International Journal of Technology 14(2) 411-421 (2023)
Received April 2022 / Revised July 2022 / Accepted November 2022

International Journal of Technology

http://ijtech.eng.ui.ac.id

# ViSTORY: Effective Video Storyboard Generation with Visual Keyframes using Discrete Cosine Transform

Ashvini Tonge[1], Sudeep Thepade[1*]

[1]*Pimpri Chinchwad College of Engineering Pune, Maharashtra, India*

**Abstract.** Nowadays, multimedia content utility is increasing rapidly. Multimedia search engines like Google, Yahoo, Bing, etc., are available just a click away to all users. There are around 500-600 hours of video uploads per unit of time to the Internet. So, among other types of multimedia content, such as text and images, video is the most complicated content for indexing, browsing, and retrieval. Videos give more scope for implementation because of their complex and unstructured nature. This paper proposes a new method of video storyboard generation with keyframe extraction in spatial and frequency domains using Discrete Cosine Transform (DCT) for video summarization. It discusses the empirical appraisal of video visual keyframes with t-test analysis in comparison with spatial and frequency domains, resulting in a quick response to customer demands by providing static storyboards. This study proposes a new performance measure as matching frames by analyzing input videos and the standard benchmarks video dataset, i.e., Open Video Project (OVP) and SumMe. Among all the keyframe extraction techniques, DCT gives higher accuracy and a better matching rate.

*Keywords:* DCT; Key frame; Spatial; Video summarization

## 1. Introduction

Along with the huge amount of multimedia data availability and its exponential growth in recent years, the use of multimedia video content is significantly increased. This gives arise to new challenges, such as storage search, and navigation issues, among others. Video frames take more random access memory space in order to process high resolution images. Therefore, issues related to access to video information need to be addressed. Among them is video content summarization which aims to generate a few clips or sets of frames that contain the most important information about the content of a video clip. In the past few years, videos become one of the most promising and strong proofs of content because it assures the inclusion of three dimensions of the image i.e. size, width and height. Nevertheless, the heavy use of video information has been creating a major problem in searching, storing, and retrieval of this type of information. Video summarization is a promising solution for this challenge. Video summarization aims at producing comprehensive and compact summaries to enable full proof browsing experience. Oftentimes, and in certain circumstances, users need faster access and quick browsing with minimum storage from a large collection of video dara sets. Efficiently and effectively

*Corresponding author's email: sudeep.thepade@pccoepune.org, Tel.: 020– 2760 0000 ; Fax.: 020 - 2760 0003
doi: 10.14716/ijtech.v14i2.5617

identifying significant (key frames) video frames is an important problem in video retrieval. Users may want to view the brief summary or abstract of the video. This may help to see the occurrences of the events in the video, wherein the concept of key frames can be used. Among the various methods of video summary generation, one of the method is to generate video summaries/storyboards by using key frame extraction. Video structure is seen as a sequence of video frames that is executed per unit amount of time. The structural analysis of a video is exhibited in Figure 1.

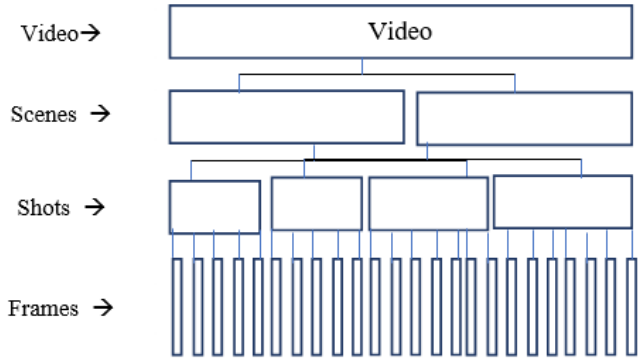



**Figure 1** Structural analysis of video data

Video is a sequence of still frames. It is an ordered set of frames. The digital input of video V with cardinality N is defined as,

$$V = \{f_1, f_2, f_3 \ldots\ldots f_N\} \tag{1}$$

Here, video data is divided into scenes, shots, and still images as frames, respectively. In these shots, there are continuous views and event-specific scenes captured by a camera. Generally, in animated videos, the story captures consumes time and costs the most (Imanullah, Yuniarno, and Sooai, 2019), hence the proposed video summarization may solve this issue. The second alternative is to find out an efficient video-searching technique for image-based video retrieval (Anayat *et al.*, 2020).

