Signal Processing: *Image Communication* 28 (2013) 34–44

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Efficient visual attention based framework for extracting key frames from videos

Naveed Ejaz, Irfan Mehmood, Sung Wook Baik*

College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 26 March 2012
Accepted 8 October 2012
Available online 17 October 2012

Keywords:
Video summarization
Key frame extraction
Visual attention model
Visual saliency

ABSTRACT

The huge amount of video data on the internet requires efficient video browsing and retrieval strategies. One of the viable solutions is to provide summaries of the videos in the form of key frames. The video summarization using visual attention modeling has been used of late. In such schemes, the visually salient frames are extracted as key frames on the basis of theories of human attention modeling. The visual attention modeling schemes have proved to be effective in video summarization. However, the high computational costs incurred by these techniques limit their applicability in practical scenarios. In this context, this paper proposes an efficient visual attention model based key frame extraction method. The computational cost is reduced by using the temporal gradient based dynamic visual saliency detection instead of the traditional optical flow methods. Moreover, for static visual saliency, an effective method employing discrete cosine transform has been used. The static and dynamic visual attention measures are fused by using a non-linear weighted fusion method. The experimental results indicate that the proposed method is not only efficient, but also yields high quality video summaries.

1. Introduction

Due to the advances in multimedia technology, the amount of the generated video data is increasing rapidly, and consumer multimedia storage devices are becoming increasingly popular [1]. This large volume of video data requires efficient video content management schemes for endowing a better overall multimedia experience to the consumer. A possible video data management scheme is to generate summaries of the videos to provide browsing capabilities to the users [2]. Apart from browsing, video summaries can also help users quickly locate a semantically relevant position in a video [3].

Video skimming and key frame extraction are the two basic methods for summarizing videos [2]. In video skimming based summarization strategies, a video of much shorter duration than the actual video is produced. The key frame extraction techniques, on the other hand, generate summaries by extracting salient frames from the videos. Generally, the video skims are more expressive and entertaining than that of the key frames. However, the key frames do not suffer from timing and synchronization issues and can be used in a variety of ways for browsing and navigation [2,4,5]. Moreover, for small devices, key frames may provide better browsing capabilities than video skims because they enable users to quickly go through the salient contents of the video in one gaze fixation without watching even a small video [2,3]. In this paper, the focus is towards extracting key frames from the videos.

The video summarization schemes must preferably be based on the high level semantic contents of the video like

* Corresponding author. Tel.: +82 02 3408 3797; fax: +82 02 3408 4339
E-mail addresses: naveed3797@gmail.com, naveed@sju.ac.kr (N. Ejaz), irfanmehmood@sju.ac.kr (I. Mehmood), sbaik@sejong.ac.kr (S. Wook Baik).

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2012.10.002

objects, events and actions. In general, the extraction of these semantic primitives is not feasible [2,3]. However, some domain specific techniques have been proposed [6–8]. The most common set of techniques for video summarization employ low level index features in a variety of ways. The usage of low level features ultimately leads to the creation of a semantic gap between actual semantics and low level features [19]. In order to approximate this semantic gap, some researchers [19–22] used visual attention based models to extract visually attractive key frames from the videos. The visual attention based key frame extraction schemes extract visually salient key frames without any additional knowledge about the video, and thus effectively bridge the semantic gap between feature space and human perception. A limiting factor in the usage of visual attention based schemes is the computational cost incurred in computing the visual attention clues. For instance, the usage of optical flow [19–22] for obtaining dynamic visual attention clues render these techniques impractical without the usage of sophisticated hardware. In this paper, an efficient visual attention based key frame extraction scheme is proposed. The proposed framework computes static and dynamic visual attention clues and then combines them non-linearly to extract key frames. The static attention model is built using the image signature based saliency detection [24]. For the dynamic attention model, temporal gradients are used to highlight the important areas of inter-frame motion. The experimental results demonstrate that the proposed scheme yields better results than some of the well-known non-visual attention based schemes [14–16,25]. Moreover, the proposed scheme is not only computationally efficient than some of the visual attention based schemes [19,21,22], but also extracts the key frames of comparable quality.

The rest of this paper is arranged as follows. Section 2 of the paper discusses the related work, Section 3 presents the framework of the proposed approach, Section 4 gives the experimental results, and finally Section 5 concludes the paper.

2. Related work

This section summarizes the major categories of the schemes presented in literature for the task of video summarization using key frame extraction. For a detailed review of the existing techniques, the readers are referred to the surveys [2,3].

There are some domain specific techniques for extraction of key frames, which use high level semantic features of the videos. For instance, Chen et al. [6] summarized basketball videos based on automatic scene analysis and camera viewpoint selection. Xu et al. [7] utilized web-casting text along with the video analysis algorithms for the summarization of sports videos. Calic and Thomas [8] selected those frames as the key frames where the salient objects merge by using the positions of areas obtained by frame segmentation. In the work of Liu and Fan [30], initially the candidate key frames were selected based on color histogram. A Gaussian mixture model (GMM) was then trained for object segmentation. The candidate key frames were refined by using the segmentation results and the trained Gaussian mixture model (GMM). The disadvantage of the high level semantic features based summarization techniques is that generally they do not work well outside the particular domain and experimental settings.

The low level index features, extracted from video frames, have been widely used for summarizing videos. The prominent schemes utilizing low level features for key frames extraction can be broadly classified into two categories [9]: (1) significant content change, and (2) clustering. In significant content change based methods, each frame is compared with the previous key frame(s) based on the differences between low level features of the frames. A new key frame is extracted only if the inter-frame difference is significant. The famous set of features include histogram differences in various color spaces [10,11,29,33], an accumulated energy function [12], compressed domain features [11], Laplacian Eigen Map features [9] and MPEG-7 visual descriptors [13]. In clustering based key frame extraction methods, the frames of the video are clustered based on low level features, and then usually one frame from each cluster is designated as the key frame. Some of the prominent clustering based techniques use color histogram [14–16], object and camera motion based features [17], and edge information [18] as features for clustering. However, regardless of how effectively low level features are used, the loss of semantic details is almost inevitable, thus leading to a significant semantic gap.

