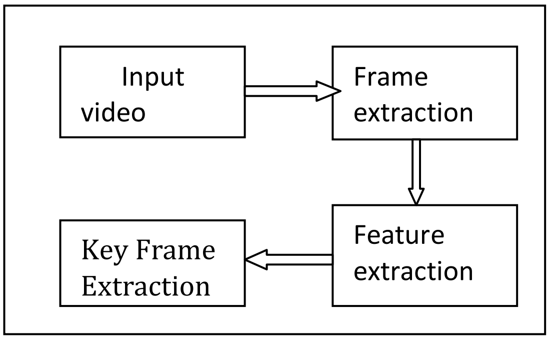

Figure 2: Key Frame Extraction

The visual features of the extracted key frames can be color, edge or motion features [1]. The low level features such as color histogram, frame correlation and edge histogram are obtained using certain frame difference measures. Then the frame difference values are calculated for all extracted frames for different videos.

Audio Features shows that for semantic and effective analysis various audio features can be embedded with low level visual features for key frame extraction. The most common audio classes in videos are speech, silence, music and the combination of later three [6].

Key frames Selection To start the extraction process, the first frame is declared as a key frame. Then the frame difference is computed between the current frame and the last extracted key frame. If the frame difference satisfies a certain threshold condition, then the current frame is selected as key frame. This process is repeated for all frames in the video.

Video retrieval Video shot segmentation and abstraction are two parts of a larger problem of content based video retrieval. Once the video has been segmented into shots which may be represented by key frames, scene transition graphs or multi-level representations, it is necessary to provide methods for similarity based retrieval. Jain et al. [15] present a method for retrieval of video clips which may be longer than a single shot. As shown in figure 3 ,the methods allow retrieval based on key frame similarity, akin to the image database approach. A video clip is then generated as a result by merging shot boundaries of those shots whose key frames are highly similar to the query specification. In another approach sub sampled versions of video shots are compared to a query clip. Image database matching techniques are extended for content based video retrieval in Ref. [16]. Fuzzy classification and relevance feedback is used for retrieval. A multi-dimensional feature vector is formed from color and motion segmentation of each video frame that is used to construct a vector and extract key shots and key frames.

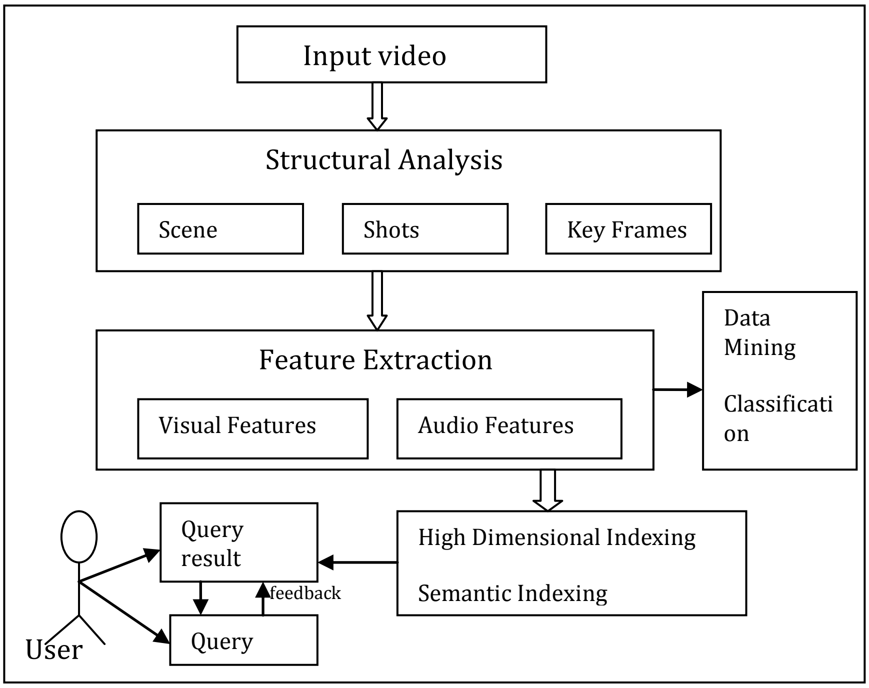

Figure 3 : Generic Framework For Visual Content-Based Video Indexing And Retrieval.

B. Extracting The Video Texture

A variety of techniques for text extraction are appeared in recent past [10]. Comprehensive surveys can be traced explicitly in [16]. These techniques can be categorized into two types mainly with reference to the utilized text features i.e. region based and texture based [14].

