IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. II (Sep – Oct. 2014), PP 52-56
www.iosrjournals.org

Comparisons of Video Summarization Methods

Dr.Jharna Majumdar[1], Spoorthy.B[2]
[1](Dean R&D, Prof. HOD of CSE (PG), Nitte Meenakshi Institute of Technology. INDIA)
[2](M Tech, Dept of CSE (PG), Nitte Meenakshi Institute of Technology. INDIA)

Abstract: *Video summarization is a process of removing the redundant frames and generating the most informative key frames of the videos. In this paper we have explained two efficient methods for video summarization and given the comparison between these two methods. The first method is Video summarization using CLD feature extraction and adaptive threshold technique for shot boundary detection. Second method is video summarization using aggregation function where three different features such as Histogram difference, correlation difference and moment of inertia difference are combined and that difference value is compared with the predefined threshold value.*

Keywords: *video summary, adaptive threshold ,key frames, aggregation function.*

I. Introduction

Internet videos have got an enormous popularity in the present world. The video contents are increasing day by day in you tube and yahoo videos .We need an efficient tool for fast video browsing of the internet videos. Usually every videos contain a lot of redundant information which can be removed to make video data more suitable for retrieval, indexing and storage. This approach of removing redundancies from the video and generating condensed versions of the video is called video summarization .The video summaries contain the most important and relevant content of a video at the same time, the original message of the video must be preserved. The video summaries can be generated in two different ways static and dynamic. Static video summaries deal with the extraction of the key frames from the video. The key frames are still frames extracted from the video which hold the most important content of the video and they are representative of the video. The dynamic video summary contains small shots that are accumulated in a time ordered sequence.

In this paper we have given two efficient methods for static video summarization. The first method is getting the video summary by extracting the color layout descriptor feature for each frame and detecting the shot boundary using the adaptive threshold technique to extract the key frames. In the second method we have aggregated the three different feature like histogram, correlation and moment of inertia and then calculated the aggregation difference, compared this difference with the predefined threshold. Finally we compared the efficiency between these two methods by using the recall and precision rates.

II. Colour Layout Descriptor

We used the color layout descriptor feature and adaptive threshold technique to extract the key frames of input video to obtain the video summary. [1][2]

Step 1: Pre-processing step: The video contains many redundant frames, our aim is to remove the redundant frames to reduce the computational time in frame comparison. We can sample the frames by splitting the input video into time segments or the very simple and commonly used method which we used in our paper is by taking every 10^{th} frames of the input video so that we get uniform samples with fixed sample rate.

Step 2. CLD frame feature extraction: Frame**s** features extraction is the method of extracting the properties of the frames. Using the frame feature we can determine the shot boundaries. The commonly used frame feature are pixel difference, histogram difference, template matching edge change etc. CLD has been designed to efficiently and compactly represent spatial layout of colors inside images[3].To obtain the CLD feature we have to undergo the following steps from step (a) to step (d).

(a) Image Partitioning: The input frames on RGB color space is divided into 64 blocks to guarantee the invariance to resolution or scaling.

(b) Representative color selection: After image partitioning stage a single representative color is selected from each blocks .Any method to select the representative color can be applied, we have used average of the pixel colors in a block as the corresponding representative color, since it is simpler and accurate. The selection results in a tiny image icon of size 8x8.This tiny image icon is converted from RGB color space to YC_bC_r color space.

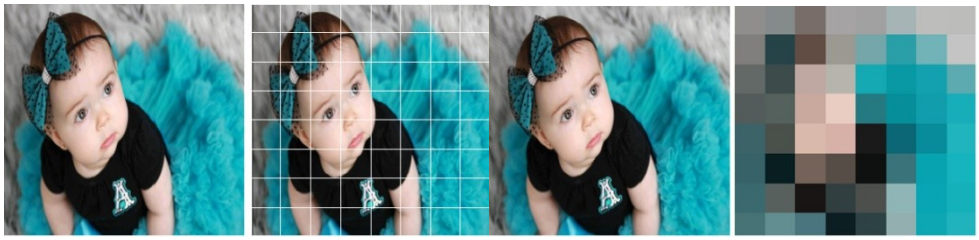

Fig 1: Image Partitioning **Fig 2: Representative color selection**

(c) DCT Transformation: The three Y,C_b,C_r 8x8 matrix is transformed by 8x8 DCT , so three sets of 64 DCT coefficients are obtained. To calculate the DCT we used the below formula.

