HCTL Open International Journal of Technology Innovations and Research (IJTIR)
http://ijtir.hctl.org
Volume 14, April 2015
e-ISSN: 2321-1814, ISBN (Print): 978-1-62951-946-3

DCT-Based Static Video Summarization

Prachi Singh[1], Suneeta Agarwal[2]

prachi.singh177@gmail.com

Abstract

Due to the recent development in the technology, large amount of video data is being produced nowadays. If any user needs to search for a video related to a particular topic, then he is unnecessarily bound to look into many videos fully to get the relevant content. Video summarization provides a concise representation of the content of a video while preserving the essential information. Many methods for summarization have been proposed so far, but most of them are time consuming and require large amount of computational space. In this paper, we present a new DCT-based static video summarization, which utilizes the DCT information of each frame. This method effectively measures the correlation between the successive frames and then selects frames having content different from their respective neighboring frames. Our results of summarization are found to be 90% closer to the user summary

Keywords

Video summarization, Discrete Cosine Transformation, Pearson Correlation coefficient, F-measure, KeyFrames.

Introduction

Due to the mushrooming of video acquisition devices, lots and lots of video data are now being produced. However, the state-of-art technology available for their storage and searching does not keep pace with the amount of video data available. This has happened due to the cheap availability of space for storage. The current need of the state is to find out efficient techniques for the efficient storage for faster searching of data. This requirement leads us to video summarization.

Video summarization, aims at extracting some important frames from a video such that those selected frames represent the overview of the video. This is helpful to the user in getting an idea about the content of the video, without watching the full video. The summary of a video can be generated in many ways.

[1]CSED,MNNIT Allahabad
[2]CSED, MNNIT Allahabad, Email: suneeta@mnnit.ac.in

Two most popular ways are static video summary and dynamic video skim. Static video summary is a small set of keyframes extracted from the video whereas dynamic video skim contains sequence of small shots while preserving the dynamic nature of video. In dynamic video skim, we also have to take care of audio synchronization. On the other hand, the keyframe sets does not have any synchronization issue and user can get the overview in less time. In this paper, our focus will be on static video summaries.

Over the past many years, various video summarization techniques have been proposed. However, these methods tend to be time consuming due to the large amount of data in a video, because even a video of 1 minute duration consists of 1800 frames, considering 30fps. However, many consecutive frames are redundant. This redundancy is required to add continuity to a video. In this paper we have come up with the idea of DCT-based static video summarization. First, pre-sampling is done to remove redundant frames to reduce our computation. Then, the frames are compressed using DCT. Then, the correlation co-efficient is calculated for each two successive compressed frames. Based on the correlation values, keyframes are selected.

Rest of the paper is organized as follows. Section 2 presents the related work. Section 3 presents DCT-based approach for video summarization in detail. Section 4 shows the Experiment and Results. Finally, conclusion and future scope is shown in Section 5.

Related Work

A comprehensive review on the various video summarization techniques along with their advantages, limitations and the methodology can be found in [1]. Some of the main static video summarization techniques are discussed next.

In [11], k-means clustering algorithm is used for video summarization. Firstly, the video is split into frames and is then only a subset of frames is selected at a predefined sampling rate. Then, the color histogram for the hue component of HSV color space is computed for each of the frame. The number of clusters (k) can be determined by calculating the Euclidean distance between the two frames, and each time the distance is larger than threshold 'τ' (0.5 in this case), k is incremented by 1.Then a frame which is similar to the cluster centroid is selected as a keyframe and then the summary generated by this method and the user summary is compared. The threshold value considered may change for different type of video. The selection of threshold value is subjective in nature.

In [4], the frames are compressed using DCT and then color histograms in HSV space is computed to get a 256-dimensional feature vector. The Zero Mean Normalized correlation (ZNCC) is used to find the similarity between the two consecutive frames. One frame from each of the frame group is selected as a keyframe. It offers customization by allowing user to adjust the value of thresholds. No pre-sampling is done, thus summarization may take larger time for larger duration of video.

In [5], three parameters, namely, inter-frame correlation of RGB channels, color histogram, moments of inertia, are used. First, the frames are divided into slots. The first parameter measures the correlation between two consecutive frames in RGB channel. The second parameter measures the frame difference in RGB channel and the third parameter captures the difference in the pattern of two frames. Then these values are aggregated to extract key frames from the video. An adaptive rule is then used to provide tolerance to lightning conditions. The selection of threshold value is subjective in nature.

