Computers, Materials & Continua
DOI:10.32604/cmc.2022.021158
Article

Tech Science Press

Effective Video Summarization Approach Based on Visual Attention

Hilal Ahmad[1], Habib Ullah Khan[2], Sikandar Ali[3,*], Syed Ijaz Ur Rahman[1], Fazli Wahid[3] and Hizbullah Khattak[4]

[1]Department of Computer Sciences, Islamia College Peshawar Khyber, Pakhtunkhwa, Pakistan
[2]Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, 2713, Qatar
[3]Department of Information Technology, The University of Haripur, Khyber Pakhtunkhwa, Pakistan
[4]Department of Information Technology, Hazara University Mansehra, Khyber Pakhtunkhwa, Pakistan
[*]Corresponding Author: Sikandar Ali. Email: sikandar@uoh.edu.pk
Received: 25 June 2021; Accepted: 23 August 2021

Abstract: Video summarization is applied to reduce redundancy and develop a concise representation of key frames in the video, more recently, video summaries have been used through visual attention modeling. In these schemes, the frames that stand out visually are extracted as key frames based on human attention modeling theories. The schemes for modeling visual attention have proven to be effective for video summaries. Nevertheless, the high cost of computing in such techniques restricts their usability in everyday situations. In this context, we propose a method based on KFE (key frame extraction) technique, which is recommended based on an efficient and accurate visual attention model. The calculation effort is minimized by utilizing dynamic visual highlighting based on the temporal gradient instead of the traditional optical flow techniques. In addition, an efficient technique using a discrete cosine transformation is utilized for the static visual salience. The dynamic and static visual attention metrics are merged by means of a non-linear weighted fusion technique. Results of the system are compared with some existing state-of-the-art techniques for the betterment of accuracy. The experimental results of our proposed model indicate the efficiency and high standard in terms of the key frames extraction as output.

Keywords: KFE; video summarization; visual saliency; visual attention model

1 Introduction

KFE and video skimming are two fundamental techniques for the summarization of videos [1]. Synoptic systems based on video skimming create a movie with a much shorter runtime than the real video. The most important edge extraction strategies produce precise results by removing prominent edges from the video. Usually, the skims of the video are extra animated and more pleasant than those of the key-frames. In any case, frames of the key have no experience of timing and synchronization problems and can be used and their behavior changed for browsing

as well as route planning. Even for little gadgets, key-frames can offer preferred viewing capacity over video flyovers, as they allow customers to experience the remarkable content of the video at a glance flxation and not even watch a little video [2].

The present paper focuses on the extraction of key frames from the videos. It must be possible to summarize video schemes at the top level of the semantic video content as well as feasible objects, occasions, and activities. As a rule, the extraction of the semantic natives isn't achievable. Nonetheless, some space specific procedures have been anticipated. To close the semantic gap, several researchers [3] used models based on visual attention, to extract visually appealing key-frames out of videos. KFE schemes that rely on visual attention extract visibly prominent key-frames with no extra attention to the video, while efficiently using visual observation-based models to remove outwardly conspicuous key edges from videos. Like the optical flow usage for acquiring dynamic visual consideration signs reduce these methods illogical without the use of advanced equipment.

This paper proposes the efficient KFE pattern that rely on visual attention. This system calculates static as well as dynamic visual viewing displays and then merges them into separate key-frames in a non-direct manner. The static viewing model uses the image signature on the basis of the salience identification [4]. Dynamic viewing model for highlighting the significant regions of the inter-frame movement using the temporal gradients and Delaunay Triangulation for frames cluster. The test results exhibit that the proposed plot yields preferred outcomes over a portion of the notable non-visual attention based plans. Thereby, the professionally presented conspiracy is not only mathematically efficient as a part of the plans based on visual attention [5], but also additionally separates the frames of the key of identical value.

Remaining part of the paper has the following structure. Section II of the paper introduces the related study, Section III outlines the proposed mechanism, section IV presents the experimental outcomes and section V presents the conclusion of the paper.

2 Related Work

There are several specific procedures for the key-frames extraction, using semantic highlights of the videos at an elevated level. For example, Chen et al, Summarized basketball clips on the basis of programmed scenario investigation and determination of the camera perspective. Calic et al. [6] picked these frames to be key frames in which the salient bodies merge by utilizing the area positions gained by the frame segmentation. Ma et al. [7] presented a working model of video skimming based on movement attention. Lai et al. [8] and Hoi et al. [9] used visual attention model for KFE. Mundar et al. [10] proposed a method based on Delaunay Triangulation to cluster the key frames [11–13]. presented plot yields preferred outcomes over a portion of the notable non-visual attention based plans. Xu et al. [14] used web throw content alongside video examination calculations for sports video summarization. A few of the dramatic clustering methods utilize the color of the histogram on the subject and features based on camera movement as characteristics of the clustering. In the research by Liu et al. [15], the applicant key frames were first chosen, which depended on the color histogram. Ding et al. [16] used attention vector with traditional LSTM while Zhang et al. [17] used integrated solution for video parsing and content based retrival and browsing. The compacted domain characteristics [18], and spectacular arrangements use the cheap level properties for KFE, which can generally be divided into 2 classes [19]: (1) clustering and (2) significant content modification. Strategies based on variation, where the content is significant, every frame is contrasted and the prior frame(s) are dissimilar between the characteristics of the low-level frames another new key-frame is removed

merely if the variation between the frames is meaningful. A well-known list of characteristics includes histogram contrast for different color spaces [20] an accumulated function of energy [21], Laplacian Eigen map characteristics [22], and visual MPEG-7 descriptors [23]. Clustering-based KFE techniques group the video images of the frames according to low-level properties, and then generally identify 1 frame from every similar data group as a key-frame. Nevertheless, the loss of semantic subtleties is almost inevitable if you pay less attention to how successfully low-level features are used, resulting in a significant semantic loophole between low level properties as well as actual semantics.

