ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 2, Issue 4, April 2014**

# Enhanced Generic Video Summarization using Large Scale Categorization

Priyamvada R Sachan[1], Keshaveni[2]

[1] Assistant Professor, Dept. of ECE, East point C.O.E & Technology, Bangalore, India

[2] Professor, Department of ECE, K V G college of Engineering, Sullia, India

**ABSTRACT**: Over the past few years, there has been a massive increase in amount of video content created. Massive growth in video content poses problem of information overload and management of content. In order to manage the growing videos on the web and also to extract efficient and valid information from the videos, more attention has to be paid towards video and image processing technologies. Video summaries provide condensed and succinct representations of the content of a video stream through a combination of still images, video segments, graphical representations and textual descriptors. Existing video summarization techniques have attempted to solve the problem of condensing the content of a video in domain specific manner. However, such domain specific summarization mechanisms do not generalize well over different genres of videos. To make video summarization scalable enough to cater to the needs of growing massive online video content, it needs to be generic and adaptable for its applicability on any category of video. This work presents a general approach towards video summarization process and proposes a two-step approach towards domain independent generic video summarization incorporating video categorization for enhanced keyframe extraction. This paper also talks about building effective mechanisms for large scale categorization over big hierarchical category tree of videos.

**KEYWORDS** - video processing, video summarization, video categorization, keyframe extraction.

## I. INTRODUCTION

Massively increasing availability of video content (boosted by user generated videos) is already creating information flood for users. For getting the best value of created video content and making it reach to the targeted audience in the most succinct way, there is a clear need for an enhanced video processing system, which can enable users to consume video content of their choice in the most effective and personalized manner. The very need for any kind of summarization arises when there is excess content and limited time/desire to consume it. Existing techniques auto summarize one large video in small one, but have emphasized less on the problem of auto summarizing content of millions of videos appearing daily for a user. In consumer driven market, video summarization to deliver its best and be able to solve the real needs of a user has to be personalized, as the very definition of summary of content is consumer driven. Hence, there is a clear need to make the video summarization process more generic and domain independent targeted toward generic user. This imposes the need of enhancement over the existing techniques. In next few sections, we present related work and keyframe based video summarization, followed by our proposal to enhance it by use of categorization on videos.

## II. RELATED WORK

The major task in video summarization is video segmentation, which can be achieved either by keyframe extraction [1] or shot boundary detection [2]. Shot-change detection is the process of identifying changes in the scene content of a video sequence so that alternate representations may be derived for the purposes of browsing and retrieval, e.g., keyframes may be extracted from a distinct to represent the summary of a original video statically or dynamically [3]. Lot of work has been carried out in the area of video summarization. Previous research has focused on content based

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 2, Issue 4, April 2014**

systems that shows the benefit of analyzing a video without user interactions, but they are monolithic, because the resulting key-frames are the same regardless of the user preferences. There are several research works on content-based keyframe extraction from videos, because a collection of still images is easier to deliver and comprehend when compared to a long video stream. Girgensohn et al. [4] found that clustering of similar colors between video scenes is an effective way to filter through a large number of key-frames. SmartSkip [5] is an interface that generates keyframes by analyzing the histogram of images every 11 seconds of the video and looking at rapid overall changes in the color and brightness. Fischlar [6] is a web-based system for capturing, storing, indexing and browsing broadcast TV material, but it only features content-based techniques. In the survey paper [7][8] a general methodology for video summarization is presented. It gives the clear picture of different levels involved in the video summarization. Again the extraction process can be based on visual or aural feature as cited in [9]. Analysis of different papers shows that shot-change detection, process of identifying changes in the scene content of a video sequence so that alternate representations may be derived for the purposes of browsing and retrieval forms the major step in summarization process. Shot detection is followed by keyframe extraction. Luthra et al. [10] proposed an unsupervised learning approach to find the frame of a video having high goodness measure value to generate the video summary. Developed system is tested over two different classes of videos viz. home-shot party and Soccer videos. The algorithm is tested using only visual features and then tested by using aural features along with visual features. The result shows that there was an improvement in the summarization techniques when aural features were embedded with the visual features.

Review of the previous work in this domain shows that very less heed is paid towards user's interest for personalized summarization techniques. Slowly with increasing data on the web and limited time to view it, people started working towards personalization techniques in a domain specific manner mainly highlighting news and sports videos. In an user attention model [11] an automatic video summarization process considering the attention of viewer's is proposed. This framework takes an advantage of computational attention models and eliminates the needs of complex heuristic rules in video summarization. Taniguchi *et al.* [12] have summarized video using a 4-D packing of "panoramas" which are large images formed by compositing video pans. A "panorama" enables a single keyframe to represent all images included in a shot with camera motion. In this work, keyframes are extracted from every shot and used for a 4-D representation of the video content. Because frame sizes were not adjusted for better packing, much white space can be seen in the summary results. Although lot of work is carried out in the domain of personalized and automatic video summarization systems, domain independent summarization is still a major challenge.