$$B_{pq} = \alpha_p \alpha_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \, COS \frac{\pi(2m+1)p}{2M} COS \frac{\pi(2n+1)q}{2N}$$

$$\alpha_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le p \le M-1 \end{cases} \qquad \alpha_q = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{\frac{2}{N}}, & 1 \le q \le N-1 \end{cases}$$

(d) Zigzag Scanning: The purpose of zigzag scanning is to group low frequency coefficients in top of vector and Zigzag scanning is performed to convert 2D array into 1D array. Zigzag scanning is performed on each of DCT matrix which we got in the last step.Finally a zigzag scanned DCT coefficents are concatenated into one feature vector by taking six Y,three C_b and three C_r values.

We have used a 12 dimentional CLD feature for video shot detection.for every pre- sampled frames we extract the CLD feature vector.f_i^{CLD}=$\left(f_{y1}^{CLD}, \ldots\ldots f_{y6}^{CLD}, f_{Cb7}^{CLD}, f_{Cb8}^{CLD}, f_{Cb9}^{CLD}, f_{Cr10}^{CLD}, f_{Cr11}^{CLD}, f_{Cr12}^{CLD}\right)$

Where f_i^{CLD} is the CLD feature vector of the i[th]frame.

Step 3: Video shot boundary detection using adaptive threshold technique.

We calculated the difference value between two successive fames and that difference value is compared with the adaptive threshold value. If the difference value is greater than the adaptive threshold value we can say that there is a shot transition in between those frames. The frame difference is calculated using Euclidean distance between Successive CLD frame features as give as $d_i^{CLD} = \left\| \frac{f_i^{CLD} - f_{i-1}^{CLD}}{\max(f_i^{CLD})} \right\|$

Adaptive threshold value is computed using the formula below: $T_i = \alpha \, . \, \mu + T_{const}$

Where α and T_{const} are constants parameter and μ represents the mean d_i^{CLD} value.

Step 4. Key frame Selection: The last frame of the shot or the frame where the shot boundary is detected is extracted as the key frame**.**

Step 5. Elimination of similar key frames: After the key frames are extracted, there can be similar key frames appeared at different temporal positions in the input video. To eliminate these similar key frames we have applied histogram difference method. First the histogram difference [3] is calculated between all the selected key frames and distance matrix is computed. This distance matrix values are compared with the predefined threshold T_{dist} .If the frames distance value is less than predefined threshold T_{dist} then one of the frames is removed from the summary. Finally the remaining key frames represent the summary of the input video.

III. Aggregation Mechanism

In this method, to calculate the frame difference we have used three frame features namely correlation of RGB color space, color histogram and moment of inertia. These three frame features measure are then combined using aggregation mechanism to extract the key frame.

Initially first frame of the video is consider as the first keyframe (KF) and the remaining frames are considered as candidate frames (CF). The feature differences are calculated as below.

(a) Correlation frame difference measure: The similarity between two frames are calculated using the correlation coefficient. Each red, green blue color channels are divided into total Ts sections of size pxq and correlation is calculated between each sections of corresponding candidate frames. Let's F(t) and F(t+1) are two frames ,the correlation coefficient for a section 's' is calculated for color channel 'c' of F(t) and F(t+1) by using the below formula.

$$r\big(F(t),F(t+1)\big)_{s,c} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\big(F_s(t)_{c,i,j} - \overline{F_{c,s}(t)_{c,i,j}}\big)\ \big(F_s(t+1)_{c,i,j} - \overline{F_{c,s}(t+1)_{c,i,j}}\big)}{\sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q}\big(F_s(t)_{c,i,j} - \overline{F_{c,s}(t)_{c,i,j}}\big)^2 \sum_{i=1}^{p}\sum_{j=1}^{q}\big(F_s(t+1)_{c,i,j} - \overline{F_{c,s}(t+1)_{c,i,j}}\big)^2}}$$

where $F_s(t)_{c,i,j}$ is the pixel value of 'c' color channel of F(t) at row 'i' and column 'j' in section 's', and $\overline{F_{c,s}(t)}$ are the mean values of the pixel values of color channel 'c' in section 's'. The correlation is computed for each section and each color channel and the mean of all the section is calculated. Later the mean is calculated from all the color channels.