Ma et al. [24] enhanced the work of Zhang et al. in order to design the open framework containing a collection of visual, auditory as well as linguistic features that are then brought together through a non-linear approach. The merged attention rate of every frame has been utilized to create an attention chart, with KFE at the vertices of the chart. Computationally, that framework is complicated by its use of top-down attention techniques. Human visual attention is generally known to be guided by both bottom-up as well as top-down attention processes [25] Furthermore, it is necessary to address the correlation among a set of visual, auditory, and linguistic characteristics at difficult. Authors [26] have made certain hypotheses regarding the visual data processing human system. Chen et al. [27] uses a method based on novel cascaded structures that show stage wise and interstage classification information. Hua et al. [28] used attention fusion function for image retrival. Authors in [29] used static and dynamic attention values were given the same priority. However, the psychological concepts of human attention entitlement that the movement component is much more significant compared to the static attention cues [30].

3 Proposed Methodology

The base-up system is animated in reaction to low-level characteristics (texture, color, movement) those differ visually from the remaining scenario. The instrument of "base up consideration" is the reflexive, autonomous task, temporary and fast. The proposed framework is given in Fig. 1. which is summarized in the next sub section.

3.1 Spatial Attention Value

A model of spatial attention is designed by calculating visual salience on the basis of a description of images known as "image signature". The signature of the image may be utilized to estimate the image of the foreground [31,32]. The essential hypothesis is the foreground of a picture, which is visually more obvious compared to the background. This plan depends on the discrete cosine transform (DCT), where just the DCT segment is kept by rejecting the amplitude.

A certain video frame "F" is first reduced to size 63∼49. Next the image tag "$IS(F_C)$" for the "c" color channel of the "F" frame shown below.

$$IS(F_c) = sign(DCT(F_c)) \tag{1}$$

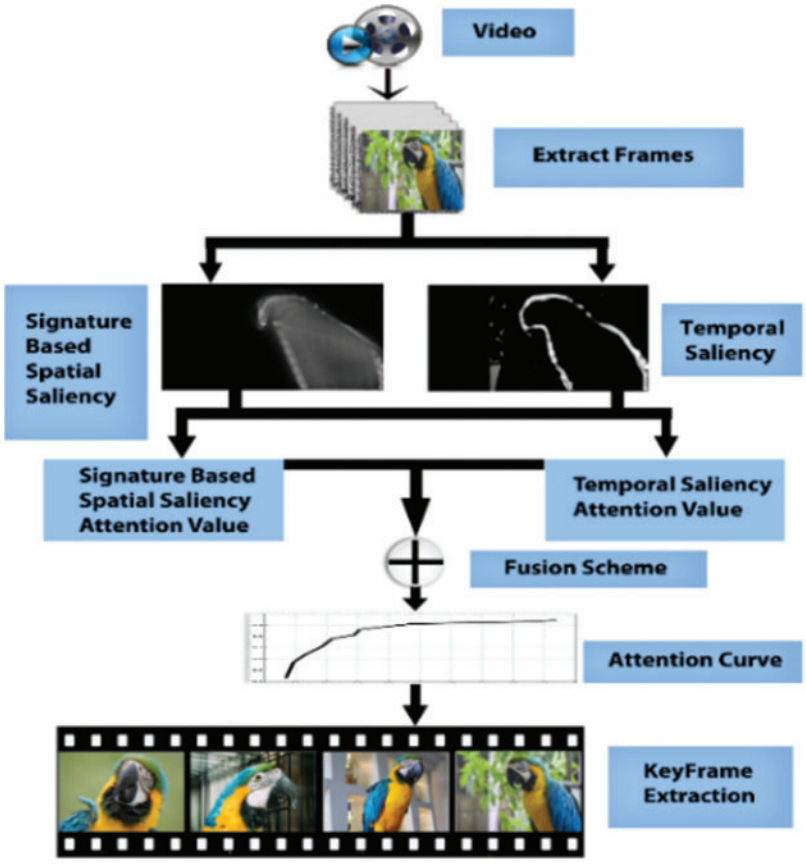


Figure 1: Framework of the proposed system

The sign (.) represents the operator of the input sign, DCT is the discrete cosine transformation and the channel color "c" of the frame "F" is represented by "Fc". The image signature is converted in the spatial range by an inverse DCT to get the recovered image "F_C".

$$\overset{'}{F}_c = IDCT(IS(F_c)) \tag{2}$$

The static salience map of "F_c", designated "$S(F_c)$", is then calculated as:

$$S(F_c) = G \times \sum_c \overset{'}{F}_c \circ \overset{'}{F}_c \tag{3}$$

Gaussian kernel "G" is utilized for the smoothing, the convolution operator "n" and the Hadamard (input) product operator is "○". Gaussian kernel in a pixel (i, j) of an image is described in the following way:

$$G(ij) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}} \tag{4}$$

"S" is the standard deviation of the distribution, whose value is assumed to be 0.045. The saliency map of every color channel is summed up linearly to obtain the total static saliency map "$S(F)$" of the frame "F".

The CIELAB color area is utilized for the selection of color channels due to its capability to effectively approximate people's vision. The saliency chart "S(F)" is then standardized from 0 to 1 by splitting every value by the highest value available in the chart. The mean of the null values in the saliency chart "S(F)" is used to get the static attention value "as" of a frame. When the value of "as" is near one, the frame is regarded as salient. Conversely, a value of "as" close to zero shows an unremarkable frame.

3.2 Temporal Attention Value

In videos, people incline to focus extra attention on the movement of things relative to one another. To get the motion data in video streams quickly, the idea of temporal gradients is used. In this manner, the movement information is subtly calculated by taking measurements of the temporal variations of the values of a pixel in adjacent frames. This feature makes them suitable for use in online systems.

