I.J. Image, Graphics and Signal Processing, **2015, 4, 60-67**
Published Online March 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2015.04.07

Design of a Video Summarization Scheme in the Wavelet Domain Using Statistical Feature Extraction

J. Kavitha
Manonmaniam Sundaranar University, Department of CSE, Tirunelveli, 627012, India
E-mail: jc_kavitha@rediffmail.com

Dr. P. Arockia Jansi Rani
Manonmaniam Sundaranar University, Department of CSE, Tirunelveli, 627012, India
E-mail: jansi_msu@yahoo.co.in

Abstract—The marine researchers analyze the behaviors of fish in the sea by manually viewing the full video for their research activity. Searching events of interest from a video database is a time consuming and tedious process. Video summary refers to representing the whole video using few frames. The objective of this work is to design and develop a statistical video summarization to perform the automatic detection of events of interest in underwater video. In this proposed work, a video is partitioned into adjacent and non-overlapping datacubes. Then, the video frames are transformed into wavelet sub-bands and the standard deviation between two consecutive frames is computed. Pixels of interest in frames are identified using threshold values. Key frames are identified using Local Maxima and Local Minima. The proposed work effectively detects even the movement of small water bodies such as crabs which is not detected using the existing methods. Finally, this paper presents the experimental results of proposed method and existing methods in terms of metrics that measure the valid of the work.

Index Terms—Video summarization, key frame, Discrete Wavelet Transform, Histogram, Discrete Cosine Transform, Local Maxima and Local Minima.

I. Introduction

Video abstraction is a short summary of the original video. This is widely used in video cataloging, indexing and retrieving. Shot, scene, frame and video are the important terms in video processing. Video is collection or sequence of frames. Frame represents a picture image. Shot represents sequence of frames in a single camera operation. Scene is a collection of consecutive shots that have semantic similarity in object, person, space and lame. There are two types of video abstraction, video summary and video skimming. Video summary, also called a still abstract, is a set of salient images (key frames) selected or reconstructed from an original video sequence. Video skimming, also called moving abstract, is a collection of image sequences along with the corresponding audio from an original video sequence. Video skimming is also called preview of an original video, and can be classified into two sub-types: highlight and summary sequence. A highlight contains the most interesting and attractive parts of a video, while a summary sequence renders the impression of the content of an entire video [1].

Due to the huge volume of video data and immense size of video but with limited manpower, the need to develop fully automated video analysis and processing tools become vital [2]. The researchers have developed many video abstraction techniques to extract still abstract or salient key frames. Key frames represent the whole video content. Extracting key frames is based on interesting events within a given video sequence. One of the possible methods to detect key frames is based on shot boundary detection by comparing the corresponding pixels of two consecutive frames [3]. In the work of Seung et al [4], shot boundary was detected using low pass filtering in histogram space and the key frame was selected using adaptive temporal sampling. Another key frame selection method is based on perceptual features such as color based selection, motion based selection and object based selection. Object based selection method computes the difference between the number of regions in the last key frame and the current frame to predict key frame using certain threshold value [5]. The motion metric of the video can be obtained by using optical flow method [6]. Another interesting technique is key frame-based video summarization using clustering. Clustering video frames is based on Delaunay Triangulation (DT) which has been used in other domains such as data mining and is widely acknowledged to be fully automatic [7].

A highly structured commercial video has short camera shots, and well- defined scene changes that aid in the automated analysis and summary of the video [8]. Liu et al. [9] developed a triangle model of Perceived Motion Energy (PME) to model motion patterns in sports, entertainment, news, and home videos. In that model, key frames are selected from the sub-segments. In [10], Mr.

Sandip et al. developed a Block based $\chi 2$ Histogram algorithm for shot boundary detection. This Shot boundary detection algorithm and key frame extraction algorithm using image segmentation for video summarization.

In Discrete Wavelet Transform (DWT) based video summarization technique, two consecutive frames are transformed using DWT and then the differences of the detail components are estimated. If the difference between a consecutive pair is greater than the threshold, then the last frame of the pair is considered as a key frame [11].

In this paper, key frame is detected using modified DWT in underwater video. The rest of the paper is organized as Follows. Section II describes the concepts behind the existing video summarization techniques and section III discusses the design of the proposed video summarization technique. In section IV, experimental results and analysis are presented. Finally, concluding remarks are furnished in section V.

