A Novel Attention-Based Key-Frame Determination Method

Huang-Chia Shih, *Member, IEEE*

***Abstract*—This paper presents a novel key-frame detection method that combines the visual saliency-based attention features with the contextual game status information for sports videos. Two critical issues of the attention-based video content analysis are addressed: 1) the visual attention characteristics when a user is watching a video clip and 2) extracting the degree of excitement about the on-going game status. First, the object-oriented visual attention map and the algorithm of determining the contextual attention are presented. The procedure of the contextual inference is used to simulate how the game status attracts the viewers. Second, a fusion methodology of visual and contextual attention analysis based on the characteristics of human excitement is introduced. In addition, the amount of key-frames is determined by using the contextual attention score, while the key-frame determination depends on integrating all the visual attention scores. In experimental results, it demonstrates the robustness of the proposed system for basketball and baseball programs.**

***Index Terms*—Adaptive key-frame rate, attention modeling, content-based analysis, contextual attention, key-frame determination, visual attention model.**

I. Introduction

THE video broadcasting has been highly prevalent application that becomes increasingly important for many usages such as video search, management, and transmission [1]–[3]. The key-frame determination is the one of the most effective mechanisms to represent the whole video using a few frames. For video retrieval application, many systems have been presented to deal with this problem [4], [5]. Most of these systems are depend on key item selection such key-frame and key shot. Pickering and Ruger [6] used boosting algorithm with the supervised learning and applied for video retrieval. It employed for weighting different features to determine the matching scores. With the viewpoint of the content semantics, the visual content can be divided into four categories according to the semantic significance which includes video clip, object, action, and conclusion [7]. Generally speaking, a good key-frame selection mechanism should be satisfied each of these categories whose required by users.

A key-frame not only stands for the whole video sequence, but also implies the volume of attractions for viewer interests. Recently, regarding the attention factors is become the considerable approach when determine user's interesting level [8], [9]. There are two cues used to understand the human attention, which are visual and contextual information. Visual cue is the most intuitive feature for monitoring the human perception system. Modeling the visual attention [10] provides a well understand of the video content. For example, the visual attention model [11] combines several representative feature maps into a single saliency map. It allocates the regions that viewer may interest. The saliency map can be used as an indication of the attention level. Moreover, the game status in sports videos is the most concerning for subscriber. Taking the advantage of prior implicit knowledge from the embedded captions, we have presented an automatic system to extract and understand the context for monitoring event occurs [12].

Another critical issue is that the number of frames within a shot should be assigned as the key-frame. Whether it is predefined or chosen dynamically, it is affected by the video content and viewer perception. Generally speaking, frames in a shot that undergo a strong visual and temporal uncertainty may complicate the problem seriously. To balance this, the selection of the number of key-frames is depends on the level of contextual attention in this paper. It implies that when the excitement of the game status is high and much more key-frames can be identified from this shot.

Recently, many human perceptual features are discovered to understand the human attention in digital image/video search applications. Ma *et al.* [13] illustrated a hybrid user attention model, which includes visual and audio features in video summarization application. Modeling the visual attention of the video can provide well understands about the content. For instance, Itti *et al.* [11] proposed a saliency-based visual attention model for scene analysis. In intelligent video applications, they combine a number of representative feature models into a single saliency map which is then allocated to those regions that are of attract to the user. The saliency feature map can be used as an indication of the attention level for determining key-frame. On the other hand, the contextual analysis, most of researchers aim to bridge the gap between the low-level features and high-level concepts using probabilistic model such as neural networks [14], hidden Markov models [15], and Bayesian networks [16], [17]. The probabilistic scheme is the one of the most useful methods to infer the uncertain semantics, which is generally difficult to access directly. To deal with the problem of the contextual attention understanding, a well-defined contextual attention modeling method based on the human perceptual characteristics is presented in this paper.

In this paper, we proposed a novel attention-based key-frame determination system by integrating the object-based visual attention maps and the contextual on-going game

Manuscript received September 26, 2012; revised May 1, 2013; accepted May 13, 2013. Date of publication August 6, 2013; date of current version August 21, 2013. This work was supported in part by the National Science Council (NSC) of Taiwan under Grant NSC 100-2628-E-155-007-MY2.