Among the visual attention based schemes, Ma and Zhang [19] presented a motion attention based model for the task of video skimming. Ma et al. [20] extended the work of [19] to develop an open framework including a set of visual, aural and linguist attention features which were then fused by a non-linear scheme. The fused attention value of each frame was used to build an attention curve and the key frames were extracted at the crests of this attention curve. This framework is computationally expensive as it employed top-down attention mechanisms. Moreover, the relationship between a combination of visual, aural and linguistic features is difficult to tackle. Peng and Xiaolin [21] initially clustered the frames based on color histograms, and then from each cluster selected the frame which was visually more salient. Because of the usage of *K*-means algorithm for clustering, the sequential order of the key frames may not be preserved. In the work of Lai and Yi [22], a time constrained clustering algorithm was used to cluster the similar frames. From each cluster, most salient frame based on motion, color and texture features was selected as key frame. The usage of time constrained clustering algorithm preserves the sequential order of the frames unlike [21]. However, both [21] and [22] use linear fusion scheme to combine visual attention features which is not representative of the complex non-linear human perception mechanism [28].

3. Methodology

In the human visual attention system, the brain and the vision systems work jointly to locate the salient

regions in images and videos. The human attention is a neurobiological notion that symbolizes the human ability of concentrating mental powers on certain areas by close observation [32]. The exact details of human attention mechanism are still not known [32]. However, the researchers have come up with some theories regarding the human system of visual information processing [32]. For instance, it is known that visual attention of humans is directed by bottom-up and top-down attention mechanisms [13]. The bottom-up mechanism is stimulated in response to low level features (color, texture, motion etc.) which appear visually distinct from rest of the scene. The bottom-up attention mechanism is reflexive, task-independent, fast and transient. The top-down attention mechanism is driven as high level cognitive features which need a voluntary attempt to shift the gaze towards the required goal. The top-down attention mechanism is goal oriented, task-specific, slow and long-lasting.

The concept of bottom up visual attention is realized in the form of visual saliency. Visual saliency measures the extent to which an area is different from its neighborhood. It is widely used in the literature as an approximation of human visual system for both static images and dynamic scenes [23]. As per the mechanism of bottom-up attention modeling, the human beings tend to concentrate more on the areas of spatial and temporal contrast [23]. For this reason, an efficient visual attention model for videos must use both spatial and temporal features for the identification of salient areas. The visual attention based framework used in this paper is based on the computation of bottom-up static and temporal visual saliencies. The top-down attention modeling has not been used because of its task-specific nature and high computational cost. Fig. 1 shows the main steps of the proposed method. The details are provided in subsequent sub-sections.

3.1. Spatial attention value

The spatial attention model is developed by computing visual saliency, based on an image descriptor called the "image signature" [24]. As shown in [24], the image signature can be used to approximate the foreground of an image. The basic assumption is that the foreground of an image is visually more conspicuous than its background. The scheme is based on discrete cosine transform (DCT) whereby only the sign of each DCT component is retained by discarding the amplitude.

A particular frame "F" of the video is initially resized to the size of 64×48. Then the image signature "$\mathrm{IS}(F_c)$" for color channel "c" of the frame "F" is defined as:

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

sign (.) is the entry wise sign operator, DCT denotes the discrete cosine transform and "F_c" is the color channel "c" of the frame "F". The image signature is transformed back to the spatial domain by taking its inverse discrete cosine transform to obtain the reconstructed image "F'_c".

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

The static saliency map of "F_c", denoted by "$S(F_c)$", is then computed as:

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

"G" is the Gaussian kernel used for smoothing, "$*$" denotes the convolution operator and "o" is the Hadamard (entry wise) product operator. The Gaussian kernel on a pixel (i, j) of an image is defined as:

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

"σ" is the standard deviation of the distribution, the value of which is taken as 0.045. The saliency map of each color channel is added linearly to get overall static

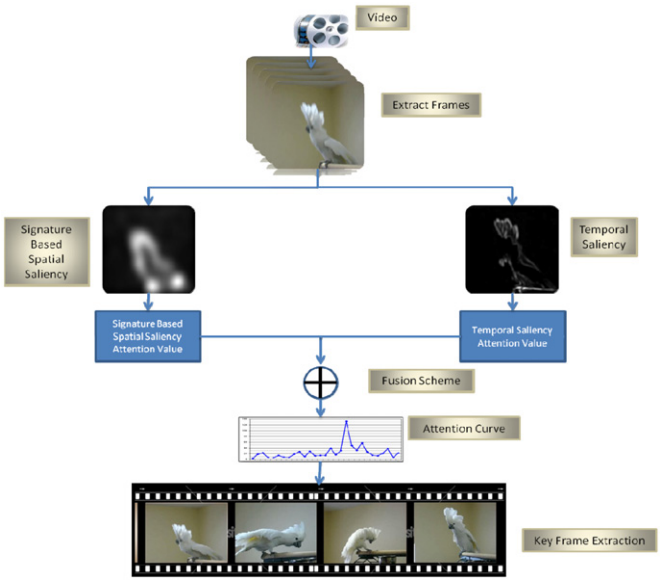


Fig. 1. Framework of the proposed system.

saliency map "$S(F)$" of frame "F". CIELAB color space is used as the choice of color channels owing to its ability to effectively approximate human vision. The saliency map "$S(F)$" is then normalized in the range of 0 to 1 by dividing each value by the maximum value in the map. The average of non-zero values in the saliency map "$S(F)$" is taken to obtain the static attention value "A_s" of a frame. If the value of "A_s" is close to 1, the frame is considered as salient. On the other hand, a value of "A_s" that is close to 0 indicates a non-salient frame.

There are multiple benefits of using the image signature based approach compared to the other schemes for saliency detection. The presence of a sufficiently sparse foreground on a sufficiently sparse background is effectively approximated by the image reconstructed from the image signature. It has been experimentally proven that the results of the saliency algorithm based on image signatures correlates well with the visually conspicuous image locations and human perception [24]. The signature based saliency detection scheme is computationally cheaper than most of the other techniques because of its holistic nature. The technique takes an average of 0.04 s/frame in a desktop environment. For further details, the reader is referred to [24].

3.2. Temporal attention value

In videos, human beings tend to concentrate more on the motion of objects relative to each other [23]. Therefore, for videos, motion is an important parameter in modeling human attention. Because of the importance of motion attention, all visual attention based video summarization techniques use motion clues in different ways [19–22]. However, most of the previously used motion modalities for visual attention modeling are fundamentally based on the computation of optical flow or motion field, which being computationally expensive hinders the applicability of these techniques in real world applications. In order to quickly obtain the motion information in video sequences, the concept of temporal gradients [26] is employed. In this way, motion information is computed implicitly by measuring the temporal changes in the values of a pixel in neighboring frames. The temporal gradients measure the change in the pixel values across frames and thus approximate the motion component of visual attention. The computation of temporal gradients is much cheaper than traditional optical flow based methods for motion estimation. This property renders them to be used in online applications.