II. VIDEO SUMMARIZATION TECHNIQUES

Histogram based image processing and transform techniques are used by many researchers in the process of video summarization

A. *Histogram Based Video Summarization*

The histogram of an image represents the relative frequency of occurrence of the various gray levels in the image [13]. An image histogram is type of histogram which acts as a graphical representation of the tonal distribution in a digital image [14]. It plots the number of pixels for each tonal value. The existing Histogram based Video Summarization (H-VS) splits video into frames. Each frame is converted into gray image. Then Histogram is generated for each frame. The difference between two adjacent frames based on pixel distribution is computed. Pixels of interest (Ƥ) are calculated using the threshold filter value. Key frames with maximum Ƥ values are identified.

B. *Transform Domain Techniques*

1) *DCT based video summarization*

The Discrete Cosine Transform (DCT) helps to separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image's visual quality). It transforms a signal or image from the spatial domain to the frequency domain [15].

This technique is also used in video summarization. First, the input video is split into frames. Each frame is converted into gray color image. Then each frame is transformed using Discrete Cosine Transform. The difference between two adjacent frames is computed. Finally, pixels of interest (Ƥ) and key frames are computed.

2) *DWT based video summarization*

Discrete Wavelet Transform (DWT) decomposes each video frame into four sub band images with different properties as shown in fig. 1. Among them LL corresponds to a smooth version of the original image. HL, LH and HH are called detail coefficients [11]. Any sudden change in the original image will affect these three sub bands.

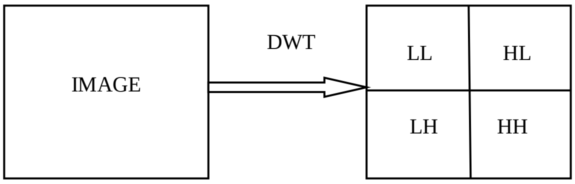

Fig. 1. Discrete Wavelet Transform

This work is implemented in four steps. In the first step, two successive frames are read and transformed using DWT. The HL, LH and HH sub-bands are used to detect key frame. For each sub-band, difference between the current frame and the next frame is calculated. In the second step, Mean and Standard Deviation are computed from the difference vectors. Then the threshold value for each sub-band is calculated by adding the Mean and Standard Deviation. In the final step, the threshold and difference values of each band are compared. If the difference value exceeds the threshold then the second frame is considered as a key frame.

III. PROPOSED TRANSFORM BASED VIDEO SUMMARIZATION

This work is the extension of DWT based summarization [10]. First, the input video is split into adjacent datacubes. Then DWT is applied to each datacube and statistical features are extracted. This result is used to select pixels of interest in each frame in the datacube. Key frames are identified by Local Maxima and Local Minima. The proposed work outperforms the existing DWT method in terms of identifying all events of interest in the input videos. Fig. 2 depicts the overall diagram of the proposed transform based video summarization.

A. *Design Procedure*

Let the input video be partitioned into adjacent and non-overlapping datacubes. Each datacube contains ten numbers of frames. Then, let the video frames be transformed into wavelet sub-bands namely ɓ$_{A,}$ ɓ$_{H,}$ ɓ$_V$ and ɓ$_D$ of size r×c. The detail sub-bands ɓ$_{H,}$ ɓ$_V$ and ɓ$_D$ are used in the video abstraction process. Compute the standard deviation (σ_m) between the two consecutive frames f_k and f_{k+1} in each detail sub-band using the formula given below.

$$\sigma_m = \sqrt{\frac{\sum_{n=1}^{N-1}(d_m(n)-M_m)}{N-1}} \quad (1)$$

Where,

$m = \{ \text{б}_{H,}\ \text{б}_{V,}\ \text{б}_{D} \}$

$\text{б}_{H,}$ - HL band of a frame

б_{V} - LH band of a frame

б_{D} - HH band of a frame

$$d_m = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(f_{k+1}(i,j) - f_k(i,j)\right) \tag{2}$$

$$M_m = \frac{\sum_{n=1}^{N-1}\sum_{i=1}^{r}\sum_{j=1}^{c}(d_m\ (n))}{N-1} \tag{3}$$

