EasyChair Preprint

№ 2753

A Stochastic Framework for Keyframe Extraction

Thangaswamy Judi Vennila and Vanniappan Balamurugan

February 22, 2020

A Stochastic Framework for Keyframe Extraction

Thangaswamy Judi Vennila
Department of Computer Science & Engineering
Manonmaniam Sundaranar University
Tirunelveli, Tamilnadu, India, 627 012
vennila@res.tech.edu

Vanniappan Balamurugan
Department of Computer Science & Engineering
Manonmaniam Sundaranar University
Tirunelveli, Tamilnadu, India, 627 012
bala_vm@msuniv.ac.in

Abstract- The sudden growth in the Closed Circuit TeleVision (CCTV) installations has paved the way for intensive video analytics. Video summarization, being a method of representing keyframes of a voluminous video, plays a major role in the video processing. Several researchers have focused on the key frame extraction since late 90's. However, several challenges still exist in keyframe extraction. The main challenge in keyframe extraction is to identify the representative frames based on the contents. Most of the existing methods adapt deterministic approaches, which involves more computational complexity and result in poor accuracy. This work aims to improve the accuracy rate by introducing a stochastic framework that uses the techniques such as binning, Markov chain, Transition Probability Matrix (TPM), and Permutation computation. The experimental results demonstrate that the proposed framework outperforms the existing methods VSUMM and VSUKFE in terms of both computational efficiency and accuracy.

Keywords – Keyframe Extraction, Video Summarization, Stochastic Framework, Markov Chain, Transition Probability Matrix, and Permutation computation

I. INTRODUCTION

The recent advancements in the video technology have resulted in abundant availability of digital visual data such as web videos, surveillance videos, social media contents, etc. These videos are not adequately analyzed on time because of the high computational complexity and time consumption. In general, the process of video summarization comprises of several activities viz. transition analysis, shot boundary detection [1], [2], [3], keyframe extraction, object detection [4], [5], object recognition [2], [4], object tracking [15], frame clustering [5], [6], etc.

The representative frames in the video summarization process are known as keyframes that are the frames of our interest. To extract the keyframes several researchers have applied low level as well as high level features of the frame. The low level features are those features of a frame which are inherent to the image or frame. The high level features include those features which are identified using the shape, colour, etc. In several cases, the features are manipulated using additional processes such as histogram representation, optical flow detection, gradient evaluation, etc. These features are also known as hand-crafted features [7].

Several research works have been carried out in the field of keyframe extraction and video summarization since 1990 and these works can be categorised based on the feature extraction techniques. Some of the features that have been widely used in the literatures are: histogram, textures, objects, and motion characteristics [8], [9]. Also, the existing works can be categorised based on the keyframe extraction methodologies such as thresholding, clustering [10], [11], interestingness measures, relevance feedback [12], etc. Though the research work has advanced in several directions, still there are challenges that need attention by the researchers in the field of keyframe extraction. Some of them are:

- Redundancy in the video frames leads to increase in computational complexity.
- As most of the methods apply thresholding techniques for keyframe extraction, estimating the appropriate threshold is difficult.
- Selecting the appropriate features is a challenging task as they may not reflect the reality.

To overcome the above challenges, a novel stochastic framework has been proposed in this paper to enhance the accuracy and to reduce the time complexity. The main advantages of the stochastic framework are:

- Use of binning techniques leads to reduction in time complexity.
- Markov chain, application of TPM, and Permutation computation improves the accuracy.

The experiments are carried out on the benchmark datasets viz. VSUMM, UCLA, WEB DATASET, VIRAT, and TVSum. The result shows that accuracy of the framework is better while comparing to the existing algorithms viz. VSUMM and VSUKFE.

The rest of this paper is structured as follows: Section II, deals with the relevant research works that have been carried out in the past on the keyframe extraction. The proposed stochastic framework for keyframe extraction is described in the section III. Section IV, explains about the experimental setup, results, and comparative analysis. Finally, the conclusion and future work are provided in section V.