Texture based methods pertain to textural properties of the text, distinguishing it from the background. The techniques mostly use Gabor filters, Wavelet, FFT, spatial variance, etc. These methods further use machine learning techniques such as SVM, MLP and adaBoost [15]. These techniques work in the top down fashion by first extracting the texture features and then finding the text regions. Texture based techniques usually give better results in complex backgrounds than region based techniques but have computationally very heavy hence not suitable for retrieval systems for hefty databases. Therefore, there is a need to improve the detection results of region-based techniques to be used for retrieval and indexing of large multimedia data

Region based approach exploits different region properties to extract text objects. This approach makes use of the fact that there is sufficient difference between the text color and its immediate background. Color features, edge features, and connected component methods are often used in this approach [14]. These techniques typically work in the bottom up fashion by first segmenting the small regions and then grouping the potential text regions. Region based techniques typically work in the bottom up fashion by initially segmenting the small regions and lately grouping the potential text regions. Region based methods are generally composed of three modules. (1) Segmenting the image into small regions which aims at segregating the character regions from its background, (2) Merging and grouping of small regions to form words and sentences (3) Differentiating between text and non text objects.

C. Presentation Of Video Summarization

After extracting the key frames of video sequence the similar key frames are clustered together and are presented in a condensed form to the user. Video summarizations are commonly presented as a set of static key frames or dynamic videos skims [8].

Static Presentation One of the most common video summarization presentation techniques is a storyboard, which is usually a static grid of extracted key frames. According to a recent study on evaluation of video summarization techniques [9], the storyboard has a capability to give an informative summary of the original video content. However, according to the user studies the storyboards lacked in their representativeness and ability to replace the original video content. **3.2 Dynamic Presentation** Dynamic video skimming is a technique that condenses the original video into a shorter version, while preserving important content with its time-evolving properties. Hence, video skims are practically short video clips cut from the original video sequence. Preservation of motion information is one of greatest advantages of video skims, in addition to aural information, which can both enhance the expressiveness of the video summary. Compared to static storyboards, dynamic videos skimming also support the recognition of objects in the content, and their representativeness is enough even for replacing the original video content. According to the user study [8] the dynamic video skims were liked especially due to the clarity in presentation and the normal pace of the moving imagery. Generation of dynamic and static video summaries has been usually carried out differently, but it is still possible to transform from one form to another. Whereas video skims can be created from key frames by joining fixed-size segments, subshots, or the whole shots they are included in, the set of key frames for static storyboard can be created from a video skim by uniform sampling or selecting one frame from each skim [8, 9].

IV. CASE STUDY

Separating the highlights and key messages from hours of raw footage requires a tool that makes for easy clipping and painless export to multiple devices. For example, suppose you want to show a series of short movie clips as part of a lesson. You can extract only the clips you want and share them as separate files. Imagine that you've identified some videos to include in your research project, but you only need to share a few minutes from each video. You can very easily extract only the parts you need

A. Soccer Ball video clip (.avi file)

Datasets:

Various sports game video files used in our experiments were collected from a wide range of sources via the Internet. After excluding those video files that either have poor digital quality or do not contain any goal scene, there are 10 video files left, with different styles and produced by different broadcasters. The proposed model accepts the input video in the form of ".avi". The avi cutter is used to cut the large avi video in to samples of certain sizes ranging from 3MB to 90 MB. The corresponding sound tracks have been extracted from the Read video and extract the number of

frames around the audio peak Create a video for the audio peak and output Repeat the steps for complete stream in fixed step size[3]. Results for such a sample is explained below,
The new method here uses human knowledge directly and in a very efficient way by a fuzzy rule base. The presented structure allows the system to process based on soccer video shots available in the database. The first phase is devoted to extracting shots from each video and making a list of features extracted from each shot. Then a fuzzy system is used to eliminate shots including insignificant events. Finally shots are classified and associated with predefined classed using a SVM. Then shots related to the class associated with the user query are provided as an answer to that query. The user may make queries on different events and concepts such as occurrence of penalties, corners or goals or team attacks throughout the data base .Figure 4 shows the detailed extraction of the video clip.

Step 1: The input for this algorithm is soccer videos, hence initially Avi soccer clip is been read.
Step 2: Key-frames are been extracted detection and 3 types of views, Far-view, mid-view and out-view is been categorized.
Step 4: Considering the frames of mid-view generated above, line features are extracted.
Step 5: Fuzzy Inference System is designed to retrieve significant events from a soccer video. The input to the fuzzy inference system is line features and grass percentage extracted from the above steps.
Step 6: Finally, the significant video is constructed around the event frame generated by fuzzy.