In the next step, compute the threshold (Ʈ) suitable to select the pixels of interest (Ƥ) in each frame,

$$\text{Ʈ}_m = M_m + \alpha\sigma_m, \tag{4}$$

Where, α is a constant

$$f(\text{Ƥ}_N) = \begin{cases} 1 & if\ (d_1(N) > \text{Ʈ}_1\ \&\ d_2(N) > \text{Ʈ}_2) \\ 1 & if\ (d_2(N) > \text{Ʈ}_2\ \&\ d_3(N) > \text{Ʈ}_3) \\ 1 & if\ (d_1(N) > \text{Ʈ}_1\ \&\ d_3(N) > \text{Ʈ}_3) \\ 0 & otherwise \end{cases} \tag{5}$$

Where, N is number of frames in video

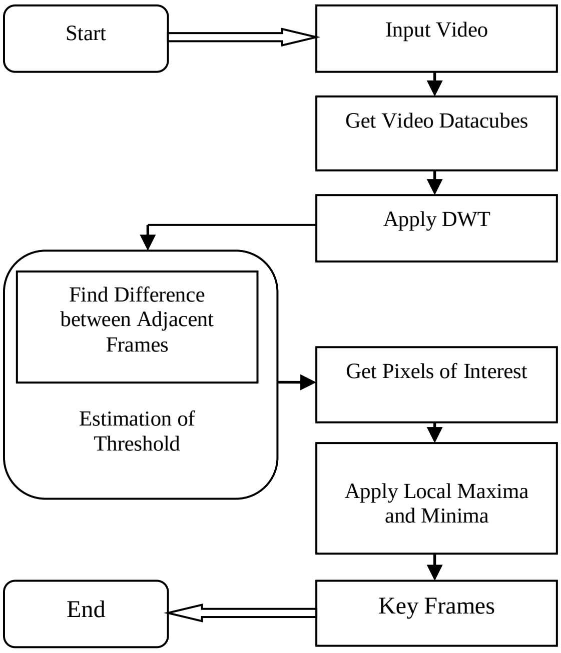

Fig. 2. Overall Diagram of Proposed Video Summarization Method

In the final step, apply local maxima and local minima (ĹĻ) to select key frames. Fig.3. demonstrates local maxima and local minima concepts clearly [12]. Local maxima and local minima may not be the minimum or maximum for the whole function, but locally it is. First, the interval values of the function are determined. Let 'f' be a function defined on an interval (a,b) and let 'p' be a point in (a,b). The height of the function at p is greater than or equal to the height anywhere else in that interval. It is called as the local maximum value of the function. The local minimum at p if f(p) is less than or equal to the values of f. The f(p) should be inside the interval, not at one end or the other.

$$f(\text{ĹĻ}) = \begin{cases} 1 & f(c) \le f(x)\ when\ x\ is\ near\ c \\ 1 & f(c) \ge f(x)\ when\ x\ is\ near\ c \\ 0 & otherwise \end{cases} \tag{6}$$

Where, c is a Ƥ in the domain [1 to 10].

The datacube contains ten numbers of adjacent frames. So's' value ranges from 1 to 10. If the s value increases then the number of key frames of the video will also decrease. So s value is proportionally inverse into the number of key frames in the video summary. The output contains key frames of a video clip with all events of interest of the input video. It also contains small movement of water animal such that crab which is not possible with the existing DWT. This modified DWT algorithm can easily and quickly detect key frames.

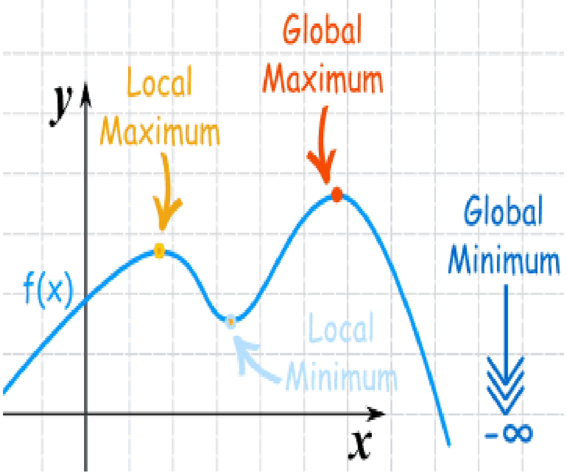

Fig. 3. Local Maxima and Local Minima

IV. Experimental Results

The implementation was done using the programming language MATLAB. The performance of the proposed work is analyzed using the metrics such as false negative, compression ratio and processing time. The main aim of this work is to detect all events of interest in the input video and to eliminate all redundant frames. To achieve this, it is desired to get a minimized false negative ratio, compression ratio and processing time. The false negative (f_n) is defined as

$$f_n = t^+ / T \tag{7}$$

Where f_n denotes the original event of interest (measured in terms of frames) that is not included in the result and T is the number of frames in the original video. Compression ratio is computed by dividing the number of key frames in the result by the number of frames in the original video.

