Visual attention based extraction of semantic keyframes

Irfan Mehmood, Muhammad Sajjad, Sung Wook Baik*

***Abstract*—**The amount of video data available on the internet and personal devices is increasing exponentially due to revolution of consumer devices, social media and web. To extract the desired information from such a huge video repository in a minimal span of time is a challenging task. Keyframe extraction is an enthusiastic research field that manages video data and provides succinct representation of videos for efficient browsing and retrieval tasks. Various existing keyframe extraction methods utilizes low-level features that results in the loss of semantic details. This paper presents a visual saliency driven framework for keyframe extraction that provides concise versions of video by extracting semantically relevant frames. The proposed visual saliency model helps to bridge the gap between low-level features and high-level information. The visual saliency model is build using static and dynamic saliency maps. The static saliency is derived from color opponent component space using center surround measure. The dynamic saliency is determined using motion intensity and its phase coherence. Then a two dimensional visual saliency curve is estimated by fusing static and dynamic saliency maps. Finally, peak points are calculated in the visual saliency curve that leads to the extraction of the keyframes. Based on different evaluation principles, experimental results demonstrate that the proposed technique successfully extracts semantically significant key frames according to the dynamics of video.

***Keywords*—** Keyframe extraction, visual attention model, video summary evaluation.

I. Introduction

Video is a composite of image sequence, audio tracks, and textual information that conveys information with their own primary elements. Recent growth of video capturing devices, data storage and transmission facilities have resulted into large video libraries. Managing such a huge amount of video data is quite difficult as compared to other forms of media like text and audio. This motivated the researchers to develop systems that are capable of tackling video data in an efficient manner [1]. A basic approach for managing video data is video summarization [2]. Video summarization aims to reduce the amount of redundant data in order to extract the most salient information known as keyframe of the video. The resultant video summary represents the most significant video content. It enables viewers to quickly figure out the important contents of video and help them in searching and retrieval tasks [3]. But to generate a video summary, a full understanding of video is required, which is very difficult for contemporary machines. In video summarization, attaining the effective content of a video is important yet an unexplored aspect. Therefore, it is desired to develop such a framework, which presents the effective elements in video data to the user by considering the semantic information [4].

In literature, many low level approaches have been used for the generation of video summaries but they are not affective as they are inconsistent with the human perception [5, 6]. We want to bridge this gap between low level features and human perception by taking into account the human visual attention. Human visual attention plays an important role in selecting and integrating vital information [7, 8]. An efficient framework designed from the point of view of human perception is provided using a biologically-inspired visual attention model in order to provide semantically relevant summaries.

II. Methodology

The bedrock of the proposed framework is the concept of visual attention modeling. Initially the static and dynamic visual saliencies are computed for each frame of the video. Then, an attention value is obtained for each of the visual attention clues. The static and dynamic attention values are fused to obtain an aggregated attention value for each frame. The aggregated attention value of each frame in the video is used to make an attention curve of the video. Finally, the key frames are extracted by finding peak points in the attention curve. Main steps of the proposed framework are shown in Figure 1.

A. Static Visual Attention Model

Consider a video V= {fi; i=1, 2, 3…, N}; where N denotes the total number of frames. These frames are in RGB color space. In human cognitive process, color plays a vital role in the analysis of video contents. Color opponent component (COC) is an efficient color space for improved video perception [9, 10]. To incorporate this cognitive property in our system, RGB color space is converted to COC space.

Irfan Mehmood is with the Digital Contents Research Institute Sejong University Seoul, Korea. (E-mail: irfanmehmood@sju.ac.kr).

Muhammad Sajjad Mehmood is with the Digital Contents Research Institute Sejong University Seoul, Korea. (E-mail: sajjad@sju.ac.kr).

Sung Wook Baik is with the Digital Contents Research Institute Sejong University Seoul, Korea. (E-mail: sbaik@sejong.ac.kr).

* Corresponding author. Tel.: +82-02-3408-3797; fax: 02-3408-4339.