All static regions distinguishable as text by humans were included in the experimentation process. Closely spaced words lying along the same alignment were considered to belong to the same text instance. Test data contains a total of 882 temporally unique artificial text instances.
Our evaluation is based on the number of text regions, including region recall region precision and region false alarm.

$$\text{Recall} = \frac{\textit{Number Of corrected dteectd text regions}}{\text{Number of det ected text regions}} \qquad (1)$$

$$\text{Precision} = \frac{\textit{Number Of corrected dteectd text regions}}{\text{Number of all ground \_ truth text regions}} \qquad (2)$$

$$\text{False alarm} = \frac{\textit{Number Of wrongly detected text regions}}{\text{Number of detected text regions}} \qquad (3)$$

Recall rate evaluates how many percents of the detected videotext regions are correct. Precision rate evaluates how many percents of all ground-truth video text regions are correctly detected. False alarm rate evaluates how many percents of the detected videotext regions are wrong as defined in (1), (2) and (3). The overall results of our method were 89,01 % Recall, 88,05 % Precision and 11,95% as false alarm. These results show that our technique is efficient, capable to locate text regions with different character sizes and styles, even in case of texts occurring within complex image background and that lay on the boundary of the video frame. Besides, our algorithm has also the high speed advantage with 2 frame pair per second in the frame size of 352 x 288 which suitable for videotext detection.