$$C_R = T_{Keyframes}/T \tag{8}$$

Where C_R is the compression ratio, and $T_{Keyframes}$ is number of key frames in the result video and T is the number of frames in the original video.

In this work, five underwater videos of various sizes are taken as input. These are downloaded from youtube video database [16]. The performance of the proposed work is analyzed using various underwater videos and the results are tabulated in Table 1. It shows the number of frames (N) and target frames (T) in the input video file. The metrics including number of detected frames (d_F), detected target frames (d_T), false negative ratio (f_N), compression ratio (C_R) and computation time (C_T) are computed. It shall be noted that the proposed work identified desired key frames. The number of detected frames (d_F) increases with the increase in input file size.

Table 1. Performance Analysis of the Proposed Work

S. No	Input Video	N	T	d_F	d_T (%)	f_N	C_R	C_T (sec)
1.	fishnew.mp4	12	3	2	67	0.08	0.167	0.4
2.	fishmp1.avi	42	3	9	100	0	0.214	1.66
3.	fishmov12.avi	49	3	7	67	0.02	0.143	5.9
4.	dataset6,mp4	100	6	21	100	0	0.21	19.9
5.	sanimal.avi	195	7	36	100	0	0.185	46

The target frames (T) of all five videos are manually identified by the user to watch the video frame by frame. The target frames are shown in the fig.4. Three target frames are identified for fishnew.mp4, fishmp1.avi and fishmov12.avi. Six and seven target frames are identified for dataset6.mp4 and sanimal.avi.

Fig.5 shows the detected target frames (d_T) of the proposed method to all five video files. Proposed method gives two detected target frames for fishnew.mp4 file and fishmov12.avi file. It presents all desirable target frames for fishmp1.avi, dataset6.mp4 and sanimal.mp4 file. The percentage of target frames detections are 67%, 100%, 67%, 100% and 100% respectively for the video fishnew.mp4, fishmp1.avi, fishmov12.avi, dataset6.amp4 and sanimal.avi files.

The performance of the proposed work is compared with the existing methods namely DWT-VS (Discrete Wavelet Transform based Video Summarization), HIST-VS (Histogram based Video Summarization), and DCT-VS (Discrete Cosine Transform based Video Summarization). Fig.6. shows the extracted key frames of proposed method, DWT-VS, HIST-VS, DCT-VS techniques and the target frames(T) for the video file fishmp1.avi (T) with $240\times320\times3$ dimensions. From fig.6 it shall be noted that the proposed method provides all desirable target frames for the fishmp1.avi video file. DWT-VS provides only one key frame but it is also undesirable. HIST-VS and DCT-VS give one desirable target frame and two desirable target frames. So it is clearly demonstrated that, the proposed work gives better result as compared to the other existing methods.

Table.2 illustrates the performance comparison of the proposed work for a video with fast moving animals. The video file fishmp1.avi with 42 frames is considered for comparative analysis. From table.2 it shall be noted that the percentage of the detected target frames (d_T) is 100% for the proposed work. DWT-VS detects no target frames (0%). H-VS detects 33% and DCT-VS detects 67% of the target frames.

Table 2. Performance Comparison of the proposed work with existing works for a video with fast moving animals (N=42; fishmp1.avi)

S. No	Method	T	d_F	d_T(%)	f_N
1	Proposed Work	3	9	100	0
2	DWT-VS	3	1	0	0.07
3	H-VS	3	8	33	0.05
4	DCT-VS	3	8	67	0.02

Similarly the proposed work clearly identifies even the smallest water animal. The video file sanimal.avi with 195 frames contains a small frog which is correctly detected by the proposed work. Table.3 depicts the detection of smallest animal movement. Proposed method detects all desirable target frames. DWT-VS and HIST-VS detect six and five target frames. DCT-VS detects only two target frames. It shall be showed that the target frame detection percentage is 100% for the proposed work. Whereas, the percentage of target frames detections are 86%, 71% & 29% respectively for the existing DWT-VS, HIST-VS and DCT-VS techniques.

