Content-Based Attention Ranking Using Visual and Contextual Attention Model for Baseball Videos

Huang-Chia Shih, *Member, IEEE*, Jenq-Neng Hwang, *Fellow, IEEE*, and Chung-Lin Huang, *Senior Member, IEEE*

***Abstract*—The attention analysis of multimedia data is challenging since different models have to be constructed according to different attention characteristics. This paper analyzes how people are excited about the watched video content and proposes a content-driven attention ranking strategy which enables client users to iteratively browse the video according to their preference. The proposed attention rank (AR) algorithm, which is extended from the Google PageRank algorithm that sorts the websites based on the importance, can effectively measure the user interest (UI) level for each video frame. The degree of attention is derived by integrating the object-based visual attention model (VAM) with the contextual attention model (CAM), which not only can more reliably take advantage of the human perceptual characteristics, but also can effectively identify which video content may attract users' attention. The information of users' feedback is utilized in re-ranking procedure to further improve the retrieving accuracy. The proposed algorithm is specifically evaluated on broadcasted baseball videos.**

***Index Terms*—Attention modeling, contextual analysis, information retrieval, interactive systems, sports videos.**

I. Introduction

NOWADAYS, a considerable amount of digital video contents are disseminated through Internet every day due to the fast progress of advanced communication framework. Effectively measuring users' attention when they observe images and videos has thus become an important task in many multimedia applications, such as multimedia information retrieval, users-content interaction, and multimedia search. The visual and contextual information are two of the most significant cues for inspecting the semantic knowledge of video content. Modeling the visual attention [1] can provide a good solution toward better understanding the video contents, and enable researchers to deal with scene analysis [2], object extraction [3], video summarization [4], and video adaptation [5]. On the other hand, in sports videos, the game status is the most concerned contextual information for subscribers. The producers usually utilize a superimposed caption box (SCB) embedded upon the screen corner to provide the real-time on-going game status information. The embedded captions in sports video programs represent digested key information of the video contents. Taking advantage of prior implicit knowledge about sports videos, we have proposed an automatic context extraction and interpretation system [6] which can be used to monitor the event occurrence. Meanwhile, how to access the useful content for video searching and retrieval mechanisms has become a prevalent research topic [7], [8]. However, the current mechanisms do not provide sufficient analysis of the video content, so that they do not offer the suitable information to user. Client users can only retrieve the desired pre-categorized full video clips through video indexing techniques. The users cannot request the brief overview of the entire video, neither can they be provided the representative summaries of the desired video. Therefore, relevance feedback mechanisms [9], which allow users to interactively choose the preferred retrieved content, have also been proposed [10], [11]. Based on the last retrieval results and the users' feedback to the retrieval system, the system can adjust its feature weights automatically and filter out the better approximations for the users.

A. Previous Work

There have been good deal of research efforts on content-based multimedia mining. Naphade *et al.* [8] proposed a video indexing system using several semantic concepts through the factor graph model. Zhang *et al.* [23] introduced an image annotation and retrieval system, based on a probabilistic semantic model, for visual features and the textual words. The VideoQ [36] developed at Columbia University for automated video object tracking and retrieval uses motion sketches and some visual cues. Doulamis *et al.* [24] attempt to temporally divide the video information into four levels: frame level, key-frame level, shot level, and key-shot level. They used the genetic algorithm (GA) to identify the key frames and the key shots. Visual cues, such as contrast and spatial frequency, have been studied extensively for the purpose of assessing the perceptual responses of different conditional colors or texture properties. Recently, more and more human perceptual features are discovered to understand the human attention in digital image/video search and retrieval systems. For instance, Itti *et al.* [2] proposed a saliency-based visual attention model for scene analysis. Ma *et al.* [4] illustrated a hybrid user attention model, which includes visual and audio features, for video summarization. On the other hand, the probabilistic approach is one of the most used methods to infer the uncertain semantics, which is normally hard to access directly. Most of researchers attempt to bridge the gap

Manuscript received April 16, 2008; revised October 06, 2008. Current version published January 16, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

H.-C. Shih is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan (e-mail: hc.shih@ieee.org).

C.-L. Huang are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, and also with the Department of Informatics, Fo-Guang University, I-Lan, Taiwan, (e-mail:huang.chunglin@gmail.com).

J.-N. Hwang is with the Information Processing Laboratory, Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: hwang@u.washington.edu).

Digital Object Identifier 10.1109/TMM.2008.2009682

between high-level concepts and low-level features by probabilistic modeling, such as (dynamic) Bayesian networks [37], [38], neural networks [39], and hidden Markov models [40].

A lot of indexing and retrieving researches focus on the sports videos applications [38], [32]–[34], [30], [41], [42] due to the continuously increasing demand from audiences. In sports videos, program producers usually attach the SCB on screen to show the updated information of the game. The occurrence of highlight is always synchronized with the caption changes in the SCB. Therefore, the caption is a useful clue for the audience to understand the game status. A general caption extraction and domain-specific text recognition system has been proposed in [33], which combines the transition model in a specific domain to improve the recognition accuracy for baseball games. Sung *et al.* [34] developed a knowledge-based numeric caption recognition system to recover the valuable information from an enhanced binary image by using a *multilayer perceptron (MLP) neural network*. The results have been verified by a knowledge-based rule set designed for a reliable output and applied for live baseball programs. Lyu *et al.* [42] proposed an approach to detect and extract the embedded texts for multilingual video, but the system is developed without the text modeling. Liu *et al.* [43] proposed a post-filtering framework to improve the accuracy of semantic concept detection using the inter-concept association and temporal analysis for concept knowledge discovery.