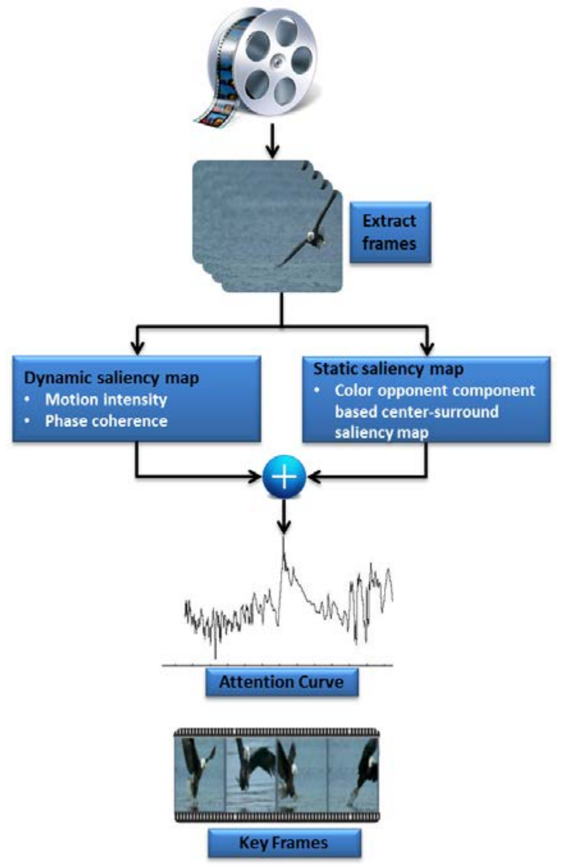


Fig. 1: Framework of the proposed keyframe extraction system

For this purpose, RGB color channels are converted into four-broadly tuned color channels as $R_i=r_i-(g_i+b_i)/2$, $G_i=g_i-(r_i+b_i)/2$, $B_i=b_i-(r_i+g_i)/2$, and $Y_i=(r_i+g_i)/2-|r_i-g_i|/2-b_i$. Then color opponent channels are estimated as:

$$RG_i = R_i - G_i \qquad (1)$$

$$BY_i = B_i - Y_i \qquad (2)$$

Intensity channel is also computed and fused with red-green and blue-yellow channels to get an aggregated image F_i as given in equations 3 and 4.

$$I_i = \left(\frac{r_i + g_i + b_i}{3} \right) \qquad (3)$$

$$F_i = RG_i + BY_i + I_i \qquad (4)$$

In this aggregated image F_i, the spatial distribution of color and its spatial position are important factor that contributes to the detection of salient regions in video frames. It has been observed in various studies that salient objects are generally surrounded by non-salient background regions that usually scatter over the entire visual scene [11]. Moreover, objects near the center of the scene are more likely to catch human's attention. These two considerations have led us to the calculation of static saliency map. First, divide F_i into n number of blocks. Now consider an image block B_i, its saliency with respect to other blocks is calculated as:

$$S_S^i = \frac{\sum_{i=1}^{n} e^{1-d_j} \times \frac{1}{D_{i,j}}}{\sum_{i=1}^{n} \frac{1}{D_{i,j}}} \qquad (5)$$

here d_j is the Euclidean distance between image block B_i and image's center. $D_{i,j}$ is the distance between image blocks B_i and B_j. This center-surround efficiently estimates saliency map by uniformly highlighting the salient objects. After the computation of the saliency map of the whole image, the average of non-zero values is taken to compute the static attention value of frame.

B. Dynamic Visual Attention Model

In case of videos, human cognitive process is more concerned about objects and motion among them[7, 12]. Therefore, in videos, motion is also a key feature in building human attention model. Furthermore, orientation is an important factor because motion of salient objects is usually associated with consistent orientation. Thus, we have built a saliency model using descriptors motion intensity and phase coherence. Motion vector field V is computed using the Horn-Schunck Method [13]. Motion intensity I_M and phase coherence O are computed from motion vector and are fused to get a dynamic saliency map as:

$$I_M = \sqrt{V_x^2 + V_y^2} \qquad (6)$$

$$O = Tan^{-1}\left(\frac{V_y}{V_x} \right) \qquad (7)$$

$$S_D = I_M + O \qquad (8)$$

where V_x and V_y represents the x and y component of the motion vector V. Final saliency map is obtained by lineally fusing the static and dynamic saliency and normalizing in the range [0 1]. Some results of final saliency maps are shown in figure 2. The average of non-zero pixel values in each saliency map is considered as a saliency value. These saliency values are then computed to from a visual saliency curve. From this saliency curve, keyframes are selected by mapping the peak saliency value to their corresponding video frames.

III. EXPERIMENT AND RESULTS

We have evaluated the proposed summarization framework on videos downloaded from a standard dataset Open Video Project[1], the detail of these videos is shown in table 1. Comparison is done with two different types of video summarization schemes.