II. RELATED WORKS

This section, briefly reviews the related works on video keyframe extraction that have been carried out in the past three decades. The works are categorized based on the features that have been used and also on the object's motion with in the frame.

To extract the keyframes from the given set of video frames, feature extraction techniques play a vital role as it reduces the execution time by avoiding irrelevant comparisons among the frames. Though it is easy to analyze frames based on the low level features it is not widely applied as it produces the false alarm. The work during the initial period was using the low level features. In 1996, Diklic et al. [12] has introduced a keyframe

extraction technique based on color space histogram features. During the first phase average color space value of each frame was estimated and it was compared with the adjacent frames. The frames with high differences of these values are considered as representative frames. Further to reduce the excessive representative frames, the variance and kurtosis of the resultant adjacent frames were evaluated. Though the method performed well, it was unable to identify the adjacent frames with minor variation. Lijie et al. [13] has introduced a method that extracts the region wise histogram by dividing the frame into 16×8 blocks. Thus the method identifies the representative frames with better accuracy even though the histogram similarity of the two successive frames is same. Though hue and saturation are useful in extracting the keyframes the hue plays a dominant role [14] in the similarity comparison. Further the role of hue component alone will lead to the reduction of the computational time.

Sandrz et al. [10] has presented a technique that extracts the keyframes using hue based color histogram. Their experimental results have proved that the computational time is reduced. However, the performance in terms of accuracy is poor as the false positives are more. As the inclusion of additional features will be helpful in improving the accuracy Naveed et al. [15] presented a keyframe extraction algorithm that combines several visual features viz. RGB color channels, color histograms, and moment of inertia. This resulted in an increase in the accuracy, with a compromise on computational time. Further, the algorithm works well for the video frames with gradual changes in the lighting conditions. Along the same line, Ying et al. [16] extracted the keyframes using multiple features such as Region of Interest (ROI), local binary pattern, histogram of gradients and color moments. Though it resulted in better performance in terms of accuracy and computational time the false positives are still present. Apart from the features discussed above several other transformation coefficients have also been used for the keyframe extraction in the recent past.

Keyframe have also been extracted using object recognition, representation and their motion estimation. The spatial coordinates [17], [2] of the target object's trajectory [8] can be considered as a feature and the keyframes can be extracted based on this. The trajectory coordinates can be obtained using object tracking and the target motion can be represented using optical flow [8], [3], dense trajectory [8], and multi objects tracking [2].

From the literature, it is inferred that the selection of feature points is important in deciding the performance of the techniques and there is a need to balance the computation time and accuracy. This paper proposes a stochastic framework that fulfils the above objective.

III. The Proposed Method

The objective of the Stochastic Framework for Keyframe Extraction (SFKE) is to reduce the computational time and to increase the accuracy of keyframe extraction. This can be achieved by the use of a subset binning method which groups the pixels into predefined range of clusters. This process can be performed, as a pre processing step reduces actual computational time to a larger extent. The flow diagram of SFKE is illustrated in Fig 1. The SFKE framework receives the sequence of video frames as input and the frames are subjected to the operations such as binning, transition probability evaluation, and similarity analysis. These processes are explained in Algo 1. The problem definition and the functionality of SFKE are described in the following sections.

Problem Formulation

In the keyframe extraction, the input video can be represented by a sequence of N frames. Therefore, the video can be expressed as a frame pool $F=[f_1,f_2,.....,f_N]\in \mathrm{R}^{D\times N}$, where f_1, f_2,.....,f_N are the individual frames and D is the number of dimensions of the frame. The main objective of the keyframe extraction is to find a set of representative frames from the given input video sequences and these frames can be represented as $F_k=[f_{k1},f_{k2},.....,f_{kM}]\in \mathrm{R}^{D\times M}$ ($\mathrm{F_k}\subset \mathrm{F}$), where k_1, k_2....k_M are the representative frames.