This paper analyzes how people are interested about the watched video content based on attention models. These models may not be adequate enough to fully reflect the degree of excitement of viewers, therefore the users' feedback should also be taken into consideration. The PageRank algorithm [12], [13] is one of relevance propagation mechanism used by Google, the most popular web search engine, to determine the relevance and importance of a webpage via an iterative algorithm contributed by incoming links (i.e., backlinks). More specifically, the PageRank belongs to connectivity-based ranking [14] and the main idea is that if there are a lot of sites linking to a particular page, that page must have the most relevant contents. Under the so-called *Surfer Random Model*, a surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue to click is controlled by a damping factor d. Assuming the surfer is likely to jump to any page p with equal probability $E(p) = 1/n$, where n is the total number of web pages, the importance of a webpage can thus be iteratively updated and it will be closer to users' actual needs. In this paper, the video content is treated as webpage and the attention rank (AR) is used to indicate the importance value. Obviously, the AR should be affected by the weights of attention characteristics and user preference.

B. Overview of the Approach

Our proposed framework consists of three modules: low-level feature extraction, attention modeling, and attention ranking/re-ranking. First of all, a group of saliency-based feature maps and interpreted context is extracted in the feature extraction module. The visual feature maps are considered first in order to avoid the distortion of the noisy background with the camera motion like zooming, panning, and tilting. We also take into account the contextual description, which denotes the context embedded in the scoreboard image appended by the video producer. This information can be used to predict the excitement score of the current game status. Secondly, visual (i.e., object and camera motion) and contextual attention modeling are utilized to approach the viewer attention with little domain knowledge involved. Finally, the excitement prediction module effectively computes the highlight distribution of on-going game status based on the values of ARs derived from a group of keyframes systematically determined by the system. To further refine the ARs, based on the feedback scores provided for the keyframes selected by users, the system can modify the weights of feature models in attention reranking phase. Reranking the excitement score of the game situation can be used to offer more reliable results to viewers. Our ultimate objective is to develop an extendable and practicable framework that can be applied in different domains of sports videos with small modifications.

To avoid the important content being ignored during the query phase, ranking the content is the more effective approach. Consequently, the highly user-interested contents are normally placed at the top of search results. In this paper, we use the AR to represent the degree of user interests. Higher AR represents the stronger support of the user interests. AR is affected by two components: intra-AR and inter-AR. In a frame-based video analysis, the intra-AR of each frame is based on its visual and contextual attention characteristics. If there are more high-attention objects contained in a frame and with a high-interest contextual description, it is highly probable that this frame has higher excitement score. With respect to an event-based analysis scenario, the inter-AR of each frame is derived based on the relevant keyframes which are located in the same event. In the re-ranking procedure, the inter-AR of each frame will be updated by the properties of keyframes based on two aspects of user feedback: 1) the score of the user's feedback and 2) the contextual similarity with the labeled frames.

C. Organization

The remainder of this paper is organized as follows: Section II addresses three important attention models used in our formulation of ARs: the object attention model, the contextual attention model, and the camera motion attention model. Section III details our proposed framework for attention ranking and re-ranking of baseball videos. Several important mechanisms are discussed, including excitement prediction, weights modification, and re-ranking of attention based on users' feedback. Section IV demonstrates the experimental results of our proposed framework followed by the conclusions in Section V.

II. Visual and Contextual Attention Modeling

Attention is a psychological reflection of a human emotion. In intelligent video analysis applications, an effective scheme to cope with this issue is visual attention modeling [2]. It combines several representative feature models into a single saliency map and then it locates the regions the user is paying attention to. Normally, most researchers use the frame-based attention model to analyze the visual contrast around the frame. However, it is possible for the contrast in the background region to

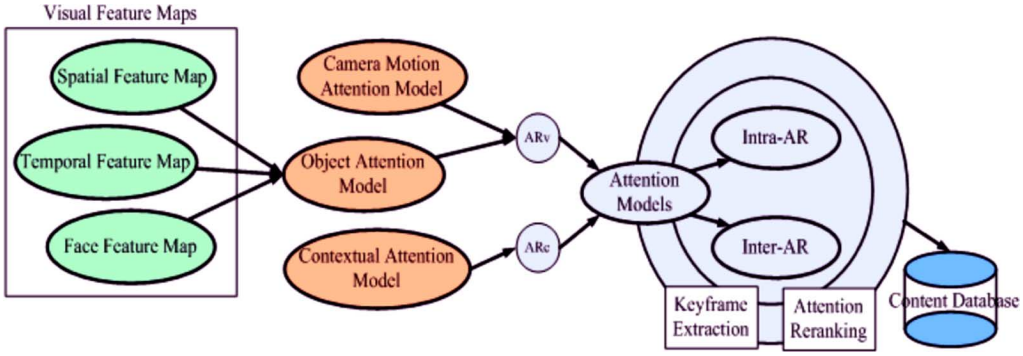

Fig. 1. The object attention model is based on the three types of visual feature map. The contextual attention model is obtained by the context description of the superimposed caption box which changes with the on-game game status. The camera motion attention model is computed by each consecutive frames pair.

be higher than in the foreground object. As a result, a lot of noise can be introduced in the attention computation. Hence, we modify the visual attention model based on the object-based approach, called the object attention model. In addition, we take into account the characteristics of the camera motion and contextual description (such as game statistics), because they both provide numerous meaningful clues of content excitement. The former one shows the excitement of the moment, the latter reflects the intenseness of the on-going game. The conceptual diagram of the attention modeling scheme is shown in Fig. 1. The frame-level attention rank (i.e., *intra-AR*) of frame f_i is quantitatively measured by using an AR-discriminant function which is defined as the weighted summation of the visual attention rank AR_v and the contextual attention rank AR_c.