[1] http://www.open-video.org/

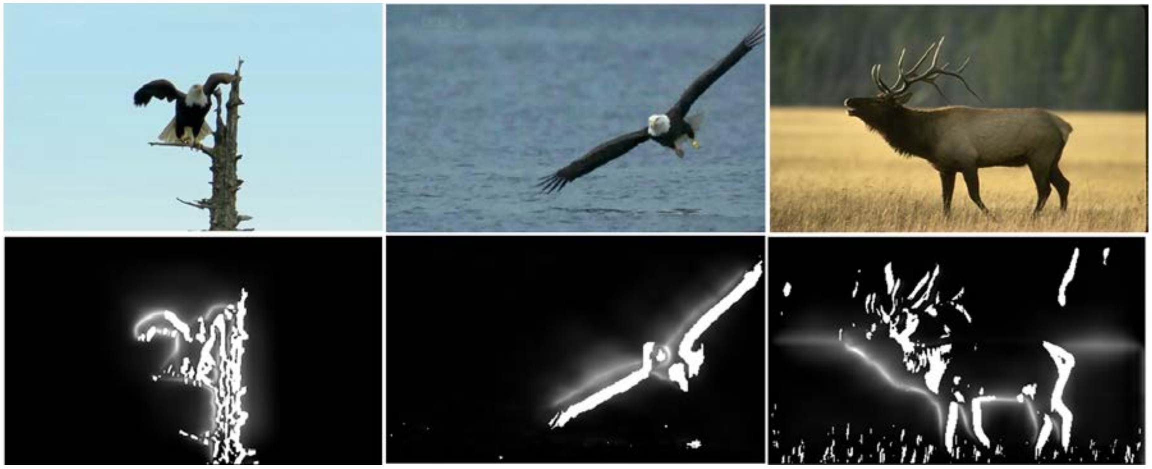


Fig. 2: First row shows original images and bottom row shows their corresponding saliency maps

Table 1: Details of Test Videos

No.	Video Name	Number of Frames
1	Wetlands Regained, segment 03 of 8	3562
2	Technology at Home: A Digital Personal Scale	3346
3	The Great Web of Water, segment 01	3279
4	The Great Web of Water, segment 02	2118
5	A New Horizon, segment 02	1797
6	A New Horizon, segment 06	1944
7	The Future of Energy Gases, segment 05	3615
8	The Future of Energy Gases, segment 09	1884
9	Drift Ice as a Geologic Agent, segment 05	2187
10	Drift Ice as a Geologic Agent, segment 10	1407

