EFFICIENT VIDEO SUMMARIZATION BASED ON VIDEO EMOTION AND AUDIO SEMANTIC ANALYSIS

J.JayaBharathy[1], V.Renu[2]
[1]*Associate Professor, Department of Computer Science Engineering, Pondicherry Engineering College, Puducherry*
[2]*Student, Department of Computer Science Engineering, Pondicherry Engineering College, Puducherry.*

***Abstract* :-Video summarization is to produce a video chunk by removing the duplications and preserving the most representative content in the original video. A Video Summary consists of a single shot with relatively consistent high-level semantics and emotional content. To extract a summary from a long sequence of video, a few representative segments are generally sufficient. These representative segments could be selected based on the segment-level semantic and emotional recognition. The input video is segmented as frames and clustered based on a similarity of frames. Each cluster is then analyzed based on semantic analysis and emotion recognition. The existing system addresses only the emotion recognition for video summary generation whereas the proposed system generates video summary based on semantic analysis and facial emotion recognition. Moreover, the proposed system also considers the most representative audio(text) for video generation. From the video human facial expression is detected and trained by GABOR filter. The detected features are classified by Relevance Vector Machine (RVM) classification technique. The proposed algorithm reduces the time complexity and provides accurate emotion detection. The experimental results prove that the video summary generated through our framework gives an efficient summary.**
***Keyboard* :Video Summarization, Support vector machine, Relevance vector machine, Gabor filter,Facial expression.**

1. INTRODUCTION

Techniques for efficiently managing video sequence are becoming increasingly important. Among many practical needs, video summarization, has gained more attention in today's era. Most existing works on video summarization focused on the professionally generated videos (PGVs), which are normally very long and contain multiple camera shots. These unique characteristics of the user-generated videos (UGVs) demand specially designed summarization solutions.In addition, as the amount of the UGVs is extremely large, efficiency is an important factor in order to ensure that the related systems can be easily deployed in real-world applications.Three important clues are considered and integrated: semantics, emotions and quality. As the UGVs contain rich semantic and emotional contents, it is important to preserve both semantic and emotional representative contents [1].To ensure a good quality of the summary, a few quality measures including both motion and lighting conditions with the semantic and emotional clues for segment selection have to be considered.

Figure 1 demonstrates the procedure that is adopted during the process of video summarization [22]. Stage1. Video clip have been taken and analyzed and is segmented into a scene or a shot of the images. 2. Each visual feature is extracted from each image for a scene or shot. The frames are grouped by an unsupervised clustering method. 3. The extracted visual features are used to identify frames with similar content. 4. One key frame per cluster is selected based on its high representativeness. This permits to identify a frame with capability to represent the visual content of all others in its cluster. 5. Finally, a static video summary is a key frame filtered to eliminate into visual feature [22].

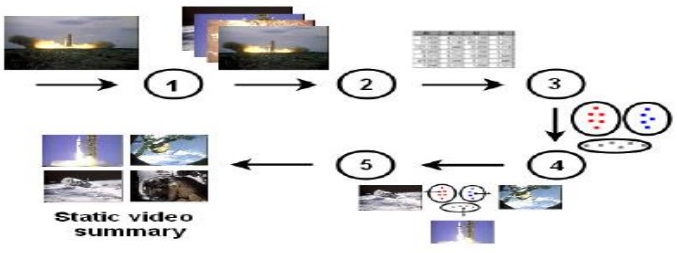

Fig. 1. Video summarization architecture.

The existing researches focus mainly on facial expression of the video sequence for summarization. In most of the video sequences audio places an important role. Considering this issue, a framework has been proposed to generate efficient video summary based on most representative video segment and audio (text). Experimental results shows that the summary generated based on our proposed framework is more efficient compared to the existing technique. The contribution of this paper is as follows. Section 2 consist of Related works are discussed and a study has carried out to understand the features that are important in video summary generation. Section 3 consist of the proposed approach and in Section 4 experimental results are presented.

2. RELATED THE WORK

This section gives the detailed discussion about the key points of video summarization and also highlights the some of the related works in recent researches.