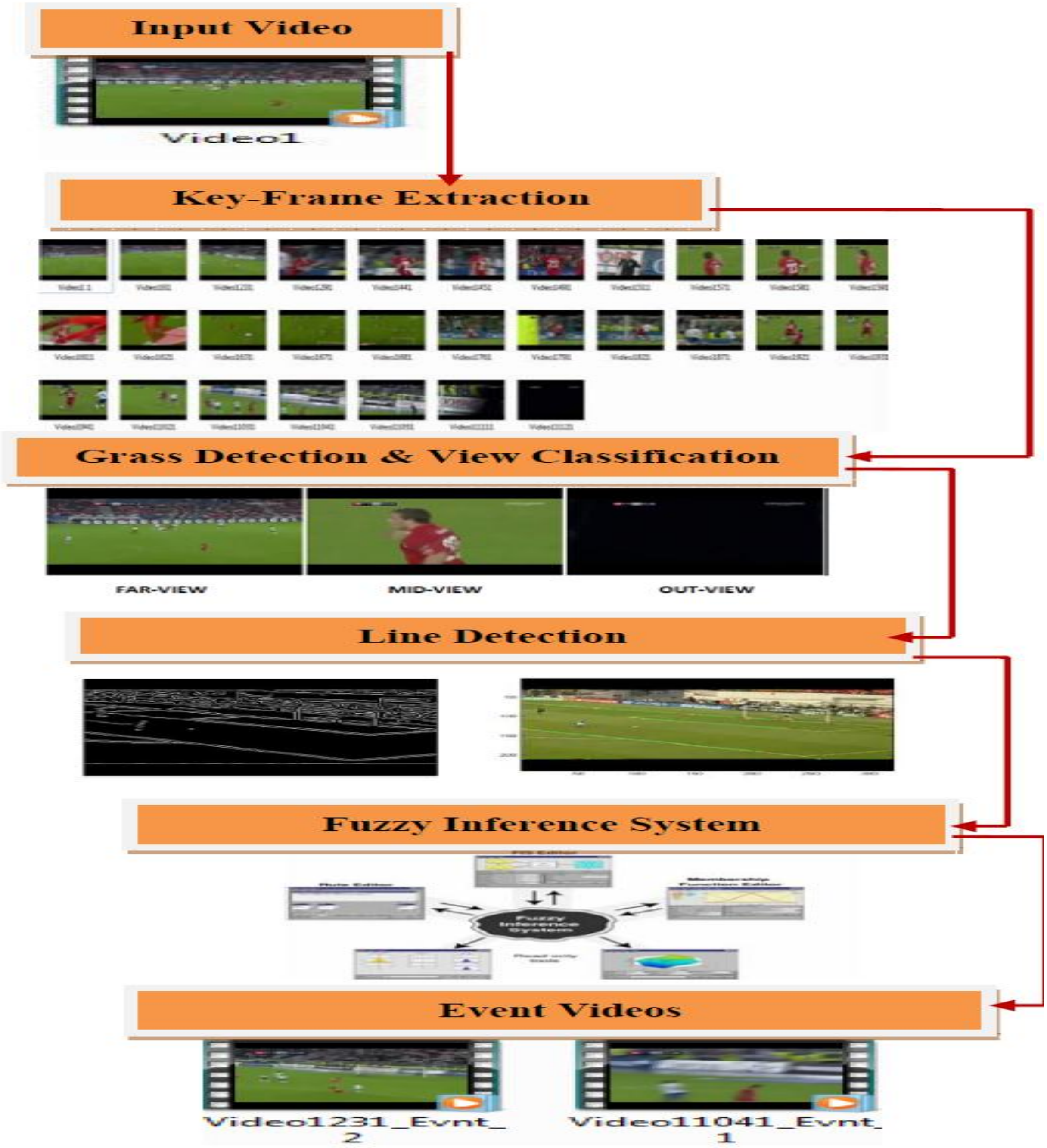

Figure 4 : Methodology For Event Retrieval.

V. CONCLUSION

The survey shows that there is a requirement of improvement of state-of-the-art techniques. Although the existing methods proved to be faster, less complex but still there is a need of an intelligent system which can automatically consider the most important key frame based on user's likes and dislikes and display the personalized video summary. The amount of research carried out in the domain of video summarization using machine learning is quite less. Also the consideration of user's requirement was not paid much importance. But the result analysis shows that the result generated by machine learning were quite effective. Hence for the better result the existing techniques can be tested in comparison with the machine learning algorithms to develop an intelligent system.

REFERENCES

[1]. Weiming Hu, Senior Member, IEEE, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. IEEE Transactions On Systems, "A Survey on Visual Content-Based Video Indexing and Retrieval" Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 6, November 2011.

[2]. Pradeep K M. Tech Dept. of CSE SITCOE, Yadrav Ichalkaranji "Significant Event Detection in Sports Video Using Audio Cues" Vol. 3 Issue 1 October 2013 Vol. 3 International Journal of Innovations in Engineering and Technology (IJIET).

[3]. Mr. Ganesh.I.Rathod and Mrs.Dipali.A.Nikam "Review on Event Retrieval in Soccer Video" International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5601-5605

[4]. Z. Xiong, X. Zhou, Q. Tian, Y. Rui, and T. Huang. Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports. IEEE Signal Processing Magazine, 23(2):18–27, March 2006.

[5]. G. Lavee, E. Rivlin, and M. Rudzsky, "Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 39, no. 5, pp. 489–504, Sep. 2009.

[6]. J. Tang, X. S. Hua, M. Wang, Z. Gu, G. J. Qi, and X. Wu, "Correlative linear neighborhood propagation for video annotation," IEEE Trans. Syst., Man, Cybern., B, Cybern., vol. 39, no. 2, pp. 409–416, Apr. 2009.

[7]. Thai V. Hoang , S. Tabbone(2010),"Text Extraction From Graphical Document Images Using Sparse Representation" in *Proc. Das*, pp 143–150.

[8]. X. Chen, C. Zhang, S. C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," IEEE Trans. Syst, Man, Cybern., C: Appl. Rev., vol. 39, no. 2, pp. 228–233, Mar. 2009.

[9]. Y. Song, X.-S. Hua, L. Dai, and M.Wang, "Semi-automatic video annotation based on active learning with multiple complementary predictors," in Proc. ACM Int. Workshop Multimedia Inf. Retrieval, Singapore, 2005, pp. 97–104.

[10]. Y. Song, X.-S. Hua, G.-J. Qi, L.-R. Dai, M. Wang, and H.-J. Zhang, "Efficient semantic annotation method for indexing large personal video database," in Proc. ACM Int. Workshop Multimedia Inf. Retrieval, Santa Barbara, CA, 2006, pp. 289–296.

[11]. Keechul Jung, "Neural Network-based Text Location in Color Images," Pattern Recognition Letters, 2001, vol. 22, pp. 1503-1515.

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

[13]. R. Lienhart and A. Wernicke, "Localizing and Segmenting Text in Images and Videos," IEEE Transactions on Circuits and System for Video Technology, 2002, vol. 12, pp. 256-268.

[14]. S. Antani, D. Crandall, and R. Kasturi, "Robust extraction of text in video," in Proc. 15th Int. Conf. Pattern Recognit, vol. 1, 2000, pp. 831- 834.

[15]. D. Chen, K. Shearer, and H. Bourlard, "Text enhancement with asymmetric filter for video OCR," in Proc. 11th Int. Conf. Image Anal.Process, 2001, pp. 192-197

[16]. M. Cai, J. Song, and M. R. Lyu, "A new approach for video text detection," in Proc. Int. Conf Image Process., Rochester, NY, Sep. 2002,pp. 117-120.