Effective and Efficient Video Summarization Approach for Mobile Devices

http://dx.doi.org/10.3991/ijim.v10i1.4827

Hesham Farouk[1], Kamal A. ElDahshan[2], Amr Abozeid[2]
[1] Electronics Research Institute, Cairo, Egypt.
[2] Al-Azhar University, Cairo,Egypt.

***Abstract*—In the context of mobile computing and multimedia processing, video summarization plays an important role for video browsing, streaming, indexing and storing. In this paper, an effective and efficient video summarization approach for mobile devices is proposed. The goal of this approach is to generate a video summary (static and dynamic) based on the Visual Attention Model (VAM) and a new Fast Directional Motion Intensity Estimation (FDMIE) algorithm for mobile devices. The VAM is based on how to simulate the Human Vision System (HVS) to extract the salient areas that have more attention values from video contents. The evaluation results demonstrate that, the effectiveness rate up to 87% with respect to the manually generated summary and the state of the art approaches. Moreover, the efficiency of the proposed approach makes it suitable for online and mobile applications.**

***Index Terms*—Video Summarization, Key Frame Extraction, Video Skimming, Visual Attention Model, Mobile Computing, Computer Vision.**

I. INTRODUCTION

The increasing processing power, camera resolution, and memory size of mobile devices have resulted in an explosive growth of video capturing and streaming experiences. Video has been an important media for entertainment and communications between mobile users. Video is a complex multimedia which is composed of a sequence of images, audio tracks, and textual information. Also, the content of the video is huge and contains a lot of redundant information [1]. On mobile devices, browsing, indexing, retrieving, streaming and storing such a huge video content is quite difficult as compared to other formats of media, like audio and text [1,2,3]. Therefore, video summarization is an important approach for quick browsing, fast streaming, efficient storage, and quick retrieval of the video content [4,5,6].

Video summarization is the process of extracting the most important information and reducing the amount of redundant information from the video. The input video must be well processed in order to extract only the most useful contents [7]. But to generate a good video summary, a full understanding of the video is required, which is still a research challenge. In literature, many video summarization approaches have been introduced [8,9,10]. Farouk et al. [11] presented an analysis and a comparative study among various techniques of mobile video summarization according to a proposed set of criteria (For example Content structure, final summary representation, summarization features, summarization speed, summarization purposes, targeted devices, adaptability and complexity). The comparative study showed that most of these approaches are based on low level features, such as color and motion, to generate the summary. Unfortunately, these approaches are not effective enough because they don't take into account the human perception of the video content. In other words, there is a gap between low-level features of the video and its semantic meaning [12,13].

Recently, the Visual Attention/saliency Model (VAM) has been widely used in computer vision and multimedia processing researches and applications. By detecting the salient content, visual attention can reflect the user interest to some content and provide user targeted applications according to their preferences. In video processing, there are several VAM based applications such as video compression, summarization, retrieval, advertising and recognition [14].

The advantages of the video summarization include, but are not limited to, enhancing browsing, streaming, storage, and quick retrieval of video content. For example, people usually use the mobile devices to capture events and celebrations then publish the captured video to social networks (e.g. Facebook) or save it using personal storage service as private or public cloud storage (e.g. Dropbox). But if the size of the captured video is large, it consumes a lot of time and bandwidth in order to transfer it across networks. In this case video summarization can be used reduce the size even more than video compression, while preserving the main content and then publish it.

In this paper, we propose an effective and efficient video summarization approach for mobile devices based on VAM. In this approach, VAM is applied to bridge the gap between the low-level video features and its semantic interpretation by the HVS. Moreover, we introduce a Fast Directional Motion Intensity Estimation (FDMIE) algorithm to calculate the motion intensity between consecutive frames. We implemented a prototype to test our approach based upon the Android platform. Any mobile device with android version 4.0 or higher can run this prototype. We carried out experiments to measure the effectiveness and efficiency of the proposed approach. The results proved that the proposed system is more effective and efficient than other related approaches.

This paper is organized as follows: Section II introduces some related work about the visual attention model. The proposed approach is presented in Section III. Section IV presents the experiments and results of our approach. Finally, section V concludes the paper and suggests future work.

II. RELATED WORK

The human brain receives a huge amount of information in every second. About, 80% of this information (up to 10 billion bits) is received by our vision system. Furthermore, the computational power of the human brain is not sufficient to perform complex analysis of all the input visual information. Therefore, the human vision system (HVS) applies a visual attention/saliency mechanism. In this mechanism, the HVS concentrates on the important visual information, is called salient information. This salient information is quickly processed with high priorities than other non-salient information using the brain to increase the processing efficiency [14].

Therefore, some researchers try to design algorithms from the visual salience mechanism to develop an intelligent and efficient application. Although, it is difficult to fully simulate the human attention mechanism, the research in this direction has significantly been ameliorated to guide computers and devices to quickly process information like the human brain [15].

The visual saliency mechanism can provide a user-targeted service according to the users' preferences and interests by emphasizing on the salient content. Recently, the visual salience mechanism has played an important role for such intelligent computer vision and multimedia applications. One of the important applications of VAM is video summarization. Ma et al. presented a user attention model to summarize the video based on the visual saliency mechanism [16]. Then this model was enhanced to be a generic framework of user attention model, including the various attention models. Such as the motion attention model, the static attention model, the face attention model, the camera attention model and the speech attention model. Then these models are merged together by a nonlinear fusion scheme [17]. Unfortunately, this framework is computationally expensive and the combinations between visual, oral and linguistic features are difficult tasks.