A. Pre-Processing

1) Segmentation of Video Object: In regard to low-level feature extraction for human attention-driven applications, dealing with the video object (VO) extraction is necessary. Generally speaking, a VO extraction scheme for content-driven applications should meet a critical criterion, that is, the objects which are semantically consistent with human perception should be segmented. Therefore, we have developed an object segmentation method [16], [17] which starts with change detection for stationary background video, followed by many heuristics based on human perception. This method exploits simple frame difference with edge detection to effectively extract shape information of moving objects in spite of the moving objects' suffering from great deal of noise due to changes of lighting conditions. We have also developed an effective video object segmentation algorithm for moving background (i.e., camera moving) cases [18] based on a target tracking scheme using the histogram back-projection refining algorithm.

2) Segmentation of the SCB: We have developed a context extraction and interpretation method [6], [19] to help viewers to catch various on-going events and highlights. Obviously, the SCB sub-image region is either stationary globally or varying locally. Therefore, we combine the color-based local dynamics and temporal motion consistency to locate the SCB from a group of frames (GoF), because there is high color correlation and low motion activity within the region of an SCB image. First, the color distance between two consecutive frames, f_k and f_{k+1}, is computed pixel-wise. The hue-saturation-intensity (HSI) color distance for pixel i in f_k and the corresponding pixel in f_{k+1} is denoted as DC_i. Then, we define the motion activity, MA_i, for every pixel i of each frame f_k using standard block-based motion estimation. We classify pixel i by using the threshold algorithm based on two histograms, *hist*(DC) and *hist*(MA), to determine whether pixel i belongs to the SCB or not. A fine-tuning process is also used to find the best thresholds. More specifically, by analyzing *hist*(DC) and *hist*(MA), we may select thresholds θ_{DC} and θ_{MA}, and obtain the potential SCB mask M_{scb} in the ith frame. Because of the transparent effect, the extracted region of the SCB image region may not be contiguous, thus we further apply the morphological (closing and opening) operations to merge all the pixels classified to the SCB, and obtain a contiguous SCB image region.

B. Calculation of Visual Attention Rank AR_v

Most frame-based visual attention models are sensitive to background clutter. The attention model with a noisy background can further be distorted by camera zooming, panning, and tilting. Therefore, instead of using the frame-based visual attention model, we adopt an object-based attention model, which provides more accurate information for the viewers. Based on their characteristics, the extracted feature maps can be classified into three types: spatial, temporal, and facial.

1) Spatial Feature Map: The static scene can bring a lot of important information. Human eyes are attracted by significant color distribution, strong contrast, and special texture. Three feature maps are chosen in this paper based on the saliency-based attention model [22], to extract the most representative information. These three spatial feature maps are derived from intensity, color contrast (red-green, blue-yellow), and orientation of the chosen keyframes. The spatial feature maps can be obtained by using each attentive feature map for each VO:

$$\bar{\mathcal{M}}_{\mathrm{spatial}}(o_k) = \sum_{(i,j)\in o_k} \delta_s(i,j)/A(o_k) \tag{1}$$

where $A(o_k)$ denotes the area of object k, and $\delta_s(\cdot)$ is a spatial observation function which is defined as

$$\delta_s(i,j) = \max_{(i_w,j_w)\in\theta}\{m(i,j) - m(i_w,j_w)\}$$
$$\text{and } m = \arg\max_{m'\in\{C,I,O\}} \Im^{m'}(i,j) \tag{2}$$

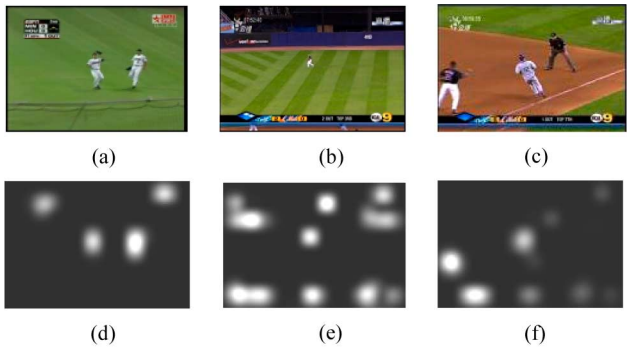

Fig. 2. Spatial feature map estimation. (a)–(c) Original video frames. (d)–(f) Spatial feature maps.

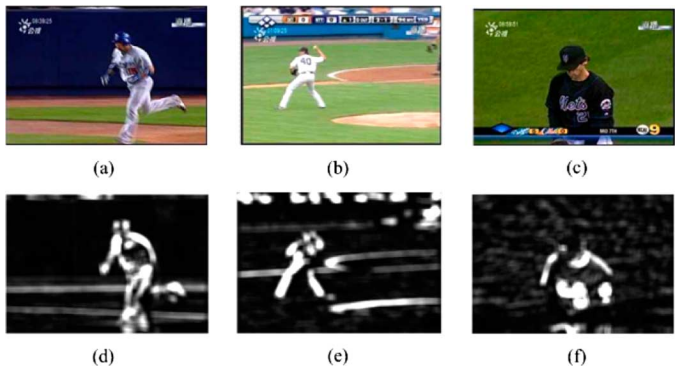

Fig. 3. Temporal feature map estimation. (a)–(c) Original video frames. (d)–(f) Images of MA.

where $\Im^m(i,j)$ denotes the value of a pixel in one of the feature maps $m = \{C : \text{Color}, I : \text{Intensity}, O : \text{Orientation}\}$ that contribute the most to the activity at the object region, and $\delta_s(i,j)$ indicates the maximum value between a pixel (i,j) and its neighboring pixels $(i_w, j_w) \in \theta$, where θ denotes a window centered at (i,j) with its size reflecting the sensitivity of contrast. Fig. 2 shows an example of the spatial feature map estimation.