The contributions of this study are i) reviewing the existing visual key frame extraction techniques and the results; ii) empirically analyzing the video visual key frame extraction techniques in spatial and frequency domains; iii) proposing a new method of key frame extraction that is employing discrete cosine transform (DCT); iv) comparing the results using performance measures, such as accuracy in terms of matched frames and keyframes, using the Open Video Project Videos and the SumMe videos.

The rest of the paper is organized as follows. Section 2 presents the overview of relevant works and discusses the techniques of key frame extraction. Section 3 discusses the proposed framework for key frame extraction using discrete cosine transform. Section 4 presents the performance measures along with the test bed used from standard video datasets. The results are discussed with t-test analysis of the proposed technique. Section 5 concludes the discussion.

## 2. Related Work

Video summarization is a mechanism to provide a short abstraction of full video data for better visualization in terms of visual content. A handful of research discusses video summarization in a variety of approaches which focuses on either image (keyframes), static video summarization (storyboards), moving images (video skims), or dynamic video summarization.

Various studies in the literature focus on good video summary creation using shots, clusters, and machine learning methods. Truong and Venkatesh (2007) define video abstraction as a method of creating a good synthetic summary of an original video in minimum time. Cayllahua-Cahuina, Cámara-Chávez, and Menotti (2012) propose an approach of static video summarization based on shot detection using color histograms. The first RGB histogram is used to find the distribution of red, green, and blue colors for a given video frame. Then, PCA (principal component analysis) is executed on feature vectors in order to reduce the dimensions of large video frames. Furini *et al*. (2010) argue that video summarization can be done by clustering the video data and detecting the redundant sample frames of the video content.

In video summarization, image, and video descriptors are essential to classify the video input frames into significant and non-significant video frames. Cayllahua-Cahuina, Cámara-Chávez, and Menotti (2012) deploy video summarization based on image descriptors. Here, color histograms are used to measure the similarity between two video frames. The video is then divided into segments and shots for cluster formation. The closest frame to each centroid is marked as the key frame and is extracted to build the storyboard.

The transform assures the reduction in the computations in video processing, so the research work presented by Badre and Thepade (2016), uses a novel method of video content summarization using Thepade's Sorted n-array Block Truncation Coding (TSBTC). In this research, the work variations of TSBTC are done with various similarity measures, and the performance measure is defined as percentage accuracy. The study by Subba, Roy, and Pradhan (2016) performs the process of static and dynamic video summarization by elaborating key frame extraction technique. A novel technique for image retrieval is discussed using color texture features using vector quantization with Kekre's fast codebook in Kekre, Sarode, and Thepade (2009).

The video qualities are evaluated in work presented by Pan *et al*. (2019) that proposes a bottom approach in which users can customize the quality of video summaries. Clip Growing concept is employed by clustering based on the similarity of video content. Jeong, Yoo, and Cho (2017) propose a method of video content summarization using content aware clustering method with keyframe selection. The selected key frames are shown in Figure 2.

V1 video frames V2 video frames V3 video frames

**Figure 2** Sample video's keyframes after video summarization

Thepade and Tonge (2014a and 2014b) introduce the video summarization method using frequency domain and frame difference. Dhagdi and Deshmukh (2012) introduce a new approach for key frame extraction using block-based histogram difference and edge matching rate. Rao and Patnaik (2014) propose the contourlet transform to extract keyframes with improved accuracy and low error rates with selected features. Shen, Tseng, and Hsu (2014) propose a model that describes boundary frames using Petri-net networks to find the cut transition and gradual transitions in a video sequence. Maharani *et al.* (2020) proposed an average image subtraction method to detect tumor detection. This image subtraction can be used for video keyframe extraction to ensure the least significant

removal from each video frame. Pribadi and Shinoda (2022) proposed the method to identify the shielded metal arc welding with the help of support vector machine. It uses the root mean square error, correlation index were extracted to recognize the hand motions and then support vector method was used to classify it into qualified and unqualified arc welding. Similarly this approach can be used to classify the keyframes and non-keyframes. Kho, Fauzi, and Lim (2022) has given the video chunk processor for processing the large image sizes in long video sequences. This chunk processor reduces the task of storing images for video processing.