Table 3. Performance Comparison of the proposed work with existing works for a video with slow moving animals (N=195; sanimal.avi)

S. No	Method	T	d_F	d_T(%)	f_N
1	Proposed Work	7	36	100	0
2	DWT-VS	7	33	86	0.005
3	H-VS	7	39	71	0.01
4	DCT-VS	7	39	29	0.03

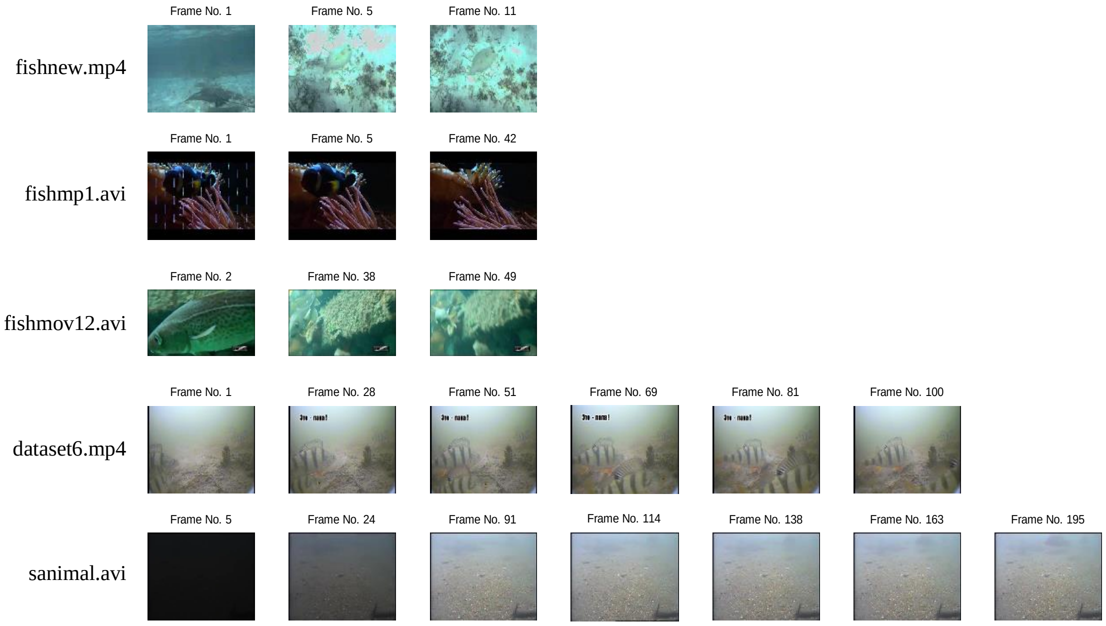

Fig. 4. Target Frames (T)