$$r\big(F(t),F(t+1)\big)_c = \frac{1}{T_s}\sum_{k=1}^{T_s} r\big(F(t),F(t+1)\big)_{k,c}$$

$$\rho\big(F(t),F(t+1)\big) = \frac{(\rho_{red} + \rho_{blue} + \rho_{green})}{3}$$

(b) Histogram frame difference measure: Histogram frame difference is calculated in two parts, calculation of histogram and finding difference. First Convert RGB values of frame into HSV, then Draw color histograms of each hue, saturation and value component.color quantization is done to reduce size of the color histogram, hue component histogram is reduced to 16 bins and other two are reduced to 8 bins each.Normalization the HSV components in the range 0-1.Combination of three histograms to form a single histogram of size 32 bin.Obtain histogram difference between corresponding sections of the frame by using Euclidean distance.

$$H\big(F(t),F(t+1)\big) = \sqrt{\sum_{i=1}^{32}(H_{F,s}(i) - H_{F+1,s}(i)}$$

(c) Moment of inertia difference: The moments of inertia is calculated by mean, variance and skewness, each color channel are used for calculation of 9 moments of each section of fame. For frame F(T) and S section and color channel c and mean variance and skewness values are computed as below.

$$\overline{F(t)}_{s,c} = \frac{1}{pXq}\sum_{i=1}^{p}\sum_{j=1}^{q} F(t)_{i,j}$$

$$\sigma^2(F(t))_{s,c} = \frac{1}{pXq}\sum_{i=1}^{p}\sum_{j=1}^{q}(F(t)_{i,j} - \overline{F(t)}_{s,c})^2$$

$$\mathrm{Y}(F(t))_{s,c} = \frac{1}{pXq}\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(F(t)_{i,j} - \overline{F(t)}_{s,c})^3}{(\sigma^2(F(t))_{s,c})^{3/2}}$$

Where $\overline{F(t)}_{s,c}, \sigma^2(F(t))_{s,c}, \mathrm{Y}(F(t))_{s,c}$ are the mean, variance, and skewness values of color channel "c" in section "s" respectively Finally, these values are combined to form a moments of inertia feature vector μ_t of frame F (t). The moments of inertia difference measure between two frames F(t)and F(t + 1) is computed by using the Euclidean distance between the respective feature vectors.

$$\mu\big(F(t),F(t+1)\big) = \sqrt{\sum_{i=1}^{9Ts}(\mu_t(i) - \mu_{t+1}(i))^2}$$

(d) Aggregation mechanism and key frame selection: The three adaptive frame difference measures ρ_n, μ_n, $H_n$are compared with pre-defined threshold $\tau_\rho, \tau_H and \tau_\mu$ respectively. As a result of this comparison, a real number called "contributing value" is obtained for each adaptive frame difference measure. ρ_n captures the amount of similarity, whereas μ_n, H_n captures the amount of difference between the current frame and the last key frame. By comparing with thresholds, the corresponding contributing values of the three measures are generated in such a way that the contributing values are high if there is a high inter-frame difference and vice versa .For instance, a value of ρ_n that is less than the threshold τ_ρ indicates significant inter-frame difference and thus results in a positive contributing value. The contributing value will be high if the difference between τ_ρ and ρ_n is high and vice versa. On the other hand, if ρ_n is greater thanτ_ρ, the result is a negative contributing value which signifies low inter-frame difference. The contributing values are calculated by

$$\mathrm{d_p} = \begin{cases} 1 + |\rho_n - \tau_\mathrm{p}| & \text{if } \rho_n < \tau_\mathrm{p} \\ -|\rho_n - \tau_\mathrm{p}| & \text{Otherwise} \end{cases}$$

$$d_H = \begin{cases} 1 + |H_n - \tau_H| & \text{if } H_n < \tau_H \\ -|H_n - \tau_H| & \text{Otherwise} \end{cases}$$

$$d_\mu = \begin{cases} 1 + |\mu_n - \tau_\mu| & \text{if } \mu_n < \tau_\mu \\ -|\mu_n - \tau_\mu| & \text{Otherwise} \end{cases}$$

The three contributing values are combined to obtain aggregate frame difference measure as below:

$D = W1 d_\rho + W2 d_H + W3 d_\mu$

Here, W1,W2,W3 are weights assigned to contributing value If a frame's aggregate comparison value is higher than threshold, it is declared as a key frame.