A. Comparison with Non-Visual Attention Based Video Summarization Techniques

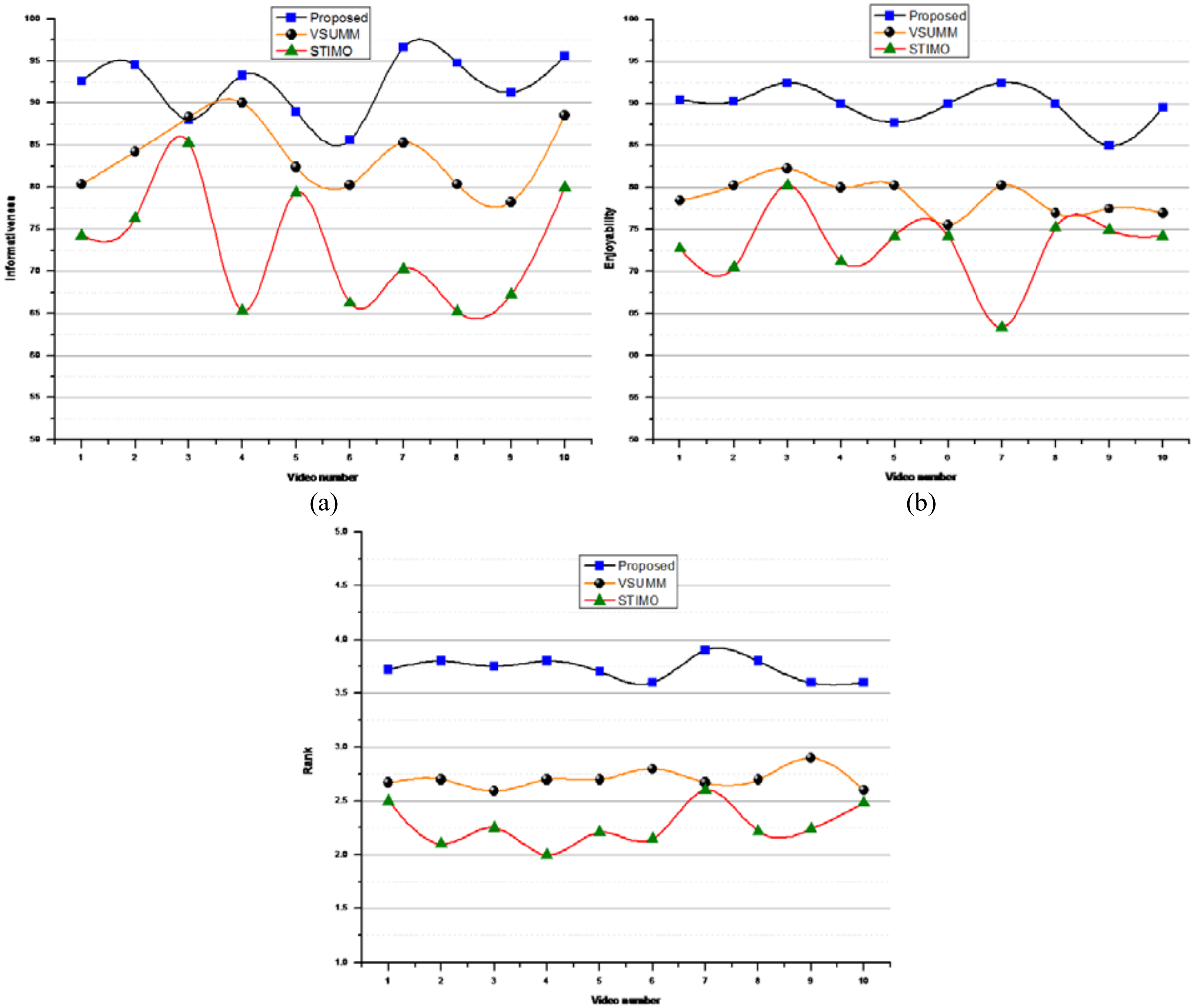
Here we have discussed the results of our proposed method with non-visual attention schemes presented in literature. These schemes are STIMO [14] and VSUMM [15]. Method we used for evaluation is based on subjective rating by a group of users. The users are asked to score each video in range [0 100] on the basis of three parameters Informativeness, Enjoyability and ranking. Informativeness measures the ability to maintain all salient content coverage by reducing redundancy; on the other hand Enjoyability measures the performance for selecting perceptually pleasing summaries for video segments. Rank is the overall satisfaction of user with summary contents, scores are assigned for each summary in the range of 0 to 5 and then these ranking score for each summary are averaged to get a single measure, results are shown in table 2. The average results of the three measures indicate that our technique outperforms other mentioned techniques. The average scores assigned by users for Informativeness and Enjoyability for the proposed method are 92.12 and 89.8, respectively that is the highest among competitors. In addition, the score for Rank is also maximum for the proposed scheme. In figure 3, graphical representation of Table 2 have been added to assist the reader in easily understanding the results.

B. Comparison with Visual Attention Based Video Summarization Techniques

In this section comparison is done between the proposed technique with two latest visual attentions modeling based key frame extraction schemes proposed by Peng and Xiaolin [16] and Ejaz et al. [2]. Initially, the results of three techniques are presented on single shot of a video. This video shows an American Bald eagle hunting a salmon on sea surface. The test sequence of frames is from 361 to 605 of the video “Fragment from BBC Nature's Great Events - The Great Salmon Run”. In this shot an eagle dives in the water to catch a salmon and successfully hunt the salmon. In this shot a key frame is one, which shows the eagle flying from surface of sea after hunting a salmon and holding it in his paws. Figure 4 shows the key frames extracted by the proposed scheme, [16] and [2] with frame 512, 384 and 472 respectively. Frame 512 is the key frame conveying more information to user as compared to frames 384 and 472. It is obvious that key frames extracted by other underlying techniques do not express the view of successful hunting of fish by the eagle; thus not semantically representative.