There are two frames of video $F(t)$ and $F(t-\tau)$ present at time "t" and "$t-\tau$" in the video. The temporal contrast between pixel "p" of frame $F(t)$ and pixel "q" of neighboring frame $F(t-\tau)$ is defined as:

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

$F_p(t)$ and $F_q(t-\tau)$ are the intensity values of frame $F(t)$ and $F(t-\tau)$ at pixel "p" and "q", respectively. Using this definition of temporal contrast between pixels of neighboring frames, the gradient value at each pixel "p" of frame $F(t)$ is computed. A 5×5 neighborhood $N_{t-\tau}(p)$ 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:

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

After computing the gradient vector for each pixel in frame $F(t)$, the temporal saliency at pixel "p" is then computed by using sum of absolute differences between the temporal gradients of 5×5 neighborhood $N_t(p)$ around pixel "p" in frame $F(t)$.

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

By computing the saliency value at each pixel, the temporal saliency map TS(F) of frame $F(t)$ is obtained. The temporal saliency map is normalized in the range of [0,1] by dividing the value of each pixel by the maximum value in the map TS(F). The values of the saliency map are then averaged to get the temporal attention value "A_T" of frame $F(t)$. Again, a high value of "A_T" indicates a salient frame and vice versa.

3.3. Fusion of attention values and key frame extraction

For the most part, linear fusion schemes have been used by researchers for the fusion of various attention values to generate an aggregated attention value [23]. If there is "n" number of attention values to be combined, the general form of linear fusion schemes looks as follows:

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

w_i is the weight of an attention value A_i, and A_L is the aggregated attention value using the linear mechanism. Usually, the linear mechanism is unable to reflect all the information possessed by the attention values of the attention components. Moreover, for the pattern recognition and related tasks, a human brain's visual section uses a non-linear processing system [27]. A linear fusion mechanism therefore is not suitable.

In the literature of visual attention based video summarization schemes, the authors of [21] and [22] used linear fusion schemes with a higher weight assigned to the motion attention values than the static ones. Ma et al. [20] used a non-linear fusion scheme, proposed in [28], with equal priority assigned to static and dynamic attention values. The psychological theories of human attention, however, claim that motion component is more important compared to the static attention clues [31]. The fusion scheme used in this paper combines the benefits of both types of techniques. The weights are calculated in such a way as to assign more weight to the temporal component of the attention. The static and dynamic attention clues are combined non-linearly [28].

The weight of the temporal attention value in frame "F", denoted by "w_T", is determined from the temporal saliency map $TS(F)$ as:

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

If there is strong motion contrast in $TS(F)$, the value of "d" will be high which leads to a high value of w_T and vice versa [20]. The weight of the static attention value "w_S" is determined as:

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

If "A_S" and "A_T" are the static and temporal attention values, respectively, the non-weighted version of fusion function is given as:

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

"γ" is a pre-defined constant, which symbolizes the significance of an attention component in the combined attention model. The value of "γ" is selected to be 0.2. If the static and dynamic attention values are represented in the form of a vector $A = [A_S,A_T]$ and the weight is defined as a vector $w = [w_S,w_T]$, the weighted fusion function is defined as:

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

"W" is defined as:

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

Next, we briefly justify the choice of the non-linear fusion scheme in comparison with linear and Max fusion schemes. Consider two set of fusion values (0.9, 0) and (0.45,0.45). A linear fusion scheme will give a fused attention value of 0.9 for both cases. However, the former set is more salient as compared to the later because of the high value of one of the attention indices. Moreover, linear fusion scheme does not fulfill following property for attention fusion schemes [28]:

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

In case of linear fusion, $F(v_1,v_2)$ is always equal to $F(v_1+\Delta,v_2-\Delta)$.

The Max fusion scheme selects the maximum of the two attention indices to be fused. The Max fusion function fulfills the property in inequality (14). However, the following trivial property of attention functions is violated by Max fusion function:

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

The used fusion scheme of Eq. (12) fulfills both the properties (inequalities 14 and 15) and thus is more effective than linear and Max fusion schemes. The parameter "γ" in Eq. (12) can be used to control the differences among left and right hand of inequalities (14) and (15). If the value of "γ" is increased, the relative difference will start decreasing. In this way, the parameter "γ" can be used to control the sensitivity of the fusion schemes towards change.

The fused attention value of each frame is then used to make an attention curve of the video which is then used for extracting key frames. If the number of key frames "n_K" is not specified by the user, then the frame having the highest attention value in each shot is selected as the key frame. If "n_K" is known, then each shot is assigned "n_{KS}" number of key frames as per the following strategy suggested in [20].

$$n_{KS} = \max(n_K \times \alpha, 1) \tag{16}$$

"α" is the ratio between the variance of attention values in a particular shot and sum of variances of attention values in all shots. If the number of desired key frames is less than the number of shots, then the candidate key frames having lower attention values are discarded.

4. Experiments and results

A variety of methods have been used by researchers for the evaluation of their key frame extraction techniques with no consensus on any standard technique. In order to properly evaluate the proposed scheme, various set of experiments were conducted. The subsequent sections provide the details of the experiments.

4.1. Significance of the proposed method

Firstly, the results of the technique have been presented on single shots of two videos downloaded from the Open Video Project (www.open-video.org).

The first test sequence is the fifth shot (frame 484 to 520) of the video ucomp03_06_m1.mpeg. In this shot, a tennis player hits the ball, and then stands and receives appreciation from the spectators. The attention curves of spatial, temporal and fused saliency attention values are shown in Fig. 2. It is evident from the fused attention curve of the proposed scheme that frame 491 has the highest attention value and is thus selected as key frame. The fused attention curve of the Lai and Yi scheme [22] suggests frame 517 as key frame. The key frames selected by [22] and proposed scheme are shown in Fig. 3. It is evident that key frame extracted by [22] does not convey the notion of 'shot playing' of the player and is thus not semantically representative. On the other hand, the key frame extracted by the proposed scheme is more highlight worthy and summarizes the shot effectively.