2.1. VIDEO SUMMARIZATION

Video summarization is considered as one of the most important feature that makes the search easier and useful than before on interest. To develop efficient indexing and search techniques to manage the huge amount of video data, new technologies need to be researched. Using this, people can get the actual idea and the important events as well as scenes could be identified without watching the full original and long videos for several hours. The developed techniques in video summarization can be used in various domains such as surveillance videos, consumer videos, movies, sports, news, etc.

Video summarization is a tool for generating a short summary of a video, as the name implies, can either be a sequence of stationary images called key frames or moving images called video skim. . The summary produced is both of the static or dynamic.Many video mining approaches have been proposed which can be roughly classified into five categories. They are: Video pattern mining, Video clustering and classification, Video association mining, Video content structure mining and Video motion mining [21]. The fact that video data are used in many different areas such as sports, medicine, traffic and education programs, shows how significant it is. The potential applications of video mining include annotation, search, mining of traffic information, event detection / anomaly detection in a surveillance video, pattern or trend analysis and detection. There are four types of videos in our daily life, namely, (a) produced video, (b) raw video, (c) medical video, and (d) broadcast or prerecorded [21]. Video summarization could be achieved based on Semantic Recognition and Emotion recognition. Table 1 briefs the recent research works in video summarization, also the methods and merits and de-merits of the methods have been highlighted

Table 1. Survey of Recent Works on Video Summarization

S.NO	TITLE OF THE PAPER	ALGORITHM /METHODS /MODEL	MERITS	DEMERITS
1.	Predicting Emotions in User-Generated Videos	Comprehensive computational framework	High –level semantic attributes	It is very competitive performance. Audio frame-work are the current framework is limited
2.	Real-Time Summarization of User-Generated Videos Based on Semantic Recognition	A simple uniform segmentation	Quality segment is used it. Both subjective and objective evaluation represented segments	Poor shooting quality segment is used it. A single long shot cannot use for deployed it
3.	Story-Driven Summarization for Egocentric Video	Novel sub shot segmentation, Unsupervised techniques.	It is uniquely for egocentric video. To select a chain of sub-shots are influential to each other	They are central objects looking at a single video
4.	Large-scale video summarization using web-image priors	Multiple summaries obtained through crowd-sourcing	Web –images based prior Unsupervised. The viewpoint cluster are largely coherent	It is randomly selected frame work.
5.	SUPER: Towards Real-time Event Recognition in Internet Videos	Speeded UP Event Recognition(SUPER) framework Quantization methods Efficient classification Extraction	Accuracy Efficiency	kernel fusion is slightly slower

6.	Real-Time Visual Concept Classification	Accelerate concept classification. Bag-of-Words algorithm	Increase of accuracy For Random Forests and K- means. High accuracy	Small increase of total computational time
7.	Audio-Based Semantic Concept Classification for Consumer Video	Single Gaussian modelling. Gaussian mixture modelling. Probabilistic latent semantic analysis	Smaller number of positive training it will lead to greater variability among the different subsets of positive	Varying kinds of sounds within the overall soundtrack duration

From the survey it is inferred that there are certain issues that are to be addressed in the area of video processing, few are listed below:

a. UGV are captured by ordinary consumers with handheld devices like mobile phones into the quality of the video diverse. The amount of the UGV is an extremely large.
b. Duration of the summaries are generated by key frame summarization is controlled in the range of 8 to 16 second.
c. A key-frame summarization techniques which generated only statics video
d. They are effective which the sparse detected based local feature can be more efficiently extracted.
e. They are highly efficient which have 1/8 of the video duration in a summary.
f. The motion-based quality measure is an extremely fast to be extracted in a both feature extraction and classification.
g. Most of the researches focus mainly on emotion based rather than semantic recognition.
h. Few researches address on dynamic video content processing.