### *2.1. Existing Technique of Key Frame Extraction*

The video summarization is divided into frame sampling and key frame extraction.

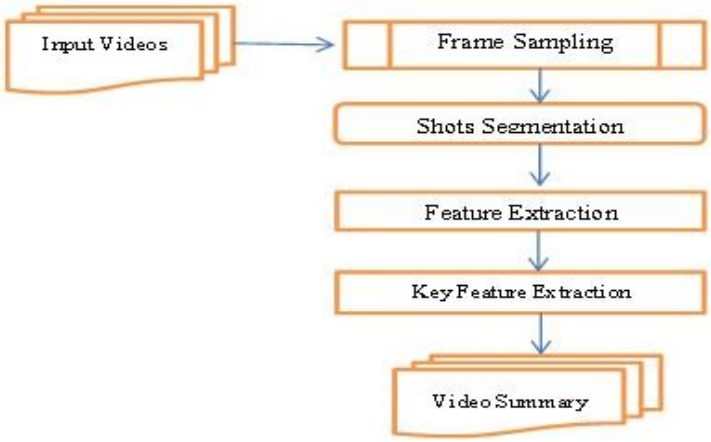



**Figure 3** Video Summarization using Keyframe Extraction

Videos consist of several frames. First, the video frames must be sampled to find the selected keyframes. This summarization process is done to split the whole video sequence into a set of meaningful frames. Key frame extraction is to choose the most informative frames from the video.

#### 2.1.1. Key frame extraction using cumulative differences

In this method, the input video is sampled into several frames, and only significant key frames are extracted. According to Liu and Fan (2005), the vital information about the scene's context is carried by keyframes only. Wu, Zhao, and Ngo (2007) claim that feature extraction is easier with keyframes when they are selected using cumulative differences.

#### 2.1.2. Video summarization using higher color moments

Video summarization can be divided into two stages, i.e., shot boundary detection and key frame extraction (Jadhav and Jadhav, 2015). These processes are carried out to find the video summaries (storyboards). Meanwhile, shot boundary detection is a process of segmenting a video into multiple shots. Here, an image histogram is used to represent a digital image. For the detection of shots, image histograms, skewness, and kurtosis metrics are used. The proposed algorithm is as follows:

Equations 2) to 6) define mean (M), standard deviation (St), skewness (Sk), and kurtosis (K) for every block (Jadhav and Jadhav, 2015)

$$M=|M(n)-Md(n+1)| \quad (2)$$

$$St=|St(n)-Sn(n+1)| \quad (3)$$

$$Sk=|Sk(n)-Sk(n+1)| \quad (4)$$

$$K=|K(n)-K(n+1)| \quad (5)$$

$$Td= M + St + Sk + K \quad (6)$$

Here, the threshold is calculated for the key frame decision.

$$\text{threshold} = \text{Td(average)} * p \quad (7)$$

Td(average) is the average of all total differences.
The threshold is based on the differences.

For each frame, the difference between two consecutive frames is calculated to find the correct shot. These differences are used further to find the distinguished frame, i.e. *key frame.* Key frame extraction is carried out using the maximum mean and standard deviation from each shot. These key frames are used for finding the static video summaries/storyboards.

2.1.3. Video Summarization using Rank-Based Approach

The key frame selection is performed in three stages. The score is calculated in the first stage. Then, keyframes are selected in the second stage based on the combined scores. Finally, near duplicates are eliminated (Srinivas, Pai, and Pai, 2016). Here, quality, attention, contrasts, representativeness, and uniformity are used for the keyframe selection. The whole process of keyframe selection is performed as shown in Figure 4. The scores are used to find the rank differences. The thresholding technique is used to decide the keyframe.