The second test video sequence is the second shot (Frames 532 to 548) of the video hcil2000_01.mpeg. In the frames under consideration, a person is standing and talking, with trees in the background. From frame 545, a subtitle starts appearing in the scene which shows the introduction of the narrator. A representative key frame of this shot must show the person and the subtitle. The respective attention curves are shown in Fig. 4. Frame 545 and 548 are the extracted key frames by [22] and the proposed scheme, respectively. Both the frames are shown in Fig. 5. In frame 545, the subtitle is not clearly legible, whereas frame 548 clearly shows the subtitle and is thus the best representation of the shot.

4.2. Comparison with other techniques

This section compares the proposed scheme with some of the prominent non-visual attention and visual attention based schemes. For the purpose of comparison, the experiments were conducted on 20 videos of various

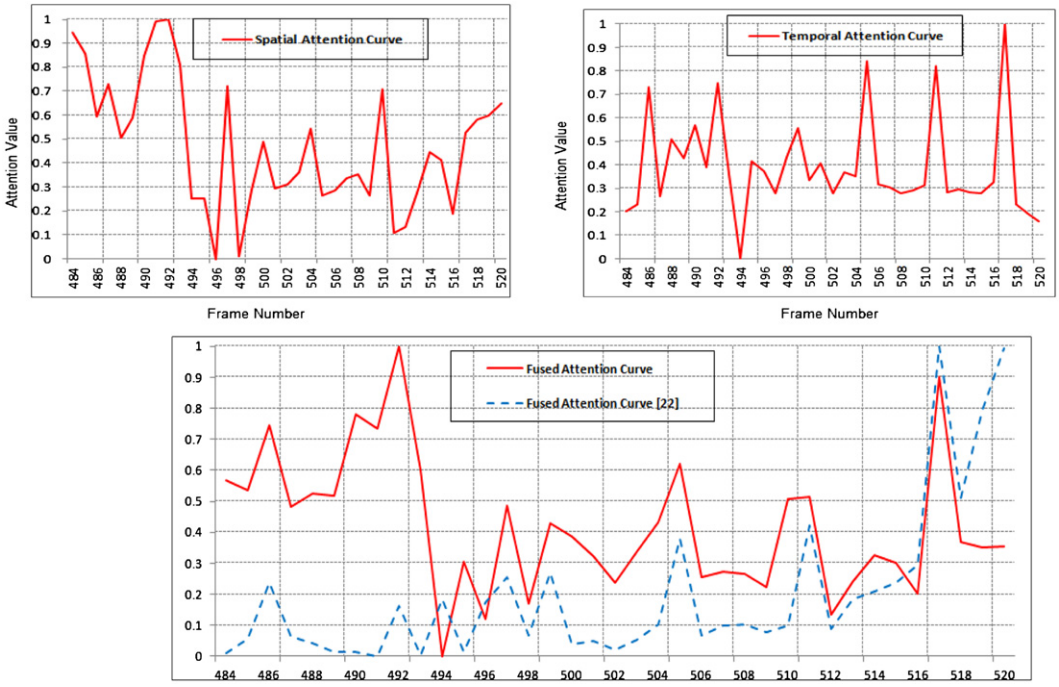
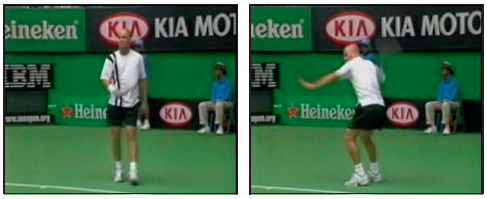


Fig. 2. Attention curves for fifth shot of the video ucomp03_06_m1.mpeg.



Frame 517 Frame 491

Fig. 3. Key frames extracted by [22] and the proposed scheme for fifth shot of the video ucomp03_06_m1.mpeg.

genres downloaded from the Open Video project. Table 1 summarizes the relevant information about these videos. For comparison, two different evaluation strategies were used. The first scheme is based on the popular metrics of Recall, Precision and *F*-measure. The second evaluation scheme is a subjective Mean Opinion Score (MOS) technique.

In the first evaluation mechanism, the key frames are first manually extracted by three human users for each video. Next, the key frames generated by a particular technique are compared with these manually extracted key frames. The two frames are considered same if they convey same semantic information. The following terms are then defined:

True Positive (T_p): A frame chosen as key frame both manually and by the technique,
False Positive (F_p): A frame chosen as key frame by the technique but not manually, and
False Negative (F_n): A frame chosen as key frame manually but not by the technique.

These terms are used to define the metrics Recall and Precision.

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

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

Recall is the probability that a relevant key frame is chosen whereas Precision is the probability that a selected key frame is relevant. There is a tradeoff between the values of Recall and Precision, whereby the gain in value of one of these parameters is attained at the cost of the other. Moreover, these two metrics are complementary to each other. A high value of one metric is usually not reliable. For instance, a high value of Precision can be achieved by selecting very few key frames and a high value of Recall can be achieved by selecting too many key frames. The high values of both Recall and Precision indicate an effective summarization. In order to get a single combined metric, the Recall and Precision are combined together using the following definition of *F*-measure:

$$F = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{19}$$

A high value of *F*-measure thus indicates a high value for both Precision and Recall. The Recall, Precision and *F*-measure scores obtained in comparison with three users' summaries are then averaged to get single values for each video. The three user summaries for each video of Table 1 are taken from the public data set of [16].