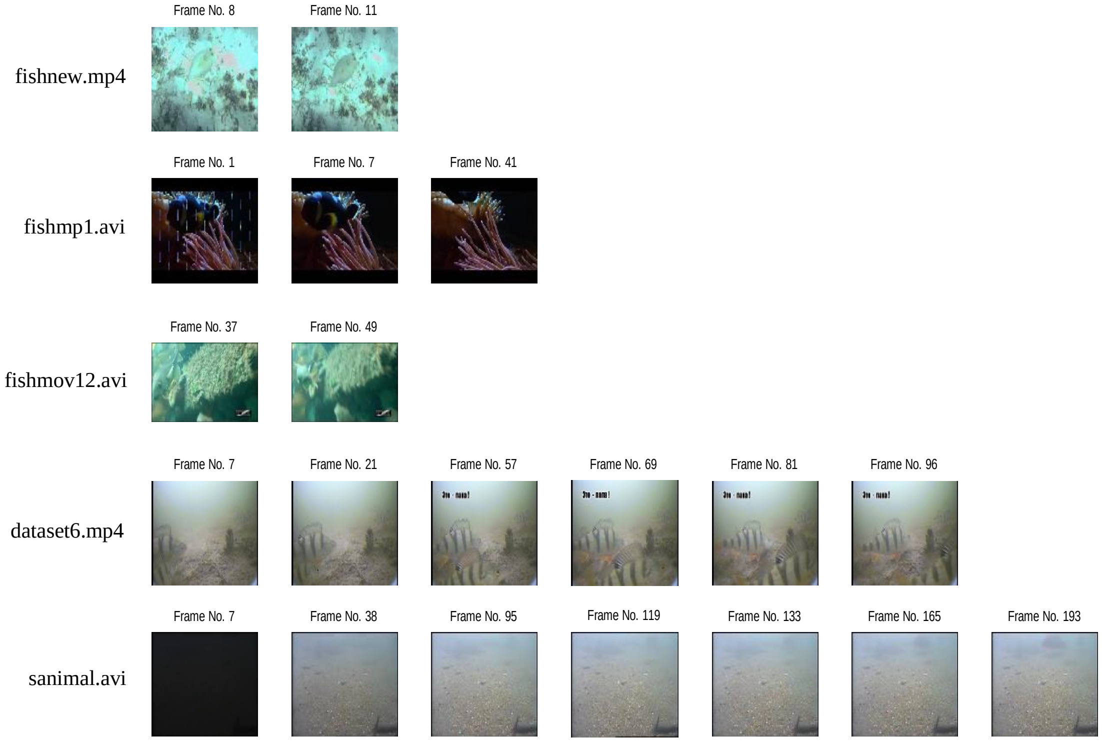

Fig. 5. Detected Target Frames (dT) of the proposed method

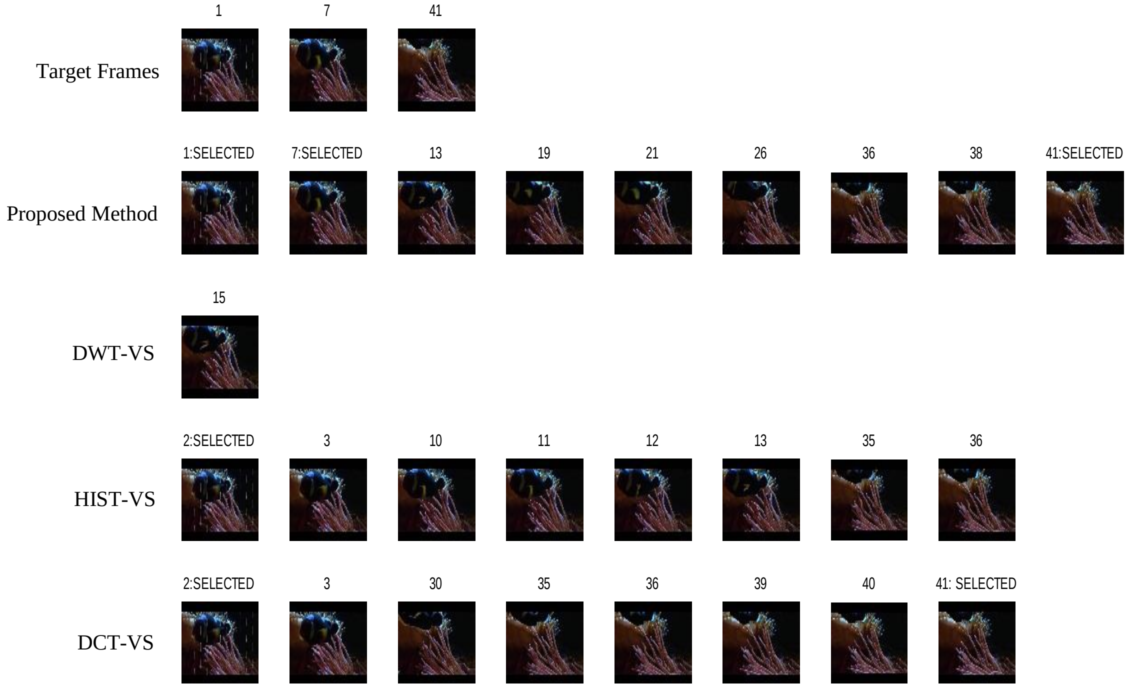

Fig. 6. Extracted Key Frames from fishmp1.avi file with fast moving animals using the proposed and existing methods

The graph in fig.7 depicts that the average false negative ratio (f_N) for the proposed work is less in comparison with that of the other existing works.