Table 2: Informativeness (I), Enjoyability (E) and Rank (R) score of different methods on video data set

	STIMO			VSUMM			Proposed		
No.	**I**	**E**	**R**	**I**	**E**	**R**	**I**	**E**	**R**
1	74.25	72.75	2.5	80.36	78.5	2.67	92.56	90.5	3.72
2	76.36	70.5	2.1	84.23	80.25	2.7	94.52	90.25	3.8
3	85.25	80.26	2.25	88.32	82.25	2.59	88	92.5	3.75
4	65.32	71.25	2	90	80	2.7	93.33	90	3.8
5	79.36	74.25	2.21	82.36	80.25	2.7	89	87.75	3.7
6	66.25	74.21	2.15	80.25	75.58	2.8	85.62	90	3.6
7	70.3	63.32	2.6	85.26	80.25	2.67	96.6	92.5	3.9
8	65.25	75.25	2.22	80.32	77	2.7	94.75	90	3.8
9	67.25	75	2.24	78.25	77.5	2.9	91.25	85	3.6
10	79.95	74.25	2.48	88.5	77	2.6	95.6	89.5	3.6
Average	72.95	73.10	2.275	83.78	78.85	2.7	92.12	89.8	3.72

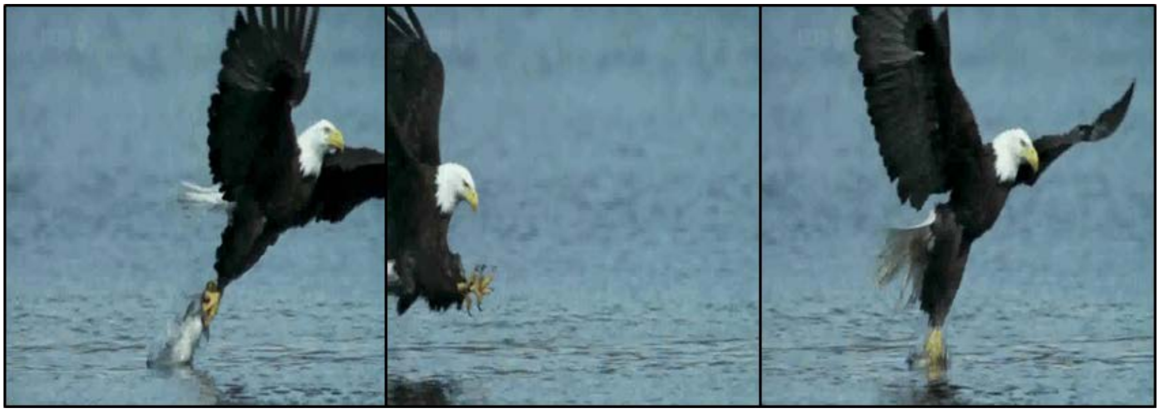


(a)

(b)

(c)

Fig. 3: Informativeness (I), Enjoyability (E) and Ranking (R) curves of different methods on video data set



Frame Number 512 Frame Number 384 Frame Number 472

Figure 4: key frames extracted by the proposed scheme, [16] and [2] for the video 'Fragment from BBC Nature's Great Events - The Great Salmon Run'

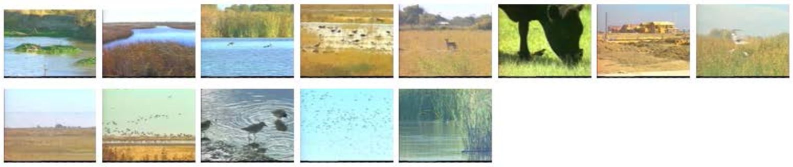
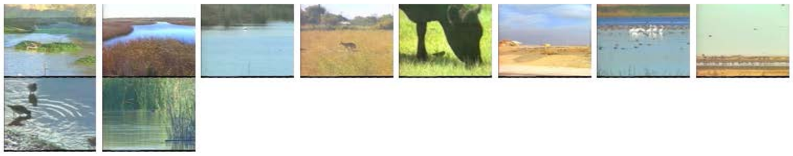
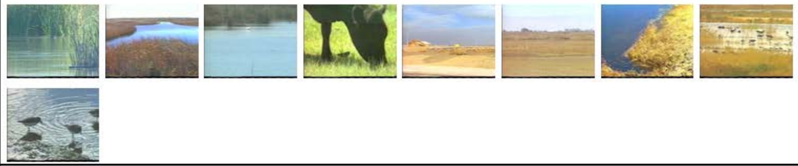
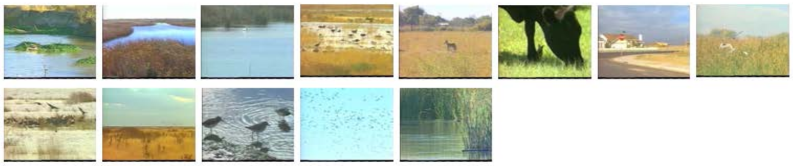
Method	Generated Key Frames
Ground Truth	
[2]	
[12]	
Proposed	

Figure 5: Comparison of key frame extraction for video 'Wetlands Regained, segment 03 of 8'

On the other hand, the key frame extracted by the proposed scheme draw attention and summarizes the shot accurately. Figure 5 shows the key frames for another video 'Wetlands Regained, segment 03 of 8'. This also shows that our scheme yields results closer to the Ground Truth key frames.

