Review of Keyframe Extraction Techniques for Video Summarization

Akshay Deshpande
UG Student
Pimpri Chinchwad College of Engineering
Pune, India

Vedang Bamnote
UG Student
Pimpri Chinchwad College of Engineering
Pune, India

Bhakti Patil
UG Student
Pimpri Chinchwad College of Engineering
Pune, India

Ashvini A. Tonge
Asst. Professor,
Department of IT
Pimpri Chinchwad College of Engineering
Pune, India

ABSTRACT

Video summarization methods tries to capture main parts, scenes and objects from specified video to have better video analysis. Two main goals of video summarization are deletion of redundant frames and extracting key frames from video. Video summarization can use different statistical measures such as mean, kurtosis and skewness for comparing each frame with next frame to calculate key frame. Key frames are crucial frames in the video. There are different methods to extract key frames from video and some of them are compared in this paper. The frame which satisfies threshold value will be considered as key frame and such key frames will then combined to form short summary.

Keywords

Video Summarization, Key Frame, Mean, Variance

1. INTRODUCTION

In today's world, information obtained from multimedia sources such as news, T.V. shows, internet contain huge amount of data. As there is limitation of time, so it is inconvenient to watch whole video. To overcome this problem video summarization comes into picture. Video summarization is nothing but creating summary which contains important information of large video, but short in time. So it makes user comfortable to watch only important content.

Video can be graphically represented as follows, which can be further divided into units as scenes, shots and frames. Thus we can consider frame as basic building block of video.

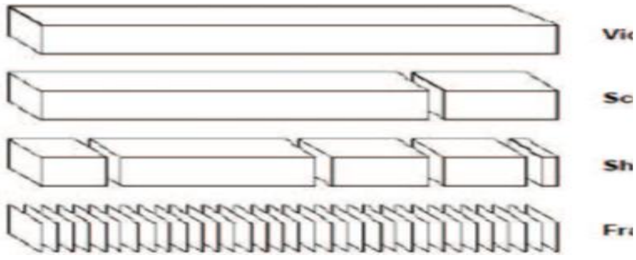

Fig 1: Anatomy of video structure

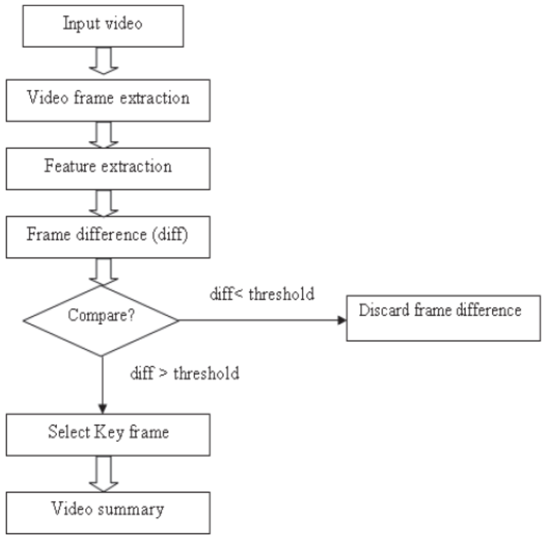

Fig 2: Work flow of proposed key-frame selection

Video summarization can be mainly divided into two parts as

- Static video summarization: This summarization uses method in which key frames are extracted from original video
- Dynamic video summarization: Dynamic video summary uses collection of video segments which are extracted from original video.

Techniques in video summarization touch various domains, such as movies, sports, news, home videos, e-learning, etc. Different techniques used for video summarization such as object base summaries, event based summaries, content based summaries, feature based summaries etc. In geographic research we can capture all moments of wild animals and then extracting only those which are significant for research. Video summarization consist of different techniques such as object based, event based, content based and feature based summaries. In this paper, section I consist of Introduction, section II consist of Related Work and section III consist of result.

2. RELATED WORK

There are many video summarization techniques. Some of them are listed below:

2.1.1 Edge Histogram Descriptor for Key frame Selection

In this method, each frame is described by Edge Histogram Desciptor(EHD) as proposed by Park et al [1].

