# Feature based Information Extraction for Generic Video Summarization

Satyabrata Maity
A.K. Choudhury School of Information Technology, University of Calcutta, Kolkata-700009, India

Amlan Chakrabarti
A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata-700009, India

Debotosh Bhattacharjee
Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India

## ABSTRACT
Video summarization plays a very significant role in navigating a video, to understand its information or to search the required event information. Our proposed research work minimizes the time required for processing each of the video frames firstly, by reducing their effective size, and then it is followed by an efficient technique for generating the summarized video. The information contained in a frame is extracted using object and motion based features where the object based feature helps to evaluate the importance of the given frame compared to its neighboring frames and the motion based feature helps to estimate the dynamism of the frame. Disturbance Ratio [DR] based measurement is used in the next step to select the shot boundary, key frame and summary generation. The results of the proposed summarization methodology show the efficiency of our algorithm, which is further supported by a comparative study of the related research works.

## General Terms
Video summarization, Shot boundary detection, Feature Extraction, Key frame detection

## Keyword
Video summarization, key frame, information extraction, frame size reduction, Disturbance Ratio

## 1. INTRODUCTION
Visual medium is the easiest way to understand any information in a most significant way. In most of the cases a raw source video generally contains a huge amount of redundant information, which cannot be perceived in a shorter amount of time for decision making purpose or for generating the semantic information of the same. Extraction of effective information from video data is an important activity for various applications like surveillance, sports, news, entertainment industry and many more. Video summarization is a process to reduce the enormous size of the raw video by removing redundant and the unnecessary information of the video while retaining the important key information, which results to a better understanding of the same within a shorter amount of time.

Video Summarization is divided into two broad categories (1) Domain Specific [2, 11, 16, 17] and (2) Non Domain Specific [1, 2, 3, 5, 4, 12, 13, 14, 15]. Domain specific summarization refers to techniques, which specifically cater to a particular domain like sports, music, news, home video etc. Focusing on a particular domain helps to reduce levels of ambiguity when analyzing the content of a video stream by applying prior knowledge of the domain during the analysis process. On the other hand in Non-Domain specific techniques solutions are presented for summarizing video content irrespective of the knowledge of domain.

The major aspect of a summarization algorithm is to follow some strategy to selectively extract the needed information frames from the given video. According to the similarity of information between the successive frames, a video can be divided into several groups called shots and the frames belonging to the same shot have a higher similarity between themselves. For the purpose of summarization the representation frame(s) for a given shot are selected by considering the information content of the frames in that shot. According to the priority of the group, numbers frames are selected to make the summary.

A survey of related research works in video summarization gives us a good insight to the present research challenges. In [1] a conceptual framework for video summarization is discussed. The authors considered the video summarization technique in three broad categories viz. internal which analyses internal information from the video stream, external which analyses external information of the video stream and hybrid which analyses both internal and external information of the video stream. The authors in [3] proposed an approach for the selection of representative (key) frames of a video sequence for video summarization by analyzing the differences between two consecutive frames of a video sequence. In [4] the authors presented a new approach for key frame extraction based on the image entropy and edge-matching rate. In [2] the author proposed an approach for video summarization that works in the compressed domain and is based on exploiting visual features extracted from the video stream. The technique extracts the DCT part of the MPEG video frame without decompressing that, which makes the algorithm faster. In our previous work [5] we have extracted the entropy and color based global information of the frames based on which we had evaluated the disturbance ratio [DR] measure of the individual frames. The DR measure was then utilized to detect the shot boundaries and selection of the key frames for generating the summarized video.

In this work we have extended our previous work [5] in terms of reduction in the processing time for generating the summarized video. We have adopted a method to reduce the size of the image frames in the video and subsequent intelligent processing of those. This resulted to a lesser time for information extraction without affecting the actual information content of the frames in a great way. In this work we have also considered the object and motion based features, which are taken to evaluate a frame for considering the dynamism of the video. This is because the object based features determines the number of component present in the frame and the motion based object determines the number of dynamic component present in the frames. Motion based and objects based feature are compared, and that gives the amount of change (single value) in the consecutive frames. The amount of change is the information which represents the response value of the frames that makes the group detection, key-frame extraction and selection of the frames for summarization easier and less time consuming.