In [3], local features are extracted from each of the frame in the form of descriptors. Then these descriptors are combined to form a global pool of keypoints.Then, the frame that best matches these global keypoints that best matches are selected as keyframes. For this they have used two criteria, namely, coverage and redundancy to ensure that no two keyframes are similar and that the number of keypoints covered by them is, maximum.

DCT-Based Approach

Given a video V to be summarized, starts a time 't', having 'n_{of}' frames is represented as :

$$V_t = \{ F(t+i) | i=0,1,\ldots\ldots\ldots\ldots, n_{of} -1\} \quad (1)$$

Equation (1), 'F' refers to a frame of a video. The aim is to extract a subset of distinct frames from this video, which represents the overview of the video with less redundancy. The steps of DCT-based summarization is as explained below:

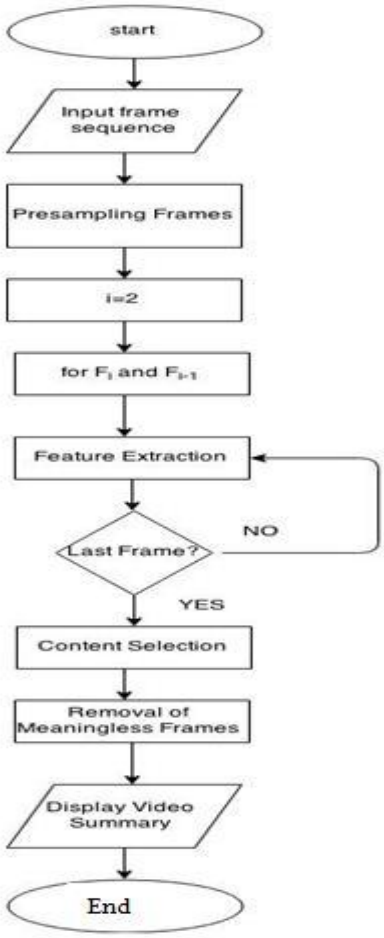

Figure 1: Flowchart for DCT-based Approach

Pre-sampling

In this step, firstly the standard deviation of each of the frames is computed and then frames are selected based on the difference between two successive frames. If the difference is very less, it means that those frames are structurally varying same and the current frame is dropped. However, if the difference is large, then the current frame is selected for further computation. This step reduces the amount of computation for summarization by a large amount.

Feature Extraction

After pre-sampling, a subset of frames is obtained. In our approach, we adopt the Pearson correlation co-efficient (PCC) [6] for measuring the similarity between the two successive frames. Each frame of $m * n$ size is divided into $8 * 8$ blocks. For each block, the DCT value is computed and the first coefficient of each DCT is stored in an array 'hashArray'. Thus, a frame of size $m * n$ is converted to a feature vector of size $(m * n)/64$ .These feature vectors of frames are then used in measuring correlation between successive frames.

Let 'D_{i-1}'and 'D_i' represent the DCT feature vectors of frame 'F_{i-1}' and 'F_i' , respectively, of a video V.

$$PCC(D^i, D^{i+1}) = \frac{\sum_i \left(D_j^{i-1} - \overline{D^{i-1}} \right) (D_j^i - \overline{D^i})}{\sqrt{\sum_{i-1} \left(D_j^{i-1} - \overline{D^{i-1}} \right) * \sum_i \left(D_j^i - \overline{D^i} \right)^2}} \quad (2)$$

Where, D_j^i represents the value of DCT value of frame 'F_i' at block 'j'. '$\overline{D^i}$', is mean of the DCT vector of Frame 'F_i'. The value for this measure varies between -1 and 1. For two similar frames, the value is 1 and for two completely different frames, value is -1.

For detecting groups of frames in which the consecutive frames do not have significant dissimilarity, pair-wise correlation of consecutive frames is computed. Figure 1 shows the example of distribution of correlation values of a video sample taken from open Video Project [6].

Content Selection

There are instants in which the correlation value is very low compared to its neighboring correlation values. These instants indicate larger difference between two successive frames. These instants are names as 'peaks'.