This descriptor is also followed by MPEG-7 standard.

The EHD is constructed as follows: The particular frame is transformed to gray scale and then divided into 4x4 sub-images. Five edge filters are applied on each sub image to calculate strength of edges in various directions. For applying the filters, each sub-image block is treated as 2x2 super-pixel image block. Let f(k) represent an edge coefficient, where k = 0,1,2,3. Let a_k(i, j) represent the gray levels for the four image sub-blocks at pixel (I, j). Then edge magnitudes m_k(i, j) for particular edge filter f at the (i, j) image block can be obtained as:

$$\mathrm{m_f(i,j)} = \left|\sum_{\mathrm{k=0}}^{3} \mathrm{a_k(i,k)} \times \mathrm{f(k)}\right| \textbf{.....(1)}$$

m_f(i,j) is calculated in five directions: m_f(i,j), mfh(i,j), mfd−45(i,j), mfd−135(i,j) and mfnd(i,j), by corresponding edge filter coefficients fv(k), fh(k), fd−45(k), fd−135(k) and fnd(k).

The edge magnitudes for the entire frame is stored an 80-dimension histogram vector ie (5 x 4 x 4 = 80):

$$H_t = [b_1, b_2, \ldots, b_{80}]$$

Where t indicates frame number and bi indicates strength of edge. A threshold can be applied on the edge magnitudes to ignore weak edges. It can also help in reducing noise.

The average of H_t is calculated. The average is taken to be the frame descriptor [2].

2.2 Localized Foreground Entropy for key-frame Selection

In this approach, for each frame a foreground object mask is obtained using foreground detection (background subtraction) algorithm. Patch-based foreground estimation method can be used for obtaining foreground masks. Each foreground mask is divided into n x n sub-regions. The information quantity in each sub-region is measured using binary entropy function. In order to describe the content from the frames, Shannon's entropy algorithm is used to measure the quantity of foreground image. Let I be the foreground mask, where I(x, y) is the value of pixel (x, y) in binary form (1 for foreground, 0 for background). We define the probability of foreground object in the k-th sub-image region I[k] as:

$$p_k = \frac{1}{r_k c_k} \sum_{x_k, y_k} (fg|I_{(x_k,y_k)}^{[k]}) \textbf{.....(2)}$$

The local entropy information of a sub-image is a binary entropy function:

$$c_k = -p_k \, log_2(p_k) - (1 - p_k) log_2(1 - p_k) \textbf{.....(3)}$$

Where $c_k = 0$ if $p_k = 0$. We define the overall entropy information for frame f as:

$$E_f = \frac{1}{m} \sum_{k=1}^{m} c_k \textbf{.....(4)}$$

Entropy values are collected over the video or over the required number of frames. Local maxima are extracted using an iterative frame deleting strategy. The remaining frames are treated as key-frame [2].

2.3 Block Truncation Coding

lock truncation coding was developed around 1979 and has many applications in different image processing domains such as Content Based Image Retrieval. BTC mainly comes under lossy type of image compression technique. This technique is well suited for grayscale images. For such CBIR different image features are used, which can also be considered as properties of focused block of BTC.

To change this original BTC to different image colour model some changes are made to this technique. Two means are calculated as m, n for each plane in RGB colour model and threshold is calculated for all three colour model as follows

$$T_R = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} R(i,j) \textbf{.....(5)}$$

$$T_G = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} G(i,j) \textbf{.....(6)}$$

$$T_G = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} G(i,j) \textbf{.....(7)}$$

After this different binary bitmaps are generated for all three planes. If particular pixel 'n' in each plane crosses threshold then 1 is assigned value to it otherwise 0 as shown in equations 4,5,6.