2) Temporal Feature Map: The extraction of motion vectors is hardly perfect when performed in complex backgrounds or nonsmoothing images. To solve this difficulty, we use the motion activity instead of the motion vector to represent the temporal attention map. Accordingly, we compute the motion activity (MA) associated with every macroblock of size $W \times H$ pixels for each object. Based on the associated location with the object boundary, a macroblock can be classified into one of three types, foreground, neighboring, and background. The MA of the background region barely attracts people, because it provides little information. On the other hand, the MA surrounding the object boundary implies the object's moving behavior with its displacement denoting the moving energy. Finally, the MA within the object region reflects the information on the texture of the corresponding object. Thus, we take different weights to assess the contributions made to the temporal attention based on the type of macroblock. Let $\text{MA}(p,q)$ indicate the motion activity of the macroblock (p,q), where $1 \le p \le W, 1 \le q \le H$, which is derived from the average magnitude of motion displacement vector $\mathbf{dv}(i)$ of a pixel i, normalized by the maximum displacement vector $\mathbf{dv}_{\max}(i)$ within (p,q). We thus define the $\text{MA}(p,q)$ as

$$\text{MA}(p,q) = \sum_{i=1}^{N_b} |\mathbf{dv}(i)| \Big/ N_b |\mathbf{dv}_{\max}(i)| \tag{3}$$

where $|\mathbf{dv}(i)|$ is the magnitude of displacement vector, assuming there are N_b pixels within a macroblock (p,q). Let $\delta_T(\cdot)$ be a temporal observation function weighted by the location function $\ell(p,q)$ which bears different values according to the related locations of the macroblock and the VO boundary: $\text{interior}(=0.8), \text{neighbor}(=1)$, and $\text{exterior}(=0.5)$.

$$\delta_T(p,q) = \ell(p,q) \times \text{MA}(p,q). \tag{4}$$

Then, the temporal feature map for object k in feature map m (object motion) can be defined as

$$\bar{\mathcal{M}}_{\text{temporal}}(o_k) = \sum_{r \in o_k} \sum_{(p,q) \in \Omega_r} \delta_T(p,q) \Big/ N_{\text{MB}} \tag{5}$$

where Ω_{r} denotes the set of attentive macroblocks r, which is located in or neighboring with the object k; N_{MB} represents the number of macroblocks associated with this VO. We assume the size of macroblocks is 8×8. Fig. 3 shows an example of temporal feature map estimation.

3) Face Feature Map: Because the photographer tends to provide the best view for subscribers to check out a region of a person's body (e.g., face) more clearly, it also implies high attention. Even though the state-of-the-art face detection techniques [25], [26] can be applied to achieve this task, they demand fairly large computation times and have large memory requirements. Consequently, we adopt the skin color feature map instead of the face detection schemes (shown in Fig. 4). However, the detected skin region is not only face, but can also be the hands of a human object and some distortion can thus be introduced in the face feature map. The face feature map for an object k can be obtained by inspecting each attentive region and normalized by the area of object

$$\bar{\mathcal{M}}_{\text{skin}}(o_k) = \sum_{(i,j) \in o_k} \delta_F(i,j) \Big/ A(o_k) \tag{6}$$

where the face observation function $\delta_F(i,j)$ in pixel (i,j) is defined as

$$\delta_F(i,j) = \begin{cases} 1, & \text{if}(i,j) \in \text{ skin tone} \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

Empirically, we set the range of skin tone in (r, g, b) color space based on the following criteria to obtain the most satisfactory results: 1) $r > 90$; 2) $(\text{r-g}) \in [10, 70]$; 3) $(\text{r-b}) \in [24, 112]$; and 4) $(\text{g-b}) \in [0, 70]$.

4) Modeling of Camera Motion Attention: There are many cinematic models [27]–[29] have been developed to help the video maker deliver the critical content to viewers. For example,

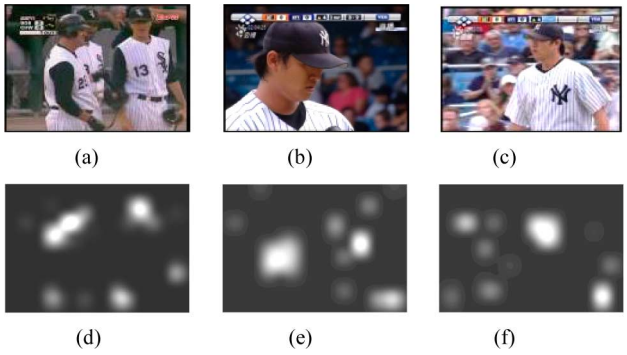

Fig. 4. Face feature map estimation. (a)–(c) Original video frames. (d)–(f) Face feature maps.

photographers can do so by changing the field of view and focal length to catch various on-going events and scenarios. When a highlight occurs, it is critical to keep tracking the key object by moving the camera orientation and changing the focal length. The slice-based motion characterization method [30], [31] is employed to extract the camera motion model. In this paper we replace the time consuming 2-D calculation with two 1-D calculations by projecting the luminance values in vertical and horizontal directions. The procedures for analyzing the 1-D camera motion are illustrated as follows.

Step 1) Assume that the frame size is $W \times H$ pixels. Let f_x be a 1-D horizontal projection profile, where f_y denotes the projection profile in vertical direction. The $f_x(i)$ and $f_y(j)$ are defined as

$$f_x(i) = \frac{1}{H}\sum_{j=1}^{H} p(i,j), \quad \text{for } i = 1,\ldots,W \tag{8}$$

$$f_y(j) = \frac{1}{W}\sum_{i=1}^{W} p(i,j), \quad \text{for } j = 1,\ldots,H \tag{9}$$

where $p(i,j)$ denotes the pixel values at (i,j).

Step 2) Divide the projection profile into small slices, each one with width N.

$$f(s) = f_x(i + (s-1)\times N), \quad \text{for } i = 1,2,\ldots,N$$

where s denotes the slice number, $s = 1,2,\ldots,S$, and $S = W/N$ or H/N.