The author is with the Department of Electrical Engineering, Human-Computer Interaction Multimedia Laboratory, Yuan Ze University, Taoyuan, Taiwan (e-mail: hcshih@saturn.yzu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2013.2265782

outcomes. The decision of the number of key-frames is determined by utilizing the contextual attention score. We apply the object-based visual attention model with the contextual attention model, it determines not only the human perceptual characteristics precisely, but also the type of video content that attracts the viewers' attention effectively.

The rest of this paper is organized as follows: Section II addresses the attention-based content modeling in terms of visual and contextual attention. Section III details our proposed framework for attention-based key-frame determination algorithm. Several important mechanisms are discussed, including attention score prediction, key-frame rate determination, and key-frame selection. Section IV demonstrates the experimental results of our proposed framework followed by the conclusions in Section V.

II. Attention-Based Content Modeling

Theoretically, attention is a psychological reflection of human emotion. When a search system considers the content attention analysis, it always enables the retrieved media more fit to user's actual needs. Here, we take both visual and contextual information into consideration to deal with the content modeling.

A. Visual Attention Modeling

Most of work for modeling visual characteristics uses the frame-based model to analyze the visual contrast of the frame. Unfortunately, the approaches of frame-based visual attention are sensitive to background region. The attention model is also suffers from the distortion of the camera motion such as zooming, panning and tilting effects. To face this problem, we adopt an object-based attention model to provide the more accurate content information. Based on the attention characteristics, the extracted attention maps are classified into three types: spatial, temporal, and facial. We take into account with the effects of the camera motion and game statistics, because they both provide numerous meaningful clues of content excitement.

First, the motion information can be extracted by using the frame difference method with the pixel-based motion energy. To accumulate the motion energy of the visual objects (VOs) in group of frames (GOF), it enables to localize the moving pixels more precisely. Second, the Sobel operators and Otsu thresholding method is used to extract the edge feature. Based on the extracted the motion and edge features, we can combine these two information to detect object by using "AND" operation. Then, the morphological operators are applied to remove the noise such as closing, opening, and region growing.

1) *Spatial Attention Map:* In static scene, human eye is usually attracted by the significant color distribution, strong contrast, and special texture in static scene. Similar to [11], three saliency maps are used to extract the most representative object-oriented spatial information. Here, the spatial attention maps include intensity, color contrast (red-green, blue-yellow), and orientation of the frames. Mathematically, the spatial attention maps for each VO can be obtained by using each attentive feature map,

$$\bar{M}_{\mathrm{spa}}(\mathrm{o_k}) = \sum_{(\mathrm{i,j})\in \mathrm{o_k}} \delta_{\mathrm{s}}(\mathrm{i,j}) \Big/ \mathrm{a}(\mathrm{o_k}), \tag{1}$$

where $\mathrm{a}(\mathrm{o_k})$ denotes the area of object k and $\delta_{\mathrm{s}}(\cdot)$ denotes a spatial observation function which is formulated by the pixels within object and its surroundings.

2) *Temporal Attention Map:* Obviously, it is difficult to extract motion vector from the image with cluttered and inhomogeneous background. To cope with this difficulty, the motion activity (MA) is applied instead of the precise motion vector to represent the temporal attention map. First of all, apart image into $W \times H$ grid units (GUs) and compute the MA associated with every GU. According to the location of the object boundary, GU can be classified into three classes, i.e., foreground, background and border. The MA of the background region barely attracts people, because it provides less information. Moreover, the MA surrounding object boundary implies the moving behavior of the object and its offset value denotes the moving energy. Obviously, the MA within foreground reflects the texture information of the corresponding object. Therefore, we assign different weights to GUs for computing the temporal attention. Suppose that MA(x, y) indicates the normalized MA of the GU(x, y), where $1 \le x \le W$, $1 \le y \le H$, and let$\delta_{\mathrm{T}}(\cdot)$ denote the temporal observation function weighted by the location of GU. The temporal attention map of object k in feature map m is defined as:

$$\bar{M}_{\mathrm{tem}}(\mathrm{o_k}) = \sum_{\mathrm{r}\in \mathrm{o_k}} \sum_{(\mathrm{p,q})\in\Omega_r} \delta_{\mathrm{T}}(\mathrm{p,q}) \Big/ N_{GU}, \tag{2}$$

where Ω_r denotes the set of GUs r, which is located in or border of the object k; N_{GU} represents the number of GUs associated with the corresponding object. The size of GUs is 8×8 pixels in this paper.