A. Binning

Binning is a process of grouping a number of contiguous pixel values into a smaller number of bins. Equal width binning [18] is applied in the SFKE to group the pixels of the input frame. If the bin length is α then the total bins T will be 255/ α. As the accuracy of the keyframe extraction directly depends on the bin size, the value α must be chosen carefully. The binning results in two dimensional vectors with $T\times T$ values or states.

Algorithm 1 SFKE Algorithm

Input: video frames, bin length α

Output: Representative frames

Step 1: Apply binning on the frame

Step 2: Apply Markov chain

Step 3: Compute transition probability

Step 4: Apply permutation computation

Step 5: Similarity analysis

Step 6: Keyframe extraction

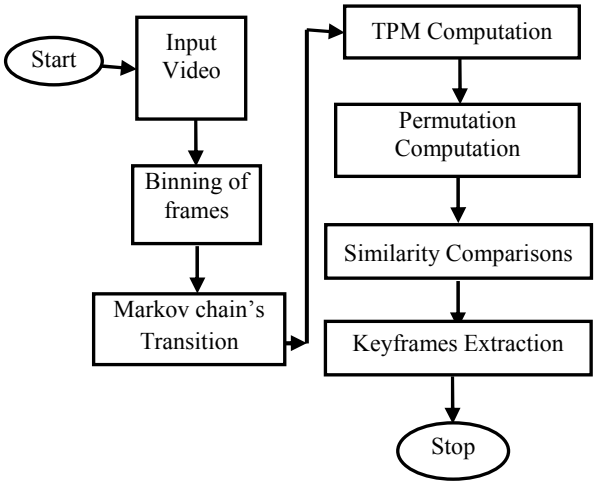

Fig. 1. *SFKE Flow Diagram*

B. Markov chain's Transition Matrix

A Markov chain is a process that moves from one state to another state depending on the previous state. The simplest Markov process is a first-order process, where the choice of the state is made only on the basis of immediate

previous state [19], [20]. As the state transition is represented as the Markov chain's transition matrix, in SFKE it is represented as matrix S of $T \times T$ dimensions. The entry *(i, j)* is the number of transitions state that takes from state i to state j. The resultant Markov chain's transition state matrix is represented in Eq. 1.

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1j} \\ s_{21} & s_{22} & s_{23} & \cdots & s_{2j} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{i1} & s_{i2} & s_{i3} & \cdots & s_{ij} \end{bmatrix}, \quad (1)$$

where $i = j = T$.

C. TPM Computations

A transition probability matrix P is defined to be a stochastic matrix with non-negative elements, in which sum of its column or row is equal to 1.

The Matrix P is derived from S and the elements are computed using the Eq. 2.

$$P_{kl} = \frac{s_{ij}}{s_{i1} + s_{i2} + s_{i3} + \cdots + s_{ij}} \quad (2)$$

As the probability distribution of states transition is represented as TPM matrix, P the entry of *(k, l)* is the probability of transition state from k to l and P has the property as mentioned in the Eq. 3.

$$p_{k1} + p_{k2} + p_{k3} + \cdots + p_{1l} = 1 \quad (3)$$

The resultant Markov chains matrix is represented as follows in the Eq. 4.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & \cdots & p_{1l} \\ p_{21} & p_{22} & \cdots & \cdots & p_{2l} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & \cdots & p_{kl} \end{bmatrix}. \quad (4)$$

D. Permutation Computation

A permutation matrix Q is a square matrix with binary values and the entries in each column and row will have exactly one entry of '1'. The matrix Q can be derived from the matrix P by deriving a vector of column with size T elements, sorting the resultant sum vector and computing the positional index of the sorted vector. To derive the matrix Q initially a null matrix of r X s where $r = s = T$ is created and the entry *(r, s)* is set to '1' if s = position index of the sorted vector and r = column index of the sum vector.