Step 3) For each pair of consecutive frames, e.g., frame f_k and frame f_{k+1}, slide each projection slice of a former frame to a latter frame, then calculate the sum of absolute difference (SAD) value with a shifted value sv for horizontal and vertical projection profiles,

$$\mathrm{SAD}(n_0,\mathrm{sv}) = \sum_{i=1}^{N} |f_k(i) - f_{k+1}(i+\mathrm{sv})|,$$
$$\text{for } -N/2 \le \mathrm{sv} \le N/2 \text{ and } \mathrm{sv} \in \mathbb{Z} \tag{11}$$

where n_0 denotes the center position index of the slice taken from the projection profile of frame f_k.

Step 4) Find the displacement vector $\boldsymbol{\mathcal{DV}}$ corresponding to the minimum SAD values for the horizontal and vertical fields $(\boldsymbol{\mathcal{DV}}_{\mathrm{H}}, \boldsymbol{\mathcal{DV}}_{\mathrm{v}})$ for each slice with center pixel n_0 by means of the following formulation:

$$\boldsymbol{\mathcal{DV}}(n_0) = \arg\min_{\mathrm{sv}}\{\mathrm{SAD}(n_0,\mathrm{sv})\}, \quad \forall n_0 > 0. \tag{12}$$

Use linear regression to acquire two consecutive displacement curves for fitting $\boldsymbol{\mathcal{DV}}_{\mathrm{H}}$ and $\boldsymbol{\mathcal{DV}}_{\mathrm{v}}$. Based on the distribution of the displacement curves, we can obtain the curve parameters such as intercept $\boldsymbol{I}$ and slope $\boldsymbol{\mathcal{S}}$ for the horizontal $(\boldsymbol{I}_{\mathrm{H}}, \boldsymbol{\mathcal{S}}_{\mathrm{H}})$ and the vertical $(\boldsymbol{I}_{\mathrm{v}}, \boldsymbol{\mathcal{S}}_{v})$.

Camera motion is mainly used in the frame-based model, which is different from the aforementioned object-based visual feature map. With reference to the related research on the user attention modeling [4], the camera motion can serve as a magnifier with a switch in the object-based visual attention model. In this paper, the camera motion attention models, in terms of magnifier function $(\mathcal{M}_{cm})$ and switch function $(\mathcal{SW}_{cm})$, are defined as functions of displacement vector and estimated line parameters, respectively

$$\mathcal{M}_{\mathrm{cm}}(f_k) = 1 + \frac{1}{2}\left(\frac{\|\boldsymbol{\mathcal{DV}}_{\mathrm{H}}\|}{\|\boldsymbol{\mathcal{DV}}_{H\_\max}\|} + \frac{\|\boldsymbol{\mathcal{DV}}_{V}\|}{\|\boldsymbol{\mathcal{DV}}_{V\_\max}\|}\right) \tag{13}$$

$$\mathcal{SW}_{\mathrm{cm}}(f_k) = (|\boldsymbol{I}_H| + |\boldsymbol{I}_V|)/N \tag{14}$$

where $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{\mathcal{DV}}_{H\_\max}$ and $\boldsymbol{\mathcal{DV}}_{V\_\max}$ are used as normalization factors and denote the unit vector times the maximal displacement vector value (i.e., $N/2$) corresponding to $\boldsymbol{\mathcal{DV}}_{\mathrm{H}}$ and $\boldsymbol{\mathcal{DV}}_{\mathrm{v}}$, respectively. The values of $\mathcal{M}_{\mathrm{cm}}(f_k)$ and $\mathcal{SW}_{\mathrm{cm}}(f_k)$ are within the intervals of $[1,2]$ and $[0,1]$ respectively. Eventually, the visual attention model of frame f_k can be emphasized/degraded by the camera motion attention through the values of $\mathcal{M}_{\mathrm{cm}}(f_k)$ and $\mathcal{SW}_{\mathrm{cm}}(f_k)$. The higher the $\mathcal{M}_{\mathrm{cm}}$ and $\mathcal{SW}_{\mathrm{cm}}$ values are, the higher the possibility that viewers may pay attention to it.

5) Modeling of Object-Based Visual Attention: Since there might be many VOs in the video scene, we first define the object-based visual attention as the average contribution from each VO, by integrating the feature maps of objects involved in the frame and weighted by a Gaussian template centered at the center of the frame. If there is no object involved, we simply take the camera motion into account. Assume frame f_i contains M feature maps, the visual attention model of frame f_i with feature map m is presented by

$$\mathcal{M}_v^m(f_i) = \left(\frac{\sum_{o_k \in f_i} \mathrm{G}_{o_k} \times \bar{\mathcal{M}}_m(o_k)}{\mathrm{N}_o}\right) \times (\mathcal{M}_{\mathrm{cm}})^{\mathcal{SW}_{\mathrm{cm}}} \tag{15}$$

where $m = 1,2,\ldots,M$, $\bar{\mathcal{M}}_m(o_k)$ denotes the contributing attention in the form of a feature map from object k to frame i; G_{o_k} denotes the Gaussian weighting function of object k generated by the position; and N_o indicates the number of objects. The camera motion attention model adjusts the visual attention

frame-by-frame emphasized or degraded by the value of $\mathcal{M}_{\rm cm}$, which is powered by the switch function $\mathcal{SW}_{\rm cm}$.

C. Calculation of Contextual Attention Rank AR_c

Different from the visual attention which varies frame-by-frame, the contextual attention updates its values per shot or per event. The problem of modeling the contextual knowledge is a domain-specific problem due to the fact that different sports contain different syntactic information. The context attention derivation for various sports videos has been reported [32]–[35]. Among various contextual information derivation, the superimposed caption box (SCB) is a popular way to provide intrinsic attributes for the audience the on-going game status.