Therefore, some improvements have been applied on Ma et al.'s [17] framework by Peng and Xiaolin [18]. These improvements are done by initially using a color histogram and the K-means algorithm to cluster the frames. Then key-frame candidates are selected from each cluster with the highest Visual Attention Index (VAI) descriptor. Because of the usage of the K-means algorithm, the outputs Keyframes don't reflect the time order and the video structure. Lai and Yi addressed this problem by using the time constrained clustering algorithm to preserve the sequential order of the video frames [12].

A Comparative study between static and dynamic saliency is introduced in [19]. There are two observations derived from this study. Firstly, the image saliency is often different from the video saliency. Secondly, the camera motions, such as zooming, panning or tilting, have a significant effect on the dynamic saliency detection.

Ejaz et al. [6,20] presented an efficient aggregated visual attention model for key frame extraction. This technique reduces the computational cost by using the temporal gradient based motion visual saliency detection instead of the traditional optical flow methods. Then, use a non-linear weighted fusion method to merge the static and dynamic visual attentions.

III. THE PROPOSED APPROACH

In this section, we give a detailed description of the proposed approach. We introduce the approach architecture in Section III A. Then, Section III B shows how to compute the static attention model. In Section III C, we show how to compute the motion attention model. Section III D describes the fusion of static and motion attention models to generate the final attention curve. Finally, Section III E discusses the extraction of static and dynamic (skims) video summary based on this attention curve.

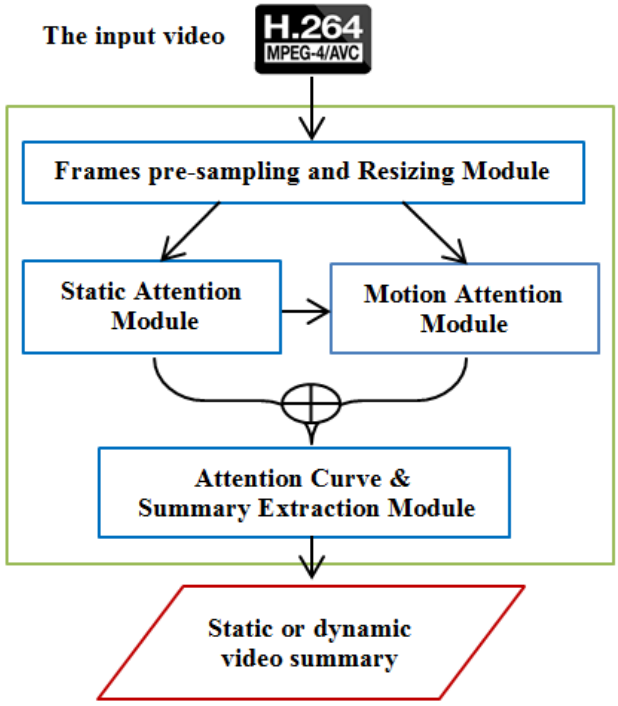


Figure 1. The proposed approach architecture

A. The Proposed Approach Architecture

The goal of this approach is to generate a video summary (static and dynamic) based on VAM for mobile devices. The proposed architecture is shown in Fig.1. It consists of four modules:

1. Frames sampling and resizing, to reduce the computation complexity of the following modules.
2. A static attention module to compute the salience map for each frame of the video samples.
3. A motion attention module to compute motion intensity for each frame of the video samples.
4. Merging both static and motion attention curves to form the final attention curve. Finally, the video summary is extracted based on the attention curve.

The aim of the pre-sampling and resizing module is to avoid the redundant frames and reduce the computational complexity to develop an efficient algorithm. The frame sampling approach is based upon the assumption of having a visual redundancy among consecutive Frames. Therefore, instead of analyzing all the video frames, only some frames are analyzed based on a predefined sampling rate. The sampling rate can be defined as a number of frames per second as in [21,22] or by a frame per a number of frames as in [23]. Based on the sampling size, the number of video frames to be analyzed is reduced. The

shorter the sampling size, the shorter the video summarization time. Nevertheless, the shorter sampling size can lead to loss of important information from the video and thus affect the quality of the summary. Therefore, the sampling size must be defined carefully to keep the important frames [21]. In our approach, the sampling rate is set to one frame per second. After that, each selected frame is resized to be w/4 × h/4 where w and h are the width and height of the original frame.

B. The Static attention module

Static areas in the video may attract the user attention as well as the motion areas. When users watch a video, the interesting static areas (salient areas) can attract them (e.g. the traffic signs on a road) [12]. Therefore, the static attention module was developed to extract the important or interesting frames from the video content. The psychological studies suggest that, HVS is sensitive to the difference between the target areas and its neighborhood. Therefore, the contrasts of color, texture, and shape features are important for visual saliency detection [12,17,24]. Consequently, we applied the generic contrast definition proposed in [17] to compute the color contrast. For each frame F_t, at a time t, the contrast value $Ct_{i,j}$ of a pixel $p_{i,j}$ is computed as in (1).