3) *Facial Attention Map:* In this study, the face appearance is also takes into consider with the attention modeling. When many faces appear in the frame, it indicates the strong support to attract viewer's attention. In this paper, we apply an efficient face detection method which is an adaboost-based algorithm with Harr-like and Gabor features undertaking the multi-pose variations [18]. We do not need to know the specific locations of face images, whereas the likelihood distribution of face appearance. The facial attention map of object k can be computed by monitoring all of the attentive regions and normalized by the corresponding object area:

$$\bar{M}_{\mathrm{face}}(\mathrm{o_k}) = \sum_{(\mathrm{i,j})\in \mathrm{o_k}} \delta_{\mathrm{F}}(\mathrm{i,j}) \Big/ \mathrm{a}(\mathrm{o_k}), \tag{3}$$

where the facial observation function $\delta_{\mathrm{F}}(\mathrm{i,j})$ is formulated as the predicted likelihood for the appearance of face in pixel (i, j).

4) *Modeling the Camera Motion Attention:* A rapid camera motion is usually implies a highlight occurs. Therefore, the camera motion plays an important role of the excitement modeling. This paper takes the camera motion into consideration,

replacing the time consuming 2-D calculation with two 1-D calculations by projecting the luminance values from the horizontal and vertical directions. A slice-based method is used to obtain the vertical and horizontal displacement vectors (i.e., τ_v^i and τ_h^i) for each pair of consecutive frames f_i and f_{i+1}. The sum of the normalized norm-2 distance for these two displacement vector is used to represent the camera motion feature M_{cam}, we have

$$M_{cam}(f_i) = \max\left(\left\|\tau_v^i\right\|/\rho_v, \left\|\tau_h^i\right\|/\rho_h\right), \tag{4}$$

where ρ_v and ρ_h denote the number of the sliding window in vertical and horizontal directions respectively. The camera motion attention model can be conducted by the normalized displacement vectors with horizontal and vertical directions. It represents the global motion for the consecutive frames. The slice-based approach enables system complexity to be reduced.

5) *Adjusting by the Center Coherency:* According to the characteristics of human perceptual system, the region closer to the center of the screen, the more attention will be attracted. For each frame, the visual attention will integrate the proposed attention maps for every involved object and weighted by a Gaussian template concentric with the center of frame. Assume that the frame f_i contains three feature maps, the visual attention model of frame f_i with attention map m can be obtained by:

$$M_{\mathrm{v}}^{\mathrm{m}}(\mathrm{f_i}) = \left[\sum_{\mathrm{o_k}\in \mathrm{f_i}} \mathrm{G}(\mathrm{o_k}) \times \bar{M}_{\mathrm{m}}(\mathrm{o_k})\right] \times M_{\mathrm{cam}}(\mathrm{f_i}), \tag{5}$$

where $m = \{spa, tem, face\}$, $\bar{M}_{\mathrm{m}}(\mathrm{o_k})$ denotes the attention contributions to frame i from the feature map m in object k, and $\mathrm{G}_{\mathrm{o_k}}$ denotes the weighting factor for object k, it can obtained by its location via Gaussian template function. The camera motion attention model M_{cam} is treated as a bias for computing the visual attention as to be emphasized or degraded.

B. Modeling the Contextual Attention

Generally speaking, the contextual analysis is a domain-specific problem because the different types of videos containing different linguistic information. It is difficult to use a generic model to solve all types of video data. Obviously, the contextual attention varies with shot-level or event-level. In this paper, the contextual attention model is defined as the probability of user's interest in particular game situation, which is formulated with the context vector. Unfortunately, it is hard to observe all statistical information from video frame. Hence, we not only adopt the observed data from the superimposed caption box (SCB), but also employe the historical statistics data. For baseball game, the context vector of SCB consists of six components which are $\Lambda = \{\lambda_i|_{i=1\sim 6}\}$ as reported in Table I.

1) *Implicit Factors:* The contextual information is divided into three classes according to the relationship between the value of the context and the degree of attracting viewers, as proportionally, specifically, or inversely tendencies. A set of implicit factors $\{\mathrm{F}_l | l = 1, 2, \ldots, l\prime, \ldots, L\text{–}1, L\}$ are used for modeling the characteristics of the human interests. Each of implicit factors can be classified as one of three classes:

$$\Omega(F_{l\prime}) \in \{\omega^p, \omega^s, \omega^i\}, \tag{6}$$

where $\Omega(.)$ denotes the classifier, ω^p, ω^s, and ω^i represent the corresponding factor as a proportional type, specific type, and inversely type respectively. In this paper, we not only use the implicit factor in the current frame f_i called as static implicit factor $\hat{F}_i$, but it also use the historic statistics called as dynamic implicit factor $\tilde{F}$. Let $\psi_c(f_i)$ indicate the contextual attention score of frame f_i, which is contributed from the static implicit factors $\hat{F}_{\mathrm{i}}^{\mathrm{l}}$, where l denotes the number of annotations available in that moment and $\tilde{F}_i$ denotes the dynamic implicit factors via historical statistics,

$$\psi_c(f_i) = \sum_{i'=1}^{L} \hat{F}_{i'}^{l} + \tilde{F}_i, \tag{7}$$

In baseball game, four static implicit factors and single dynamic implicit factor are used in determining the contextual attention score.

As shown in Table II, we employ the exponential kernel to model the contextual attention. In order to indicate the attention score into probability density formation, we apply the exponential kernel function in our study. It is also efficient and simply to extend to different types of sports game with score-board broadcasting. Four static implicit factors are proposed, and the α_1–α_5 denote the normalization terms. First, the score difference between teams greatly affects viewers' attention. For example, when scoring run is the same or very close, it indicates that game is very intense. Second, the number of ratio between the strikes and balls can be applied for modeling the viewers' attention. The number of the balls is repeatedly shows from 0 to 3, meanwhile the number of the strikes and outs are repeatedly show from 0 to 2. When λ_5 reaches 3 and λ_6 reaches 2, it indicates that game is received high attention due to the current player may be struck out or get to walk at soon. Third, when less number of remaining innings, it indicates the more attention will be attracted. Therefore, we design the implicit factor λ_1 which represents the number of the inning. Normally, the maximum number of innings is 9. Finally, the high number of the outs infers the strong excitement will be appeared.

In this paper, we concern the historic game statistics. Bill James [19] who is a baseball writer, historian and statistician whose achievement has been widely influential on the baseball

TABLE I
THE CONTEXTUAL ANNOTATION AND THE ASSOCIATED RELATIONSHIPS IN BASEBALL GAME

Annotation	Semantic meanings
λ_1	The current inning
λ_2	The base-occupied situation
λ_3	The current score difference
λ_4	The number of the outs
λ_5	The number of the balls
λ_6	The number of the strikes

TABLE II
THE PROPOSED IMPLICIT FACTORS FOR BASEBALL GAME

implicit factors		Semantic meanings	Models
Static	$\hat{F}_i^1$	The score difference	$\exp(-\alpha_1\lambda_3)$
	$\hat{F}_i^2$	The number of the balls-strikes pairs	$\exp[-\alpha_2(\lambda_5-3)]*\exp[-\alpha_3(\lambda_6-2)]$
	$\hat{F}_i^3$	The number of the inning being played	$\exp[-\alpha_4(\lambda_1-9)]$
	$\hat{F}_i^4$	The number of the outs	$\exp[-\alpha_5(\lambda_4-2)]$
Dynamic	$\tilde{F}_i$	The expected runs scored in the remaining inning	As shown in Eq. (8)

TABLE III
EXPECTED RUNS SCORED IN THE REMAINDER OF THE INNINGS

		Base-occupied situation							
{B3, B2, B1}		**{000}**	**{001}**	**{010}**	**{011}**	**{100}**	**{101}**	**{110}**	**{111}**
#outs	0	0.49	0.85	1.11	1.30	1.39	1.62	1.76	2.15
	1	0.27	0.51	0.68	0.94	0.86	1.11	1.32	1.39
	2	0.10	0.23	0.31	0.38	0.42	0.48	0.52	0.65

0(empty), 1(occupied). e.g.: {1, 1, 1} denotes 'base full' case.

field as well as the field of statistics, the game of baseball is one of the most statistical games in sports. The base-occupied situation usually attracts much user interest. Coaches tend to change their strategy considering the base-occupied situation. Jim Albert [20] collected case studies and applied statistical and probabilistic aspects to the baseball game. He conducted a thorough statistical analysis of the data from the National League for the 1987 season, the play-by-play data of which is downloadable from [21]. The statistics of the expected runs scored in the remaining inning for 24 possible pairs (λ_2, λ_4) are shown in Table III. The possible scoring under different base-occupied and number of out scenarios using the historic game statistics are reported. Based on the statistics, the implicit factor can be adjusted by weighting sum of the past attention score as λ_4, i.e., the pre-trained probability $p(\hat{F}_i^4|\lambda_4)$ and the expected runs scored by looking up from Table 3.

$$\tilde{F}_i = \beta_1 P(\hat{F}_i^4|\lambda_4) * \beta_2 LUT(\lambda_2, \lambda_4), \tag{8}$$

where β_1 and β_2 indicate the weights, and LUT(.) denotes the look-up-table function which normalized by the maximal value for the corresponding outs.