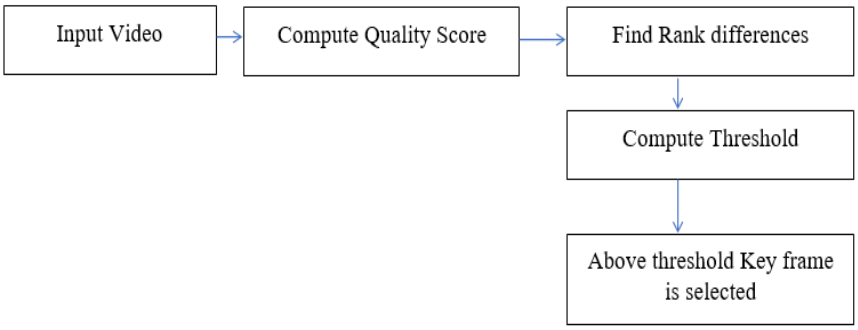



**Figure 4** Flow chart for rank-based video summarization

## 3. Proposed Method of Key Frame Extraction

In this proposed method of keyframe extraction, a discrete cosine transform is used to retrieve more keyframes for storyboard generation. Here, features are extracted using discrete cosine transform coefficients, and keyframes are selected by comparing sequential frame differences for video summarization. The results are summarized using Open Video Project videos. The details are given with video length, frame details, etc., with accuracy as the performance measure. The algorithm flow is shown in Figure 5 below.

Consider a video V, with 'n' number of frames, and it extracts a set of key frames K for input video V.

1. For each frame 1 to n, read a video for each frame.
2. Apply DCT for feature extraction.
3. Then, consecutive frame differences are calculated.
4. Calculate the mean and standard deviation
5. Calculate *threshold* = *p* * *std*
6. For all frames, *if* (*diff* (*n*) >*threshold* ), Output $n^{th}$ frame to a set of keyframe K.
7. Create a new video with selected key frames K.

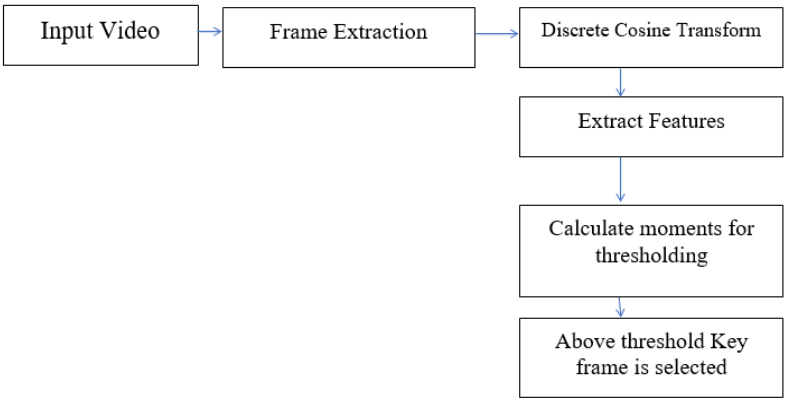



**Figure 5** Proposed method of key frame extraction using discrete cosine transform

The DCT works in different parts of an image using different DC coefficients. During quantization, less significant information is eliminated. The video frames are divided into blocks of 8 by 8, the DCT coefficients are quantized. Then, inverse DCT is applied for obtaining an original image.

## 4. Implementation and Performance Measures

The proposed system is implemented on a basic computer system of Intel core 2duo with 4GB RAM. It is implemented using MATLAB 2015a. A few of the selected video samples are shown in Figure 6. The performance evaluation of this proposed system is measured using completeness, i.e. the number of extracted frames that matches the total number of frames in the video.

### *4.1. Performance Evaluation*

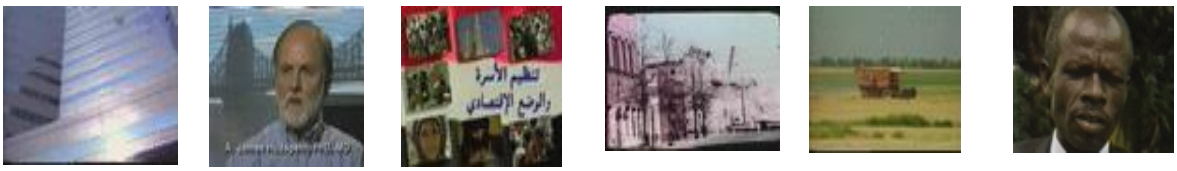
To evaluate the performance of the proposed system, the percentage of accuracy is used. It is calculated as per the following equation.

$$Accuracy = \frac{ActualExtractedKeyFrames}{ExpectedNumberofKeyFrames} \quad (8)$$

Generally, precision and recall are used to find the relations between retrieved and relative frames. Rather than precision and recall, accuracy is the most prominent performance evaluator for key frame extraction. The related equation for accuracy is given in equation 8).