$$BMR = \begin{cases} 1 \; if \; R(i,j) \geq T_R \\ 0 \; if \; R(i,j) < T_R \end{cases} \textbf{.....(8)}$$

$$BMG = \begin{cases} 1 \; if \; G(i,j) \geq T_G \\ 0 \; if \; G(i,j) < T_G \end{cases} \textbf{.....(9)}$$

$$BMB = \begin{cases} 1 \; if \; B(i,j) \geq T_B \\ 0 \; if \; B(i,j) < T_B \end{cases} \textbf{.....(10)}$$

After calculating binary bitmaps for three components two means as Upper mean and Lower mean for pixel greater than or equal to threshold and less than threshold respectively. Upper mean UM=(RM,GM,BM) and Lower mean LM=(RM',GM',BM')

$$RM = \frac{1}{\sum_{i=1}^{m} \sum_{j=1}^{n} BMR(i,j)} \sum_{i=1}^{m} \sum_{j=1}^{n} BMR(i,j) * R(i,j) \textbf{.....(11)}$$

$$RM' = \frac{1}{m*n - (\sum_{i=1}^{m} \sum_{j=1}^{n} BMR(i,j))} \sum_{i=1}^{m} \sum_{j=1}^{n} \{1 - BMR(i,j) * R(i,j)\} \quad \textbf{.....(12)}$$

Upper and lower mean provide us way to get feature vector and using this feature vector key frames can be extracted [3].

2.4 Thepade's Ternary Block Truncation Coding

Dr Sudeep D. Thepade, Dr H.B.Kekre, Anil T. Lohar introduced advanced version of Block Truncation Coding as Thepade's Ternary Block Truncation Coding. In this technique they have considered 'm' as its height and 'n' as its width, then thresholds are calculated for R,G,B planes same as of BTC

$$T_R = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} R(i,j) \textbf{.....(13)}$$

$$T_G = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} G(i,j) \textbf{.....(14)}$$

$$T_B = \frac{1}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} B(i,j) \textbf{.....(15)}$$

Then they have proposed that mean of all three threshold is calculated as

$$T = \frac{T_R+T_G+T_B}{3}\textbf{.....(16)}$$

Then two thresholds values for R, G, B planes are calculated which are upper and lower level and this are shown as in this equation

$$T_{shrh} = T_R + n|TR - T| \textbf{.....(17)}$$

$$T_{shgl} = T_G - n|TG - T| \textbf{.....(18)}$$

$$T_{shgh} = T_G + n|TG - T| \textbf{.....(19)}$$

$$T_{shbl} = T_B - n|TG - T| \textbf{.....(20)}$$

$$T_{shbh} = T_B + n|TG - T| \textbf{.....(21)}$$

Two mean threshold values for each plane red, green and blue component are produced as upper and lower thresholds. Ternary bitmaps are produced for all 3 planes are computed. If pixel 'n' is greater than or equal to respective higher threshold or less than equal to 255 then value is given as 1 and then 0 and -1 are assigned as value to that pixel with respective conditions.

$$TMr(i,j) = \begin{cases} 1 \ \ if \ Tshrh < R(i,j) \le 255 \\ 0 \ \ if \ Tshrl < R(i,j) \le Thsrh \\ -1 \ \ if \ 0 \le R(i,j) \le Tshrl \end{cases} \textbf{.....(22)}$$

$$TMg(i,j) = \begin{cases} 1 \ \ if \ Tshgh < R(i,j) \le 255 \\ 0 \ \ if \ Tshgl < R(i,j) \le Thsgh \\ -1 \ \ if \ 0 \le R(i,j) \le Tshgl \end{cases} \textbf{.....(23)}$$

$$TMb(i,j) = \begin{cases} 1 \ \ if \ Tshbh < R(i,j) \le 255 \\ 0 \ \ if \ Tshbl < R(i,j) \le Thsbh \\ -1 \ \ if \ 0 \le R(i,j) \le Tshbl \end{cases} \textbf{.....(24)}$$