3. PROPOSED SYSTEM

This section deals about the proposed framework for video summarization. Figure 2 shows the framework for analyzing the video for summarization. The main objective of the proposed system is to develop an integrated framework for video summarization. This could be achieved by the following steps:

1. Get the input video sequence, this video is fed as input to three major phases that are i). Audio processing ii). Video processing iii). Video summarization processes
2. ***Audio Processing***
   a) Audio is extracted from the input video.
   b) HMM is adopted to convert the audio to text format
   c) From the text, through the CPLNVN summarization technique [16] the most prominent sentences are extracted.
3. ***Video Processing***
   a) Input video is fragmented and stored. The audio associated with the fragment is extracted and converted to text.
   b) Vector table is maintained to store the video fragment and associated audio text.
   c) Apply Semantic similarity measure [18] to identify the duplicate text in the vector table.
   d) Delete the fragments with duplicate audio texts.
   e) RVM and Gabor filter technique are used to detect the facial emotions of the images. Classify the images according to the expression. This is given as an input for the video summarization phase.
4. This phase plays a major role in the process of video summarization.
   a) The most representative sentences form the Audio processing phase is received as an input in this phase.
   b) Secondly, the vector identified from the video processing phase is considered for the generation of video summarization.
   c) Video summary is generated based on the keyword sentences, and frames that match with the vector table.

The following sub sections details about the techniques that are adopted for the generation of video summarization.

3.1 AUDIO PROCESSING

3.1.1 Hidden Markov Model (HMM) [17]

Speaker recognition has major application associated to audio and/or video document processing, like information retrieval. It is well known that audio/video recordings consists of conversion between on or many speakers, like telephone conversations, news relay broadcasting, TV shows, movies, expert meeting, any domain specific videos like lectures, question answering session etc. [17]. Hidden Markov Model (HMM) has been adopted to convert the audio to text process. The converted documents are collected and grouped to form text documents discarding non textual information such as images, tags etc

3.1.2 CPLNVN Summarization Technique [16]

This phase takes the converted text from HMM model for summary creation. The sentence which gives the maximum similarity considering the concepts is taken as representative sentence of the document. Likewise sentences are extracted from each audio text. This technique includes some of the sentence characteristics like Centriod, Length, Position, Noun-Verb and Numerical data for the creation of the multi-document summary. Calculating the sum of all features gives the score for each sentence and then ranks are assigned to the sentences according to their scores. The score is increased by 1, if the sentence consists of numerical data as the news documents with date should be given high importance. The system also takes

care of eliminating redundancy in summary based on concepts. The redundant sentence would be discarded from adding into summary, if the same concept sentence already exists in the summary.

3.1.2.1Semantic Similarity [18]

Semantic-based similarity is the similarity measure used. This similarity measure is a function of the following factors:

- The number of matching concepts, m, in each document (d)
- The total number of the labeled verb-argument structures, v, in each sentence s
- The ct fi of each concept ci in s for each document d where (i = 1, 2, ...,m)
- The c fi of each concept ci in each document d where (i = 1, 2, ...,m)

The semantic-based similarity between two documents d1 and d2 is calculated by**:**

$$\mathrm{sim_s}\left(\mathrm{doc_p}, \mathrm{d_j}\right) = \sum_{\mathrm{i-1}}^{\mathrm{mc}} \mathrm{weight_{i1}} * \mathrm{weight_{i2}} \quad (2)$$

$$\mathrm{weight_i} = \mathrm{cf\,weight_i} + \mathrm{ctf\,weight_i} \quad (3)$$

$$\mathrm{cf\,weight_i} = \mathrm{cf_{ij}}/(\sum_{\mathrm{j=1}}^{\mathrm{cn}} \left(\mathrm{cf_{ij}}\right)^2)^{1/2} \quad (4)$$

$$\mathrm{ctf\,weight_i} = \mathrm{ct\,f_{ij}}/(\sum_{\mathrm{j-1}}^{\mathrm{cn}} \left(\mathrm{ct\,f_{ij}}\right)^2)^{1/2} \quad (5)$$

3.2 VIDEO PROCESSING

In this phase the Input video is fragmented and stored using matlab. The audio associated with the fragment is extracted and converted to text. Vector table is maintained to store the video fragment and associated audio text.

3.2.1Emotion Recognition

This emotion recognition phase is adopted to identify the visual feature of the input video. The input video is fragmented as each scene and shot. The input video is segmented as frames and clustered based on a similarity of frames. This enables faster browsing of large video collections and also more efficient content indexing and access.RVM and Gabor filter technique are used to detect the facial emotions of the images and classified the images according to the expression. This is given as an input for the video summarization phase.