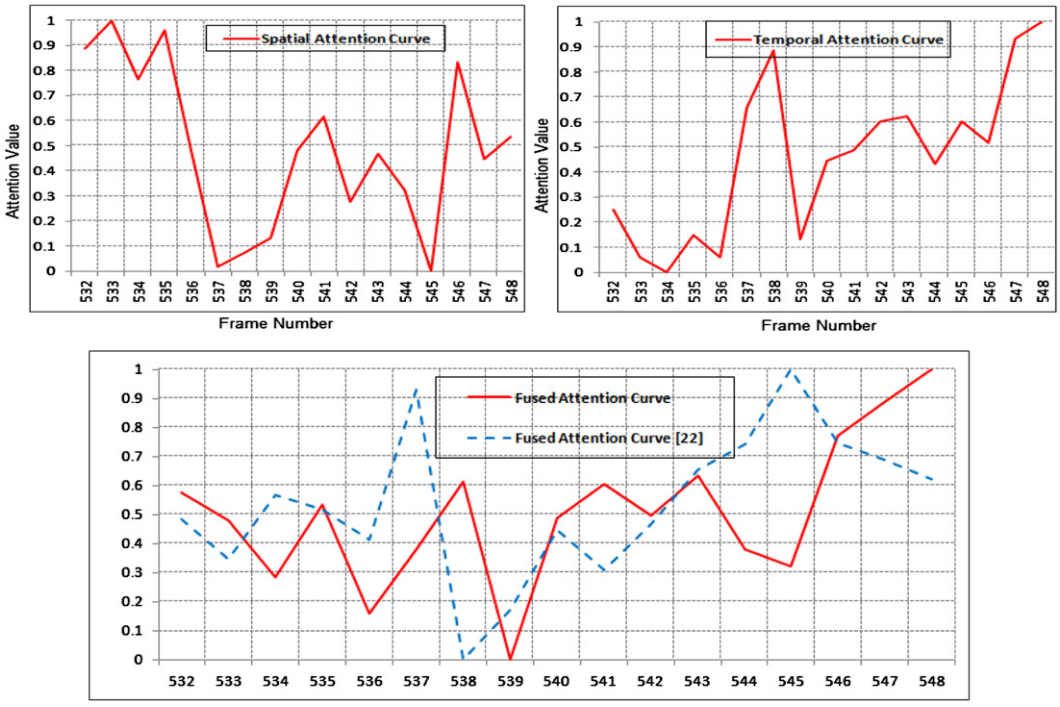


Fig. 4. Attention curves for second shot of the video hcil2000_01.mpeg.

Frame 545 Frame 548

Fig. 5. Key frames extracted by [22] and proposed scheme for second shot of the video hcil2000_01.mpeg.

In the second evaluation strategy, the data set of Table 1 was evaluated based on the Mean Opinion Score (MOS) criteria [15]. In this assessment mechanism, three users are asked to rate the quality of each summary on a scale of 0 (minimum) to 5 (maximum) after watching the full video and the corresponding summaries generated by all techniques. The scores of all users for a particular video are then averaged to obtain the MOS of a video.

In Sections 4.2.1 and 4.2.2, the comparison of the proposed technique has been performed with some of the non-visual attention and visual attention based schemes, respectively.

4.2.1. Comparison with non-visual attention based techniques

In this section, the comparison of the proposed scheme is performed with four prominent non-visual attention schemes: OV [25], DT [14], STIMO [15] and VSUMM [16]. All of these techniques are based on the extraction of low level features from the video frames. OV [25] is a significant content change based scheme, while DT [14], STIMO [15] and VSUMM [16] are based on clustering. Table 2 shows the Recall, Precision and *F*-measure values for each of the 20 videos of the data set. The proposed methodology clearly dominates the other techniques by consistently achieving higher scores for all three measures. However, there are some exceptions. For instance, for video 5, the DT scheme achieves a high value of Precision. However, for this video, DT selects only one key frame and thus values of T_p and F_p are both 1 which

Table 1
Details of test videos.

No.	Video name	No. of Frames
1	Wetlands Regained, segment 03 of 8	3562
2	Technology at Home: A Digital Personal Scale	3346
3	Introduction to HCIL 2000 reports	2454
4	Ocean floor Legacy, segment 05 of 14	4665
5	The Great Web of Water, segment 01	3279
6	The Great Web of Water, segment 02	2118
7	The Great Web of Water, segment 07	1745
8	A New Horizon, segment 01	1806
9	A New Horizon, segment 02	1797
10	A New Horizon, segment 06	1944
11	A New Horizon, segment 08	1815
12	Exotic Terrane, segment 04	4797
13	The Future of Energy Gases, segment 05	3615
14	The Future of Energy Gases, segment 09	1884
15	Oceanfloor Legacy, segment 01	1740
16	Oceanfloor Legacy, segment 02	2325
17	Oceanfloor Legacy, segment 09	2106
18	Hurricane Force—A Coastal Perspective, segment 03	2310
19	Drift Ice as a Geologic Agent, segment 05	2187
20	Drift Ice as a Geologic Agent, segment 10	1407

Table 2
Recall (R), Precision (P) and F-measure (F) of different techniques on the video data set.

No.	OV [25]			DT [14]			STIMO [15]			VSUMM [16]			Proposed		
	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F
1	0.59	0.63	0.61	0.53	0.59	0.56	0.76	0.51	0.61	0.57	0.59	0.58	0.84	0.80	0.82
2	0.56	0.83	0.67	0.58	0.73	0.65	0.72	0.56	0.63	0.66	0.65	0.65	0.85	0.82	0.83
3	0.82	0.49	0.61	0.36	0.63	0.46	0.72	0.59	0.65	0.58	0.67	0.62	0.83	0.82	0.82
4	0.66	0.70	0.68	0.49	0.49	0.49	0.69	0.61	0.65	0.61	0.57	0.59	0.83	0.80	0.81
5	0.48	0.49	0.49	0.63	1.00	0.77	0.67	0.41	0.51	0.67	0.51	0.58	0.82	0.75	0.78
6	0.78	0.67	0.72	0.30	0.33	0.32	0.77	0.48	0.59	0.62	0.67	0.64	0.90	0.85	0.87
7	0.72	0.58	0.64	0.52	0.59	0.55	0.59	0.43	0.50	0.90	0.59	0.71	0.81	0.80	0.80
8	0.70	0.70	0.70	0.49	0.52	0.50	0.67	0.67	0.67	0.79	0.70	0.74	0.85	0.80	0.82
9	0.37	0.93	0.53	0.46	0.50	0.48	0.42	0.89	0.57	0.71	0.90	0.79	0.87	0.85	0.86
10	0.47	0.73	0.57	0.63	0.71	0.67	0.52	0.67	0.59	0.78	0.72	0.75	0.83	0.81	0.82
11	0.54	0.81	0.65	0.42	0.62	0.50	0.63	0.90	0.74	0.70	0.88	0.78	0.75	0.75	0.75
12	0.92	0.44	0.59	0.52	0.67	0.58	0.50	0.42	0.45	0.92	0.75	0.82	0.83	0.80	0.82
13	1.00	0.55	0.71	0.59	0.75	0.66	1.00	0.64	0.78	1.00	0.71	0.83	0.80	0.75	0.77
14	0.69	0.50	0.58	0.86	0.73	0.79	0.92	0.67	0.77	0.92	0.72	0.81	0.80	0.75	0.77
15	0.22	0.50	0.30	0.50	0.78	0.61	0.61	0.33	0.43	0.41	0.55	0.47	0.85	0.80	0.82
16	0.64	0.89	0.74	0.48	0.80	0.60	0.64	0.76	0.69	0.64	0.76	0.69	0.70	0.66	0.68
17	0.71	0.72	0.71	0.58	0.75	0.66	0.62	0.79	0.69	0.62	0.79	0.69	0.88	0.85	0.86
18	0.65	0.72	0.69	0.67	0.78	0.72	0.73	0.76	0.75	0.73	0.76	0.75	0.86	0.89	0.87
19	0.74	0.89	0.81	0.92	0.83	0.87	0.90	0.90	0.90	0.90	0.90	0.90	0.93	0.91	0.92
20	0.80	0.73	0.77	0.72	0.67	0.69	0.85	0.80	0.82	0.87	0.82	0.84	0.90	0.87	0.88
Avg.	0.65	0.67	0.64	0.56	0.67	0.61	0.70	0.64	0.65	0.73	0.71	0.71	0.84	0.81	0.82