### *4.2. Test Bed*

To evaluate the performance of the above algorithms Open Video Project dataset and SumMe datasets are used, which include more than 1000+ videos with a variety of scenes. Open video dataset contains a variety of categories like documentaries, historical, television, lecture notes, etc. SumMe dataset contains videos of air forces, cooking shows, etc. A few of the video sample frames from the above dataset are given below in Figure 6.



a) Set of video frames from the test bed Open Video Project (Open-video.org).

b) Set of video frames from the SumMe video dataset (gyglim.github.io).

**Figure 6** Few sets of video frames from the test bed Open Video Project (OVP)

*4.3. Results and Discussions*

In this study, the output of the proposed system is the number of keyframes received from each video from the test bed. The performance of the system is measured in percentage accuracy. This performance is compared with the performances of the existing system for video key frame extraction and is shown in Table 1 and Table 2. The result is tested using t-test method and found true for all methods.

**Table 1** Key frame extraction using higher color moments with OVP videos

| | Average Open video key frames | Average key frames using VSHCM (Jadhav and Jadhav, 2015) | Average Matched key frames [OVP] | Average Accuracy (%) |
|---|---|---|---|---|
| NASA | 14.2 | 52.7 | 11.6 | 81.69 |
| Seg- 1 | 15.9 | 43.2 | 13.1 | 82.39 |
| Seg. 2 | 16.3 | 52.0 | 13.6 | 83.44 |
| Family | 17.1 | 49.3 | 14.4 | 84.21 |
| Satellite | 17.2 | 47.6 | 15.2 | 88.37 |
| Sports | 13.6 | 47.8 | 11.7 | 86.03 |
| TV | 13.9 | 53.7 | 16.3 | 88.59 |
| Empirical | 18.1 | 50.5 | 15.8 | 87.29 |
| News | 17.3 | 57.4 | 15.4 | 89.02 |
| Historical | 16.2 | 53.2 | 14.1 | 87.04 |
| Average | 15.98 | 50.74 | **14.12** | **85.81** |

Here few of the videos from Open Video Project and a few of the videos show approximately 89% of accuracy. For the given videos the average matching frame rate is 14.12 with reference to open video project summaries i.e. 15.98. Table 1 shows the results obtained using OVP.

**Table 2** Key frame extraction using Ranked based Approach with OVP videos.

| | Average Open video key frames | Average key frames using Rank Based Approach (Srinivas, Pai, and Pai, 2016) | Average Matched key frames [OVP] | Average Accuracy (%) |
|---|---|---|---|---|
| NASA | 14.2 | 55.32 | 11.3 | 79.58 |
| Seg- 1 | 15.9 | 23.54 | 11.5 | 72.33 |
| Seg. 2 | 16.3 | 29.32 | 11.7 | 71.78 |
| Family | 17.1 | 47.21 | 10.4 | 60.82 |
| Satellite | 17.2 | 39.56 | 11.5 | 66.86 |
| Sports | 13.6 | 52.78 | 9.4 | 69.12 |
| TV | 13.9 | 48.45 | 11.4 | 82.01 |
| Empirical | 18.1 | 39.78 | 10.2 | 56.35 |
| News | 17.3 | 51.63 | 11.4 | 65.90 |
| Historical | 16.2 | 56.24 | 10.3 | 63.58 |
| Average | 15.98 | 44.38 | 10.91 | 68.83 |

Here, the videos from Open Video Project and SumMe video dataset are used for the experimentation. As shown in Table 1 and Table 2, only a few videos generate approximately 86.39 % accuracy.