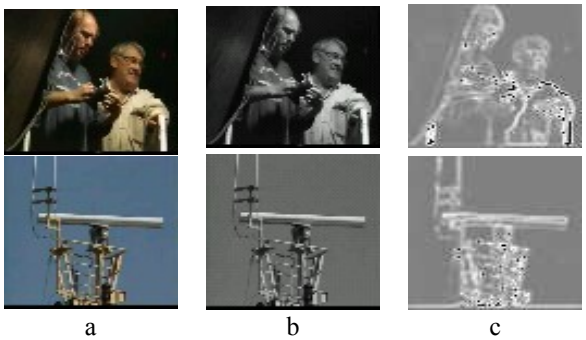
$$Ct_{i,j} = \sum_{q \in N_8(p_{i,j})} d\big(D(p_{i,j}), D(q)\big) \qquad (1)$$

Where $i \in [0, W]$, $j \in [0, H]$ and $W \times H$ is the frame size. The symbol $D(p_{i,j})$ denotes the descriptor at the pixel $p_{i,j}$ (Such as color value) and q is the pixel belongs to 8-neighborhood of $p_{i,j}$ ($N_8(p_{i,j})$). The distance measure (d) between two pixels may be any suitable distance measure. In this approach, d is computed as the Euclidean distance.

The HVS is more sensitive to luminance (gray level) than color [25], and to reduce the complexity. We consider the luminance value of the pixel $p_{i,j}$ as the descriptor. After normalizing all the contrasts at each pixel to [-128, 127], a saliency map is created, as shown in Fig. 2 (c).

A saliency map is a gray image which contains attended/salient areas (bright areas) and unattended/non-salient areas (dark areas). The attended areas usually attract the user attention. In order to extract the attended areas of the saliency map, we use the following method.

Each Saliency Map (*SM*) is divided into non-overlapping Macroblocks (*MB*), each MB is a 2- dimensional vector with size $m \times n$ and represented as ($MB_{i,j} \in R^m \times R^n$, $i \in [0, W]$, $j \in [0, H]$). Where $m \times n$ is the number of pixels in each MB and $W \times H$ is the frame size. Each $MB_{i,j}$ has a location (i, j) defined by the location of the upper left pixel of $MB_{i,j}$ in the *SM*. Accordingly, each *SM* is represented by two sets (*A and U)*. The set *A* is the set of all non-overlapping attended blocks (areas). Similarly, *U* is the set of all non-overlapping unattended blocks (areas). The two sets *A and U* are defined as in equations (2) and (3), respectively.



a b c

Figure 2. Constructed saliency map: (a) original frame, (b) gray level frame and (c) saliency map

Algorithm 1: Static attention detection algorithm

Input: F_t // the input frame at a time t
Output: $P_A(SM_t)$ // the probability of attended areas *A* in SM_t
Start
1. Initialize $A = U = \emptyset$
2. Compute SM_t for F_t
3. Loop for each $MB_{i,j}$ in the SM_t
4. If ($C(MB_{i,j}) \geq \varepsilon^{SM}$) then
5. Add $MB_{i,j}$ to the attended set *A*
6. Else
7. Add $MB_{i,j}$ to the unattended set *U*
8. End loop
9. $P_A(SM_t) = \frac{|A|}{|A|+|U|}$

End

$$A = \{MB_{i,j} \mid C(MB_{i,j}) \geq \varepsilon^{SM}\} \qquad (2)$$

$$U = \{MB_{i,j} \mid C(MB_{i,j}) < \varepsilon^{SM}\} \qquad (3)$$

Such that $C(MB_{i,j})$ denoted the average gray level (brightness) of the pixels in the block $MB_{i,j}$. A threshold ε^{SM} is used to control the membership of $MB_{i,j}$ to a set A or U. According to the algorithm 1, a saliency map is computed and the probability of attended areas *A* in each *SM* for a given threshold ε^{SM} is obtained by (4). Where |A| denotes the cardinality of the set *A*.

$$P_A(SM_t) = \frac{|A|}{|A| + |U|} \qquad (4)$$

After normalizing the value of $P_A(SM_t)$ for each frame to [0, 1], a static attention curve (SC) is obtained, as shown in Fig 3. The horizontal parts on the curve mean that the corresponding frames having the same attended areas probability and almost contain the same information. In the other hand, sudden changes in the curve mean that there is a difference in the content of the corresponding frames. The complexity of the visual static attention detection algorithm (algorithm 1) for each frame is $O\left(\frac{W}{4} \times \frac{H}{4}\right) + O\left(\left(\frac{W}{4} \times \frac{H}{4}\right) \times \frac{1}{D}\right)$, where D is the macroblock size.

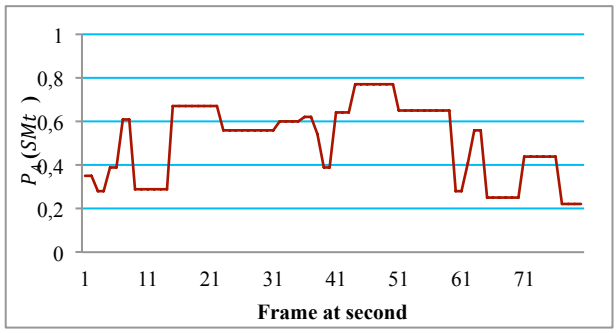


Figure 3. Static attention curve of "The Great Web of Water, segment 01" video

C. The Motion Attention Module

The motion feature is important. It often increases the intensity of users' attention and keeps them locked on significant features and objects [6]. Therefore, most of the visual attention based video summarization approaches are based on the motion attention in different ways [6,12,20,26].