There are two frames of the video $F(t)$ and $F(t-\tau)$ which are present in the video at the moment "t" and "$t-\tau$". The temporal difference between the pixel "p" of frame $F(t)$ and the pixel "q" of the adjacent frame $F(t-\tau)$ is determined as:

$$TC_{p,q}(t) = |F_p(t) - F_q(t-\tau)| \tag{5}$$

$F_p(t) - F_q(t-\tau)$ are the intensity values of frame F(t) and $F_q(t-\tau)$ at pixel "p" and "q", respectively. By means of this description of temporal contrast between pixels of neighboring frames, the gradient value at each pixel "p" of frame F(t) is computed. $A\,5 \times 5$ neighborhood $Nt-\tau$ (p) is defined corresponding to the pixel "p" of frame $F(t)$ in neighboring frame $F(t-\tau)$. The temporal gradient vector at pixel "p" is then defined as:

$$T_p(t) = \left\{ TC(t)_{p,r\tau} \right\},\ \forall\gamma \in N_{t-r}P \tag{6}$$

After calculating the gradient vector for every pixel in frame $F(t)$, the temporal saliency at pixel "p" is then figured through the sum of absolute differences among the temporal gradients of $A5 \times 5$ neighborhood $\mathrm{N_t(p)}$ around pixel "p" in frame F(t).

$$TS_p = \sum_{s=1}^{N_t(p)} |T_p(t) - T_s(t)| \tag{7}$$

By calculating the salience value at every pixel, the temporal salience map $TS(F)$ of the image $F(t)$ is gained. The map of temporal expression is standardized in the range [0, 1] by dividing every pixel value through the maximum value in the map $TS(F)$. To obtain the temporal attention value "A_T" of frame $F(t)$, the saliency map values are averaged. Here too, a maximum value of "A_T" shows a salient frame and vice versa.

3.3 Fusion of Attention Values and KFE

In most cases, researchers have used linear fusion schemes for the fusion of numerous attention values to make an overall attention value. Given that there are "n" number of attention values

that need to be combined, the general shape of linear fusion schemes is as follows:

$$A_L = \sum_{i=1}^{n} w_i.A_i \ \ where \ \ \sum_{i=1}^{n} w_i = 1 \tag{8}$$

W_i is the weight of an attention value A_i, and A_L is the accumulated attention value utilizing the linear scheme. Normally, the linear procedure is not capable of reflecting all the information that the attention values of the attention resources have. Furthermore, the visual part of the human brain uses a nonlinear processing system for pattern recognition and related tasks. Therefore, a linear fusion procedure is not appropriate.

In the literature on visual attention-based video summary patterns, the authors have utilized key-frame-based video summaries utilizing visual attention cues and linear fusion schemes with a greater weight of movement attention scores compared to static schemes.

Within frame "F", denoted by "W_T", the weight of the temporal attention value is obtained from the temporal saliency map $TS(F)$ as:

$$W_T = d \ e^{1-d}, \ \ d = \max(TS(F)) - \min(TS(F)) \tag{9}$$

If the motion contrast in $TS(F)$ is strong, the value of "d" will be higher, resulting in a higher value of W_T and vice versa. The weight of the static attention value "W_S" is given as:

$$w_s = 1 - w_T \tag{10}$$

In case "A_T" and "$A_S A_T$" represent the temporal and static attention values, respectively, the unweight portion of the fusion function is considered to be assumed:

$$F_A(A_S, A_T) = \frac{1}{2}\left[(A_S + A_T) + \frac{1}{1+\gamma}|A_s - A_T|\right], \ \ where \ \ \gamma > 0 \tag{11}$$

"γ" is a constant, which represents the importance of a component of attention in the model of combined attention. "γ" is equal to 0.2. With static and dynamic values of attention expressed as a vector $A = [A_S, A_T]$ as well as weight expressed as a vector $W = [W_S, W_T]$, the function of weighted fusion is described as:

$$F_{AW}(A_S, A_T) = \frac{\underline{w.A} + \frac{1}{2(1+\gamma)}(|2w_s A_s - \underline{w.A}| + |2w_T A_T - \underline{w.A}|)}{W} \tag{12}$$

"W" is determined as

$$W = 1 + \frac{1}{2(1+\gamma)}(|1 - 2w_s| + |1 - 2w_T|) \tag{13}$$

Then, we justify briefly the selection of the nonlinear fusion scheme compared to the Max and linear fusion schemes. Let us ponder 2 groups of fusion values (0.45, 0.45) and (0.9, 0). In both cases, the linear fusion pattern will result in a merged attention value of 0.9. Nevertheless, the first group is much more conspicuous than the second, which is due to the high value of the attention index. Furthermore, the linear fusion scheme doesn't satisfy the following characteristic for attention fusion schemes:

$$F(v_1, v_2) < F(v_1 + \Delta, \ v_2 - \Delta), \ \ where \ \ 0 < \Delta \le v_2 \le v_1) \tag{14}$$

$F(v1, v2)$ is equivalent to $F(v1 + \Delta,\ v2 - \Delta)$ for a linear fusion.

Max's fusion model chooses the highest two attention indexes to be merged. MAX's fusion model satisfies the feature of inequality (14). Yet this trivial feature of attention functionality is infringed by the Max Fusion:

$$F(v_1, v_2) < F(v_1 + \Delta, v_2),\ \ where\ \ \Delta > 0 \tag{15}$$

The fusion scheme of Eq. (12) used fulfils both of these characteristics (inequalities 14 and 15) and is, therefore, more efficient compared to linear as well as Max fusion Model. However, the "g" parameter in Eq. (12) allows handling the differences in inequalities between the right and left hands (14) and (15). As "g" rises, the absolute gap starts to decrease. The parameter "g" thus allows managing the sensibility of the fusion scheme to changes.