3.2.1.1 Relevance Vector Machine

The Relevance vector machine (RVM) is used for identical functional form to the support vector machine, but provides probabilistic classification. The RVM classifier is used for determining the facial expression for image segmentation. The Emotion recognition will identify for expression of the face and the movement of the body direction. RVM classifier is used to separate images into a frame. The Emotion recognition will be an expression of face and the movement of the body direction.

3.2.1.2GABOR FILTER [12]

The GABOR filter is used for band pass filters which are used in image processing for feature extraction, texture analysis. These functions in two dimensions, it is possible to create filters, which are selective for orientation. The RVM classifiers and GABOR filter are used to detect the facial expression The Segment analysis is used for collecting the entire image segment. This reduces the time, complexity and provides accurate emotion detection. Figure 3 shows the sample image detection using Gabor filter.

Gabor filter generator

Coding Algorithm: Gabor filter generator

Inputs (parameters): Image, Variances along x and y-axis, Frequency of the sinusoidal function and the orientation of Gabor filter.

Outputs: output filters (G) and output filtered image (GaborOut)

```
function [G,GABOUT] = Gaborfilter (I,S,F,W,P);
if is a (I,'double') =1
I=double (I);
end
size = fix (1.5/5); % exp(-1.5^2*pi) <0.1%
k=1;
for x= -size: size
for y= -size: size
G(size+x+1, size+y+1 )=k*exp(-pi*S^2*(x*x+y*y))*. . .
(exp(j*pi*F*(x*cos(W))+P))-
exp(-pi*(F/S)^2+j*P));
end
end
GABOUT=conv2(I, double(G),'same');
```

The technique used for GOBAR filter to separate the video images and audio from the video frames. The RVM classifier is used to detect the facial expression of the video frame. These techniques produce effective and clear expression.

Figure 3: Face Detection Using Gabor Filter

4. EXPERIMENTAL RESULTS

41. Dataset

The dataset used for the experimental setup contains the medical talk show videos that are collected from various websites. It includes the different types of medical related videos about cancer, diabetes, allergy, eye-care, diet and anti-aging TV talk show etc. About 110 videos are taken into account the average duration of those videos can be 5-10 minutes in .avi format. These talk shows are taken into consideration for the following experimental analysis.The videos are converted, pre-processed and summarized for the given video talk shows dataset.

The second dataset taken for the experimental setup is 100 videos from Youtube with duration ranging from 5 to 10 minutes. This video mainly potrays the emotions like Angry, Disgust, Fear , Happy , Neutral , Sad and Surprise of the actors are selected.

4.2 Evaluation Metrics

Accuracy, Coverage and F- Measure are the evaluation metrics considered to evaluate the quality of the summary generated using our proposed algorithm and the existing algorithm.

A. Accuracy [19]

Given a set of data points from a series of measurements, the set can be said to be *precise* if the values are close to the *average value* of the quantity being measured, while the set can be said to be *accurate* if the values are close to the *true value* of the quantity being measured. The two concepts are independent of each other, so a particular set of data can be said to be either accurate, or precise, or both, or neither.

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (6)$$

Where, TP – True Position, TN –True Negative, FP–False Positive, FN – False Negative.

B. F-Measure combines the precision and recall [20]. The precision and Recall is measured using the following parameters.

- ***Correct****-the number of video frames extracted by the system as well as by the human*
- ***Wrong****-the number of video frames extracted by the system but not by the human*
- ***Missed****-the number of video framesextracted by the human but not by the system*

Precision and Recall is computed as:

$$\text{Precision}=\frac{\text{Correct}}{\text{Correct+Wrong}} \quad (7)$$

$$\text{Recall}=\frac{\text{Correct}}{\text{Correct+Missed}} \quad (8)$$

F-Measure is computed by using the formula given below

$$\text{F-Measure}=\frac{2x\ \text{Precision x Recall}}{\text{Precision +Recall}} \quad (9)$$

C. Coverage [1]: the summary contains sufficient information to understand the full story with little content redundancy.