The organization of the paper is as follows. Section (2) gives an insight to our proposed technique of feature based information

extraction and video summarization. A detailed analysis of the results is presented in Section (3) and the concluding remarks in Section (4).

# 2. PROPOSED METHOD

A video must be divided into some groups, where the frames in the groups are of similar in nature. The frame(s) which can represent the message of the group is called key frame(s). It is very difficult to understand the total information from the extracted key frames only. Thus the proposed method incorporates some supporting frames (from the group) with each of the key frames to maintain continuity. Supporting frames are selected according to the priority given by the information index profile. The size of the frame is reduced to save the processing time to extract amount of information of each frame. In our proposed application we have focused on the global information and the amount of information is defined as the weight of the corresponding frame. Our work is divided into four major steps as shown in Figure 1 below:

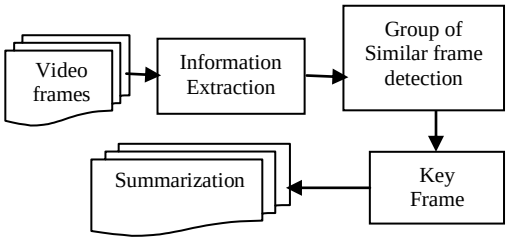


**Fig 1: Steps of video summarization**

- Information Extraction: The amount of information that is desirable for the application is extracted in this step. This is the most basic but important part of video summarization because the effectiveness of this step can control the efficiency of the successive steps.
- Group of similar frames detection: Frames which are similar to a certain extent are the frames of the same group i.e. shot boundary detection.
- Representative frame selection: Key frame or the representative frame of a certain group is the frame that contains the most of the information of the group. Sometimes there may be more than one key frames of a group.
- Summary generation: Summary is not only the collection of all the representative frames but it also hold three basic properties as depicted in [2] i.e. continuation to understand the message carried out by video, priority to include the main concern, and no-repetition to reduce the redundant areas of the original video.

Proposed approach is a generic video summarization on feature-based model that utilizes the similarity measurement among consecutive frames for determining the suitable frames for the summarized video. In this case we considered the object and motion based features, which are taken to evaluate a frame for considering the dynamism of the video as the object based features determines number of component present in the frame and the motion based object determines the number of dynamic component present in the frames.

## 2.1. Information extraction

In the proposed approach information is extracted in the following steps.

*2.1.1 Reduction of the frame size*

The reduction in the image frame leads to an overall reduction of the computation time as it helps to minimize time to analyze the given frame. The proposed technique divides each frame into grids of cell size *kXk* pixels. The reduced frame contains one value (the mean of the cell) from each of the cell. The resulting frame is reduced by *kXk* times of the original size. The effective size is reduced following the algorithm shown in Figure 2 and Table-1 shows that the Edge Pixel Ratio (EPR) and Entropy (Figure 4 shows the Entropy profile of the original and reduced frames of the video of [8]) between original and the reduced image are very similar which assures that the information content of the reduced image is not much changed compared to the original ones.

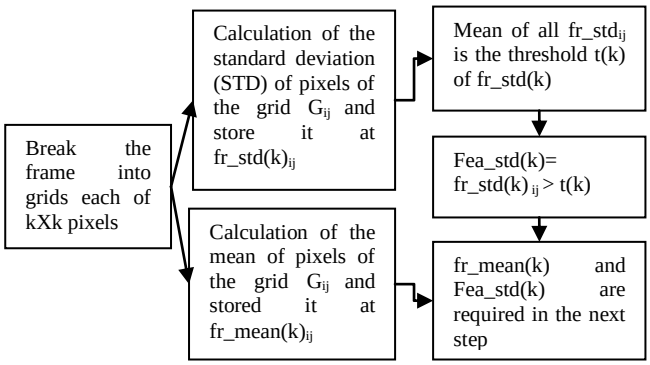


**Fig 2: Frame size reduction**

In Figure 2 first the frame is divided into grids of *k*k* pixels, then the standard deviation and mean of each cell is calculated and stored into *fr_std* and *fr_mean* respectively. The threshold *t(k)* is the mean of all values stored in *fr_std* and *Fea_std* takes the values in the corresponding position greater than *t(k)*. *Fea_std* and *fr_mean* are two feature matrix calculated in the step.