leads to a Precision value of 1. For this video, DT has a sufficiently low value of Recall because of having a high value of F_N. Similarly, OV has the highest value of Recall for Video 13, whereas the Precision value is significantly low. The proposed methodology consistently achieves higher values for both Recall and Precision and thus is more effective than other methods.

Table 3 shows the MOS of all the techniques under consideration, for the 20 videos of the data set. It is observed that the proposed scheme consistently achieves the highest MOS value among all the techniques. This subjective evaluation based on direct opinion of the human users, indicates that the proposed scheme is more highlight intensive and user friendly compared to the other non-visual attention techniques.

Table 3
MOS test scores for various techniques.

No.	OV [25]	DT [14]	STIMO [15]	VSUMM [16]	Proposed
1	4.03	3.09	2.50	4.53	4.44
2	3.86	2.89	3.63	4.11	4.47
3	4.19	2.41	3.81	3.49	4.06
4	4.00	2.75	3.21	2.72	3.91
5	4.02	3.09	3.38	2.72	4.08
6	4.47	3.81	3.65	3.30	4.39
7	3.82	3.50	3.69	3.56	4.25
8	2.80	3.50	3.84	3.56	4.28
9	3.29	3.39	3.13	3.15	4
10	3.84	3.35	3.45	3.43	4.15
11	4.05	3.95	3.78	3.69	4.16
12	4.13	3.10	3.35	2.75	3.99
13	3.02	2.38	2.80	3.00	4.16
14	3.54	3.18	3.29	3.41	4.31
15	3.75	2.75	3.74	3.65	4.25
16	3.17	2.38	2.88	3.69	4.06
17	3.98	3.51	3.68	4.06	4.19
18	12.00	3.62	2.50	2.31	4.19
19	3.23	3.66	3.29	3.15	4.31
20	3.21	3.25	3.13	3.24	4.4
Avg.	4.11	3.17	3.33	3.37	4.2

4.2.2. Comparison with visual attention based techniques

In this sub-section, a comparison of the proposed mechanism has been performed with some of the visual attention based schemes including Ma and Zhang [19], Peng and Xiaolin [21], and Lai and Yi [22]. Table 4 lists the basic properties of some prominent visual attention schemes used for the task of video summarization. It can be seen that most of the schemes are based on modeling of bottom up component of visual attention. This is because the task oriented top down mechanism is difficult to generalize and hard to compute. Among the techniques under consideration, only Ma et al. [20] employed top-down attention modeling. The framework in [20] used face detection, audio saliency, and linguistic attention; which rendered it as an expensive framework, hard to be used in practical systems. Because of these facts, we did not compare our scheme with that of [20].

Table 5 lists the Recall, Precision, and F-measure values for the proposed technique and visual attention schemes [19], [21], and [22]. It can be observed, that generally the results of all visual attention based schemes are better than those of the low level features based non-visual attention based schemes (compare with Table 2). Moreover, within the visual attention based schemes, the results of the proposed mechanism are comparable with the rest of the other techniques. The same conclusions can be drawn by looking at Table 6 which displays the MOS score for all the visual attention based schemes under question.

Since, the quality of the results is comparable with rest of the techniques; the advantage of reducing the computation cost becomes prominent. Next, the running time of the proposed mechanism is compared with that of other

Table 4
Properties of some visual attention based schemes.

	Ma & Zhang [19]	Ma et al. [20]	Peng & Xiaolin [21]	Lai & Yi [22]	Proposed
Visual attention modeling	Bottom up	Top Down, Bottom Up	Bottom up	Bottom up	Bottom up
Fusion scheme	Linear	Non-Linear	Linear	Linear	Non-Linear
Weighted fusion	No	No	Yes	Yes	Yes
Real time processing	No	No	No	No	Yes
Suitability for web scenarios	No	No	No	No	Yes

Table 5
Recall (R), Precision (P) and F-measure (F) of visual attention based techniques on video data set.

	Ma & Zhang [19]			Peng & Xiaolin [21]			Lai & Yi [22]			Proposed		
No.	**R**	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
1	0.83	0.75	0.79	0.82	0.70	0.75	0.81	0.80	0.80	0.84	0.80	0.82
2	0.75	0.75	0.75	0.73	0.73	0.73	0.90	0.85	0.87	0.85	0.82	0.83
3	0.85	0.70	0.77	0.80	0.72	0.76	0.86	0.82	0.84	0.83	0.82	0.82
4	0.86	0.85	0.85	0.83	0.85	0.84	0.83	0.80	0.81	0.83	0.80	0.81
5	0.83	0.75	0.79	0.80	0.75	0.77	0.80	0.83	0.81	0.82	0.75	0.78
6	0.88	0.73	0.79	0.83	0.70	0.76	0.85	0.75	0.80	0.90	0.85	0.87
7	0.75	0.75	0.75	0.83	0.72	0.77	0.80	0.70	0.75	0.81	0.80	0.80
8	0.83	0.75	0.79	0.81	0.75	0.78	0.83	0.72	0.77	0.85	0.80	0.82
9	0.88	0.82	0.84	0.85	0.82	0.83	0.85	0.78	0.81	0.87	0.85	0.86
10	0.82	0.80	0.81	0.80	0.82	0.81	0.83	0.80	0.81	0.83	0.81	0.82
11	0.78	0.83	0.80	0.75	0.80	0.77	0.82	0.80	0.81	0.75	0.75	0.75
12	0.80	0.80	0.80	0.80	0.80	0.80	0.85	0.83	0.84	0.83	0.80	0.82
13	0.75	0.83	0.79	0.75	0.82	0.78	0.73	0.81	0.77	0.80	0.75	0.77
14	0.78	0.80	0.79	0.80	0.80	0.80	0.85	0.79	0.82	0.80	0.75	0.77
15	0.75	0.90	0.82	0.78	0.70	0.74	0.73	0.75	0.74	0.85	0.80	0.82
16	0.80	0.75	0.77	0.80	0.73	0.76	0.85	0.80	0.82	0.70	0.66	0.68
17	0.83	0.83	0.83	0.83	0.80	0.81	0.80	0.81	0.80	0.88	0.85	0.86
18	0.85	0.88	0.86	0.83	0.75	0.79	0.82	0.82	0.82	0.86	0.89	0.87
19	0.93	0.85	0.89	0.80	0.83	0.81	0.90	0.88	0.89	0.93	0.91	0.92
20	0.90	0.80	0.85	0.88	0.80	0.84	0.83	0.85	0.84	0.90	0.87	0.88
Avg.	**0.82**	**0.79**	**0.81**	**0.80**	**0.77**	**0.78**	**0.83**	**0.80**	**0.81**	**0.84**	**0.81**	**0.82**