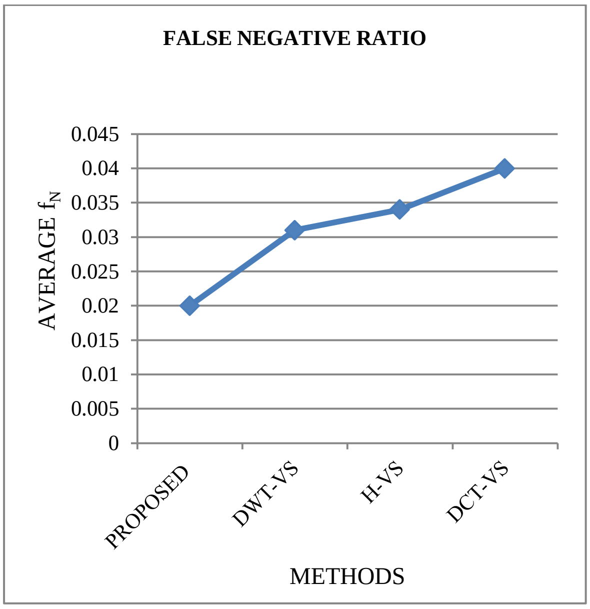

Fig. 7. False Negative Ratio

The graph in fig.8 clearly demonstrates that the proposed work consumes less computation time in comparison with that of the other existing works.

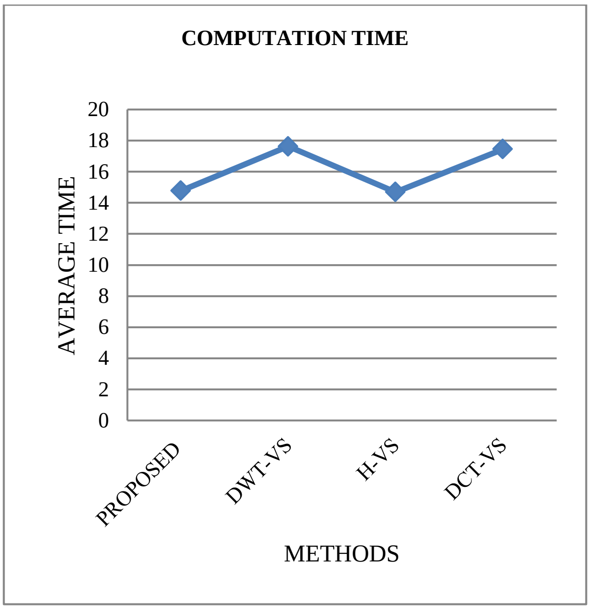

Fig. 8. Computation Time

Compression ratio represents the number of key frames in the video summary. It is directly connected with the‘s’ value in the local maxima and local minima. If the s value increases then the compression ratio of the video will also decreases. So ‘s’ value is proportionally inverse into the compression ratio. Based on the experimental results, the value of s is eight. It gives less number of key frames and the key frames should be target frames. The graph in fig.9 clearly demonstrates that the proposed method has less compression ratio in comparison with that of the other existing works.

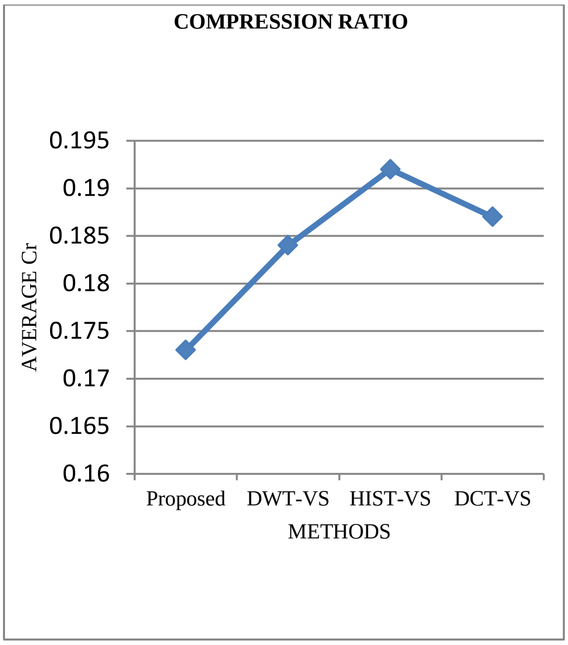

Fig. 9. Compression Ratio

V. CONCLUSION

A new threshold based statistical video summarization scheme has been designed and developed for the automatic detection of events of interest in underwater video. Due to automated process this scheme easily detects smallest water animals which are not possible with the existing schemes. Also the role of statistical analysis enables the proposed scheme to detect even fast moving animals efficiently. In future, we will do research for summary of specific events of interest such as particular fish motion activity, fish availability, etc. The decimation techniques will reduce video processing time. So, another future step would be to apply decimation techniques to reduce video processing time.