*2.1.2 Extraction of the differential information between any two consecutive frames*

The difference in *fr_mean(i+1)* and *fr_mean(i)* of the $i^{th}$ and $(i+1)^{th}$ frame has been calculated in this phase which helps to detect the dynamic information that exists between the consecutive frames as the mean of the cell, to reduce the noise and the effective size of the dynamic object. The difference is taken as 1 when it is greater than the threshold, otherwise 0 is taken and is stored in $diff(i)_{jk}$ where *i* is the corresponding frame number and *(j,k)* represents the corresponding cell number. The threshold is calculated for each frame by considering the change in the difference values between the frames which makes the thresholding technique adaptive.

*2.1.3 Merging of two features for getting response value*

From the above two feature matrix *fr_std(i)* which gives the number of component present in the frame and *diff(i)* which gives the dynamic information of the corresponding frames are merged for calculating the amount of change (response value) in the corresponding frames and are stored in the weight vector *W(i)* which contains the response value of each frame.

## 2.2 Group of Similar frame Detection

Similarity group has been detected with respect to the changing of weight as shown in Figure 3 (amount of information assessed in previous step) between the frames. If the amount of change between $i^{th}$ and $j^{th}$ frame is less than the calculated threshold ε and all the frames between $F_i$ and $F_{j-1}$ is in the same group $g_k$, then $F_j$ must be in $g_k$. The threshold is calculated by

taking the standard deviation of the weight vector (containing the weight of the corresponding frames).

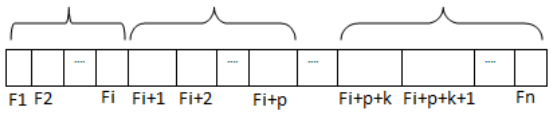


**Fig 3: Similarity groups feet**

All the frames between $F_i, F_{j-1} \in g_k$ (j>i) and If $|F_i - F_j| \leq \varepsilon$; then $F_j \in g_k$, where (i,j)=1 to n numbers of frames and ε is the threshold and k=1 to m number of groups

## 2.3 Representative frame selection

Disturbance Ratio (DR): measures the dissimilarity between the respective frames. It calculates intra-frame dissimilarity of a group so that the amount of changes can be reflected. Standard deviation reflects the disparity of the elements of any group.

$$DR(G_i) = \frac{STD(G_i)}{G_{STD}} \times \frac{G_\delta}{\delta_{G_i}}. \quad (1)$$

$G_{STD}$ is the globally calculated Standard Deviation, $\delta_{G_i} = |Max(G_i) - Min(G_i)|$ $G_\delta$ = Global (Max)-Global (Mean) and STD ($G_i$) is the standard deviation of group $G_i$

Standard deviation (STD) of the information profile reflects the disparity of the elements of the groups and the difference(δ) between the maximum valued and the minimum valued element defines the range of the disparity. If δ is less than a certain threshold ($T_{\delta 1} = G_{STD}/G_\delta$), then the group is static irrespective of the STD. Alternatively, when δ greater than a certain threshold ($T_{\delta 2} = G_{STD}*\sqrt{2}$), then there is a cut (or fade) between the frames in the video segment and thus we can detect a group. Then we go ahead in measuring the DR of each group. For the group having δ less than $T_{\delta 1}$, its DR value becomes $T_{\delta 1}$ ( the lowest value) as there is no dynamic information in the group.

STD is inversely proportional to the similarity i.e. if standard deviation increases, similarity among the elements must be decreased and vice-versa. So standard- deviation is a parameter to measure the DR. Difference between the minimum-valued and maximum-valued element has a big role to measure the DR as it gives the highest deviation in a group. So it is another parameter to calculate DR.