Lower, Medium and Higher means per colour components are calculated and for Red plane can be given as

$$LR = \frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n} 1, iff \ TMr \ \ (i,j)=-1}\sum_{i=1}^{m}\sum_{j=1}^{n} R(i,j)\,, iff \ TMr = -1 \quad \textbf{...(25)}$$

$$\mathrm{MR} = \frac{1}{\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}} 1,\mathrm{iff\ TMr}\ \ (\mathrm{i,j})=-1}\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{R(i,j)},\ iff \ TMr = \ 0 \quad \textbf{...(26)}$$

$$\mathrm{HR} = \frac{1}{\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}} 1,\mathrm{iff\ TMr}\ \ (\mathrm{i,j})=-1}\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{R(i,j)},\ iff \ TMr = \ 1 \quad \textbf{...(27)}$$

Then features for each component are calculated and feature vector with nine values are calculated as [LR, MR, HR, LG, MG, HG, LB, MB, HB] are used to find out key frames from that video[3].

2.5 Key Frame Extraction Based On Frame Blocks Differential Accumulation

First the video is divided into number of shots. These shots are used to extract key frames. The key frames are extracted using visual features such as colour, texture, shape. Based on these features some techniques are included for key frame extraction. They are:

1. Shot-Based method: This method is applicable only for static video. It is simple method. This video is divided into number of shots and In each shot first or last frame is considered as key frame[5].

2. Motion-analysis based Method: In this method, key frames were extracted based on motion of camera or object in video. Y.Z.Ma et al detected according to acceleration moving objects in video. M.Guironnet et al proposed a method based on camera motion, were motion type of camera are detected & then key frames are extracted based on order of motion types. Its disadvantage is that it requires large calculations[6].

3. Clustering-based Method: In this, similar frames are clustered to same category. Clustering can be used in single shot. When it used in shots, sub shots are generated. This method is proved as effective [7].

To overcome disadvantages of above techniques new method is proposed named as Key Frame extraction based on frame blocks differential accumulation with two thresholds. Consider dimension of each image frame in video as M*N. Initially assume first frame as reference frame. Then, image frames are partitioned into equal size sub images. The current frame is compared with reference frame to detect whether the contents have changed or not. The number of blocks with content changes a lot in current frame is counted. If the number is greater than threshold predicts that current frame content changes a lot. And now consider this frame as keyframe [4].

The main algorithm is divided in main 4 steps:1. Consider first frame as reference frame. Then divide image frames in video into equal sized sub-images. Then calculate color mean in RGB color space for each frame and reference frame also. The color mean in m row and n column for current frame is represented as

$$R_i(m,n), G_i(m,n), B_i(m,n)$$

The color mean in m row and n column for reference frame is represented as -

$$R_r(m,n), G_r(m,n), B_r(m,n)$$

2. Calculate the color mean differences of corresponding blocks between the current frame & reference frame by equation 28,29,30.

$$D_R = R_i(m,n) - R_r(m,n)\textbf{.....(28)}$$

$$D_G = G_i(m,n) - G_r(m,n) \textbf{....(29)}$$

$$D_B = B_i(m,n) - B_r(m,n)\textbf{.....(30)}$$

Equation 28 to 30 calculate the difference of red, green, blue components respectively.

$$Diff_i(m,n) = \sqrt{\frac{D_R^2+D_G^2+D_B^2}{3}}\textbf{.....(31)}$$

Equation 31 will give color mean difference between current frame and reference. Here m is m^th^ row and n is n^th^ column.

3. Suppose T_1 is the block threshold, if $Diff_i(m,n) > T_1$ means this block changes a lot, if $Diff_i(m,n) < T_1$ means this block is similar to reference frame. Now count the number of blocks which changes a lot by equation 32.

$$C_i = \begin{cases} C_i \ Diff_i(m,n) < \ T_1 \\ C_i \ Diff_i(m,n) > T_1 \end{cases}\textbf{.....(32)}$$

In formula (32), C_i is the number of blocks whose contents change a lot in current frames. Its initial value is 0.

4. Suppose T_2 is the global threshold, compare C_i with T_2, if $C_i > T_2$, it means that this frame changes a lot in relation to the reference frame. If $C_i < T_2$, it means that the current frame is similar to the reference frame. Take the next frame as the current frame and repeat steps above until the last frame.