The fused attention score of every frame is utilized to create the attention graph representing a video and then utilized for KFE. When the key frame number "n_k" is not given by the user, the user selects the frame containing the maximum attention value in every frame as the key frame. When "n_k" is given, then "n_{ks}" will assign a number of key frames to each shot according to the proposed approach in below.

$$n_{ks} = \max(n_k \times \alpha, 1) \tag{16}$$

"α" shows the relationship between the variance of attention scores in a given recording and the total variance of attention scores in all recordings.

4 Experiments and Results

First, the technology outcomes were displayed on a single shot, that was taken from the Open Video Project (www.open-video.org).

The 1st series of tests is the 5th recording (frames 484–520) of the ucomp03_06_m1.mpeg video. Tennis player strikes the ball, stands up, and gains the credit of the crowd. Fig. 2 shows the attention graphs of the temporal, spatial, and merged salience values. From the merged attention graph of your proposed pattern, it is seen the 491 frame holds the maximum attention score and was therefore chosen as the key frame. Lai and Yi's fused attention curve proposes frame 517 as key-frame. The key-frames chosen by and the suggested system are illustrated in Fig. 2. It is obvious that the key-frame as extracted from doesn't communicate the idea of the game of shooting of the player and is therefore not representative semantically. Furthermore, the KEF through the proposed system is extra high light and effectively summarizes the recording.

Sequence of 2nd video test is the 2nd shot (frames 532–548) of the hcil2000_01.mpeg video. A subject speaks and stands in the frame under consideration, with the surrounding trees. From frame 545, a subtitle appears in the scenario to show an introduction by the narrator. A key frame that is representative of the scene has to display the people and the caption. The attention graphs are illustrated in Fig. 3. Key frames 545–548 are extracted according to and the proposed model, respectively. Two frames are illustrated. Frame 545 has a caption that is unclear to read, while the 548 frame displays the caption clearly and is, therefore, the most accurate demonstration of the recording.

4.1 Comparison with Other Techniques

Here, this unit matches the proposed system with several of the outstanding schemes relying on non-visual attention as well as visual attention. For comparison purposes, the experiment was carried out with twenty videos of different types, which were downloaded directly from the Open Video Project.

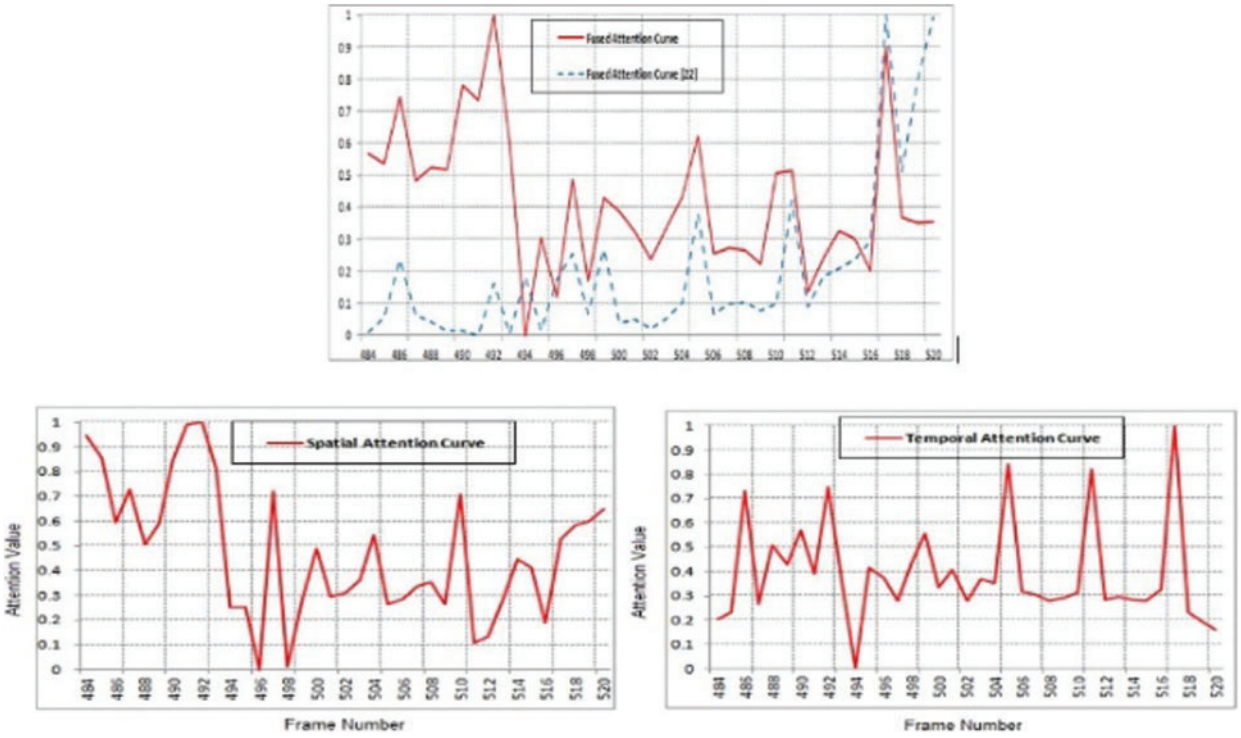


Figure 2: Curves of attention for the fifth recording of the video ucomp03_06_m1.mpeg

Multiple approaches were used to compare the results. One is based on the well-known measurement categories F-measure, precision, and recall. And another one is the subjective MOS method (Mean Opinion Score) for evaluation pattern.