The frames in a group must have certain amount of similarity, that means the information that they hold have certain amount of redundancy. So the frame(s) that hold highest information according to the information profile is selected the representation frame of the group. In case of much dissimilarity among the frames so that one frame cannot enough to represent the group, more frames are selected.

## 2.4 Summary generation

In this phase, we select the frames for summary generation by maintaining three basic properties of video summarization.

*Continuation*: A summarized video can be represented by its key frames only but though a collection of key frames may represent the optimized information of the original video, but it is not enough to understand the original information of the video as a minimum amount of time is required to understand the abbreviated information. Taking some supporting frames with the key frame(s) of each group we can make the message understandable. Groups are sub-divided according to the information changing, and supporting frames are selected from each subgroup. As the groups having higher DR values are more dynamic, more frames selected from that area. Supporting frames are selected from same group and in chronological order to make the continuation. Continuation assures that the information included in the summarized video is in chronological order. We divide the video into chronological segments (groups of similar frames) and select the frames in that order. Thus it maintains the continuity.

*Priority*: The information variation in the consecutive frames leads dynamism in the group. The DR value of the group increases according to the variation of the group and the number of frames selected from the group with respect to their DR values. Thus the priority scheme comes into action.

*No-repetition*: Frames are taken according to the DR of the group. DR must be decreased when the similarity among the frames is increased. Thus DR helps to reduce the redundancy in summarization.

## 2.5 Post processing

Post processing collects the selected frames in the previous steps and constructs a video which is very smaller in size but containing all effective information. After getting the summarized video if it is needed to increase or decrease the size of the video, then changing the selection of the number of required frames is done by initiating the summary generation phase once again. Number frames to create the summary can be controlled and it can be reduced to the number of selected key frames at the most.

# 3. RESULTS AND DISCUSSIONS

As discussed in section 2.1.1, a frame reduction results show that the reduced frame contained similar amount of information like the original frame, but it helps to save considerable amount of processing time. Table 1 shows the changes in different properties of original and reduced image. The size of the original frame was 326*484 and the size becomes 40*60 of each frame after reduction as the value of k is 8 in this case. "Figure 4" shows original images in the first row, edge of the respective images in second row, reduced images and their corresponding edges in third row.

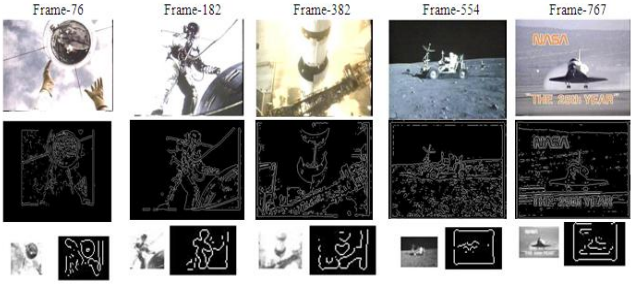


**Fig 4: The figure shows original images in the first row, edge of the respective images in second row, reduced images and their corresponding edges in third row.**

**Table 1. The change in different properties of the original & reduced image.**

| Img Idx: | Entr_Org | Entr_Red | EPR_Org | EPR_Red |
|---|---|---|---|---|
| Fr_76 | 3.77 | 4.05 | 0.0135 | 0.024 |
| Fr_182 | 2.42 | 2.85 | 0.012 | 0.025 |
| Fr_382 | 3.76 | 3.65 | 0.013 | 0.0275 |
| Fr_554 | 4.59 | 4.62 | 0.015 | 0.025 |
| Fr_767 | 4.58 | 4.60 | 0.016 | 0.030 |