Motion estimation is a classical problem and has a long research history. The two key algorithms for motion estimation are: Optical Flow and Block Matching Algorithms (BMA), which received attention by the researchers because of their simplicity and efficiency. The BMAs are usually less complex than the optical flow algorithms. This is because, the optical flow algorithms are based on pixel processing technique while BMAs are based on a block processing technique [27]. Yaakob et al [28] introduced a comparative study among several BMAs in term of their efficiency and quality. These algorithms are: Full Search (FS), Three Step Search (TSS), Four Step Search (4SS), Diamond Search (DS), HEXagonal Block Search (HEXBS), Multi Directional Gradient Descent Search (MDGDS) and Fast Directional Gradient Descent Search (FDGDS). They concluded that, the FDGDS is a balanced algorithm which produces a high prediction quality and has a low computational cost.

In this approach, we introduce a Fast Directional Motion Intensity Estimation (FDMIE) algorithm. FDMIE is an adapted version of the FDGDS algorithm [29] and was introduced to detect the Motion Intensity (MI) between the consecutive frames. In general, motion estimation is an intensive computation task, especially if it performed for all regions in each frame of a video sequence. However, There are two ways of improving the efficiency of the motion estimation algorithm, one is to decrease the matching points and the other is to choose an efficient blocking matching measure to reduce the complexity [30].

Therefore, in this approach, the motion intensity estimation has been applied to the regions in each frame that could potentially attract users attention due to the motion (i.e. attended areas), hence, decreasing the computational cost significantly. Also, the Sum of Absolute Differences (SAD) is used to determine the matching between two blocks. The SAD is more used because it has a higher quality precision and involves lower computational cost [30,31].

Let $P \in SM_{t-1}$ and $Q \in SM_t$ be two MBs, where SM_t is the current saliency map and SM_{t-1} is the previous one. The SAD between P and Q is defined as in (5).

$$SAD(P,Q) = \sum_{x=0,y=0}^{m,n} |P_{xy} - Q_{xy}| \qquad (5)$$

According to the FDMIE algorithm (algorithm 2), the motion intensity between the saliency maps is computed. For each block in SM_{t-1}, FDMIE computes the current minimum (C_{MIN}) distortion between this block and the corresponding block in SM_t by the equation 5. Then, FDMIE searches the eight directions around the target block (shown in Fig. 4) for the directional minimum (D_{MIN}) distortion. The Relative Distortion Ratio (RDR) between D_{MIN} and C_{MIN} is defined as in (6).

$$RDR(D_{MIN}, C_{MIN}) = \frac{D_{MIN}}{C_{MIN}} \qquad (6)$$

A threshold ε^D is used in FDMIE to control the convergence speed of the algorithm. If RDR is lower than ε^D then other directional searches will be skipped and a new round of search will be started. The FDMIE output is a numeric value that represents the motion intensity of the frame F_{t-1}. After normalizing the motion intensity value for each frame to [0, 1] a motion attention curve (MC) is obtained, as shown in Fig 5. The complexity of FDMIE algorithm (algorithm 2) for each frame is $O\left(|A| \times \left(\frac{W}{4} \times \frac{H}{4}\right) \times \frac{1}{D}\right) \cong O(m)$.

Algorithm 2: FDMIE Algorithm

Input: $A_{t-1}, A_t, SM_{t-1}, SM_t$
Output: MI_{t-1}
Start
For each $P \in SM_{t-1}$, $P \in A_{t-1}$
1. Initialize flag=false
2. Compute $C_{MIN} = SAD(P_{i,j}, Q_{i,j})$
3. **For each** direction around the point with C_{MIN}
   a. Compute $D_{MIN} = SAD(P_{i,j}, Q_{i+di,j+dj})$
   b. If $D_{MIN} < C_{MIN}$
      If $RDR(D_{MIN}, C_{MIN}) < \varepsilon^D$
      Then $C_{MIN} = D_{MIN}$ and go to Step 5.
      Else flag = true
   End for
4. If flag = true then D_{MIN}s are compared. The lowest one is set as C_{MIN} and update the corresponding position, go to step 1.
5. Add the final $MI_{i,j}$ pointing to the position with the C_{Min}, to MI_{t-1}
End For
Return MI_{t-1}
End

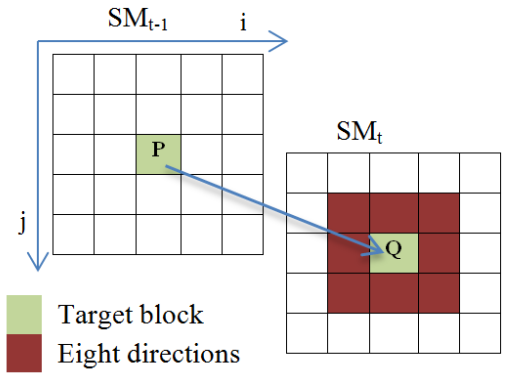


Figure 4. Eight directional searches

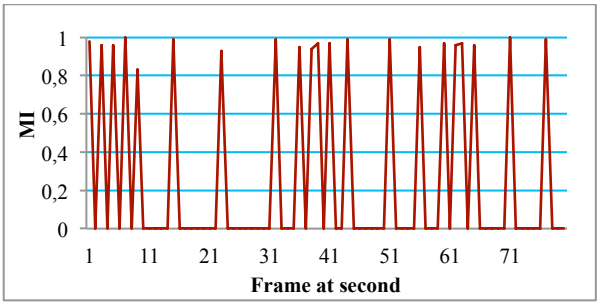


Figure 5. Motion attention curve of "The Great Web of Water, segment 01" video

D. Attention Curve and Summary Extraction

After the static and motion curves are obtained separately, the two curves need to be merged in a meaningful way to construct the final attention curve (AC). In this approach, the final attention curve was constructed based on the linear merged scheme that is defined as in (7).

$$AC = w_s \times SC + w_m \times MC \quad (7)$$

Where SC and MC are normalized [0-1] static and motion attention modules, respectively. w_s and w_m are the weight values for linear combination which satisfy the two conditions in (8).

$$w_s, w_m \geq 0, \quad w_s + w_m = 1 \quad (8)$$

Since the human vision system is more sensitive to motion information than static information [19,32], we chose $w_s = 0.4$ and $w_m = 0.6$. Figure 6 shows an example of the final attention curve that has been created by our approach during the experiments stage.