4.3 Methods used for analyzing the performance

1. **Uniform Sampling :** : Uniform sampling is most frequently used technique to extract key frames from the summary without considering any semantic relevance between the frames. Usually every ith frame is selected as samples for the video summary[24].
2. **Video based summary [5]:** the frame which contains important and significant information are seleted to form the summary considering only video images.
3. **Audio based summary** : Extracted the audio alone from the video the most significant sentences are extracted along with the video to make summary.
4. **Human generated summary**: Summary generated by the human judges considering video and audio representatives;
5. **Proposed video summarization** : Hybrid technique which generates video summary based on facial emotion recognition and audio based on semantic analysis.

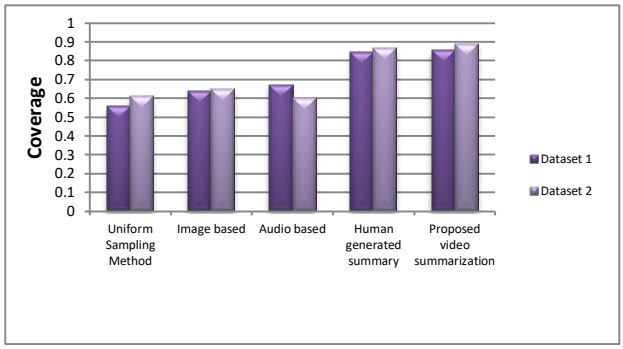

Figure 4: Video Summarization Comparison using Coverage

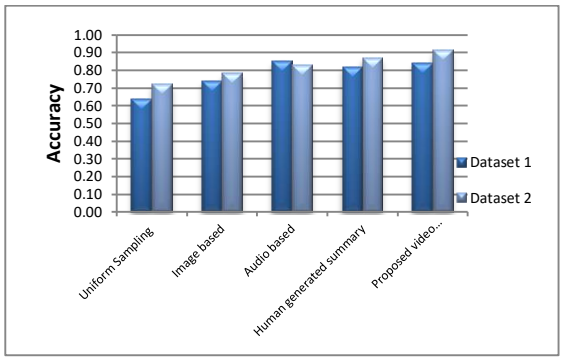

Figure 5: Video Summarization Comparison using Accuracy

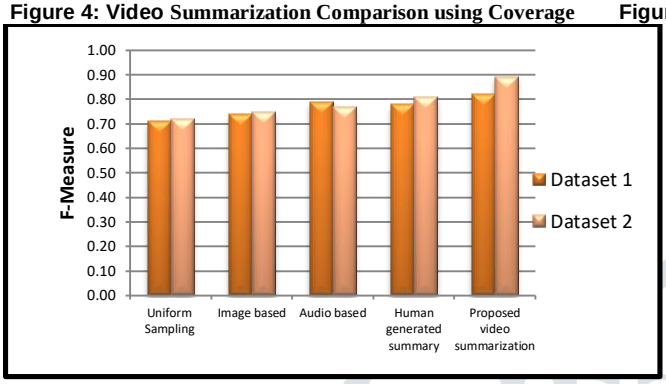

Figure 5: Video Summarization Comparison using F-Measure

4.4 Result Analysis

The first data set input video is categorized according to the content like diabetes, cancer etc and is presented to 10 human judges (each were given 11 videos) and requested them to generate summary based on the importance of video and audio, most representative clippings. These clippings are extracted manually through MATLAB and summary is generated. For evaluating the quality of the proposed system, the summary generated by the human judges, uniform sampling method, audio based summary, video based summary is evaluated by the experts for Coverage, F-Measure and Accuracy. Fig 4,5 & 6 depicts the performance of the existing and proposed algorithm for the above metrics. Our proposed method has shown significant improvement compared to uniform sampling, audio and video based methods and the quality and accuracy is on par with the human generated summary. The length of the video generated by human judges is promintely high compared to our proposed system. Since uniform sampling method doesn't address the emotion and semantic features our proposed summary gives better results equivalent to human generated video summary. Considering audio semantic and video semantics alone also does not yield better accuracy and coverage compared to the proposed algorithm. Time taken to generate the summary based on our proposed algorithm is comparely 30% lesser than the human generated summary. Compared to dataset 1 our algorithm outperforms dataset 2 because emotion based video clippings contribute more in the generation of better summary compared to videos of normal conversation. We could infer that the videos which has more conversation may adopt video summarization based on audio text content and the proposed algorithm gives better results for video clippings with more emotion.