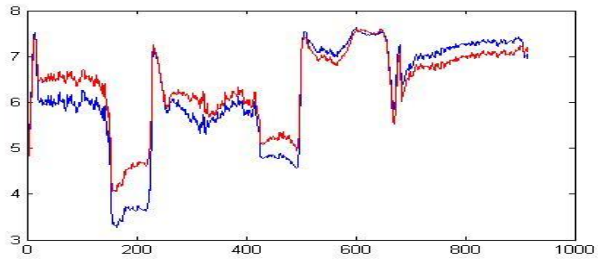

**Fig 5: The figure shows the entropy profile (blue) of a video of original sized frames and the entropy profile (red) of anni001 video of reduced sized frames**

Img Idx is the index of the image as in Table 1, Size_Org, Size_Red, Entr_Org, Entr_Red, EPR_Org, and EPR_Red represent size of the original image, Size of the reduced image, entropy of the original image, entropy of the reduced image, Edge Pixel Ratio (EPR) of original image and EPR of reduced image respectively. The results show that the difference is negligible in case of entropy of both the images. But as incase of EPR, the noise is reduced in the reduced image, the difference is little more in compare to the previous properties. Figure 5 shows the plot of entropy value versus frame numbers and Figures 6 and 7 show the plot of the *W(i)* versus frame numbers for the different experiments performed. In Figure 5, the entropy profile of the original frames of the video is plotted in blue, and the entropy profile of the reduced size frames is plotted in red. Both the plots show similar variation and hence it can be inferred that the nature of two profiles are reasonably similar.

The results in Figure 6 are the extracted information *W(i)* from frames following the techniques as discussed in Section 2.1. The entropy based information profile as in Figure 5 and the information profile in Figure 6 are different in nature. This is because, our proposed technique works on the motion and the object based features that provide reasonable better information of dynamism and presence of objects in the frame instead of entropy based information which we had proposed earlier [5]. Figure 6 also locates 6 shot cut points that imply 7 shots are present in the video. Figure 7 shows the selected frames containing key-frames from anni001 video [8] based on the information profile. The key-frames from anni001 [8] and basketball [9] are shown in Figure 8 and Figure 9 respectively.

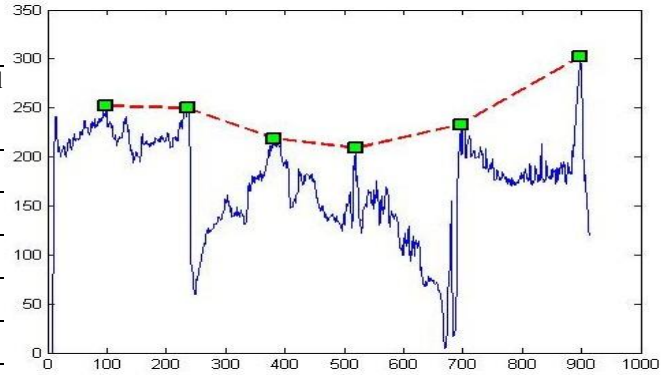

**Fig 6: Shot boundary detection of Anni001 video**

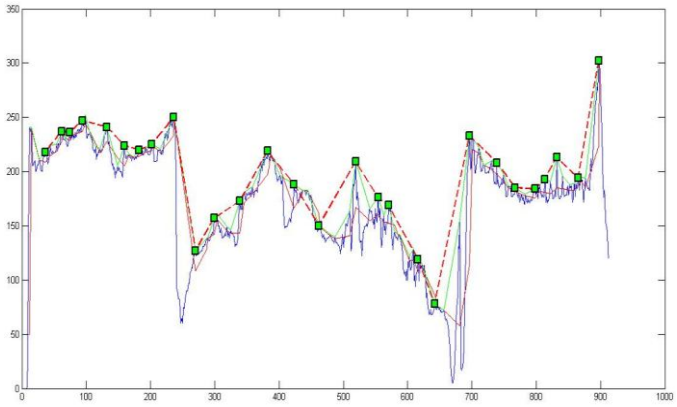

**Fig7: Key-frames of Anni001 video**

Table 2 includes a comparative study between our proposed approach and previous approach in [5]. *V_id* is the id. of the video used for the experiment, NOG is the number of groups extracted by the respective methods, NKF is the total number of extracted key frames from the video, NSFS is the number of selected frames for the summary of the video.