Users unintentionally embed their understanding of the video content in their interaction with computers. This valuable knowledge, which is difficult for computers to learn autonomously, can be utilized for video summarization process. Yu, Bin, et al. presents an intelligent video browsing and summarization system that utilizes previous viewers' browsing log to facilitate future viewers [13]. Specifically, a novel ShotRank notion is proposed as a measure of the subjective interestingness and importance of each video shot. A ShotRank computation framework is constructed to seamlessly unify low-level video analysis and user browsing log mining. The resulting ShotRank is used to organize the presentation of video shots and generate video skims. Experimental results from user studies have strongly confirmed that ShotRank indeed represents the subjective notion of interestingness and importance of each video shot, and it significantly improves future viewers' browsing experience.

Further enhancement in the existing techniques is reflected by incorporating different machine learning (ML) algorithms for categorizing the large data on web in view of condensed and succinct representations of the data [14][15][16].

## III. General Methodology For Keyframe Extraction

Keyframes are the most informative frames of a video. Keyframe forms the foremost step of any video summarization process. As depicted in fig.1, the overall process of keyframe extraction is as follows:

### *A. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

*B. 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.

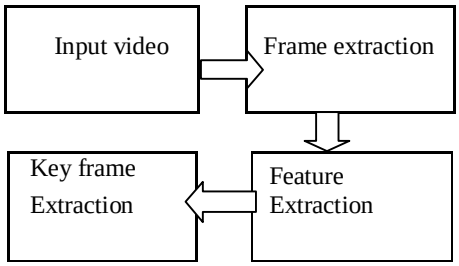


Fig. 1. Keyframe Extraction

*C. Feature Extraction*

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

1) *Visual Features*: The visual features of the extracted 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.

*a) Color histogram:* The color histograms have been commonly used for key frame extraction in frame difference based techniques. This is because the color is one of the most important visual features to describe an image. Color histograms are easy to compute and are robust in case of small camera motions. The idea behind histogram based approaches is that two frames with unchanging background and unchanging (although moving) objects will have little difference in their histograms. The color histogram difference d(Ii, Ij) between two consecutive frames Ii and Ij can be calculated as in Eq. 1

$$d(I_i, I_{i+1}) = \sum_{j=1}^{n} \frac{|H_i(j) - H_{i+1}(j)|^2}{H_{i+1}(j)} \qquad (1)$$

Where '$H_i$' and '$H_{i+1}$' stands for the Histogram of Ii and Ij frames respectively and 'n' represents the total number of frames. Another approach can be similarity measures as Eq. 2. between the frames or set of frames to categorise them into different shots and then extract the key frames [17].

$$\sum_{h=1}^{16}\sum_{s=1}^{8} \min(Hi(h,s), Hj(h,s)) \qquad (2)$$

Equation 2 can be normalized in order to limit the similarity between two frames i and j within [0, 1]. Eq. 3. is the normalized equation of Eq. 2.

$$S_{i,j} = \frac{\sum_{h=1}^{16}\sum_{s=1}^{8} \min(Hi(h,s), Hj(h,s))}{w \times h} \qquad (3)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Where "w" is the width of the frame (image), and "h" is the height of the frame. S i,j denotes the similarity between two images and $0 \leq Si,j \leq 1$.

*b) Edge detection*: Edge detection is one of the commonly used operations in image analysis. An edge is the boundary between an object and the background, and indicates the boundary between overlapping objects. This means that if the edges in an image can be identified accurately, all of the objects can be located and basic properties such as area, perimeter, and shape can be measured. Edges define the boundaries between regions in an image, which helps with segmentation and object recognition. The edge matching rate is used to match the edges of adjacent frames to eliminate redundant frames.