III. KEY-FRAME DETERMINATION

In this paper, the contextual attention score is adopted to determine the best number of key-frames from each key moment, meanwhile, the key-frames being determined based on the visual attention score.

A. SCB Segmentation and Modeling

Here, the SCB can be extracted by using the well-defined template construction method [12], cooperating with digital object labeling mechanism, the contextual game status can be obtained. To model the SCB, we apply color distribution to construct a representative template model h_R. The candidate SCB region in the ith frame will be obtained by using the similarity measurement which is formulated as the *Mahalanobis* distance measurement in this paper.

B. Key-Frame Rate Determination

The problem in selecting the sufficient set of frames from a shot as key-frames has become critical issue. Generally speaking, a shot with the more exciting event, it requires the more key-frames. The contextual information is normally represented by shot units. Therefore, it appropriates to determine the key-frame rate according to the contextual attention score. Here, the contextual attention score is obtained by combining a set of implicit factors. Different context combinations reflect the different level of game excitement. Let T_p denote the predefined percentage of the accumulated attention score be required for determining key-frame, N_s denotes the total frames within the shot I, and R_i denotes the key-frame rate of the shot i, then we have

$$R_i = \arg\min_{\hat{R}_i} \left\{ \sum_{j=1}^{\hat{R}_i} \psi_c(f_j) \geq \sum_{k=1}^{N_s} \psi_c(f_k) \times T_p\% \right\}. \tag{9}$$

C. Key-Frame Selection

Basically, the criterion of the key-frame selection is based on two rules: On one hand, the key-frame must be visually significant. On the other hand, the key-frame must be temporally representative. Combining all attention feature maps can meet the first rule. In this study, the camera motion characteristics are treated as the balancing coefficient to support the second rule. Here, the frame-level attention score is quantitatively obtained by the weighting mean of the visual attention which is defined as the combination of the all visual saliency feature maps. A numerical score is derived from the visual characteristics of the all objects in current frame. The denominator of each part is the normalization term, which is the sum of the attention maps for all frames belonging to the same video shot. Based on the pre-defined number of key-frames, the R_i

TABLE IV
THE EXAMPLES WITH DIFFERENT ATTENTION SCORES

#Frame	Integral Visual	Spatial	Temporal	Camera Motion		Facial
				τ_v^i	τ_h^i	
573	**0.243**	0.053	0.688	**144**	**241**	0.107
650	0.093	0.051	**0.110**	**0**	**11**	0.039
931	**0.691**	**0.506**	0.672	82	242	**0.562**
1716	0.051	0.069	**0.041**	**0**	**11**	0.091
3450	0.186	0.038	0.253	**24**	**129**	0.038
4001	**0.286**	0.056	0.394	26	41	0.047

key-frames $\{F_k^*\}$ with the maximal visual attention score ψ_v, we have

$$F_k^* = \bigcup_{i=1}^{R_i} \left\{ \arg\max_{f_i} \left[\psi_v(f_i) \right] \right\}, \tag{10}$$

where

$$\psi_v(f_i) = \sum_{j=1}^{m} \gamma_j M_v^j(f_i), \tag{11}$$

where $M_v^j(f_i)$ denotes the visual attention model for frame i with feature map m, referring to (5), and γ_j denotes the weighting coefficients for the visual feature maps j.

IV. EXPERIMENTAL RESULTS

A. Preliminaries

We collected about 83,430 frames in 1,012 shots from six video sequences. The video streams are AVI format digitized at 10 frames/s, and the resolution of each image frame is $352 \times 240 \times 24$ bits in true color. In order to evaluate our system, the rule-complicated sports videos such as baseball and basketball games are employed. The video frames used in our experiments were captured by a digital recorder via the TV cable broadcasting. We adopted the sports programs including National Basketball Association (NBA) and Major League Baseball (MLB).