**Table 2. Comparative study of results of our proposed approach and [5]**

| V_id | NOG | | NK F | | NSF S | |
|---|---|---|---|---|---|---|
| | PM | PM[5] | PM | PM[5] | PM | PM[5] |
| Vid_1 | 7 | 7 | 18 | 28 | 90 | 88 |
| Vid_2 | 5 | 6 | 15 | 16 | 47 | 38 |

Some essential properties which are needed to be included in the video summarization to prove its efficiency are shown in Table 3 and a comparison with the related research works [3] is also presented. In any video, certain number of frames is similar with respect to information content. *Similarity group detection* is the procedure to accumulate those frames together to form similarity groups. *Key frame(s)* represents the message of the group, automatic key frame selection confirms to select these frames automatically. Sometimes one frame form each group is not enough to represent the message of the group, variable number of key frames is selected to reflect the actual message. The number of frames selected from each group according to their *DR* measure, *controls the total number of frames* according to the requirement and that will be distributed according to the *DR* of the corresponding groups that holds the *priority scheme.*

**Table 3. The properties includes in the proposed approach and in reference [3]**

| | MP [3] | CF [3] | Our approach |
|---|---|---|---|
| Automatic key frame selection | Y | Y | Y |
| Variable number of key frames | N | Y | Y |
| Similarity group detection | N | N | Y |
| Can control the number of frames for summary | N | N | Y |
| Any priority scheme | N | N | Y |

# 4. CONCLUSION AND FUTURE WORK

This paper explains the proposed methodology of video summarization for non-domain specific video application and hence it can work for any arbitrary video information. The size of the effective frame is reduced that noticeably minimizes the time requirement for information extraction. Information extraction has been done through motion and object based information that includes activity measurement in the scene. Our method tracks most active areas of the video based on DR measure to keep the originality of the information and also to maintain the quality parameters of the summarized video sequences. This method is novel in the sense that most of the existing works are domain specific but our proposed approach is non-domain specific and still generating a good quality of summarized video. In future we would try to modify intelligence rules to incorporate some complex object scenarios like back-ground estimation, knowledge based estimation etc.

# 5. ACKNOWLEDGEMENT

This research work is done through the funding provided by the Department of Science and Technology, Govt. of India under the Inspire Fellowship Scheme.

# 6. REFERENCES

[1] G. Money and H. Agius, "Video summarization: A conceptual framework and survey of the state of the art", Journal, ELSEVIER, April,2007

[2] J. Almeida, R. da S. T and N. J. Leite, "Rapid Video Summarization on Compressed Video" IEEE International Symposium on Multimedia, 2010.

[3] G. Ciocca1 and R. Schettini , "An Innovative Algorithm for Key Frame Extraction in Video Summarization", Journal Real Time Image Processing,2006,69-88

[4] L. Ren, Zhiyi Qu, Weiqin Niu, Chaoxin Niu and Yanqiu Cao, "Key Frame Extraction Based on Information Entropy and Edge Matching Rate", ICFCC, June, 2010

[5] S. Maity, A.Chakrabarti and D.Bhattacharjee; "An Innovative Technique for Adaptive Video Summarization", SPRINGER, ICIP, Bangalore, August,2011Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[6] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. Huang, "A Unified Framework for Video Summarization, Browsing, and Retrieval". Book, Elsevier Inc,2006

[7] M.K. Hu,"Visual pattern recognition by moment invariants," IRE Trans. on Information Theory, 8, pp. 179–187, 1962.

[8] www.youtube.com/user/2000turtle

[9] www.ivl.disco.unimib.it/temp/video.zip

[10] V. Khanna, P.Gupta and C.J. Hwang, "Finding Connected Components in Digital Images by Aggressive Reuse of Labels" Image and Vision Computing 20(Science Direct), 2002,557-568

[11] N. Benjamas, N. Cooharojananone and C. Jaruskulchai, "Flashlight and player detection in fighting sport for video summarization" Proceedings of the IEEE International Symposium on Communications and Information Technology (ISCIT 2005), vol. 1, Beijing,China, 12–14 October 2005, pp. 441–444.