schemes. For this purpose, the length of the videos to be summarized was varied from 1000 frames to 6000 frames. In the proposed mechanism, if the computation cost is to be reduced further, an optional step of pre-sampling can be used. In this way, only a subset of video frames can be selected for processing based on a pre-defined sampling rate. However, the sampling rate must be chosen cautiously to avoid any loss of information. Fig. 6 shows the time taken by [19,21,22], and the proposed scheme with a sampling rate of 20 frames, for the videos of various lengths. The results were obtained on a general purpose computer (Intel core 2 Duo 1.6 Hz equipped with 2 GB RAM). It can be observed that scheme of [19] takes the maximum time. This is because this scheme not only computes optical flow based motion field, but also employed expensive steps of calculating spatial and temporal orientation coherence. The computational cost for [21] and [22] are also higher than the proposed scheme because of employing non-holistic spatial attention schemes and optical flow based dynamic attention schemes. If the sampling step is used, the computational cost is further reduced. The computation time depends heavily on the used hardware; however the presented results are beneficial in understanding the

Table 6
MOS test scores for various visual attention based techniques.

No.	Ma & Zhang [19]	Peng & Xiaolin [21]	Lai & Yi [22]	Proposed
1	4.25	4.19	4.25	4.44
2	4.34	4.1	4.29	4.47
3	4.23	4.04	4.2	4.06
4	4.15	4.15	4.29	3.91
5	4.25	4.07	4.65	4.08
6	4.14	4.29	4.38	4.39
7	4.26	4.25	4.24	4.25
8	4.14	4.1	4.14	4.28
9	4.06	4.35	4.5	4
10	4.29	4.16	4.1	4.15
11	4.08	4.25	4.03	4.16
12	4.13	4.41	4.1	3.99
13	4.4	4.06	4	4.16
14	4.15	4.24	4	4.31
15	4.1	4.09	4.1	4.25
16	3.99	4.11	4.1	4.06
17	4.15	4.06	4.13	4.19
18	4	4.14	4.11	4.19
19	4.18	4.16	4	4.31
20	4.3	4.22	4.13	4.4
Avg.	**4.18**	**4.17**	**4.19**	**4.20**

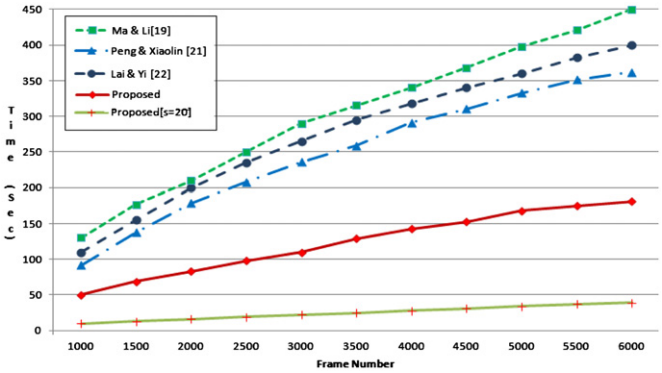


Fig. 6. Time taken by proposed scheme, [19], [21], and [22] for extracting key frames.

Fig. 7. Comparison of key frame extraction for video 'Hurricane Force—A Coastal Perspective, segment 03'.

relative difference in summary generation time between various techniques.

Finally, the key frames extracted by various schemes are presented visually (in Fig. 7) for the documentary video 'Hurricane Force—A Coastal Perspective, segment 03'. The video initially shows the first researcher of the US Geological Survey explaining the importance of knowledge concerning the geology of coastal regions along with background scenes. Then, the second researcher discusses a survey on the catastrophic effects of extreme weather on coastal erosion. The ground truth key frames and the extracted key frames by various techniques for this video are shown in Fig. 7. It can be observed that there are some important frames missing from the OV and DT summaries. Moreover, in both summaries, the subtitles introducing the researchers are missing (Key frame 2 and 6 of OV, Key frame 2 and 4 of DT). STIMO also misses important frames, including the frame showing the first researcher. The subtitle for second researcher is also missing (Key frame 5 of STIMO). Even though it generates reasonably good summary, VSUMM misses the subtitle for the frame showing second researcher (Key frame 8 of VSUMM). The results of all visual attention based schemes [19,21,22] and the proposed scheme are visually comparable. However, both [21] and [22] miss the subtitles showing introduction of the researchers (Key frame 2 and 6 of [21], Key frame 3 and 9 of [22]). The summary of [19] misses subtitle details of the first researcher (Key frame 2 of [19]). It can be

observed, that the key frames generated by the proposed scheme cover all the missing aspects and are closer to the manually extracted key frames in comparison to the other techniques.

5. Conclusions

In this paper, an efficient visual attention based framework for key frame extraction from videos is proposed. The technique not only yields effective results, but is also suitable to be used in small devices. The usage of temporal gradients provides an efficient replacement of the previously used traditionally optical flow oriented features. The usage of non-linear weighted fusion scheme combines the benefits of previously used schemes. Overall, the framework takes much less time than the latest visual attention based schemes [19,21,22]. The experimental results based on a number of criteria shows that the extracted key frames using the proposed scheme are semantically relevant and more highlight oriented than those generated by the other techniques to which they are compared.