2.5.1 Selection of Thresholds

There are two thresholds in this method, the block threshold T_1 and the global threshold T_2 . The global threshold is self-adaptive.

Calculation procedure is as follow:

Step 1: Compute the mean differences of the all blocks in each frame by equation (33):

$$mValue = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{Diff\,(i,j)}{(m \times n)} \ldots..\textbf{(33)}$$

Step 2: Compute T_1 and T_2 .

$$T_1 = a \times mValue \ldots..\textbf{(34)}$$

$$T_2 = b + (m \times n) \times 60\% \ldots..\textbf{(35)}$$

Different users can change the parameters a and b in equation (34) and (35) to achieve satisfactory results. Where $a \in$ [0,1], $b \in$ [-10,10].

3. OBSERVATIONS

Comparison of all the above techniques is calculated using precision and recall.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{\text{retrieved documents}}$$

$$precision = \frac{|\{relevant\ document\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

Table 1. Comparison of techniques used

Technique	Features	Precision	Recall
Edge Histogram Descriptor	Strength of edge in various direction is calculated using EHD	0.580	0.840
Thepade's Block Truncation Coding	Advanced method of BTC with additional statistical features provide better result	0.637	0.638
Localized foreground entropy	Shannon's entropy function used for measuring quantity of foreground image	0.682	0.948
Differential Accumulation	Threshold is calculated two times to have more accurate frame.	0.583	0.596

4. CONCLUSION

Key frame extraction finds its use in many video-based applications. This may be indexing, retrieval and browsing applications. The comparison concludes that key frame extraction can be done based on various different features for video summarization. The performance for these methods changes depending upon the concerned video. Methods such Edge histogram descriptor and Localized foreground entropy are found out to be comparatively better in terms of performance. In addition, the recall is found out to be more when retrieval rate is more and the precision rate is found out to be more when retrieval rate is less.

5. REFERENCES

[1] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in Proceedings of ACM Workshops on Multimedia, 2000, pp. 51–54.

[2] Yan Yang, Farhad Dadgostar, Conrad Sanderson, Brian C. Lovell, "Summarisation of Surveillance Videos by Key-frame Selection," The University of Queensland, School of ITEE, QLD 4072, Australia.

[3] Dr.Sudeep.D.Thepade, "Novel Method for Keyframe Extraction using Block Truncation Coding and Mean Square Error", Pimpri Chinchwad Colllege of Engg, Pune, India.

[4] Chanquing Cao, Zehua Chen, Gang Xie, Shaoshuai Lei, "Key Frame Extraction Based on Frame Blocks Differential Accumulation," Taiyuan University of Technology, Taiyuan, 030024, China.

[5] S. Behzad, C. D. Gibbon. Automatic generation of pictorial transcripts of video programs, Proc. SPIE. 512-518, 1995.

[6] Y. Z. Ma, Y. L. Chang, H. Yuan. Key-frame extraction based on motion acceleration, Optical Engineering, Vol.47, No.9,2008.

[7] L. Pan, X. J. Wu, Video shot segmentation and key frame extraction based on clustering, Infared and Laser Engineering, Vol.34, No.3, 341-344, 2005.

[8] S. R. Badre and S. D. Thepade, "Summarization with key frame extraction using thepade's sorted n-ary block truncation coding applied on haar wavelet of video frame," 2016 Conference on Advances in Signal Processing (CASP), Pune, 2016, pp. 332-336.

[9] A. A. Tonge and S. D. Thepade, "Key frame extraction for video content summarization using orthogonal transforms and fractional energy coefficients," 2015 International Conference on Information Processing (ICIP), Pune, 2015, pp. 642-646.

[10] S. D. Thepade and P. H. Patil, "Novel visual content summarization in videos using keyframe extraction with Thepade's Sorted Ternary Block truncation Coding and Assorted similarity measures," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, 2015, pp. 1-5