*c) Block correlation:* In the block motion compensation (BMC), the frames are partitioned in blocks of pixels (e.g. macro blocks of 16×16 pixels in MPEG). Each block is predicted from a block of equal size in the reference frame. The blocks are not transformed in any way apart from being shifted to the position of the predicted block. This shift is represented by a motion vector. To exploit the redundancy between neighbouring block vectors [4] (e.g. for a single moving object covered by multiple blocks) it is common to encode only the difference between the current and previous motion vector in the bit-stream. The result of this differencing process is mathematically equivalent to global motion compensation capable of panning. It is possible to shift a block by a non-integer number of pixels, which is called sub-pixel precision. The in between pixels are generated by interpolating neighbouring pixels. Commonly, half-pixel or quarter pixel precision (used by H.264 and MPEG- 4/ASP) is used. The computational expense of sub-pixel precision is much higher due to the extra processing required for interpolation and on the encoder side, a much greater number of potential source blocks to be evaluated. Block motion compensation divides up the current frame into non-overlapping blocks, and the motion compensation vector tells where those blocks come from (a common misconception is that the previous frame is divided up into non-overlapping blocks, and the motion compensation vectors tell where those blocks move to). The source blocks typically overlap in the source frame. Some video compression algorithms assemble the current frame out of pieces of several different previously-transmitted frames. Frames can also be predicted from future frames. The future frames then need to be encoded before the predicted frames and thus, the encoding order does not necessarily match the real frame order. Such frames are usually predicted from two directions, i.e. from the I- or P-frames that immediately precede or follow the predicted frame [5]. These bidirectional predicted frames are called B-frames. A coding scheme could, for instance, be IBBPBBPBBPBB.

2) *Audio Features*: The study 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]. These classes can be well distinguished by using Short Time Energy (STE), Zero Crossing Rate (ZCR) and Fundamental Frequency functions. The Short Time Energy function (STE) basically distinguishes speech and music and Short Time Zero Crossing Rate (STZCR) is used to separate voiced speech from unvoiced one. Whereas, Short Time Fundamental Frequency (STFF) separates audio into harmonic and non-harmonic classes. This way all the speech components are well distinguished using these three features. Both STE and STZCR are calculated for every overlapping window of 511 samples of the audio signal with an overlap of 35 samples at either end of the window at a sampling rate of 44100 samples/s. The STFF of the audio segment is estimated over an overlapping window of 2048 samples with an overlap of 284 samples. When no fundamental frequency is estimated, the STFF is set to zero. Once these features have been extracted, different audio classes are characterized using statistical property of variance over overlapping windows of 140 feature samples with an overlap of 40 samples at either end of window. Thus we obtain a feature with a sample for every second of the audio.

*D. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

## IV. Problem Statement

Existing video summarization techniques attempt to do the summarization in a very domain specific manner, which makes them ineffective beyond the scope of domain. With the massive outburst of online video content, there is need to develop automatic video summarization techniques, which are end user oriented and can scale across various domains in a generic way.

## V. Our Proposal

The summary of a video is often obtained by identifying keyframes in the video. This identification can be done by the analysis of low-level features extracted from an individual frame or temporally adjacent frames. However, the visual saliency in the temporal domain is not directly related to content-wise importance in the video sequence, which makes it extremely difficult to perform the task fully automatically. Although it is not straightforward for computer to find the semantically useful information from a different category of video, it can capture visually similar frames and cluster them together based on low-level feature analysis. This makes the task domain specific.

Machine learning models (SVM, RFs, etc.) have been found to automatically capture semantic model from relationships between low level features. Also, very fine-grained video categories can very well define the domain of the videos and are also being used by users for browsing videos of their choice. This work get motivated from the above 2 facts and proposes use of category as a highly weighted feature along with other video features to build automatic models for keyframes extraction using machine learning. Use of automatic machine learning based models also enhances its applicability on video processing on large-scale.

The idea here is to build a ML based model for the problem of **keyframes identification** by relying on an enhanced feature set that includes '*frame-position*', '*subtitles'* and '*fine-grained video category*' along with existing audio/visual features. Also, for getting the fine-grained video category, the work proposes to pre-process the video through a large-scale hierarchical categorizer. The proposed features are as explained below:

- *Category*: This represents the fine-grained category of the video in a very large category tree(taxonomy) for the videos. Category of a video is supposed to encode domain specific information for the content and is helpful in inducing any domain specific processing in a generic way.
- *Frame-Position*: This basically represents the sequential order of a frame in the video and is expressed as the time at which it occurs during the run of video.
- *Subtitles*: These are text associated with each frame and often contain important semantic information. They help in leveraging from advanced text summarization techniques for associated video summarization.