The resultant Permutation computation matrix is represented as shown in the Eq. 5.

$$Q = \begin{bmatrix} q_{11} & q_{12} & q_{13} & \cdots & q_{1x} & \cdots & q_{1T} \\ q_{21} & q_{22} & q_{23} & \cdots & q_{2x} & \cdots & q_{2T} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{y1} & q_{y2} & q_{y3} & \cdots & q_{yx} & \cdots & qyT \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q_{T1} & q_{T2} & q_{T3} & \cdots & q_{T1} & \cdots & q_{TT} \end{bmatrix} \quad (5)$$

E. Keyframe Extraction

Keyframes are extracted by computing the similarity distance between the respective permutation matrixes of successive frames. In SFKE, Euclidean distance measure as mentioned in Eq. 6 is applied for computing the similarity.

$$\text{Distance} = \sqrt{\sum_{x=1}^{T} \sum_{y=1}^{T} (q1_{xy} - q2_{xy})^2}, \quad (6)$$

where *q1* and *q2* are elements of the respective consecutive permutation matrices *Q1, Q2* and *x, y* are the corresponding coordinates. Finally, thresholding is applied on the similarity distance to extract the keyframes.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

The experiment on SFKE is carried out using Matlab version 13. Five Benchmark datasets viz. VSUMM [21], [22], UCLA [23], WEB dataset [24] [15], VIRAT [25], and TvSum [21], [26] are used for experimentation and their descriptions are illustrated in the Table 1.

TABLE 1
DATASET DESCRIPTION

Dataset Name	Video Nature	Total videos	Frame Rate	Video length
VSUMM	Surveillance	15	29	1 to 5 secs
UCLA	Surveillance	13	28	1 to 10 secs
WEB	Youtube	16	28 - 30	1 to 20 secs
VIRAT	Surveillance	35	29-30	1 to 9 secs
TvSum	Commercial	50	28-30	1 to 15 secs

To evaluate the performance of the SFKE, the accuracy rate is estimated using the eqn. 7 and 8.

$$CUS_A = \frac{n_{mAS}}{n_{us}} \quad (7)$$

$$CUS_E = \frac{n_{m'AS}}{n_{US}}, \quad (8)$$

where *CUS* is the Comparisons of User Summary, CUS_A is the Accuracy Comparisons of User Summary, CUS_E is the Error rate Comparisons of User Summary, n_{mAS} is the number of matching key-frames from automatic summary, n_{us} is the number of key frames from user summary, and $n_{m'AS}$ is the number of non-matching keyframes from automatic summary [4].

TABLE 2
EXPERIMENTAL RESULTS

Dataset wise Performance				
Dataset	**Accuracy**	**Precision**	**Recall**	**MSE**
VSUMM	97.50	96.43	87.10	0.025
UCLA	97.00	89.29	89.29	0.030
WEB	94.50	90.41	94.29	0.055
VIRAT	97.00	62.50	100.00	0.030
TvSum	94.00	62.07	94.74	0.060

The experimental results of the keyframe extraction on the five datasets are furnished in the Table 2. The accuracy is measured on extracted keyframes against the ground truth keyframes. The performance measures such as

precision [27], recall [27] and Mean Square Error (MSE) [27] are used to measure the accuracy.

TABLE 3
FRAME SIZE VARIATIONS

Accuracy wise Analysis				
Name of the Datasets	**50 Frames**	**100 Frames**	**150 Frames**	**200 Frames**
VSUMM	98.04	98.04	98.00	96.02
WEB	78.04	96.08	96.20	86.00
VIRAT	96.24	99.02	99.90	99.90
UCLA	98.04	96.08	99.90	94.00
TvSum	72.55	96.08	80.39	98.00

The experimental results for varying length of video inputs are listed in Table 3, where VSUMM, VIRAT, and UCLA are the surveillance datasets and WEB and TvSum are the WEB datasets. The result shows that the performance in terms of accuracy is consistent for the surveillance datasets as the background of these video is static when compared to the web datasets. Therefore the framework is suitable for the surveillance video datasets.