The attention curve peaks indicate the corresponding video frames which most likely attract users attentions [17]. Based on this curve, static and dynamic (skims) video summary are extracted around the curve peaks. If the length of static summary (number of keyframes) "L" is specified by the user, then the L frames having the highest attention values from the sorted candidate keyframes are selected. If L is unknown then a percentage equal to 5-15 % from the set of candidate frames having the highest attention value are selected.

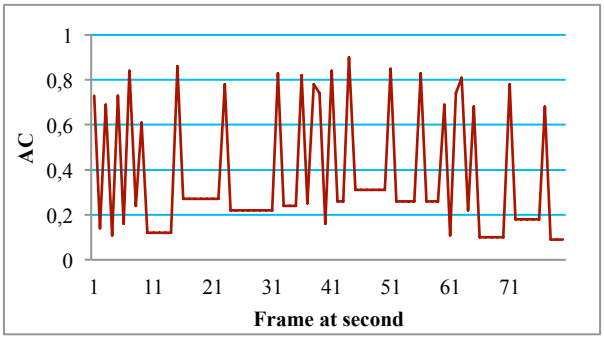


Figure 6. Final attention curve of "The Great Web of Water, segment 01" video

The dynamic video skimming problem can be defined as selecting an optimal set of clips that minimize the distortion between the original video and its skimming [33]. Based on the attention curve, dynamic video skimming generation also becomes much simpler [17]. Given a skimming length or ratio, skim clips are selected around the peaks attention curve. If we have Z pre-sampled frames then the total complexity of our approach is computed as in (9).

$$Z \times \big(O(n) + O(m)\big) \cong O(N) \quad (9)$$

IV. EXPERIMENTAL RESULTS

The Quality of Service (QoS) requirements are essential to multimedia and mobile applications. QoS is commonly defined as the capability of a system to provide better service to users with high degree of a satisfaction. There are several metrics used to evaluate and measure the QoS. They include delay, jitter, packet loss ratio, throughput, error rates and service availability [34,35].

This section presents the experiments of the proposed approach in term of quality and efficiency with a discussion of the results. More QoS metrics will be considered in a subsequent paper.

A. Data set and Testing Devices

This experiment carried out on 5 video files from the standard data set used by many authors and available at the VSUMM web site [36]. The descriptions of these videos are listed in Table I. All videos are in MPEG-1 format with resolution 352×240. Because of the input format to our approach is H.264/AVC, each video is firstly transcoded to H.264/AVC format with resolution of 320×240 to match the standard format of mobile videos.

TABLE I.
DESCRIPTION OF TEST VIDEOS

Video no.	Video name	Duration	#Frames
1	The Great Web of Water, segment 01	00:01:50	3279
2	The Great Web of Water, segment 02	00:01:11	2118
3	Ocean floor Legacy, segment 01	00:00:58	1740
4	Drift Ice as a Geologic Agent, segment 10	00:00:46	1407
5	Exotic Terrane, segment 04	00:02:40	4797

We implemented a prototype to test our approach using an Android platform. Any mobile device with android version 4.0 or higher can run this prototype. Table II, shows the characteristics of the mobile devices that were used in this experiment.

TABLE II.
CHARACTERISTICS OF TESTING MOBILE DEVICES

Mobile phone	CPU	Memory / RAM	Display Resolution	OS
Samsung Galaxy Core Prime	Quad-core 1.2 GHz	8 GB/ 1 GB	480 x 800 pixels	Android, v4.4.4
Samsung Galaxy Grand I9082	Dual-core 1.2 GHz	8 GB/ 1 GB	480 x 800 pixels	Android, v4.2.2

B. Evaluation Strategy

The evaluation strategy is based on the popular metrics of Recall (R), Precision (P) and F-measure (F) [6]. In this strategy, the quality of the automatically generated summary by the approach is compared with the users' (three different users) generated summary of the same video. Then, compute the metrics of R, P and F as in equations (10), (11) and (12) respectively.

$$R = \frac{n_{TM}}{n_{TM} + n_{FN}} \quad (10)$$

$$P = \frac{n_{TM}}{n_{TM} + n_{FP}} \quad (11)$$

$$F = 2 \times \frac{R \times P}{R + P} \quad (12)$$

Where the number of true match frames (n_{TM}) is the number of frames that chosen as key frames both manually and automatically using the new approach. The number

of false positive frames (n_{FP}) is the number of frames that have been chosen as key frames by the approach but not manually. The number of false negative frames (n_{FN}) is the number of frames that have been chosen as key frames manually but not by the approach. The recall metric represents the probability of a relevant key frame to be selected by the approach. Whereas, the precision metric represents the probability that an extracted key frame is relevant. Both recall and precision are complementary metrics and the highest summary quality was achieved when high values for both metrics are achieved. So that, F-measure is the averages of recall and precision metrics, the highest value of F-measure led to the highest summary quality.