The initial assessment procedure involves manually extracting the key frames for each video using 3 human users. The two frames are assumed to be identical if they carry identical semantic content. The terms below are then described:

T_P **(True Positive):** Frame that was elected manually as well as by the technology as a key frame, F_P **(False Positive):** Frame elected by technology, not manually, as a key-frame, and

F_n **(False Negative):** Frame elected manually, not by technology as key frame.

These words are used for the definition of the metrics Precision and Recall.

$$Recall = \frac{T_p}{Tp + F_n} \tag{17}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{18}$$

To obtain a combined single metric, both Precision and Recall are merged using the following f-measure definition:

$$F = 2 \times Precision \times Recall / Precision \times Recall \tag{19}$$

In the 2nd assessment strategy, the dataset of Tab. 1 was analyzed on the basis of the criteria of the MOS. All users' ratings for a given video are next aggregated to get the video's MOS. In sections 1 and 2, the proposed technique was compared with several of the schemes based on non-visual attention as well as visual attention respectively.

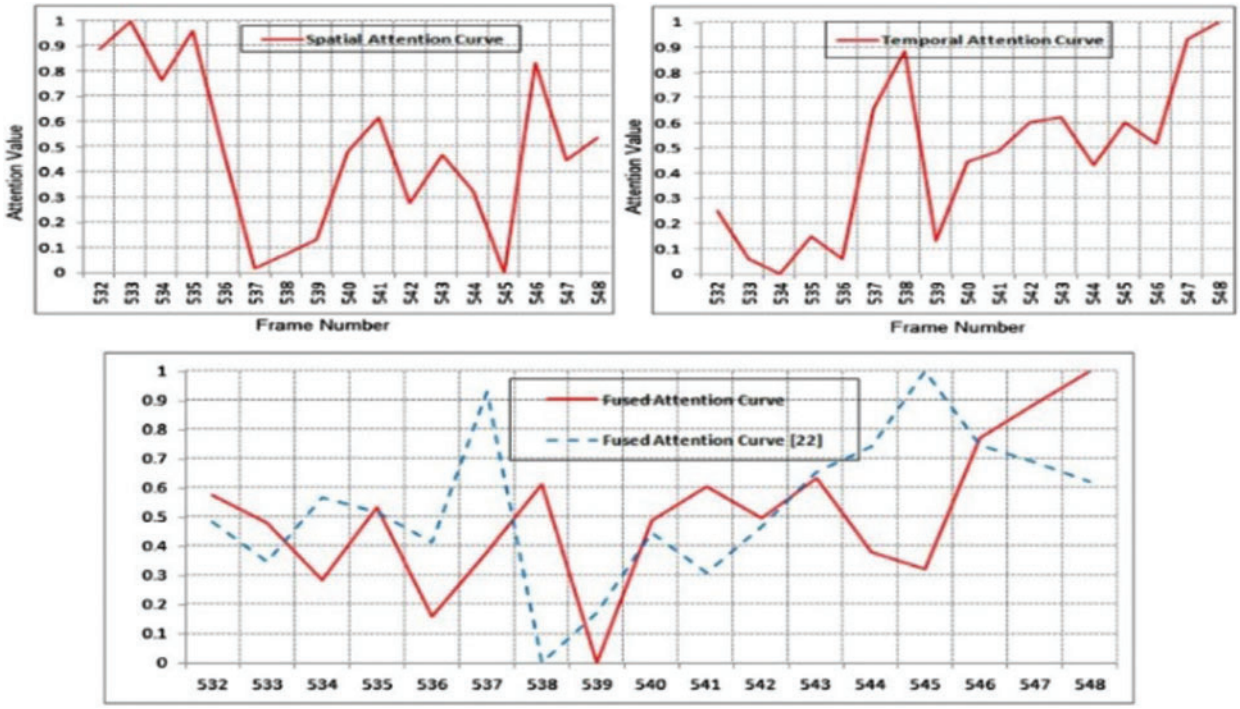


Figure 3: Attention curves for the second shot of the video hcil2000_01.mpeg

Table 1: Facts of query videos

S.no	Name of the video	Total number of frames
01	From shots 03 of 8 in wetlands regained	3563
02	A Digital personal scale in technology at home	3345
03	Outline toward HCIL "2000" reports	2453
04	From shots 05 of 14 in ocean floor legacy	4664
05	In shot 01 of the great web of water	3278
06	In shot 02 of the great web of water	2117
07	In shot 07 of the great web of water	1744
08	In shot 01 of a new horizon	1805
09	In shot 02 of a new horizon	1796
10	In shot 06 of a new horizon	1943

4.2 Comparisons with Methods Based on Non-Visual Attention

This section provides a comparison between the proposed approach and 4 outstanding non-visual attention patterns: DT [11], VSUMM [12], OV [13] and STIMO [31]. All these technologies rely on low-level features extracted from the video frames. OV [13] is a scheme based on meaningful content variation, whereas DT [11], VSUMM [12], and STIMO [31] are clustering based. Tab. 2 displays the recall, precision, and F measurement readings for every one of the ten videos included in the dataset. The proposed technique is clearly dominating other methods by attaining higher values consistently across all 3 measurements.