An instant 't' is considered as 'peak' if the correlation value at that instant is smaller than a threshold 'τ'.

The threshold value is calculated as:

$$\tau = \frac{\overline{PCC} + \min(PCC)}{2} \quad (3)$$

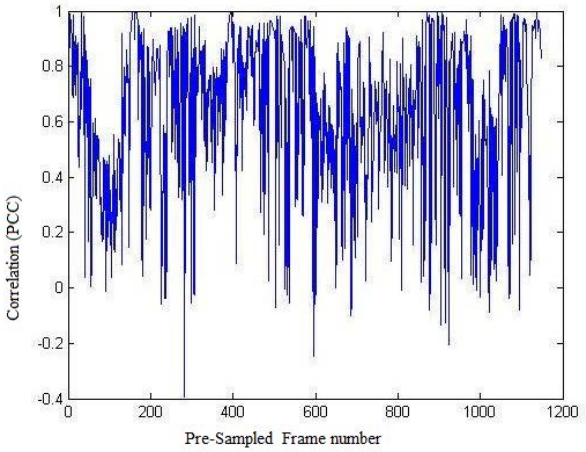

Figure 2: Pair-wise Correlation of between frames of a Video from open video Project [6]

If the duration of a group of similar frame, is below a limit 'α', then that group of frames is discarded. Let 'Gavg' be the average size of all the cluster of frames so obtained.

$$\alpha = \frac{1}{2} * (Gavg) \quad (4)$$

Otherwise, one frame from the middle of the group of frames is selected as a keyframe.

Removal of Meaningless Frames

The above algorithm may take meaningless frames, such as, completely black frames, as a keyframes. This may happen due to fade-in or fade-out or due to use of flashes. In this step, these meaningless frames are removed. This takes very negligible amount of time. This can be done by calculating the standard deviation of each frame. If the standard deviation comes out to be zero, then it is discarded.

Experiments and Results

Evaluation of video summarization is subjective in nature due to the lack of an objective ground-truth. In this work, a subjective evaluation method to compare our results with the user summary, known as Comparison of User Summaries (CUS) [7] is used. In such method the user manually generate summary after watching the full video and this user summary is used as ground-truth.

For comparison of frames from different summaries, the correlation method as described in section (3) is used. The standard measures, precision and recall is used to evaluate the quality of our summary. Precision is the ratio of number of frames matched to the total number of frames in summary, obtained by the proposed method and recall is the ration of number of frames matched to the total number of frames in user. To assess the quality, F-measure is used F-measure is the harmonic mean of precision and recall:

$$F - measure = \frac{2*precision*recall}{Precision+Recall} \quad (5)$$

Figure 3 shows the comparison of user summary with the summary generated with our proposed method. The number of frames in the user summary is 7 and the number of frames in the summary generated by our method is 6 and the number of matched frames is 6.
DCT-based approach is evaluated on the set of 30 videos downloaded from YouTube and from the Open Video project [6]. All the videos are in MP4 format (30 fps, 240*320 pixels). These videos are distributed among various genres (sports, education, documentary).

Conclusion

In this paper, we presented DCT-based static video summarization. This approach is based on partitioning the video into frames, pre-sampling the original frames, feature extraction from pre-sampled frames, similar frames are grouped. Then the middle frame each group of similar frames is selected as a key-frame.

Future work includes elimination of redundant frames in generated summary. Redundant frames may occur because of the repetition of shots or may be due to lightening effects.
Also, this work may be augmented to generate dynamic video skim, since static video summarization is a prerequisite for dynamic video skim.

S. No.	Approach	Mean F-measure
1	OV [8]	0.67
2	DT [9]	0.61
3	STIMO [10]	0.65
4	VSUMM [2]	0.72
5	VISON [4]	0.75
6	DCT-based Video Summary	0.79

Table 1: Mean F-measure achieved by different video summarization approaches.