C. Quality Evaluation

In order to evaluate the quality of the proposed approach, we compare it with other static video summary approaches. The compared approaches include Video SUMMarization (VSUMM) [21], and STIll and MOving video storyboard for the web scenario (STIMO) [37] which are non-visual attention based video summarization approaches. Also, we compare the proposed approach with other visual attention based video summarization approaches. They include Lai and Yi [12] and Ejaz et al [6].

The comparative results are provided in Table III and an example is shown in Fig. 7. The results demonstrate that, the proposed approach achieved an average F-measure of 0.87 with respect to the manually generated summary. Moreover, the results indicated that, the proposed approach has high values for both R and P metrics in comparison with the other approaches.

TABLE III.
THE RECALL (R), PRECISION (P) AND F-MEASURE (F) OF DIFFERENT TECHNIQUES

No	*STIMO*			*VSUMM*			*Lai &Yi*			*Ejaz et al*			*Proposed*		
	R	***P***	***F***	***R***	***P***	***F***	***R***	***P***	***F***	***R***	***P***	***F***	***R***	***P***	***F***
1	0.67	0.41	0.51	0.67	0.51	0.58	0.8	0.83	0.81	0.82	0.75	0.78	0.8	0.89	0.85
2	0.77	0.48	0.59	0.62	0.67	0.64	0.85	0.75	0.8	0.9	0.85	0.87	0.83	1	0.91
3	0.61	0.33	0.43	0.41	0.55	0.47	0.83	0.8	0.81	0.83	0.8	0.81	0.75	1	0.86
4	0.85	0.8	0.82	0.87	0.82	0.84	0.83	0.85	0.84	0.9	0.87	0.88	0.83	1	0.91
5	0.5	0.42	0.45	0.92	0.75	0.82	0.85	0.83	0.84	0.83	0.8	0.82	0.76	0.87	0.81
Average	**0.68**	**0.49**	**0.56**	**0.7**	**0.66**	**0.67**	**0.83**	**0.81**	**0.82**	**0.86**	**0.81**	**0.83**	**0.79**	**0.95**	**0.87**

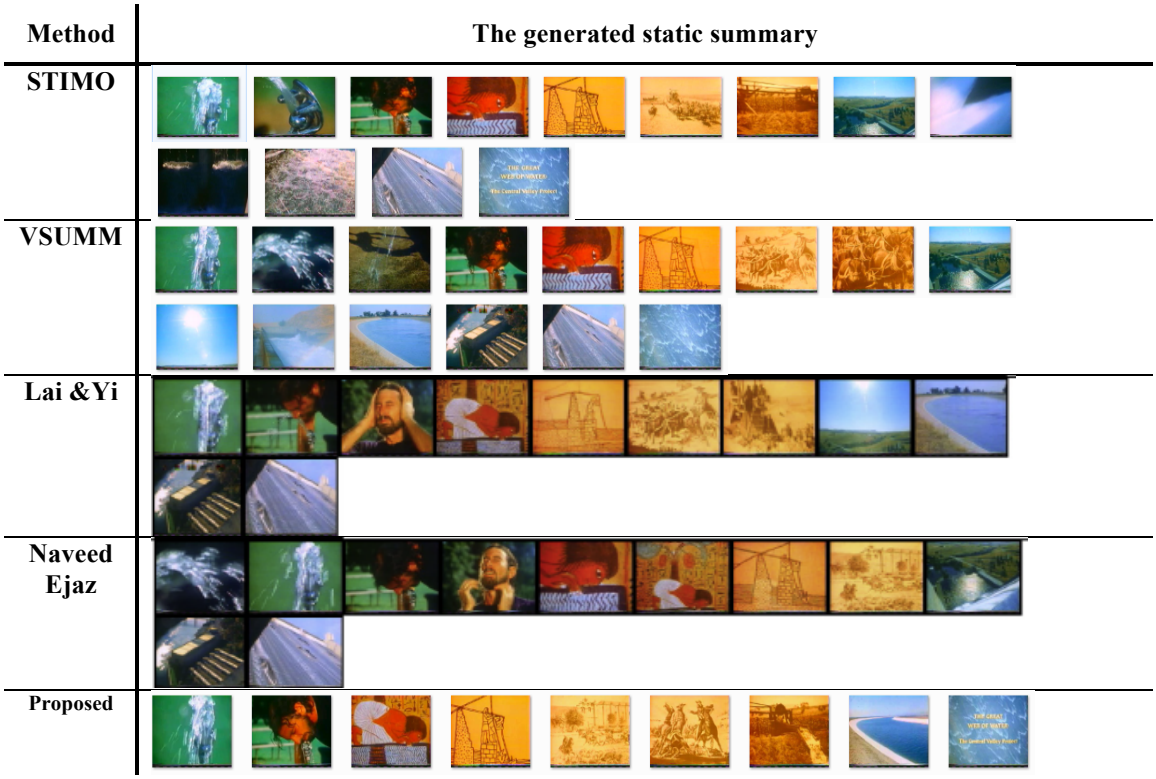


Figure 7. Comparison of static summary extraction for video "The Great Web of Water, segment 01"

D. *Efficiency Evaluation*

Efficiency evaluation is an important issue when comparing similar approaches. The source codes of most of the video summarization approaches are not available, and the time complexity required for producing a video summary (static or skimming) depends on a particular hardware and the adopted features, it is almost impossible to produce a fair evaluation in terms of efficiency among these approaches [11]. Therefore the efficiency of the proposed approach is evaluated by counting the number of frames that can be processed per second. This includes the partial decoding/encoding time of each frame. This study was carried on the first mobile phone described in Table I and on all mentioned videos in Table II. According to those experiments, the proposed approach can process an average of 14 FPS, as shown in Table IV. For online applications, based on a maximum waiting time of 39s [38]. The proposed approach can process an average 546 frames in 39s. With sampling rate equal to 1 FPS. Our approach can be used for videos of duration up to 9 min (about 16200 frames at 30 FPS). Therefore, the proposed approach can be used for online applications with video segmentation and initial small delay. It is important to note that those results depend on the computational power of the target mobile device.