Table 2: Overall precision, Recall, and F-measure of various methods on a video dataset

S.no	In [7]			In [13]			In [8]			Proposed Method		
	P	R	F	P	R	F	P	R	F	P	R	F
1	0.75	0.83	0.79	0.70	0.82	0.75	0.80	0.81	0.80	0.80	0.84	0.82
2	0.75	0.75	0.75	0.73	0.73	0.73	0.85	0.90	0.87	0.82	0.85	0.83
3	0.70	0.85	0.77	0.72	0.80	0.76	0.82	0.86	0.84	0.82	0.83	0.82
4	0.85	0.86	0.85	0.85	0.83	0.84	0.80	0.83	0.81	0.80	0.83	0.81
5	0.75	0.83	0.79	0.75	0.80	0.77	0.83	0.80	0.81	0.75	0.82	0.78
6	0.73	0.88	0.79	0.70	0.83	0.76	0.75	0.85	0.80	0.85	0.90	0.87
7	0.75	0.75	0.75	0.72	0.83	0.77	0.70	0.80	0.75	0.80	0.81	0.80
8	0.75	0.83	0.79	0.75	0.81	0.78	0.72	0.83	0.77	0.80	0.85	0.82
9	0.82	0.88	0.84	0.82	0.85	0.83	0.78	0.85	0.81	0.85	0.87	0.86
10	0.80	0.82	0.81	0.82	0.80	0.81	0.80	0.83	0.81	0.81	0.83	0.82
Avg.	**0.76**	**0.83**	**0.79**	**0.76**	**0.81**	**0.78**	**0.78**	**0.83**	**0.80**	**0.81**	**0.84**	**0.82**

However, several exceptions exist. For example, the DT scheme for Video 5 attains a high level of precision. Yet for that video, DT chooses only 1 key-frame, so the values of F_p and T_p are both 1, resulting in a precision score of 1. DT has an enough low recall value for this video because it has a high F_N score. Likewise, OV has Recall's highest value for Video Thirteen, whereas the value of precision is also significantly low.

4.3 Comparison with Techniques Based on Visual Attention

Tab. 2 displays the recall, precision, and F-measurement scores for the proposed method and visual attention systems. It is observed that, in general, all visual attention-based patterns have good outcomes than the non-visual attention patterns based on low-level features. However, with results of comparable quality to the other methods, the benefit of lowering calculation costs is clear. The duration for the proposed technique is then compared to the duration of alternative techniques.

Moreover, within the schemes based on visual attention, the outcome of the proposed technique is comparable to the rest of the other mechanisms. Similar findings can be extracted from Tab. 3, showing the MOS score for all the possible schemes on the basis of visual attention.

Therefore, it was changed from 1,000 to 6,000 frames for the duration at which the videos to summarized. In the proposed methodology, an optional pre-sampling step can be used if the computational effort is to be further reduced. Fig. 4 illustrates the time taken by [8,9] and the proposed pattern with a sampling rate of twenty frames for videos of various durations. The findings were achieved on a general-purpose PC (Intel Core 2 Duo 1.6 Hz, equipped with 2 GB RAM). It is observed that the pattern of [8] requires the maximum time.

Table 3: Results of MOS tests for different techniques based on visual attention

S.no	In [7]	In [13]	In [8]	Our method
01	4.08	4.25	4.03	4.16
02	4.13	4.41	4.1	3.99
03	4.4	4.06	4	4.16
04	4.15	4.24	4	4.31
05	4.1	4.09	4.1	4.25
06	3.99	4.11	4.1	4.06
07	4.15	4.06	4.13	4.19
08	4	4.14	4.11	4.19
09	4.18	4.16	4	4.31
10	4.3	4.22	4.13	4.4
Avg.	**4.15**	**4.17**	**4.07**	**4.20**

Finally, the key-frames extracted through different patterns are displayed visually in Tab. 4 for the documentary movie ‘Hurricane Force-A Seaside View, segment 3’. First, the video presents the initial researcher of the United States Geological Survey, who explains the significance of knowledge about the geology of marine areas together with contextual scenarios.

The proposed technique is evaluated based on the F-measure, recall, and precision. The formula used for these measurements are similar to conventional VS techniques.

Compression on the basis of the OV dataset with alternative available techniques is assessed with the F-measure, recall, and precision. Precision uses for the accuracy of a technique and computes the count of wrong extraction key-frames. The recall value displays the possiblility of all key-frames that are available in the basic truth. We performed validation of our technique utilizing 2 benchmark video datasets by comparing our findings to the prior art of VS techniques.

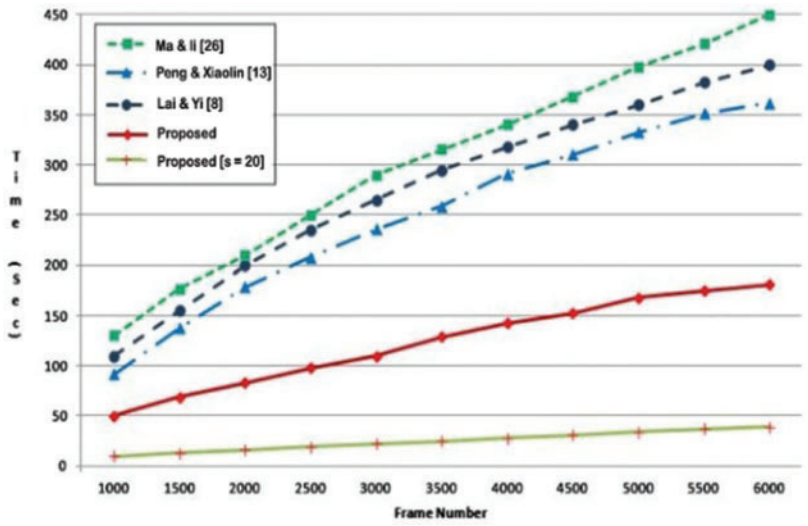
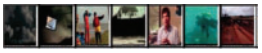
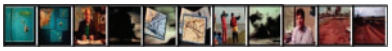


Figure 4: The total time of frame duration in [8,9,14] for mining key frames

Table 4: Comparison of KFE for video 'hurricane force - a seaside view, segment 3'

Method	Generated Key Frames
Ground Truth	
[12]	
[10]	
[31]	
[11]	
[7]	
[13]	
[8]	
Our Method	

4.3.1 Open Video Project Dataset Evaluation

Dataset is called OV (Open Video Project) and comprises videos with a standard RGB form with 30 FPS and 352 × 240 pixels. This dataset includes different types of videos, e.g., documentaries, surveillance videos, educational videos, historical videos, ephemeral, and lectures videos [33]. Total time of the record is 75 mins, where each video has a time of 1 min to 4 mins. Our methodology is compared to five techniques (OV [12], DT [10], STIMO [31], VSCAN [33], and VSUMM [11]) utilizing the original assessment measurements for F measurement, precision and recall based on the data set.