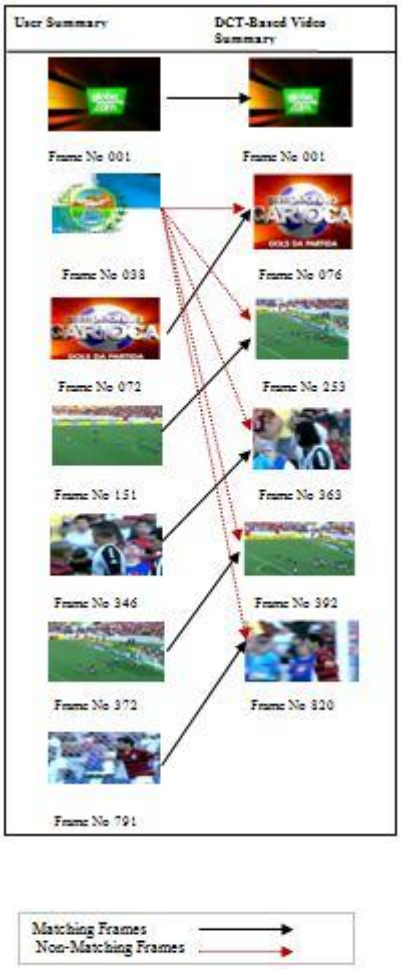

Figure 3: Evaluation of summary generated by DCT-based approach

References

[1] Muhammad Ajmal, Muhammad Husnain Ashraf, Muhammad Shakir, Yasir Abbas, and Faiz Ali Shah "Video summarization: Techniques and classification,'Computer Vision and Graphics. Springer Berlin Heidelberg', pp. 1-13, (2012). [Online] Available at http://link.springer.com/chapter/10.1007%2F978-3-642-33564-8_1

[2] Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes , Antonio da Luz Jr., Arnaldo de Albuquerque Araújo, 'Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method', Pattern Recognition Letters 32(1), 56–68 (2011). [Online] Available at http://laplace.dcc.ufmg.br/npdi/uploads/96a40be5-db2d-f171.pdf

[3] Genliang Guan; Zhiyong Wang; Shiyang Lu; Deng, J.D.; Feng, D.D., 'Keypoint-Based Keyframe Selection' ,Circuits and Systems for Video Technology, IEEE Transactions on , vol.23, no.4, pp.729,734, (2013). [Online] Available at http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6279461

[4] Jurandy Almeida, Neucimar J. Leite, Ricardo da S. Torres, VISON: VIdeo Summarization for ONline applications, Pattern Recognition Letters, Volume 33, Issue 4, Pages 397-409 (2012). [Online] Available at http://dl.acm.org/citation.cfm?id=2109377

[5] Naveed Ejaz, Tayyab Bin Tariq, Sung Wook Baik, 'Adaptive key frame extraction for video summarization using an aggregation mechanism', Journal of Visual Communication and Image Representation, Volume 23, Issue 7, Pages 1031-1040,(2012). [Online] Available at http://r32.ir/wp-content/uploads/2013/10/Adaptive-key-frame-extraction-for-video-summarization.pdf

[6] Open video project. [Online] Available at http://www.open-video.org/

[7] Guironnet, M., Pellerin, D., Guyader, N., Ladret, P, Video summarization based on camera motion and a subjective evaluation method. EURASIP J. Image Video Process.. Article ID 60245, 12 p (2007). [Online] Available at https://hal.archives-ouvertes.fr/hal-00164602/document

[8] D. DeMenthon, V. Kobla, and D. Doermann, 'Video summarization by curve simplification', In Proceedings of the sixth ACM international conference on Multimedia, pages 211–218. ACM,(1998). [Online] Available at http://r32.ir/wp-content/uploads/2013/10/Adaptive-key-frame-extraction-for-video-summarization.pdf

[9] P. Mundur, Y. Rao, and Y. Yesha. 'Keyframe-based video summarization using Delaunay clustering', International Journal on Digital Libraries, 6(2), pp 219–232(2006). [Online] Available at http://www.csee.umbc.edu/~yongrao1/IJDL2005.pdf

[10] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, 'Abstracting digital movies automatically', Journal of Visual Communication and Image Representation, 7(4):345–353,(1996). [Online] Available at https://ub-madoc.bib.uni-mannheim.de/791/1/TR-96-005.pdf

[11] Chheng Tommy, 'Video Summarization using clustering', Department of Computer Science University of Calfornia, Irvine (2007). [Online] Availableathttp://www.ics.uci.edu/~dramanan/teaching/ics273awinter08 /projects/tchheng_tommy_chheng.pdf