B. Shot Boundary Detection

In this work, the shot boundary is determined by the visual semantic units [22]. The main problem for segmenting a video sequence into shots is the distinguishing ability between the scene breaks and the normal object moves. These changes result from the dynamic objects or the camera motions (e.g., zooming and panning). In this paper, the camera motion attention model is provided, thus we employ this feature to deal with shot boundary detection. Satisfactory results can be obtained to manage the semantic information in the video shot.

C. Evaluations of System Responsiveness

To show the system responsiveness, Table IV shows the attention scores for different examples of the six representative frames shown in Fig. 1.

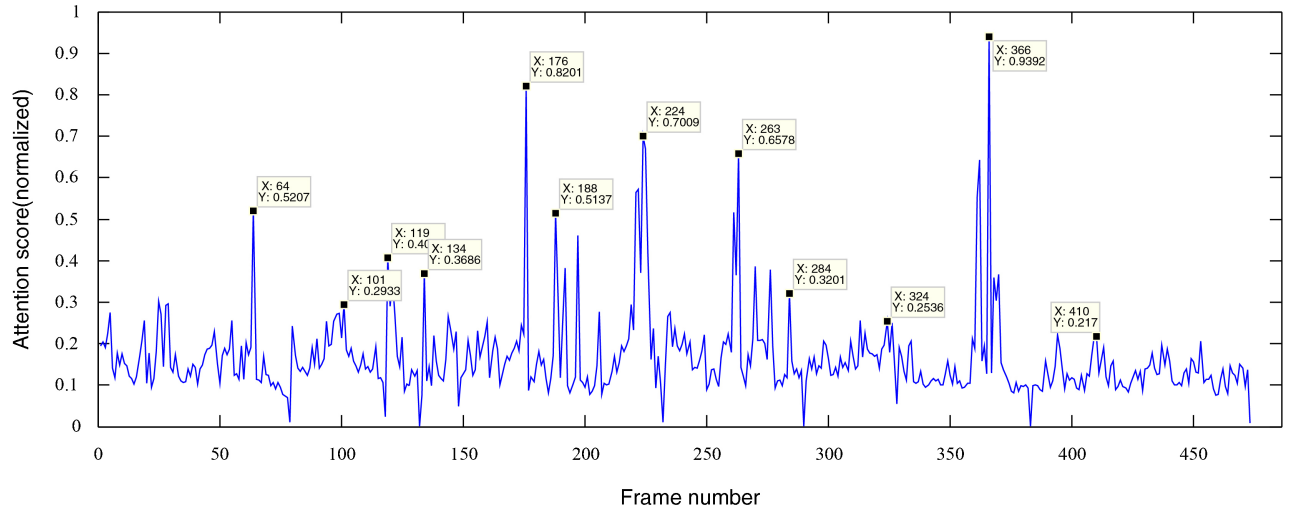
The testing frames #650 and #1716 are static view, they receive the low temporal attention scores and low camera

Fig. 1. The six representative frames.

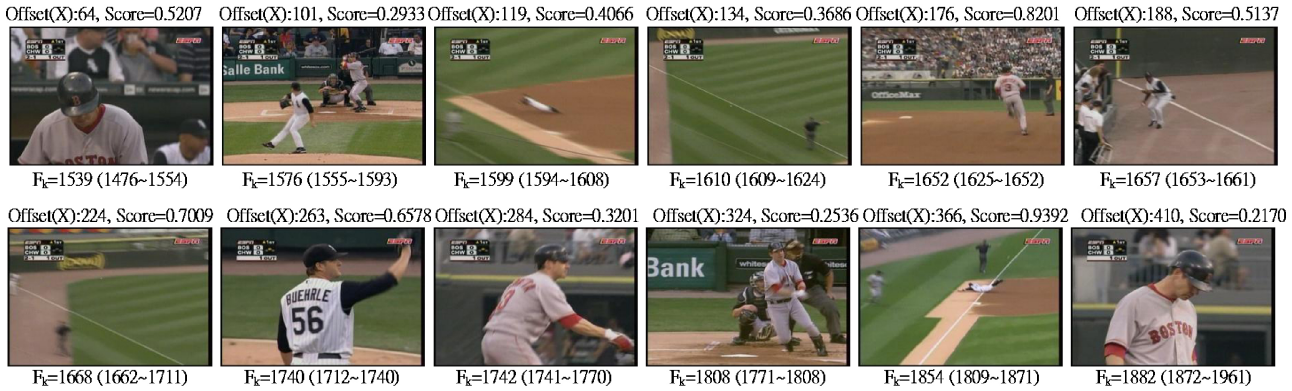
Fig. 2. The results of the key-frame determination method.

motions, resulting in the low visual attention. The testing frame #931 represents a close-up view with zoom-in camera motion, however the object is stable. Therefore, it performs the high visual attention. The testing frame #4001 shows a mid-distance view with local motion and the zooming effect. However, the camera motion is zoom-in and face is located near the center of the frame, it produces high attention score. The testing frame #3450 stands for a mid-distance view with middle size of face. The camera panning effect produces the high attention score. The testing frame #573 receives a high attention due to the rapid panning.