6. CONCLUSION

Video Summarization which generates a short video summary chunck based on a Long Input Video. This application was mainly used for efficient search result browsing. Video summaries are especially useful on mobile platforms, because users can preview the critical contents before downloading the entire video, so that the Internet bandwidth may be conserved. One of the most important goals of video summarization is to produce a video clip as short as possible, while preserving the most representative content in the original video. Proposed Audio semantic and video emotion video summarization used for integrating multiple clues. High-level semantics and emotions are recognized first and simple scoring functions are proposed to select both semantically and emotional representative segments to form a video summary. The RVM Classifier has a Gaborfilterwhich was used to identify the Facial expressions from the video summary.

Acknowledgements

My sincere gratitude to University Grant Commission (UGC),. New Delhi for funding the grant under the scheme of Minor Research Project.

REFERENCES

1. Fast Summarization of User-Generated Videos Using Semantic, Emotional and Quality Clues, BaohanXu, Xi Wang, Yu-Gang Jiang, IEEE mutlimedia journals,2015 .
2. Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence, 2014.*
3. X. Wang, Y.-G. Jiang, Z. Chai, Z. Gu, X. Du, and D. Wang. Realtime summarization of user-generated videos based on semantic recognition. In *Proceedings of ACM Multimedia, 2014.*
4. Z. Lu and K. Grauman. Story-driven summarization for egocentricvideo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013*
5. A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image

priors. In *Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition, 2013*

6. Y.-G. Jiang. SUPER: towards real-time event recognition in internet videos. In Proceedings of the ACM International Conference on Multimedia Retrieval, 2012.
7. Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user generated videos. In Proceedings of the AAAI Conference on Artificial Intelligence, 2014.
8. Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In Proceedings of the ACM International Conference on Multimedia Retrieval, 2011
9. A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
10. K. Lee and D. P. Ellis. Audio-based semantic concept classification for consumer video. IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1406–1416, 2010
11. Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
12. https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
13. Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In Proceedings of the European Conference on Computer Vision, 2008.
14. P. Over, A. F. Smeaton, and P. Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. In Proceedings of NIST TRECVID Workshop, 2007.
15. Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. IEEE Transactions on Circuits and Systems for Video Technology, 15(1):52–64, 2005.
16. Jayabharathy. J, Kanmani. S and Sivaranjani. N, "Correlation Based Multi-Document Summarization for Scientific Articles and News Group", in the proceedings of ACM -International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1093-1099, August 3-5, 2012
17. Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille ; Gerald Friedland ; OriolVinyals. 2012. Speaker diarization: A review of recent research, *IEEE Transactions on Audio, Speech, and Language Processing* ,Vol 20 , No 2 , 356 – 370.
18. Shaban, K, "A Semantic Approach for Document Clustering", Journal of Software, Academy Publisher, Vol. 4, No. 5, pp. 391-404, 2009.
19. https://en.wikipedia.org/wiki/Accuracy_and_precision as on June 2017.
20. Steinbach, M., Karypis, G. and Kumar, V. 2000. A Comparison of Document Clustering Techniques,*In the Proceedings of Workshop on Text Mining, 6th ACM SIGKDD International Conference on Data Mining (KDD'00)*, pp.109–110.
21. V. Vijayakumar · R. Nedunchezhian, "A study on video data mining", International Journal of Multimed Information Retrival, Springer, Vol 1, pp-153-172,2012.
22. Sandra E. F. de Avila , Antonio da Luz Jr.†§, and Arnaldo de A. Araújo, "VSUMM: A Simple and Efficient Approach for Automatic Video Summarization "
23. Jharna Majumdar , Spoorthy.B, "Comparisons of Video Summarization Methods", IOSR Journal of Computer Engineering (IOSR-JCE) Vol 16, Issue 5,2014.