## VI. Top Down Approach For Large Scale Hierarchical Video categorization

Since, video contents are generated for almost every domain, the category tree (taxonomy) associated with videos is very large (> 10000 nodes) and ever increasing. Since, a video could belong to multiple categories due to behavioural overlap, it makes the problem a multi-class and multi-label classification. Modeling this classification task as classical l–vs-all or 1-vs-1 multi-class classification problems is not feasible as the number of classes (nodes of taxonomy) is quite high leading to increase in number of inconsistent decisions inherent with these techniques [18][19]. Also, these classical techniques pose huge performance challenge while scaling as they require classification of each record through k or k(k-1) binary classification models, where k being the size of category tree. Here, we present a hierarchical approach for classification that divides the problem of big classification into smaller classification steps utilizing the hierarchical nature of the tree.

The overall classifier is trained over the baseline category-tree to generate a SVM [20] model per node of the category tree. Each model does binary classification only to tell whether a record belongs to the category node or not along with the confidence behind its decision. For each category node, training set is gathered from all the records under the subtree rooted at that node. The classification of a new incoming record happens in a hierarchical fashion starting from the root of the tree and descending down till leaf nodes as in Fig. 5. At each step, classification confidence of parent is combined with current confidence and compared with a threshold to determine the final classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

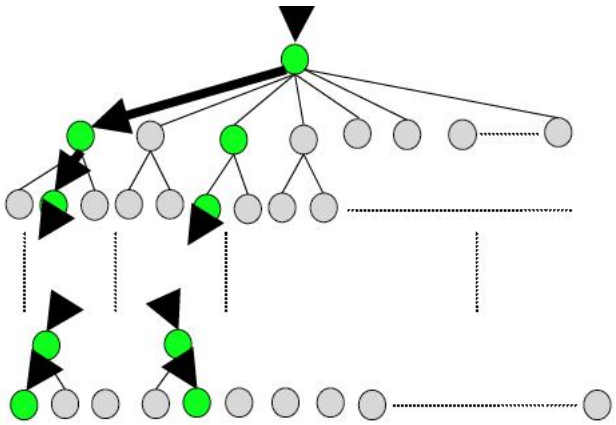

Fig. 5. Top Down Hierarchical Classification

output. At any step, if none of child models of a parent are able to confidently classify a record positively, it is assigned to a virtual 'Miscellaneous' child category of the parent. This mechanism offers the possibility to categorize a record into multiple categories at the same time avoiding evaluations over subcategories that don't match. On an average number of attempted classifications per record will be O (category-tree depth).

## VII. EXPERIMENTAL RESULTS

The performance of the proposed technique is evaluated in the form of Recall and Precision. Experiments for key frame extraction are conducted as follows: using (i) existing features (Base) and (ii) enhanced features (Proposed features embedded with Base feature). Color histogram is used as a Base feature for this experiment. See table 1 for result. Also the test is conducted on a dataset of 200 videos extracted from YouTube and other sources to evaluate the performance of Top Down approach for hierarchical video categorization. Features used for top down approach are : "bag-of-words over video title, subtitles and video metadata", "video-length". Refer table 2 for result.

**For Keyframe Extraction:**
*Recall = (no. of correctly detected keyframes)/ (no. of true keyframes*)

*Precision= (no. of correctly detected keyframes)/ ( no. of totally detected keyframes)*

**For Hierarchical Categorization:**
*Recall = (no. of desired categories present in categorizer's output category set)/ (total no. of desired categories*)

*Precision= (no. of desired categories present in categorizer's output category set)/ ( total no. of categorizer's output categories)*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

TABLE 1: Performance of Keyframe Extraction Technique

| Videos | Base Features | | Enhanced Features | |
|---|---|---|---|---|
| | Recall | Precision | Recall | Precision |
| Movie | 0.73 | 0.61 | 0.82 | 0.72 |
| soccer | 0.79 | 0.65 | 0.86 | 0.77 |

TABLE 2: Performance of Top Down Approach for Hierarchical Video Categorization

| Training Dataset (no .of videos) | Category Tree | | | Training Set break-up | | | Classification algorithm | Results | |
|---|---|---|---|---|---|---|---|---|---|
| 200 | Avg. depth | Width | Leaf nodes | Training | Testing | Tuning | SVM | Recall | Precision |
| | 5 | 12 | 60 | 60% | 30% | 10% | | 85.12% | 75.87% |

## VIII. Future Directions

The current work can be extended by experimenting on various audio/video features and different ML algorithms (LR, RFs) for categorization accuracy improvements. We would also like to explore creation of user specific models for both categorization & keyframes identification shall lead to ultimate goal of personalized video summaries. Additionally, inclusion of more video low level features and meta features for overall video is expected to boost semantic modeling of video content. Finally, we would like to explore Hadoop for implementation of above mentioned techniques on large scale.