To test another long video sequence, the results of key-frame determination are shown in Fig. 2. It shows that the proposed scheme is capable of extracting a suitable number of key-frames. According to the attention score, the extracted key-frames are highly correlated with human perceptual feelings. However, two redundant key-frames were extracted as shown in the second and fourth rows of Fig. 2. It is because that the detected shot boundaries are wrong due to the fast panning and tilting camera motion. Apparently, all frames within the shot show the identical distribution with the attention scores. The proposed key-frame determination algorithm can be applied for slow-motion replay clips such as the last key-frames in Fig. 2. We assume that each shot will select at least a key-



(a)



(b)

Fig. 3. Results of the key-frame detection method: (a) the curve of the attention score and (b) the key-frame with the shot boundary and frame number.

frame from a video shot. It is reason why a few key-frames look like a normal play.

D. Key-Frame Determination

To evaluate the distributions of the attention score, it is identical with the game excitement. Fig. 3 shows that a hit event followed by the slow motion replay. When the hit event occurs, the rapid scene change is used for providing the most content about the exciting views such as the duration between the offset positions #101 and #134. In offset position #176, the runner was attempting to occupy the second base. In offset position #263, the camera was tracing the pitcher with a close-up view, thus the attention score is relatively high. The duration between the offset positions #284 and #366 represents the replay scene with slow motion. Therefore, the distribution of the attention scores in this period is apparently flat. The highest attention score occurred at the offset position #366, it is because that the zooming camera motion has been used for capturing the most exciting moment when the third base defender missed the fast rolling ball. The last key-frame is a stable runner close-up view. Therefore, the attention scores within this shot are very similar.

E. Subjective Evaluation of the System

To evaluate the system responsiveness, we attempt to model the contextual information for basketball video, the implicit factors can be constructed as $\{\lambda_5$: quarter number, λ_6: score difference, λ_7: minute being played, λ_8: second being played, λ_9: shot clock left, λ_{10}: total team fouls, λ_{11}: shooting duration$\}$, which denote {the number of the quarter being played, the current score difference, the time remaining for minutes, the time remaining for seconds, the shot clock, the number of team fouls, and shooting distance}. In this experiment, the contextual attention is integrated by the following models,

1) $\hat{F}_i^1 = \exp[-\alpha_6(\lambda_5 - 4)]$,
2) $\hat{F}_i^2 = \exp(-\alpha_7\lambda_6)$,
3) $\hat{F}_i^3 = \exp[-\alpha_8\ (\lambda_7 - 15)]$*$\exp[-\alpha_9\ (\lambda_8 - 60)]$,
4) $\hat{F}_i^4 = \exp[-\alpha_{10}(\lambda_9 - 24)]$,
5) $\hat{F}_i^5 = \exp[-\alpha_{11}(\lambda_{10} - 25)]$
6) $\hat{F}_i^6 = \exp(-\alpha_{12}\lambda_{11})$,

where $\alpha_6 = 0.2$, $\alpha_7 = 0.2$, $\alpha_8 = 0.15$, $\alpha_9 = 0.15$, $\alpha_{10} = 0.1$, $\alpha_{11} = 0.1$, $\alpha_{12} = 0.1$.

We have evaluated the performance of the proposed key-frame extraction algorithm for baseball game. However,

TABLE V
SUBJECTIVE EVALUATION ON KEY-FRAME RANKING (%)

	HR	R	N	D	HD	Avg. ranking
Baseball	25.93	52.86	18.18	2.53	0.51	1.27
Basketball	17.52	60.20	21.26	0.68	0.34	1.11
Average	**21.72**	**56.53**	**19.72**	**1.60**	**0.42**	**1.19**

there are no benchmarking or ground truth results for key-frame determination algorithms so far, we do not perform any comparison between the proposed algorithm and others. Therefore, we attempt to evaluate the robustness of the system by subjective user studies. We invited ten subjects to review the testing video sequence in advance and assess each selected key-frame as {highly representative (HR), representative (R), neutral (N), redundant (D), and highly redundant (HD)} with the quantified score: {3, 1, 0, –1, –3}. Table V demonstrates the score ranks of the subjective evaluation. The average scores are reported in the right columns of Table V. These results indicate that the proposed attention model and key-frame determination algorithm are consistent with human perception.