IV. Results And Comparisions Of Both The Methods

We have implemented both the algorithm in Visual Studio 6.0, VC++. We have used the different input videos downloaded from the open-video web site and experimented by giving different α and T_{const} values till we get a satisfactory key frames video summary. Fig 4 shows the results obtained from CLD method for 3 different videos and the Fig 5 shows the results obtained from the aggregation function for the same three different videos.

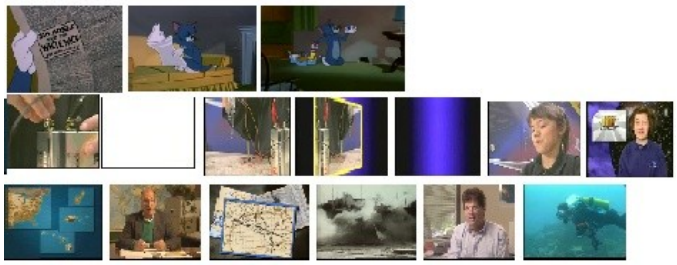

Fig 4: Result of CLD method for 3 different videos (one video per row).

Fig 5: Result of Aggregation method for 3 different videos (one video per row).

Comparison: We have calculated the precision and Recall values for both the methods for the output of 5 different videos of different frame number and calculated the time complexity.

According to the observations the CLD method is fast as it takes less time to execute when compared to aggregation method because in CLD method the frame is converted into 64 coefficients and these 64 coefficients are used in the computation. But in aggregation function the whole frame size is used for computation. As the total number of frames increases, the efficiency of the aggregation function becomes less compared to CLD method.

The advantage of CLD method is it uses adaptive threshold technique for shot boundary detection so there is no need of selecting predefined threshold values but in aggregation method we have to give the predefined threshold value according to the input videos.

Input Videos	Total no of frames	CLD		Aggregation Function	
		Recall	Precision	Recall	Precision
Cartoon	406	100	100	100	100
Movie	622	70	100	75	75
Exotic terrane	2938	81.25	81.25	71.42	66
Hurricane force	2390	80	80	70	73
Drift Ice	1816	92.30	80	78	73

Table 1: Recall and precision values for both the method.

V. Conclusion

In this paper two efficient techniques for video summarization are being explained. First is Color layout descriptor for frame feature extraction, using the adaptive threshold technique and the second is aggregation mechanism which combines three feature difference measure to extract the key frames which represents the input video. Both the methods gives good informatics key frames. We compared the results obtained from both the methods by using the recall and precision values and finally we concluded that CLD method is more efficient and fast compared to aggregate method. Future work is to apply adaptive threshold for aggregation mechanism as the predefined threshold doesn`t produce accurate results for all the input videos.

References

[1] "Efficient use of MPEG-7 Color Layout and EdgeHistogram Descriptors in CBI Systems"; Balasubramani R Dr.V.Kannan; GJCST , 2011.
[2] "Multimedia Content Filtering, Browsing, and Matching using MPEG-7 Compact Color Descriptors"; Santhana Krishnamachari , Akio Yamada , Mohamed Abdel-Mottaleb;In Proceedings of the Fourth Intl' Conf. On Visual Information Systems, Nov. 2000, France.
[3] "Image Retrieval System Based on Color Layout Descriptor and Gabor Filters"; Hamid A. Jalab;2011 IEEE Conference on Open Systems (ICOS2011).
[4] "An Algorithm for Shot Boundary Detection and Key Frame Extraction Using Histogram Difference";Ganesh. I. Rathod , Dipali. A. Nikam;International Journal of Emerging Technology and Advanced Engineering-2013.
[5] "Key frame extraction using color histogram method.";Miss.A.V.Kumthekar, ;Prof.Mrs.J.K.Patil; IJSRET-2013
[6] "Key Frame Extraction using Edge Change Ratio for Shot Segmentation "; Azra Nasreen , Dr Shobha G ; International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2013.
[7] "Video summarization using color features and efficient adaptive threshold technique" ; Stevica CVETKOVIC, Marko JELENKOVIC, Sasa V. NIKOLIC.
[8] "Adaptive Key Frame Extraction for Video Summarization Using an AgregationMechanism";Naveed Ejaz , Tayyab Bin Tariq ,Sung Wook Baik;ELSEVEVIER -2012.
[9] "Key Frame Extraction Using Features Aggregation"; B. F. Momin, S. B. Pawar ; International Journal of Recent Development in Engineering and Technology -2012.
[10] "Video summarization using clustering ",Tommy chheng.
[11] " www.open-video.com".