B. Result Analysis

Though the SFKE performs well for all the five datasets the accuracy rates of the framework in case of the VSUMM, VIRAT and UCLA datasets are better as they belong to the category of surveillance videos. In general, unlike traditional videos, the surveillance video differs in the sense that the variations in the background are very minimum. Since the SFKE applies the Markov chain's process which considers the previous state where the framework performs better in case of surveillance videos.

C. Comparatives Analysis

The performance of the SFKE is computed with VSUMM [4], VSUKFE [15] and the results are shown in Table 4. The experiment on six videos of WEB datasets reveals that SFKE outperforms the other benchmark algorithms.

TABLE 4
COMPARATIVE ANALYSIS

Input Video	**VSUMM [4]**		**VSUKFE [15]**		**SFKE**	
	CUS_A	**CUS_E**	**CUS_A**	**CUS_E**	**CUS_A**	**CUS_E**
Cartoon	0.87	0.22	0.83	0.2	0.90	0.18
Commercial	0.93	0.06	0.90	0.10	0.91	0.05
TV-Shows	0.91	0.33	0.83	0.30	0.94	0.31
Home	0.85	0.23	0.82	0.22	0.84	0.21
Sports	0.76	0.65	0.75	0.48	0.77	0.50
News	0.88	0.32	0.85	0.25	0.87	0.27

V. CONCLUSION AND FUTURE WORK

A novel SFKE framework for video keyframe extraction is introduced in this paper. As the framework applies the stochastic model the keyframe extraction is effective in terms of accuracy. Further, the experimental results show that the framework performs well in the case of surveillance videos. The framework can be extended to detect either the gradual or abrupt shot boundary. Also, it can be used for summarizing the video by integrating any of the frame clustering algorithms.

REFERENCES

[1] Amir H. Meghdadi, and Pourang Irani, "Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization," IEEE Trans. Visu. and Comp. Grap., vol. 19, pp. 2119 – 2128, Dec. 2013.

[2] Qian Xie, Oussama Remil, Yanwen Guo, Meng Wang, Mingqiang Wei, Jun Wang, "Object Detection and Tracking Under Occlusion for Object-level RGB-D Video Segmentation," IEEE Trans. on Multimedia, vol. 14, pp. 1 – 13, Aug. 2015.

[3] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection: in CVPR, pp. 1039 – 1048, 2016.

[4] Sandrz Eliza Fontes de Avita, Ana Paula Brandao Lopes, "VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method," Elsevier, Pattern Reco. Let., pp. 56 – 68, Aug. 2010.

[5] Anastasios D. Doulamis, Nikolaos D. Doulamis, Stefanos D. Kollias, "A fuzzy video content representation for video summarization and content-based retrieval," Elsevier, Signal Proc., pp.1049-1067, Nov. 1999.

[6] Milan Kumar Asha Paul, Jeyaraman Kavitha and P. Arockia Jansi Rani, Key-Frame Extraction Techniques: A Review: Recent Patent, Comp. Sci., pp. 3-16, 2018.

[7] Hung Vu, TuDinh Nguyen and Dinh Phung, Detection of Unknown Anomalies in Streaming Videos with Generative Energy-based Boltzmann Models: Pattern Reco., pp. 1-14, 2018.

[8] Yunzuo Zhang, Ran Tao, Yue Wang, "Motion – State – Adaptive Video Summarization via Spatio-Temporal Analysis," IEEE Trans. Circ. and Syst. for Video Tech., vol. 27, pp. 1340-1352, Feb. 2016.

[9] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li and Leonid Sigal, "Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization," IEEE Trans. on Affective Comp., vol. 9, pp. 1 – 16, Feb. 2016.