**Table 3** Video summarization using discrete cosine transform (DCT) using open video project videos.

| Open videos | Open video key frames | Average key frames using DCT | Average Matched key frame | Average accuracy |
|---|---|---|---|---|
| NASA | 14.2 | 59.6 | 10.50 | 73.94 |
| Seg- 1 | 15.9 | 40.9 | 10.3 | 64.78 |
| Seg. 2 | 16.3 | 60.2 | 11.9 | 73.01 |
| Family | 17.1 | 56.8 | 11.5 | 73.46 |
| Satellite | 17.2 | 43.1 | 13.9 | 80.81 |
| Sports | 13.6 | 80.1 | 10.6 | 77.94 |
| TV | 13.9 | 10.6 | 13.60 | 97.84 |
| Empirical | 18.1 | 72.8 | 12.20 | 68.89 |
| News | 17.3 | 55.1 | 12.00 | 70.93 |
| Historical | 16.2 | 54.1 | 11.90 | 73.46 |
| Average | 15.98 | 53.33 | 11.84 | 75.51 |

Table 3 shows the result of the proposed method of key frame extraction for the Open Video Project dataset. Table 4 shows the comparison of various existing video summarization methods, i.e. sequential differences, VSHCM, ranked-based approach, ranked plus entropy, and Discrete Cosine Transform (DCT). As can be seen, 75.51% is the average accuracy with the given set of OVP key frames using discrete cosine transform. The average matching rate here is 11.83.

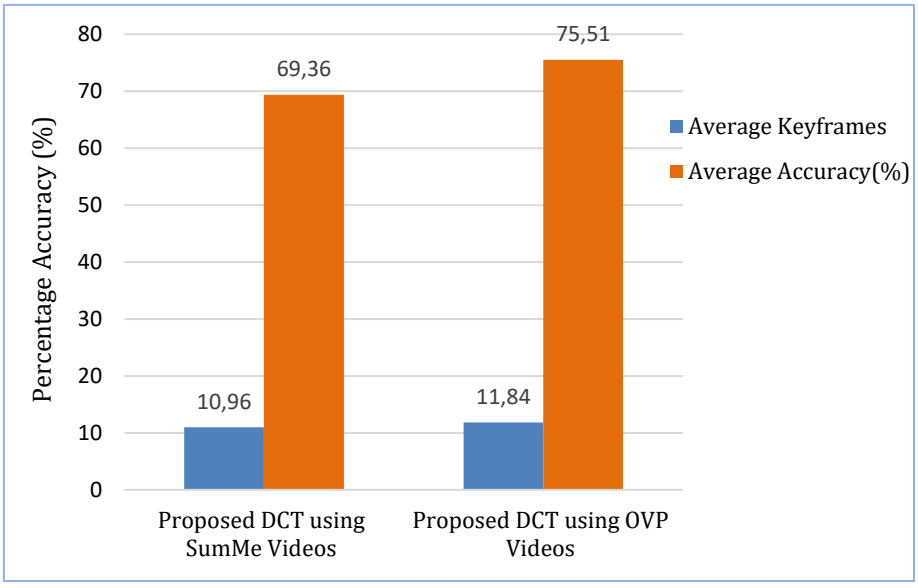



**Figure 7** Keyframe extraction using proposed discrete cosine transform (DCT) with Open Video Project and SumMe video dataset.

**Table 4** Comparison of performances obtained sequential differences, ranked-based approach, rank + entropy, DCT with percentage.

| Method | Average Open Video Given key frames | Average Matched key frames | Average Accuracy (%) |
|---|---|---|---|
| Sequential Frame Differences | 15.8 | 10.17 | 64.37 |
| Higher Color Moments | 15.8 | 13.65 | 86.39 |
| Ranked Based Approach | 15.8 | 11.57 | 68.83 |
| Discrete Cosine Transform | 15.8 | 11.93 | 75.51 |
| Average Accuracy | 15.8 | 11.83 | 73.78 |