4.3.2 Youtube Dataset Evaluation

There are 50 videos in this dataset of various types, i.e., surveillance and sports videos, animated videos, TV home videos, and commercials videos with a total duration of one to ten minutes. We compared the outcome with VSUMM, five-user summaries, and the concept of Fei et al. [34] for this dataset and hesteroscopy videos [35].

They segment the recordings using a perceptual hashing technique, which is insufficient for monitoring streams and has limited performance.

4.3.3 MOS Based Comparison

Besides the quantitative measurement, it is essential to evaluate achievement on the basis of a subjective or qualitative assessment.

The MOS is a subjective assessment metric utilized to evaluate the compiled results of various VS techniques. The MOS returns the opinion of the users directly and displays their areas of concern. Tabs. 5 and 6 shows the MOS score of our pattern and alternative techniques for twenty videos with an average score in the final row.

Table 5: Comparing our technique with other methods with MOS-score

Video No.	In [7]	In [13]	In [8]	In [12]	In [10]	In [31]	In [11]	In [36]	In [37]	Our method
1	4.26	4.18	4.24	3.08	2.50	4.52	4.04	4.04	4.50	4.60
2	4.35	4.2	4.28	2.88	3.62	4.12	3.86	4.47	4.49	4.50
3	4.24	4.05	4.3	2.42	3.82	3.48	4.19	4.06	4.22	4.50
4	4.16	4.6	4.28	2.76	3.20	2.71	3.4.00	3.91	4.22	4.50
5	4.26	4.08	4.66	3.08	3.37	2.72	4.02	4.08	4.44	4.45
6	4.15	4.28	4.39	3.82	3.64	3.30	4.47	4.39	4.51	4.60
7	4.27	4.26	4.25	3.50	3.68	3.57	3.82	4.25	4.35	4.40
8	4.15	4.2	4.15	3.50	3.83	3.56	2.80	4.28	4.48	4.19
9	4.07	4.36	4.6	3.38	3.12	3.14	3.29	4	4.18	4.19
10	4.28	4.17	4.2	3.36	3.46	3.44	3.84	4.14	4.02	4.05
Avg.	**4.22**	**4.24**	**4.33**	**3.18**	**3.42**	**3.46**	**3.81**	**4.16**	**4.34**	**4.40**

4.3.4 Analysis of Computational Intricacy

Computing intricacy is an essential measurement for assessing VS techniques and specifically for monitoring video gathered in a limited resource environment. With this in mind, we have evaluated the duration of our strategy and compared the intricacy of our approach with similar methods. For this aim, we have looked at various videos with frames from 1,000 to 6,000. The mean running time for VS representative techniques is 304.61, 249.27, 277.84 and 123.97 s for [36–38]. In comparison to these methods, our technique obtains the most efficient results in time intricacy by calculating the existing range of individual frames in only 112.87 s.

Table 6: Evaluation of our technique with altered methods with MOS-Score

Shot segmentation	In [7]	In [13]	In [8]	In [11]	In [35]	In [36]	In [37]	In [38]	Proposed Method
Features	Motion attention model	Visual attention model	Visual attention model	Color features	Visual saliency	Motion attention model	Deep features and aesthetics	Object motion	Deep features with high accuracy
Fusion scheme	Linear	Linear	Linear	None	Non-Linear	None	Hierarchical	None	Non-Linear
Weighted fusion	×	✓	✓	×	✓	×	✓	×	✓
Post-processing for redundancy elimination	×	×	×	✓	×	×	✓	✓	✓
Processing in real time	×	×	×	×	✓	×	✓	×	✓
Suitability for surveillance	×	×	×	×	×	✓	✓	×	✓

5 Conclusion

In this paper, we recommend an effective frame-work that relies on visual attention for KFE from video. The method not only delivers effective outcomes but is also appropriate for usage in small devices. Using temporal gradients offers an effective substitute for the traditional flow-oriented optical characteristics used so far. Using a nonlinear weighted fusion pattern adds all the advantages of the earlier used patterns. In general, the framework requires far less time than the recent patterns on the basis of visual attention. The experimental outcomes, on the basis of a set of criteria, indicate that the extracted key-frames utilizing the proposed pattern are related semantically and more strongly focused on highlighting than the ones produced by the alternative methods with which it is evaluated.

Acknowledgement: This work was supported by the Qatar National Library, Doha, Qatar, and in part by the QU Internal Grant Qatar University Internal under Grant IRCC-2021-010.

Data Availability: The data collected during the data collection phase will be provided upon request to the authors.

Funding Statement: This work was supported in part by Qatar National Library, Doha, Qatar, and in part by the Qatar University Internal under Grant IRCC-2021-010

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," in *ACM Transactions on Multimedia Computing, Communications, and Applications*, New York, NY, USA, vol. 3, no. 1, pp. 3, 2007.