IV. CONCLUSION

In this paper, we have presented a method for static and dynamic visual saliency computation and investigate the impending of their fusion for generating videos summaries. This work uses a biological inspired model of saliency which considers different important features such as color contrast, motion intensity and motion orientation between consecutive frames. We believe that the proposed video summarization leads to a more informative and enjoyable summaries for the users. A simple fusion method is used for combining static and dynamic attention values and from this attention curve; key frames are extracted by finding the highest saliency points between two consecutive frames. The proposed evaluation method shows that the extracted key frames are semantically important and closer to the human perceptions as compared to other techniques. In future, we will consider audio features to create video skimming to provide more attractive, natural and informative video summaries.

ACKNOWLEDGMENT

This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2012904).

REFERENCES

[1] C. Chen, C.-Y. Zhang, Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data, Information Sciences, (2014).

[2] N. Ejaz, I. Mehmood, S. Wook Baik, Efficient visual attention based framework for extracting key frames from videos, Signal Processing: Image Communication, 28 (2013) 34-44.

[3] Q. Xu, Y. Liu, X. Li, Z. Yang, J. Wang, M. Sbert, R. Scopigno, Browsing and exploration of video sequences: A new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence, Information Sciences, 278 (2014) 736-756.

[4] B. Lu, G. Wang, Y. Yuan, D. Han, Semantic concept detection for video based on extreme learning machine, Neurocomputing, 102 (2013) 176-183.

[5] N. Ejaz, T.B. Tariq, S.W. Baik, Adaptive key frame extraction for video summarization using an aggregation mechanism, Journal of Visual Communication and Image Representation, 23 (2012) 1031-1040.

[6] N. Ejaz, S.W. Baik, Video summarization using a network of radial basis functions, Multimedia systems, 18 (2012) 483-497.

[7] M. Guo, Y. Zhao, C. Zhang, Z. Chen, Fast Object Detection Based on Selective Visual Attention, Neurocomputing, (2014).

[8] J.L. Orquin, S. Mueller Loose, Attention and choice: a review on eye movements in decision making, Acta psychologica, 144 (2013) 190-206.

[9] R. Shapley, M.J. Hawken, Color in the cortex: single-and double-opponent cells, Vision research, 51 (2011) 701-717.

[10] L. Dong, W. Lin, Y. Fang, S. Wu, H.S. Seah, Saliency detection in computer rendered images based on object-level contrast, Journal of Visual Communication and Image Representation, 25 (2014) 525-533.

[11] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33 (2011) 353-367.

[12] S. Spotorno, B.W. Tatler, S. Faure, Semantic consistency versus perceptual salience in visual scenes: Findings from change detection, Acta psychologica, 142 (2013) 168-176.

[13] B.K. Horn, B.G. Schunck, "Determining optical flow": a retrospective, Artificial Intelligence, 59 (1993) 81-87.

[14] M. Furini, F. Geraci, M. Montangero, M. Pellegrini, STIMO: STIll and MOving video storyboard for the web scenario, Multimedia Tools and Applications, 46 (2010) 47-69.

[15] S.E.F. de Avila, A.P.B. Lopes, VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters, 32 (2011) 56-68.

[16] J. Peng, Q. Xiao-Lin, Keyframe-based video summary using visual attention clues, IEEE MultiMedia, 17 (2010) 0064-0073.

Irfan Mehmood
received his BS degree in Computer Science from National University of Computer and Emerging Sciences from Pakistan. He is currently pursuing his PhD degree at Sejong University, Seoul, Korea. His research interests include video summarization, prioritization of medical images and brain tumor segmentation.

Muhammad Sajjad
received his MS degree in Computer Software Engineering from National University of Sciences and Technology, Pakistan. He is currently pursuing PhD course in Sejong University, Seoul, Korea. His research interests include super-resolution and reconstruction, Sparse coding, video summarization and mixed reality.

Sung Wook Baik
is a professor in the College of Electronics and Information Engineering at Sejong University. His research interests include Computer vision, Pattern recognition, and Data mining. He has a PhD in Information Technology and Engineering from George Mason University.