TABLE IV.
TIME EFFICIENCY EVALUATION

Video no.	Duration in second	# of Frames	# of Samples	Total time(s)	FPS
1	110	3279	110	8	13.75
2	71	2118	71	5	14.2
3	58	1740	58	4	14.5
4	46	1407	46	3	15.33
5	160	4797	160	14	11.43
Average					**13.84**

V. CONCLUSIONS

This paper proposes an effective and efficient video summarization approach which is suitable for mobile device usage and online applications. This approach is summarized as follows. Firstly, the static attention module is applied to generate a static attention curve. Secondly, we introduce a Fast Directional Motion Intensity Estimation (FDMIE) algorithm to calculate the motion intensity between consecutive frames. Then, the motion intensity values are used to construct a motion attention curve. Thirdly, the static and motion attention curves are merged together to form a final attention curve. Finally, static and dynamic video summary is extracted based on this attention curve. Our evaluation is experimental. We measure the quality and efficiency of our approach and compare it with other similar approaches. It is shown that our approach has a high quality (up to 87%) and efficiency with respect to the similar approaches.

In the future, we intend to build a content aware video summarization and streaming based on the proposed approach. For this, the QoS requirements will have to be taken into consideration.

REFERENCES

[1] W. Gao, Q.-m. Huang, S.-q. Jiang, and P. Zhang, "Sports video summarization and adaptation for application in mobile communication," *Journal of Zhejiang University SCIENCE A,* vol. 7, no. 5, pp. 819-829, 2006. http://dx.doi.org/10.1631/jzus.2006.A0819

[2] N. V. Uti, and R. Fox, "The Challenges of Compressing and Streaming Real Time Video Originating from Mobile Devices," *Multimedia Services and Streaming for Mobile Devices: Challenges and Innovation*, pp. 1-24: IGI Global, 2011.

[3] W. Wang, and M. R. Lyu, "Automatic generation of dubbing video slides for mobile wireless environment," in Proceedings of the International Conference on Multimedia and Expo (ICME'03), IEEE, 2003. http://dx.doi.org/10.1109/icme.2003.1220905

[4] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, "Video summarization: techniques and classification," *Computer Vision and Graphics, Springer,* vol. 7594, pp. 1-13, 2012. http://dx.doi.org/10.1007/978-3-642-33564-8_1

[5] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang, *A unified framework for video summarization, browsing & retrieval: with applications to consumer and surveillance video*: Academic Press, 2006. http://dx.doi.org/10.1016/b978-012369387-7/50009-5

[6] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication,* vol. 28, no. 1, pp. 34-44, 2013. http://dx.doi.org/10.1016/j.image.2012.10.002

[7] H. Karray, M. Ellouze, and A. Alimi, "Indexing video summaries for quick video browsing," *Pervasive Computing*, pp. 77-95: Springer, 2010.

[8] 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. http://dx.doi.org/10.1016/j.jvcir.2007.04.002

[9] B. T. Truong, and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP),* vol. 3, no. 1, pp. 37, 2007. http://dx.doi.org/10.1145/1198302.1198305

[10] R. Pal, A. Ghosh, and S. K. Pal, "Video Summarization and Significance of Content: A Review," *Handbook on Soft Computing for Video Surveillance*, pp. 79-102: CRC Press, 2012. http://dx.doi.org/10.1201/b11631-5

[11] H. Farouk, K. ElDahshan, and A. Abozeid, "The State of the Art of Video Summarization for Mobile Devices: Review Article," *Graphics, Vision and Image Processing GVIP,* vol. 14, no. 2, pp. 37-50, 2014.

[12] 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. http://dx.doi.org/10.1016/j.jvcir.2011.08.005

[13] I. Mehmood, M. Sajjad, and S. W. Baik, "Visual attention based extraction of semantic keyframes," *Advances in Information Science and Applications,* vol. 1, 2014.

[14] J. Li, and W. Gao, *Visual Saliency Computation: A Machine Learning Perspective*, p.^pp. 1, 215: Springer Publishing Company, Incorporated, 2014.

[15] S. Filipe, and L. A. Alexandre, "From the human visual system to the computational models of visual attention: a survey," *Artificial Intelligence Review,* vol. 39, no. 1, pp. 1-47, 2013.

[16] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in Proceedings of the tenth ACM international conference on Multimedia, 2002, pp. 533-542. http://dx.doi.org/10.1145/641007.641116

[17] 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," *Multimedia, IEEE Transactions on,* vol. 7, no. 5, pp. 907-919, 2005. http://dx.doi.org/10.1109/TMM.2005.854410

[18] J. Peng, and Q. Xiao-Lin, "Keyframe-based video summary using visual attention clues," *IEEE MultiMedia*, no. 2, pp. 64-73, 2010.