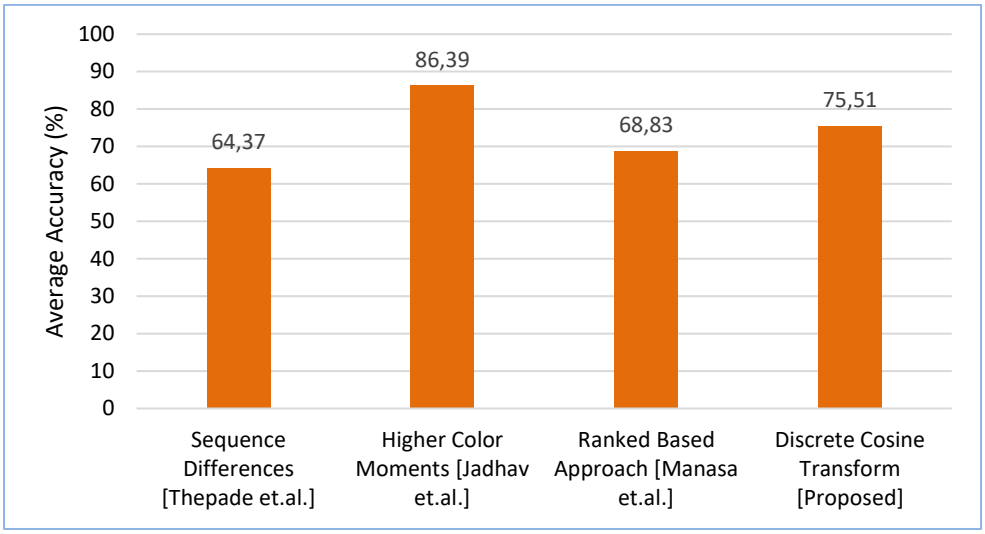



**Figure 8** Comparison of video summarization using spatial and frequency domain using Open video project (OVP).

Table 4 and Figure 8 show that higher color moments obtain better accuracy of 86.39 %, while discrete cosine transform gives a second better accuracy of 75.51%.

*4.4. Results Validation*

The statistical validation of the obtained result is tested by comparing the proposed method with the existing methods using a 2-variable t-Test. Assuming Equal Variance, the hypothesis is defined as:

$H_0$: There exists no significant difference between the proposed and the existing methods.

$H_a$: There exists a significant difference between the proposed method and the existing methods.

**Table 5** The result validation using two variable t-Test assuming equal variance

| | Proposed DCT vs HCM | Proposed DCT vs Rank | Proposed DCT vs Sequential Differences |
|---|---|---|---|
| t-Stat | 3.5717 | 2.2000 | 4.3579 |
| t-Critical 1 Tail | 1.7340 | 1.7340 | 1.7340 |
| t-Critical 2 Tail | 2.1000 | 2.1000 | 2.1000 |
| P-1 tail | 0.0012 | 0.0205 | 0.0001 |
| P-2 tail | 0.0021 | 0.0411 | 0.0003 |

The obtained t-Stat, t-Critical and P-value are as given in Table 5. The proposed method is compared with other existing key frame extraction methods, with t- critical for (1-tail and 2-tail) being less than t-stat for alpha (0.05), rejecting the null hypothesis ($H_0$) and proving that the proposed method is more significant than other methods.

## 5. Conclusions

The work presented in this paper focuses on video storyboard generation using key frames based on spatial and frequency domain features for static video content summarization. In this study, an efficient video storyboard generation technique is proposed using discrete cosine transforms. The result shows that the average matching rate accuracy of higher color moments and discrete cosine transform is 86.39 and 75.51, respectively, i.e. key frames extracted are similar and match the actual content of the videos in OVP and SumMe storyboards. The video storyboards speed up the responses while

searching for offline and online videos. The feature extraction using discrete cosine transform gives better accuracy and it is validated using t-test analysis. The video summarization can also be extended with the fusion of some other features, such as SIFT, SURF, TSBTC, etc., to improve the reliability and accuracy of storyboard generation.