Acknowledgment

This research is supported by (1) The Industrial Strategic technology development program, 10041772, (The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the Ministry of Knowledge Economy (MKE, Korea), and (2) The MKE (The Ministry of Knowledge Economy), Korea, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0502-12-1013).

References

[1] J. Son, H. Lee, H. Oh, PVR: a novel PVR scheme for content protection, IEEE Transactions on Consumer Electronics 57 (1) (2011) 17–177.
[2] B.T. Truong, S. Venkatesh S, Video abstraction: a systematic review and classification, ACM Transactions Multimedia Computing, Communications and Applications 3 (1) (2007).
[3] A.G. Money, H. Agius, Video summarisation: a conceptual framework and survey of the state of the art, Journal of Visual Communication and Image Representation 19 (2) (2008) 121–143.
[4] A. Girgensohn, J. Boreczky, L. Wilcox, Keyframe-based user interfaces for digital video, IEEE Computer 34 (9) (2001) 61–67.
[5] S. Uchiashi, J. Foote, A. Girgensohn, J. Boreczky, Video manga: generating semantically meaningful video summaries, In ACM Multimedia (ACMMM'99), Florida (1999) 383–392.
[6] F. Chen, D. Delannay, C. Vleeschouwer, An autonomous framework to produce and distribute personalized team-sport video summaries: a basketball case study, IEEE Transactions on Multimedia 13 (6) (2011) 1381–1394.
[7] C. Xu, J. Wang, H. Lu, Y. Zhang, A novel framework for semantic annotation and personalized retrieval of sports video, IEEE Transactions on Multimedia 10 (3) (2008) 421–436.
[8] J. Calic, B. Thomas, Spatial analysis in key-frame extraction using video segmentation, in: Proc. Workshop Image Anal. Multimedia Interactive Services, Portugal, 2004.
[9] R.M. Jiang, A.H. Sadka, D. Crookes, Hierarchical video summarization in reference subspace, IEEE Transactions on Consumer Electronics 55 (3) (2009) 1551–1557.
[10] H. Zhang, J. Wu, D. Zhong, S. Smoliar, An integrated system for content-based video retrieval and browsing, Pattern Recognition 30 (4) (1997) 643–658.
[11] E.K. Kang, S.J. Kim, J.S. Choi, Video retrieval based on scene change detection in compressed domain, IEEE Transactions on Consumer Electronics 45 (3) (1999) 932–936.
[12] X.D. Zhang, T.Y. Liu, K.T. Lo, J. Feng, Dynamic selection and effective compression of key frames for video abstraction, Pattern Recognition Letters 24 (9–10) (2003) 1523–1532.
[13] J.H. Lee, G.G. Lee, W.Y. Kim, Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder, IEEE Transactions on Consumer Electronics 49 (3) (2003) 742–749.
[14] P. Mundur, Y. Rao, Y. Yesha, Keyframe-based video summarization using Delaunay clustering, International Journal on Digital Libraries 6 (2) (2006) 219–232.
[15] M. Furini M, F. Geraci, M. Montangero, M. Pellegrini M, STIMO: STIll and mOving video storyboard for the web scenario, Multimedia Tools and Applications 46 (1) (2010) 47–69.
[16] S.E.d. Avila, A.B.P Lopes, L.J. Antonio, A.d.A Araújo, VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters 32 (1) (2011) 56–68.
[17] G. Yue, W. Wei-Bo, Y. Jun-Hai Yong, "A Video Summarization Tool using Two-Level Redundancy Detection for Personal Video Recorders", IEEE Transactions on Consumer Electronics, 54, 2, 521–526.
[18] P.P.K. Chan, H. Yu, W.W.Y Ng, Yeung DS, "A novel method to reduce redundancy in adaptive threshold clustering key frame extraction systems", Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, vol 4, pp. 1637–1642, 2011.
[19] Y.-F. Ma, H.-J. Zhang, "A model of motion attention for video skimming", in: Proc. of International Conference on Image Processing, vol. 1, pp. I-129-I-130, 2002.
[20] Y.F. Ma, X.S. Hua, L. Lu, H.J. Zhang, A generic framework of user attention model and its application in video summarization, IEEE Transactions on Multimedia 7 (5) (2005) 907–919.
[21] J. Peng, Q. Xiaolin, Key frame based video summary using visual attention clues, IEEE Transactions on Multimedia 17 (2) (2010) 64–73.
[22] L.J. Lai, Y. Yi, Key frame extraction based on visual attention model, Journal of Visual Communication and Image Representation 23 (1) (2012) 114–125.
[23] A. Toet, "Computational versus Psychophysical Bottom-Up Image Saliency: A Comparative Evaluation Study", IEEE Transactions on Pattern Analysis and Machine Intelligence, 33, 11, 2131–2146.
[24] H. Xiaodi, H. Jonathan, C. Koch, Image signature: highlighting sparse salient regions, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (1) (2012) 194–200.
[25] D. DeMenthon, V. Kobla, D. Doermann, "Video summarization by curve simplification". in: Proc. ACM Internat. Conf. on Multimedia. NY, USA, pp. 211–218, 1998.
[26] Y.-T. Chen, C.-S. Chen, Fast human detection using a novel boosted cascading structure with meta stages,, IEEE Transaction on Image Processing 17 (8) (2008) 1452–1464.
[27] D. Sanchez D, J.C. Martinez, M.A. Vila, Modelling subjectivity in visual perception of orientation for image retrieval, Information Processing & Management 39 (2) (2003) 251–266.
[28] X.S. Hua, H.J. Zhang, "An Attention-Based Decision Fusion Scheme for Multimedia Information Retrieval", Pacific Rim Conference on Multimedia, Japan, 2004.
[29] 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 (7) (2012) 1031–1040.
[30] L.J. Liu, G.L. Fan, Combined key-frame extraction and object based video segmentation, IEEE Transactions on Circuits and Systems for Video Technology 15 (7) (2005) 869–884.
[31] D. Gao, V. Mahadevan, N. Vasconcelos, On theplausibility of the discriminant center-surround hypotheis for visual saliency', Journal of Vision 8 (7) (2008) 1–18.
[32] E.A. Styles, The Psychology of Attention, Psychology Press, U.K, 1997.
[33] N. Ejaz, S.W. Baik, Video summarization using a network of radial basis functions, Multimedia Systems, in press (corrected proof, Online First, May 2012), 〈http://dx.doi.org/10.1007/s00530-012-0263-3〉.