[12] A.M. Ferman and A.M. Tekalp, "Two-stage hierarchical video summary extraction to match low- evel user browsing preferences" IEEE Transactions on Multimedia 5 (2) (2003) 244–256.

[13] X. Zhu and X. Wu, "Sequential association mining for video summarization" Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '03), vol. 3, Baltimore, MD, USA, 6–9 July, 2003, pp. 333–336.

[14] W. Cheng and D. Xu, "An approach to generating two-level video abstraction" Proceedings of the 2nd IEEE International Conference on Machine Learning and Cybernetics, vol. 5, Xi-an, China, 2–5 November, 2003, pp. 2896–2900.

[15] Z. Cernekova, I. Pitas and C. Nikou, "Information theory-based shot cut/ fade detection and video summarization" IEEE Transactions on Circuits and Systems for Video Technology 16 (1) (2006) 82–91.

[16] A. Ekin, A.M. Tekalp and R. Mehrotra, "Automatic soccer video analysis and summarization" IEEE Transactions on Image Processing 12 (7) (2003) 796–807.

[17] H. Shih and C. Huang, "MSN: statistical understanding of broadcasted baseball video using multi-level semantic network" IEEE Transactions on Broadcasting 51 (4) (2005) 449–459.

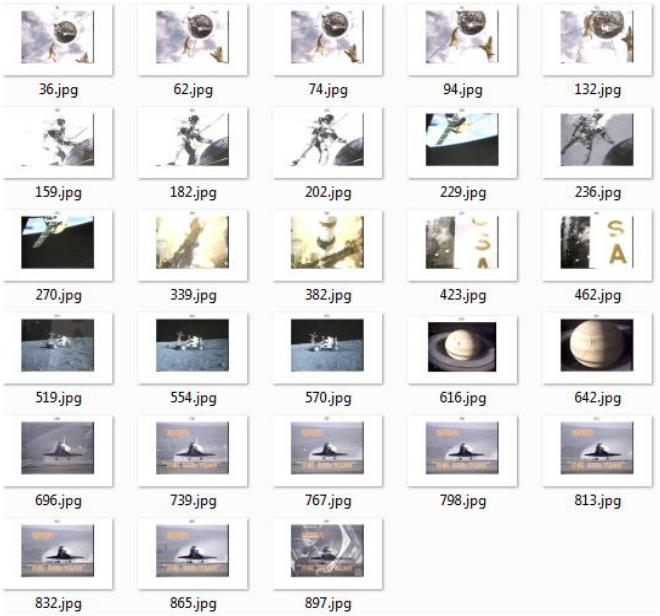


**Fig 8: Key-frames of Anni001 video**

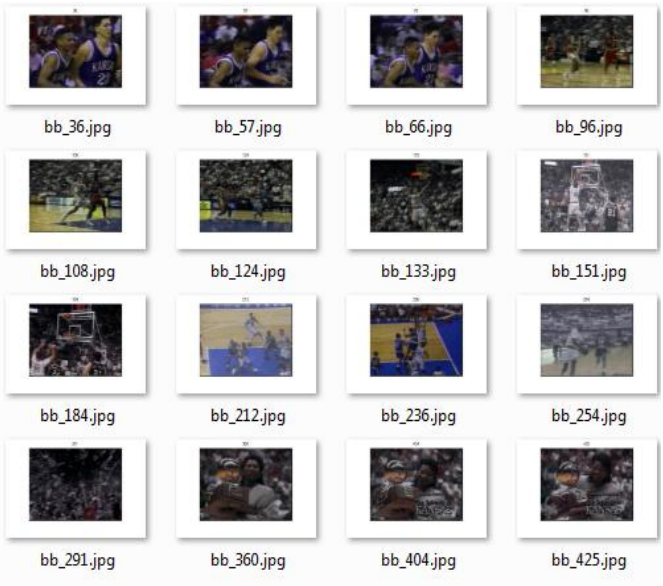


**Fig 9: Key-frames of basketball video**