REFERENCES

[1] Video Data Management and Information Retrieval, by Sagarmay Deb (Author).

[2] Y. Li, T. Zhang, D. Tretter, "An Overview of Video Abstraction Techniques", HP Laboratories Palo Alto, Tech. Report No. HPL-2001-191, July, 2001.

[3] Ardizzone, E., & Cascia, M.," Automatic video database indexing and retrieval". Multimedia Tools and Applications, 4, 29-56, 1997.

[4] Seung Hoon Han, Kuk Jin Yoon and In So Kweon "A new technique for shot detection and key frame selection in histogram space" Workshop on Image Processing and Image Understanding, 2000, pp 305-310.

[5] Kim, C., & Hwang, J., " An integrated scheme for object-based video abstraction", Proceedings of ACM Multimedia 2001, Los Angeles, CA, 303-309.

[6] Wolf, W., "Key frame selection by motion analysis", Proceedings of IEEE International.Conference on Acoustics, Speech, and Signal Processing, 1996, Atlanta, GA, 1228-1231.

[7] Padmavathi Mundur, Yong Rao, and Yelena Yesha " Keyframe-based Video Summarization using Delaunay Clustering", Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County 1000 Hilltop Circle, 2005.

[8] T. Liu and J. Kender, "Rule-based semantic summarization of instruc- tional videos," in International Conference on Image Processing, vol. 1, 2002, pp. 601–604.

[9] T. Liu, H.-J. Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 10, pp. 1006 –1013, Oct.2003.

[10] Mr. Sandip T. Dhagdi, Dr. P.R. Deshmukh "Key frame Based Video Summarization Using Automatic Threshold & Edge Matching Rate" International Journal of Scientific and Research Publications, Volume 2, Issue 7, July 2012.

[11] Khin Thandar Tint, Dr. Kyi Soe, " Key Frame Extraction for Video Summarization Using DWT Wavelet Statistics", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, No 5, May 2013.

[12] http://www.mathsisfun.com/algebra/functions-maxima-minima.html.

[13] Tinku Acharya and Ajoy K. Ray, "Image Processing Principle and Application," John Wiley & Sons, Inc., Hoboken, New Jersey, Canada, 2005.

[14] E. Sutton. "Histograms and the Zone System". Illustrated Photography.

[15] Nageswara Rao Thota, and Srinivasa Kumar Devireddy, "Image Compression Using Discrete Cosine Transform", Georgian Electronic Scientific Journal: Computer Science and Telecommunications 2008|No.3 (17).

[16] http:/www.youtube.com.

Authors' Profiles

J. Kavitha, She is doing her Ph.D in Computer Science and Engineering in Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. She obtained her M.C.A Degree in Computer Science and Applications from Indira Gandhi National Open University, New Delhi, India in 2004 and M. E Degree in Computer Science and Engineering from J. J. College of Engineering and Technology, Anna University, Tamil Nadu, India in 2007. She has more than five years of teaching experience. Her research interest includes content based video retrieval.

Dr. P. Arockia Jansi Rani, graduated B.E in Electronics and Communication Engineering from Government College of Engineering, Tirunelveli, Tamil Nadu, India in 1996 and M.E in Computer Science and Engineering from National Engineering College, Kovilpatti, Tamil Nadu, India in 2002. She has been with the Department of Computer Science and Engineering, Manonmaniam Sundaranar University as Assistant Professor since 2003. She has more than ten years of teaching and research experience. She completed her Ph. D in Computer Science and Engineering from Manonmaniam Sundaranar University, Tamil Nadu, India in 2012. Her research interests include Digital Image Processing, Neural Networks and Data Mining.

How to cite this paper: J. Kavitha, P. Arockia Jansi Rani,"Design of a Video Summarization Scheme in the Wavelet Domain Using Statistical Feature Extraction", IJIGSP, vol.7, no.4, pp.60-67, 2015.DOI: 10.5815/ijigsp.2015.04.07