[19] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan, "Static saliency vs. dynamic saliency: a comparative study." pp. 987-996.

[20] N. Ejaz, I. Mehmood, and S. W. Baik, "Feature aggregation based visual attention model for video summarization," *Computers & Electrical Engineering,* vol. 40, no. 3, pp. 993-1005, 2014. http://dx.doi.org/10.1016/j.compeleceng.2013.10.005

[21] S. E. F. de Avila, and A. P. B. Lopes, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters, Elsevier* vol. 32, no. 1, pp. 56-68, 2011. http://dx.doi.org/10.1016/j.patrec.2010.08.004

[22] E. Asadi, and N. M. Charkari, "Video summarization using fuzzy c-means clustering," in 2012 20th Iranian Conference on Electrical Engineering (ICEE), 2012, pp. 690-694. http://dx.doi.org/10.1109/iraniancee.2012.6292442

[23] S. CVETKOVIC, M. JELENKOVIC, and S. V. NIKOLIC, "Video summarization using color features and efficient adaptive threshold technique," *Przegląd Elektrotechniczny,* vol. 89, 2013.

[24] L. Xu, H. Li, L. Zeng, and K. N. Ngan, "Saliency detection using joint spatial-color constraint and multi-scale segmentation," *Journal of Visual Communication and Image Representation,* vol. 24, no. 4, pp. 465-476, 2013. http://dx.doi.org/10.1016/j.jvcir.2013.02.007

[25] I. E. Richardson, *The H. 264 advanced video compression standard*: John Wiley & Sons, 2011.

[26] A. B. Mejía-Ocaña, M. de-Frutos-López, S. Sanz-Rodríguez, O. del-Ama-Esteban, C. Peláez-Moreno, and F. Díaz-de-María, "Low-complexity motion-based saliency map estimation for perceptual video coding," in Telecommunications (CONATEL), 2011 2nd National Conference on, 2011, pp. 1-6.

[27] J. T. Philip, B. Samuvel, K. Pradeesh, and N. Nimmi, "A comparative study of block matching and optical flow motion estimation algorithms," in Emerging Research Areas: Magnetics, Machines and Drives (AICERA/iCMMD), 2014 Annual International Conference on, 2014, pp. 1-6.

[28] R. Yaakob, A. Aryanfar, A. A. Halin, and N. Sulaiman, "A Comparison of Different Block Matching Algorithms for Motion Estimation," *Procedia Technology,* vol. 11, pp. 199-205, 2013. http://dx.doi.org/10.1016/j.protcy.2013.12.181

[29] L.-M. Po, K.-H. Ng, K.-W. Cheung, K.-M. Wong, Y. M. S. Uddin, and C.-W. Ting, "Novel directional gradient descent searches for fast block motion estimation," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 19, no. 8, pp. 1189-1195, 2009. http://dx.doi.org/10.1109/TCSVT.2009.2020320

[30] Q. YANG, C. LI, and Z. LI, "Motion Navigation System Estimation Algorithm in Mobile Phone Video Learning System⋆," *Journal of Computational Information Systems,* vol. 10, no. 16, pp. 7187-7194, 2014.

[31] M. Santamaria, and M. Trujillo, "A comparison of block-matching motion estimation algorithms," in Computing Congress (CCC), 2012 7th Colombian, 2012, pp. 1-6. http://dx.doi.org/10.1109/colombiancc.2012.6398002

[32] B. Wu, L. Xu, L. Zeng, Z. Wang, and Y. Wang, "A unified framework for spatiotemporal salient region detection," *EURASIP Journal on Image and Video Processing,* vol. 2013, no. 1, pp. 1-12, 2013. http://dx.doi.org/10.1186/1687-5281-2013-16

[33] C. T. Dang, and H. Radha, "Heterogeneity image patch index and its application to consumer video summarization," *IEEE TRANSACTIONS ON IMAGE PROCESSING,* vol. 23, no. 6, pp. 2704-2718, 2014. http://dx.doi.org/10.1109/TIP.2014.2320814

[34] H. Luo, and M.-L. Shyu, "Quality of service provision in mobile multimedia-a survey," *Human-centric computing and information sciences,* vol. 1, no. 1, pp. 1-15, 2011.

[35] V. P. H, and G. S. N. , "A Survey on Quality of Service Provision in 4G Wireless Networks," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 3, no. 7, 2014.

[36] "VSUMM (Video SUMMarization)," 6, 2015; https://sites.google.com/site/vsummsite.

[37] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario," *Multimedia Tools and Applications,* vol. 46, no. 1, pp. 47-69, 2010. http://dx.doi.org/10.1007/s11042-009-0307-7

[38] J. Almeida, N. J. Leite, and R. d. S. Torres, "Online video summarization on compressed domain," *Journal of Visual Communication and Image Representation,* vol. 24, no. 6, pp. 729-738, 2013. http://dx.doi.org/10.1016/j.jvcir.2012.01.009

AUTHORS

Hesham Farouk is with Computers and Systems Dept., Electronics Research Institute, Cairo, Egypt.

Kamal A. ElDahshan is with Dept. of Mathematics, Computer Science Division, Faculty of Science, Al-Azhar University, Cairo, Egypt.

Amr Abozeid is with Dept. of Mathematics, Computer Science Division, Faculty of Science, Al-Azhar University, Cairo, Egypt.

Submitted 24 June 2015. Published as resubmitted by the authors 03 September 2015.