[10] Yue Gao · Wei-Bo Wang· Jun-Hai Yong · He Jin Gu, "Dynamic Video Summarization Using Two-Level Redundancy Detection," Springer, Multimedia Tools Appl., pp. 233 – 250, Nov.2009.

[11] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, Tat-Seng Chua, "Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification," IEEE Trans. Multimedia, vol. 14, pp. 975 – 985, 2012.

[12] D. Diklic, D. Petkovic, and R. Danielson, "Automatic Extraction of Representative Keyframes Based on Scene Content," IEEE Conf. Rec. Signals, Syst. and Comp., pp. 877 – 881, Aug. 1998.

[13] Lijie Liu, Guoliang Fan, "Combined Key -Frame Extraction and Object-Based Video Segmentation," IEEE Trans. Circ. and Syst. for Video Tech., vol. 15, pp. 869 – 884, Jul. 2005.

[14] Ahmet Ekin, A. Murat Tekalp, and Rajiv Mehrotra, "Automatic Soccer Video Analysis and Summarization," IEEE Trans. Image Proc., vol. 12, pp. 796 - 807, Jul. 2003.

[15] Naveed Ejaz a, Tayyab Bin Tariq b, Sung WookBaik, "Adaptive Key Frame Extraction for Video Summarization using an Aggregation Mechanism," Elsevier, J. Vis. Commun. Image R., pp. 1032 – 1040, Feb. 2012.

[16] Ying Zhang and Roger Zimmermann, "Efficient Summarization from Multiple Geo-referenced User-Generated Videos," IEEE Trans. Multimedia, vol. 18, pp.1 – 30, Jan. 2016.

[17] Ciocca Gianluigi Æ Schettini Raimondo, "An Innovative Algorithm for Key Frame Extraction in Video Summarization," Springer, Real-Time Image Proc., pp. 69 – 88, May 2006.

[18] R. Sasikumar, A. Sheik Abdullah, "Stock Market Forecasting Using Time Invariant Fuzzy Time Series Model," Research & Reviews: Journal of Statistics, vol. 7, pp. 104-111, 2018.

[19] Cheng Lu, Mark S. Drew & James Au, An Automatic Video Classification System Based on a Combination of HMM and Video Summarization: Int. Jour. of Smart Eng. Syst. Design, pp. 1 – 14, 2010.

[20] R. Sasikumar, A. Sheik Abdullah, "Forecasting The Stock Market Values Using Hidden Markov Model," Int. Jour. of Business Analytics and Intel., vol. 4, pp. 17-21, 2016.

[21] Cheng Huang and Hongmei Wang, "A Novel Key - frames Selection Framework or Comprehensive Video Summarization," IEEE Trans. Circ. and Syst. for Video Tech., pp. 1 – 13, 2019.

[22] VSUMM Dataset. [Online]. Available: https://sites. google. com/site/vsummsite/download

[23] UCLA Dataset. [Online]. Available. http://web.cs.ucla. edu/ ~yuanluxu/research/reid.html

[24] WEB Dataset. [Online]. Available: http:/ /homepages. inf. ed. ac. uk/rbf/CVonline/Imagedbase.htm

[25] VIRAT Dataset. [Online]. Available: https://gitlab. kitware.com/ viratdata/viratannotations

[26] TvSum Dataset. [Online]. Available: https://github.com / yalesong/tvsum

[27] Yang Xian, Xuejian Rong, Xiaodong Yang, and Yingli Tian, "Evaluation of Low-level Features for Real-World Surveillance Detection," IEEE Trans. Circ. and Syst. for Video Tech., vol. 27, pp. 1-11, Oct. 2017.

Thangaswamy Judi Vennila is a research scholar in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India. She has received her M.Tech. degree in Information Technology from Manonmaniam Sundaranar University, Tirunelveli. Her research interests include image and video processing.

Dr. Vanniappan Balamurugan is working as a professor in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India. He has published more than 30 papers in the reputed journals and conferences. His research interests include Pattern Recognition, Video Processing and Natural Language Processing.