1) Extraction of Contextual Information: A robust contextual information extraction scheme is proposed in [19]. The text embedded in an SCB is located first. Then the extracted texts are classified into characters or digits. More specifically, the extracted characters from baseball SCBs indicate the names of the team and the innings. Whereas the extracted digits present the current scores of the game, or the number of the balls/strikes. The texts are grouped as a semantic unit called the annotative object, whereas the digits are grouped as the digit object. Normally, the digit object comes with certain annotative objects. In baseball videos, the digit object indicating the current inning is usually followed by an annotative object. Finally, the semantic interpretation of annotative or digit objects can be carried out by using the following principles: 1) the digit and the annotative objects are distinguished based on whether the in-between distance is near enough; 2) the digit objects and their associated annotative objects are always on the same horizontal line or vertical line; and 3) two digit objects may sandwich the same annotative object, but two independent digit objects cannot correspond to the same annotative object. Here, we transform the semantic interpretation problem to the object labeling problem by using Relaxation Labeling [20], which assigns each segmented character or digit a label, i.e., an annotative or digit object.

2) Modeling of Contextual Attention: In this paper, we focus on dealing with the baseball video by using the accessible captions of the SCB as shown in Table I. The context vector $\boldsymbol{C}$ can be constructed as $\{\mathrm{INNS}, \mathrm{RUNNERS}, \mathrm{RUNS}_H, \mathrm{RUNS}_V, \mathrm{OUTS}, \mathrm{BALLS}, \mathrm{STRIKES}\}$, which denotes {inning, the base-occupied situation, the score of the home team, the score of the visiting team, the number of outs, the number of balls, the number of strikes}. We attempt to approximate viewer attention behavior in different contextual situations from some observations. For instance, the less innings remaining in a game, the more exciting the situation is, and the smaller the score difference the more it increases viewer attention. The contextual information is divided into three classes which are based on the relationship between the value of the context and the degree of the viewer being excited, i.e., *proportionally, specifically*, and *inversely*. After classifying the contextual description, we use a group of contextual matrices to model the human excitement characteristics.

Let us consider a context vector $\boldsymbol{C}$ comprised of v classes, which represent the semantic meaning from the SCB. Each element of a context vector $\boldsymbol{C} = [c^1, c^2, \ldots, c^n, \ldots, c^v]$ represents a specific annotation. According to the relationship between the semantic meaning of the annotation and the viewer's excitement, each context c^n can be categorized into one of the three classes $\{\boldsymbol{c}_p, \boldsymbol{c}_i, \boldsymbol{c}_s\}$, indicating *proportionally*, *inversely*, and *specifically*, respectively. For example, if $c^n \in \boldsymbol{c}_i$, it denotes that the label of context c^n is inversely proportional to the viewer's excitement. In particular, when a specific number of contexts or a certain ratio of paired contexts appears, it denotes that an exciting event occurs and strongly supports the viewer's attention. The contextual attention model is constructed from a set of contextual matrices $\{\mathcal{M}c_1, \mathcal{M}c_2, \ldots, \mathcal{M}c_m, \ldots, \mathcal{M}c_k\}$, where each contextual matrix $\mathcal{M}c_m$ is defined in terms of a group of composite relations from the context vector $\boldsymbol{C} = \{c^1, c^2, \ldots, c^m\}, 1 \le m \le \nu$. More specifically

$$\begin{aligned}\mathcal{M}c_m &= T(c^1, c^2, \ldots, c^m | c'_p, c'_i, c'_s) \\ &= \frac{T[(c'_p)]}{T[(c'_i)]} \times T[(c'_s, c_{key})] \end{aligned} \tag{16}$$

where T is a transform function denoting the excitement from three contexts classes $\{c'_p, c'_i, c'_s\}$, which is a subset of $\boldsymbol{C}$. Note that $c_{\rm key}$ is the key value of context c'_s that supports additional excitement when the context reaches the corresponding value $c_{\rm key}$. Assume that there are k' contextual matrices obtained from composite contexts with a corresponding normalization term $\{\omega_i | i = 1 \sim k'\}$, then mixed contextual model based on these k' contextual matrices is

$$\boldsymbol{M}_{\rm CAM} = \bigcup_{i=1}^{k'} (\omega_i, \mathcal{M}c_i). \tag{17}$$

Indeed, we only need an approximate distribution of a viewer's excitement through contextual annotation with the context vectors. The first norm exponential kernel is adopted for integrating the contextual information provided by $\boldsymbol{M}_{\rm CAM}$. The reason of employing one-norm exponential function in this paper is because the exponential distribution is not only easy to compute and modify, but also because it is one of the most popular models in signal processing. Empirically, the exponential function also shows the best performance when dealing with the human attention model. Therefore, the contextual attention is integrated by the following contextual matrices

$$\mathcal{M}c_1 = \exp[-\omega_1(|\mathrm{RUNS_H} - \mathrm{RUNS_V}|)] \tag{18}$$

$$\mathcal{M}c_2 = \exp[-\omega_2((3 - \mathrm{BALLS}) + (2 - \mathrm{STRIKES}))] \tag{19}$$

$$\mathcal{M}c_3 = \exp[-\omega_3(7 - \mathrm{RUNNERS})], \tag{20}$$

$$\mathcal{M}c_4 = \exp[-\omega_4(\mathrm{OUTS})]. \tag{21}$$

TABLE I
CONTEXTUAL MEANINGS AND THE ASSOCIATED RELATIONSHIPS FOR BASEBALL GAME

Annotation	Contextual meanings	Relationship	Notes
INNS	The current inning	P, S=9	P: proportionally
RUNNERS	The base-occupied situation	P, S=7 [1]	S: specifically
$RUNS_H$, $RUNS_V$	The current score difference	I, S=0	I: inversely
OUTS	The number of the OUTs	P, S=2	
BALLS	The number of the BALLs	P, S=3	
STRIKES	The number of the STRIKEs	P, S=2	

[1]Base-occupided situation{B3,B2,B1} can be binarized by 1(occupied) or 0(empty). For example: {1,1,1}=7 denotes 'base full' case.