V. CONCLUSION

In this paper, a novel key-frame detection method was proposed by integrating the object-based visual and contextual attention models. The result of key-frame determination indicated the proposed attention model and key-frame determination algorithm were consistent with human perception. Moreover, the key-frame rate determination was indicated successfully using the contextual attention score, meanwhile, the key-frames were determined from the all visual attention scores. Integrating the object-based visual attention model and the contextual attention model not only produced the more precisely human perceptual characteristics, but it also effectively determined the type of video content that attracted much more of the viewers' attention. The proposed algorithm can be easily extended to the other commercial sports videos broadcasting with embedded SCB.

REFERENCES

[1] M. Naphade, I. Kozintsev, and T. S. Huang, "A factor graph framework for semantic video indexing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 1, pp. 40–52, Jan. 2002.

[2] H. C. Shih and C. L. Huang, "MSN: Statistical understanding of broadcasted baseball video using multi-level semantic network," *IEEE Trans. Broadcast.*, vol. 51, no. 4, pp. 449–459, Dec. 2005.

[3] H. C. Shih, J.-N. Hwang, and C. L. Huang, "Content-based attention ranking using visual and contextual attention model for baseball videos," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 244–255, Feb. 2009.

[4] A. Graves and M. Lalmas, "Video retrieval using an MPEG-7 based inference network," in *Proc. ACM SIGIR*, 2002, pp. 339–346.

[5] D. Zhong and S. F. Chang, "Spatio-temporal video search using the object-based video representation," in *Proc. ICIP*, 1997, pp. 21–24.

[6] M. J. Pickering and S. Ruger, "Evaluation of key-frame based retrieval techniques for video," *Comput. Vis. Image Und.*, vol. 92, no. 2, pp. 217–235, Nov.–Dec. 2003.

[7] H. C. Shih and C. L. Huang, "Content-based multi-functional video retrieval system," in *Proc. IEEE Conf. Consum. Electron. (ICCE)*, 2005, pp. 383–384.

[8] H. C. Shih, "Key-frame extraction and key-frame rate determination using human attention modeling," in *Proc. ICME*, Jul. 2011, pp. 1–4.

[9] H. C. Shih, "A novel attention-based keyframe detection method," in *Proc. Int. Conf. DICTAP*, Dijon, France, Jun. 2011.

[10] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modeling visual-attention via selective tuning," *Artif. Intell.*, vol. 78 no. 1–2, pp. 507–545, 1995.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[12] H. C. Shih and C. L. Huang, "Content extraction and interpretation of superimposed captions for broadcasted sports videos," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 333–346, Sep. 2008.

[13] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[14] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, "A neural network approach to interative content-based retrieval of video databases," in*Proc. ICIP*, 1999, pp. 116–120.

[15] M. Bicego, M. Cristani, and V. Murino, "Unsupervised scene analysis: A hidden markov model approach," *Comput. Vis. Image Und.*, vol. 102, no 1, pp. 22–41, Apr. 2006.

[16] M. Naphade and T. Huang, "A probabilistic framework for semantic indexing and retrieval in video," in *Proc. ICME*, 2000, pp. 475–478.

[17] C. L. Huang, H. C. Shih, and C. Y. Chao, "Semantic analysis of sports video using dynamic Bayesian network," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 749–760, Aug. 2006.

[18] H.-Y. Chen, C. L. Huang, and C.-M. Fu, "Hybrid-boost learning for multi-pose face detection and facial expression recognition," *Pattern Recognit.*, vol. 41, no. 3, pp. 1173–1185, Mar. 2008.

[19] B. James, *The New Bill James Historical Baseball Abstract*. New York: Simon & Schuster, 2003.

[20] J. Albert, "Using play-by-play baseball data to develop a better measure of batting performance," Bowling Green State Univ., Bowling Green, OH, Tech. Rep., 2001.

[21] *Retrosheet* [Online] Available: http://www.retrosheet.org

[22] H. C. Shih, "Multi-level video segmentation using visual semantic units," in *Proc. IEEE ICCE*, Jan. 2013.