## IX. Conclusion

It has been realized that the existing techniques for video summarization are effective in domain specific manner. In order to cater the needs of users over massively growing online video content in view of limited time and desire, these techniques need to be generic and domain independent. Experimental result shows, incorporating the proposed techniques with the existing keyframe extraction techniques for video summarization works effectively in cross domain. Also large scale hierarchical categorization helps in identifying the category of video, which is the foremost step in our proposal for keyframe extraction.

## References

[1] A.V.Kumthekar, Prof. J.K. Patil “Key frame extraction using color histogram method”, International Journal of Scientific Research Engineering & Technology (IJSRET) Volume 2 Issue 4 pp 207-214 , ISSN 2278 – 0882, july-2013

[2] Baber, Junaid, et al. "Shot boundary detection from videos using entropy and local descriptor." Digital Signal Processing (DSP), 2011 17th International Conference on pp 1-6, IEEE, 2011.

[3] Taskiran,C. and E. Delp. "Video summarization" Digital Image Sequence Processing, Compression, and Analysis : 215-231, 2005

[4] Girgensohn A, Boreczky J, Wilcox L, “ Keyframe based user interfaces for digital video”. Computer 34(9):61-67, 2001

[5] Drucker SM, Glatzer A, De Mar S, Wong C (2002) SmartSkip: consumer level browsing and skipping of digital video content. In: Proceedings of the SIGCHI conference on human factors in computing systems: changing Our world, changing ourselves (Minneapolis, Minnesota, USA). CHI ’02. New York, NY: ACM. pp 219-226, , April 20–25, 2002

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 2, Issue 4, April 2014**

[6] Li FC, Gupta A, Sanocki E, He L, Rui Y, Proceedings of the SIGCHI conference on human factors in computing systems (the Hague, the Netherlands, April 01–06, 2000). CHI '00, New York, NY: ACM. pp 169-176, April 01-06, 2000.

[7] Truong, B. T. & Venkatesh, S. "Video Abstraction: A Systematic Review and Classification". ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 3, No. 1(3), 2007

[8] Sachan Priyamvada R, Keshaveni, "A survey on automatic video summarization techniques", International Journal of Electronics, Electrical and Computational System (IJEECS) ISSN 2348-117X Volume 3 Issue 1,pp 1-5, April 2014.

[9] Liu, Zhu, et al. "Audio feature extraction and analysis for scene classification." Multimedia Signal Processing, First Workshop on IEEE, 1997.

[10] Luthra, Varun, Jayanta Basak, Santanu Chaudhury, and K. A. N. Jyothi. "A Machine Learning based Approach to Video Summarization", 2008

[11] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li, "A user attention model for video summarization", In Proceedings of ACM Multimedia, pp. 533-542, 2002.

[12] Y. Taniguchi, A. Akutsu, Y. Tonomura. "Panorama Excerpts:" Extracting and Packing Panoramas for Video Browsing, Proc ACM Multimedia 97. pp. 427-436, 1997.

[13] Yu, Bin, et al. "Video summarization based on user log enhanced link analysis. " Proceedings of the eleventh ACM international conference on Multimedia. ACM, 2003

[14] Lavesson, Niklas, and Paul Davidsson. "Quantifying the impact of learning algorithm parameter tuning." AAAI. Vol.6. 2006.

[15] Chapelle, Olivier, et al. "Choosing multiple parameters for support vector machines." Machine learning 46.1-3, 131-159. 2002.

[16] Koster, Cornelis HA, and Jean G. Beney. "On the importance of parameter tuning in text categorization." Perspectives of Systems Informatics. Springer Berlin Heidelberg, 270-283,2007.

[17] http://www-scf.usc.edu/~taibaixu/document

[18] Wang, Hua, Chris HQ Ding, and Heng Huang. "Multi-Label Classification: Inconsistency and Class Balanced K Nearest Neighbor." AAAI. 2010.

[19] Tax, David MJ, and Robert PW Duin. "Using two-class classifiers for multiclass classification." Pattern Recognition, Proceedings, 16th International Conference on. Vol. 2. IEEE, 2002.

[20] http://en.wikipedia.org/wiki/Support_vector_machine

## BIOGRAPHY

Priyamvada R Sachan (Gold Medalist in M.Tech ) is pursuing Ph.D in the ECE Department, Visvesvaraya Technological University, East point C.O.E & Technology, India. She received Master of VLSI Design & Embedded systems degree in 2011, VTU, Karnataka, India. B.E (Instrumentation), 2004, Mumbai University, Mumbai, India. Her research interests are Video processing, Image processing & Machine learning.