[2] A. Girgensohn, J. Boreczky and L. Wilcox, "Keyframe-based user interfaces for digital video," *Computer (Long. Beach. Calif)*, vol. 34, no. 9, pp. 61–67, 2001.
[3] S. Uchihashi, J. Foote, A. Girgensohn and J. Boreczky, "Video manga: Generating semantically meaningful video summaries," in *Proc. of the Seventh ACM Int. Conf. on Multimedia*, New York, NY, USA, pp. 383–392, 1999.
[4] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008.
[5] F. Chen, D. Delannay and C. De Vleeschouwer, "An autonomous framework to produce and distribute personalized team-sport video summaries: A basketball case study," *IEEE Transaction on Multimedia*, vol. 13, no. 6, pp. 1381–1394, 2011.
[6] J. Calic and B. Thomas, "Spatial analysis in key-frame extraction using video segmentation," in *Proc. of the 5th International Workshop on Image Analysis for Multimedia Interactive Services*. Lisboa, Portugal, April 21–23, 2004.
[7] Y. F. Ma and H. J. Zhang, "A model of motion attention for video skimming," in *IEEE Int. Conf. on Image Processing*, vol. 1, pp. I–129, 2002.
[8] J. L. Lai and Y. Yi, "Key frame extraction based on visual attention model," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, pp. 114–125, 2012.
[9] X. Hou, J. Harel and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.
[10] P. Mundur, Y. Rao and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 219–232, 2006.
[11] S. E. F. De Avila, A. P. B. Lopes, A. Da Luz and A. De Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," in *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.
[12] D. DeMenthon, V. Kobla and D. Doermann, "Video summarization by curve simplification," in *Proc. of the 6th ACM Int. Conf. on Multimedia, MULTIMEDIA*, Bristol, United Kingdom, pp. 211–218, 1998.
[13] J. Peng and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," *IEEE Multimedia*, vol. 17, no. 2, pp. 64–73, 2010.
[14] C. Xu, J. Wang, H. Lu and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," in *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 421–436, 2008.
[15] L. Liu and G. Fan, "Combined key-frame extraction and object-based video segmentation," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 15, no. 7, pp. 869–884, 2005.
[16] S. Ding, S. Qu, Y. Xi and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Generation Computer Systems*, vol. 93, pp. 583–595, 2019.
[17] H. J. Zhang, J. Wu, D. Zhong and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, pp. 643–658, 1997.
[18] E. K. Kang, S. J. Kim and J. S. Choi, "Video retrieval based on scene change detection in compressed streams," *IEEE Transaction on Consumer Electronics*, vol. 45, no. 3, pp. 932–936, 1999.
[19] N. Ejaz, T. Bin Tariq and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *Journal of Visual Communication and Image Representation*, vol. 23, no. 7, pp. 1031–1040, 2012.
[20] X. D. Zhang, T. Y. Liu, K. T. Lo and J. Feng, "Dynamic selection and effective compression of key frames for video abstraction," *Pattern Recognition Letters*, vol. 24, no. 9–10, pp. 1523–1532, 2003.
[21] R. M. Jiang, A. H. Sadka and D. Crookes, "Hierarchical video summarization in reference subspace," *IEEE Transaction on Consumer Electronics*, vol. 55, no. 3, pp. 1551–1557, 2009.
[22] J. H. Lee, G. G. Lee and W. Y. Kim, "Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder," in *IEEE Transactions on Consumer Electronics*, vol. 49, no. 3, pp. 742–749, 2003.
[23] Y. Gao, W. B. Wang and J. H. Yong, "A video summarization tool using two-level redundancy detection for personal video recorders," *IEEE Transaction on Consumer Electronics*, vol. 54, no. 2, pp. 521–526, 2008.

[24] Y. F. Ma, X. S. Hua, L. Lu and H. J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transaction on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.
[25] E. A. Styles, *In The Psychology of Attention*, London: Psychology Press, 2006.
[26] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2131–2146, 2011.
[27] Y. T. Chen and C. S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Transaction on Image Processing*, vol. 17, no. 8, pp. 1452–1464, 2008.
[28] X. -S. Hua and H. -J. Zhang, "An attention-based decision fusion scheme for multimedia information retrieval," *Advances in Multimedia Information Processing-PCM*, vol. 3332, pp. 1001–1010, 2004.
[29] D. Gao, V. Mahadevan and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 13–13, 2008.
[30] M. Furini, F. Geraci, M. Montangero and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," *Multimedia Tools and Application*, vol. 46, no. 1, pp. 47–69, 2010.
[31] K. Muhammad, J. Ahmad, M. Sajjad and S. W. Baik, "Visual saliency models for summarization of diagnostic hysteroscopy videos in healthcare systems," *Springerplus*, vol. 5, no. 1, pp. 1495, 2016.
[32] K. M. Mahmoud, M. A. Ismail and N. M. Ghanem, "VSCAN: An enhanced video summarization using density-based spatial clustering," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8156 LNCS, no. PART 1, pp. 733–742, 2013.
[33] M. Fei, W. Jiang and W. Mao, "Memorable and rich video summarization," *Journal of Visual Communication and Image Representation*, vol. 42, pp. 207–217, 2017.
[34] K. Muhammad, M. Sajjad, M. Y. Lee and S. W. Baik, "Efficient visual attention driven framework for key frames extraction from hysteroscopy videos," *Biomedical Signal Processing and Control*, vol. 33, pp. 161–168, 2017.
[35] N. Ejaz, I. Mehmood and S. Wook Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 34–44, 2013.
[36] S. Zhang, Y. Zhu and A. K. Roy-Chowdhury, "Context-aware surveillance video summarization," *IEEE Transaction on Image Processing*, vol. 25, no. 11, pp. 5469–5478, 2016.
[37] K. Muhammad, T. Hussain, S. Member and M. Tanveer, "Cost-effective video summarization using deep CNN with hierarchical weighted fusion for IoT surveillance networks," *IEEE Internet of Things Journal*, vol. PP, no. c, p. 1, 2019.
[38] Y. Zhang, R. Tao and Y. Wang, "Motion-state-adaptive video summarization via spatiotemporal analysis," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 27, no. 6, pp. 1340–1352, 2017.