## References

Anayat, S., Sikandar, A., Rasheed, S.A., Butt, S., 2020, A deep analysis of image-based video searching techniques. *International Journal of Wireless and Microwave Technologies,* Volume 10(4), pp. 39–48

Badre, S.R., Thepade, S.D., 2016. Novel video content summarization using Thepade's Sorted n-arry block truncation coding. Procedia Computer Science, Volume 79, pp. 474–482

Cayllahua-Cahuina, E.J.Y, Cámara-Chávez, G., Menotti, D., 2012, A static video summarization approach with automatic shot detection using color histograms. *In*: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)

Dhagdi, M.S.T., Deshmukh, P.R., 2012. Keyframe-based video summarization using automatic threshold and edge matching rate, *International Journal of Scientific and Research Publications,* Volume 2(7), pp. 1–12

Furini, M. Geraci, F., Montangero, M., Pellegrini, M., 2010, STIMO: still and moving video storyboard for the web scenario. *Multimedia Tools and Application,* Volume 46, pp. 47–69

Imanullah, M., Yuniarno, E.M., Sooai, A.G., 2019. A novel approach in low-cost motion capture system using color descriptor and stereo webcam. *International Journal of Technology,* Volume 10(5), pp. 942–952

Jadhav, M.P., Jadhav, D.S., 2015, Video summarization using higher order color moments (VSUHCM). *Procedia Computer Science,* Volume 45, pp, 275–281

Jeong, D.J, Yoo, H.J., Cho, N.I., 2017, A static video summarization method based on the sparse coding of features and representativeness of frames, *EURASIP Journal on Image and Video Processing*, Volume 2017, pp. 1–14

Kekre, H.B., Sarode, M.T.K., Thepade, S.D., 2009, Image retrieval using color-texture features from DCT on VQ code vectors obtained by Kekre's fast codebook generation, *International Journal on Graphics, Vision, and Image Processing,* Volume 9(5), pp. 1–8 (Not Found in The Text)

Kho, D.C.K., Fauzi, M.F.A., Lim, S.L., 2022. Video chunk processor: low-latency parallel processing of 3 x 3-pixel image kernels. *International Journal of Technology,* Volume 13(5), pp. 1045–1054

Liu, L., Fan, G., 2005, Combined key-frame extraction and object-based video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 15(7), pp. 869–884

Maharani, R., Edison, R.E., Ihsan, M.F., Taruno, W.P., 2020. Average subtraction method for image reconstruction of brain using ECVT for tumor detection. *International Journal of Technology,* Volume 11(5), pp. 995–1004

Pan, G., Zheng, Y., Zhang, R., Han, Z., Sun, D., Qu, X., 2019, A bottom-up summarization algorithm for videos in the wild. *EURASIP Journal on Advances in Signal Processing,* Volume 2019, p. 15

Pribadi, T.W., Shinoda, T., 2022. Hand motion analysis for recognition of qualified and unqualified welders using 9-DOF IMU Sensors and Support Vector Machine (SVM) approach. *International Journal of Technology*, Volume 13(1), pp. 38–47

Rao, P.C., Patnaik, M.R., 2014, Contourlet transforms based shot boundary detection. *In*: *International Journal of Signal Processing, Image Processing and Pattern Recognition* Volume 7(4), pp. 381–388

Shen, V.R., Tseng, H.Y., Hsu, C.H., 2014, Automatic video shot boundry detection of news stream using a High level fuzzy petri-net. *In*: 2014 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1342–1347

Srinivas, M., Pai M.M., Pai, R.M., 2016, an improved algorithm for video summarization –a rank-based approach. *In*: Twelfth International Multi-Conference on Information Processing, Volume 2016, pp. 812–819

Subba, T., Roy, B., Pradhan, A., 2016, A study on video summarization. *International Journal of Advanced Research in Computer and Communication Engineering*, Volume 5(6), pp. 738–741

Thepade, S.D., Tonge, A.A., 2014a. An improved approach of key frame extraction for content-based video retrieval. *In*: CPGCON, National Symposium Post Graduate Conference in Computer Engineering, Savitribai Phule Pune University Pune India

Thepade, S.D., Tonge, A.A., 2014b. An optimized keyframe extraction for detection of near duplicates in content-based video retrieval. *In*: 2014 International Conference on Communication and Signal Processing, pp. 1087–1091

Truong, B.T., Venkatesh, S., 2007. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications*, Volume 3(1), p. 3

Wu, X., Zhao, W.L., Ngo, C.W., 2007, Near-duplicate keyframe retrieval with visual keywords and semantic context. *In*: Proceedings of the 6th ACM international conference on Image and video retrieval, pp. 162–169