3) The Similarity Measure of the CAM: The contextual attention model is formed based on the context vector $\boldsymbol{C} = [c^1, c^2, \ldots, c^v]$, which includes v classes. From the point of view of the similarity measure, the intrinsic contextual information of different keyframes has different impacts on a viewer's attention. Consequently, a similarity measure between two games with two corresponding context vectors needs to be defined. The Cosine correlation measurement is adopted to compute the contextual similarity between f_l and f_k with the corresponding context vectors $\boldsymbol{C}_l$ and $\boldsymbol{C}_k$ as follows:

$$\begin{aligned} & d(\boldsymbol{M}_{\mathrm{CAM}}(f_l), \boldsymbol{M}_{\mathrm{CAM}}(f_k)) \\ & \quad = \mathrm{cosine}(\boldsymbol{C}_I, \boldsymbol{C}_k) \\ & \quad = \begin{cases} \frac{\boldsymbol{C}_l \cdot \boldsymbol{C}_k}{\|\boldsymbol{C}_l\| \times \|\boldsymbol{C}_k\|}, & \text{if } \boldsymbol{C}_1 \neq \emptyset \cap \boldsymbol{C}_k \neq \emptyset \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{22}$$

III. The Proposed Attention Ranking Framework

A. Excitement Prediction

Intuitively, when a video shot contains higher AR keyframes, it indicates the corresponding video shot is worthy to be looked at. Each shot captured by a video producer always contains certain information to be conveyed. Therefore, there should be at least one frame selected as the keyframe in each video shot for measuring the attention. The frame-level attention rank (i.e., intra-AR) of frame f_i can be quantitatively measured by using a discriminant function which is defined as integrating the visual attention rank AR_v and contextual attention rank AR_c with a bias factor β (set as 0.65 in our simulations)

$$\mathrm{AR}_{\mathrm{intra}}(f_i) = \beta \times \mathrm{AR}_v(f_i) + (1 - \beta) \times \mathrm{AR}_c(f_i) \tag{23}$$

where AR_v can be derived by transforming visual features (including spatial, temporal, and face features) to human attention rank [21]. More specifically, the visual attention rank of frame k, i.e., f_k, can be established by the weighted sum of M different types of attention maps (e.g., $M_{\mathrm{Spatial}}, M_{\mathrm{temporal}}$, and M_{face}, *etc.*)

$$\mathrm{AR}_v(f_k) = \frac{\sum_{m=1}^{\mathcal{M}} \lambda_m M_v^m(f_k)}{\sum_{f_k, f' \in S_i} \sum_{\mathrm{m}=1}^{\mathcal{M}} \lambda_m M_v^m(f')} \tag{24}$$

where λ_m denotes the weights of the m-th attention map$M_v^m(f_k)$, which is a numerical value derived from the visual characteristics of several segmented objects within frame k. The denominator is the normalization term, which is the sum of attention maps for all the frames $\{f^{'}\}$ belonging to the same shot S_i.

On the other hand, let $\mathrm{AR}_c(f_k)$ indicate the contextual attention rank of frame k, which is directly affected by context vector $\boldsymbol{C}$. The contextual attention rank, which corresponds to AR_c, of frame k is normally obtained by combing the allk' contextual matrices, which somewhat reflect the excitement via different context combinations

$$\mathrm{AR}_c(f_k) = \sum_{j=1}^{k'} \omega_j \times \mathcal{M}c_j(C) \tag{25}$$

where $\mathcal{M}c_j$ is the j-th contextual matrix, and ω_j is the weight of corresponding $\mathcal{M}c_j$ determined by user preference. The keyframes can be selected from a collection of J shots $\{S_j, j = 1, \ldots, J\}$ of video sequences by selecting the best P_k frames of the highest AR whose sum of ARs is larger than N percentages of the total attention rank AR_T, within these J shots

$$K_j = \min_{P_k} \left\{ \bigcup_{i=1}^{P_k} f_i \left| \sum_{\mathrm{i}=1}^{P} \mathrm{AR}_{\mathrm{intra}}(f_i) > \mathrm{AR}_T \times N\% \right. \right\} \tag{26}$$

where

$$\begin{aligned} & \mathrm{AR}_{\mathrm{intra}}(f_i) \geq \mathrm{AR}_{\mathrm{intra}}(f_j), \\ & \qquad \forall j \in S_j, j \neq i \text{ and } 1 \leq j \leq J. \end{aligned}$$

B. Weights Modification

The results of the keyframe extraction can be ranked by the score of AR-discriminant function. In addition to the keyframes being systematically selected based on the calculated AR-discriminant function, we can further modify the attention rank value to guarantee the results to be more similar to the user's preference by using the feedback keyframes $\boldsymbol{KF}^*_{\mathrm{FB}} = \{f^*_{\mathrm{r}} | 1 \leq r \leq N_{\mathrm{RT}}\}$, which are manually chosen by the user. It is critical to predict what attention features are preferred by the user. Here, we introduce two scenarios,

Scenario 1) *if the values of the attention model $\mathcal{M}_v^m$ of the keyframes chosen by the users are quite different from each other, then $\mathcal{M}_v^m$ is not a good indicator*. All the attention models of the frames of $\mathbf{KF}^*_{\mathrm{FB}}$ are used to compute the variance vector $\boldsymbol{\Phi} = \{\sigma_1, \sigma_2, \ldots, \sigma_m, \ldots, \sigma_M\}$. These variances are further normalized by using the Gaussian normalization method to create $\{\sigma_m^{'}\}$

$$\sigma'_m = (\sigma_m - \bar{\mu})/3\bar{\sigma} \tag{27}$$

where $\bar{\mu}$ and $\bar{\boldsymbol{\sigma}}$ indicate the mean and the standard deviation of the variance vector $\boldsymbol{\Phi}$ respectively. Based on the $3 - \sigma$ rule [9], the probability of an entry being in the range of $[-1, 1]$ is approximately 99%. The weights λ_m used in (24) can thus be updated accordingly as

$$\lambda_m^{\mathrm{new}} = \lambda_m^{\mathrm{old}} + \Delta_{\max} \times \left(\frac{1}{\sigma'_m} \right) \tag{28}$$

where $\Delta_{\max}$ denotes an adjustment factor computed by factor 2.

Scenario 2) *if all the attention rank values of the keyframes chosen by the user are persistently high, which means that the proposed attention model calculation is a good indicator reflecting the user's interest*. Therefore, the factor$\Delta_{\max}$ can be defined as

$$\Delta_{\max} = \max\{R_{\mathrm{score}}(\forall f_i \in \mathrm{KF}^*_{\mathrm{FB}})\} \tag{29}$$

Note that $\Delta_{\max}$ denotes the confidence of the feedback score by the users for each selected keyframe, and R_{score}: $\{+3, +1, 0, -1, -3\}$ denotes { highly representative, representative, neutral, redundant, highly redundant}. We assume that

the user will choose at least one image frame that is relevant or highly relevant so that the values of $\boldsymbol{\Delta}_{\max}$ and $\boldsymbol{\sigma}'_m$ will not be zero. Finally, the attention model weights are normalized by

$$\tilde{\lambda}_m^{\text{new}} = \lambda_m^{\text{new}} / \lambda_m^T \tag{30}$$

where $\lambda_m^T = \sum_{m=1}^{M} \lambda_m^{\text{new}}$.

C. Attention Reranking

The AR-discriminant function has been used to deal with the self-significance of the attention for each frame based on its visual and contextual characteristics. Additionally, we have attempted to extend the prediction of user interests to the event-level. Based on the user feedback, the ranking procedure is affected by the relevant keyframes located in the same event, the so-called *inter-AR*, also denoted as AR_{inter}. From the viewpoint of the event-level, the calculation of the *inter-AR* for frame f_i in event E_i summarizes the user interests of the game status from three aspects.

1) The resulting score of the keyframes located in E_i from the user's evaluation.
2) The original AR of f_i calculated by the AR-discriminant function.
3) The similarity of the game status supported by the contextual information between the current frame f_i and the corresponding frame f'.

From these observations, we built the re-ranking model in *inter-AR* as follows:

$$\begin{aligned}\text{AR}_{\text{inter}}(f_i) &= e \times \sum_{f' \in (\text{KF}_{\text{FB}}^*, E_i)} \frac{R_{\text{score}}(f') \times \text{AR}(f_i)}{1 + d(\mathbf{M}_{\text{CAM}}(f_i), \mathbf{M}_{\text{CAM}}(f'))} \\ &\quad + (1-e) \times R_{\text{SM}}\end{aligned} \tag{31}$$

where R_{SM} is similar to the *Random Surfing Model* in Google's PageRank algorithm [12,13]. We assume that $R_{\text{SM}} = 1/\text{N}_{\text{E}}$, where N_{E} is the frame number of the event E_i, and e is a damping factor which is empirically set to be $e = 0.85, \text{KF}_{\text{FB}}^*$ denotes the set of user selected keyframes in E_i, and these keyframes should have a higher attention rank and help promote the AR of the other frames within the same event. Basically, at least one frame will be selected as the keyframe in a gradually moving video shot S_i. For event E_i, there are several keyframes collected. Thus, the new AR can be updated by adding the *inter-AR*

$$\text{AR}_{\text{new}}(f_i) = \text{AR}_{\text{old}}(f_i) + \text{AR}_{\text{inter}}(f_i) \quad \forall f_i \in E_i. \tag{32}$$

The re-ranking system gathers the relevant keyframes determined by the AR-discriminant function, and then, users can retrieve the video sequence based on the new ranking score (i.e., AR_{new}). The most interesting video clips and events can be found through few interactions. The proposed effective attention models can help the system to speed up the retrieving procedures, not only the frame-level retrieval, but also the shot-level and event-level.

TABLE II
TESTING SEQUENCE

No.	Teams	#Shots	#Frames
I	DET vs. NYY	240	17,370
II	BOS vs. CHW	177	10,347
III	LAD vs. NYM	194	13,644

Fig. 5. Representative frames of extracting attention model.

TABLE III
SCORES OF DIFFERENT ATTENTION MODEL FOR REPRESENTATIVE FRAMES

#Frame	Visual AR	Spatial	Temporal	Face	Camera Motion	
					DV_{v}	DV_{H}
78	0.073	0.048	**0.0047**	0.102	**0**	**0**
606	**0.365**	**0.182**	0.189	**0.260**	166	251
1106	**0.189**	0.081	0.174	0.158	**176**	**278**
4753	**0.155**	**0.182**	0.044	**0.215**	57	232
5850	0.078	0.129	**0.018**	0.108	66	107
6697	0.096	0.054	0.0057	**0.484**	**0**	**0**

IV. EXPERIMENTAL RESULTS

A. Experimental Dataset and Setup

We collected three video sequences, for a total of about 41 361 frames in 610 shots as shown in Table II. The video streams are AVI format digitized at 10 frames per second, and the resolution of each image frame is $352 \times 240 \times 24$ bits in true color. We chose quite complicated sports videos, baseball games, for the performance evaluation. The video frames used in our experiments were captured by a SONY RDRGX315 DVD recorder from the TV broadcasting programs of the *Major League Baseball* (MLB) in the 2006 season.

B. Measurement of Attention Model

Each VO contributes the attention level to the corresponding frame in many ways. Our proposed object-based attention rank is used as an effective scheme to measure this score. Table III shows the attention scores in different modules for six frames shown in Fig. 5 selected from testing video sequence III. The values of each column denote the score of attention model. Obviously, it can faithfully reflect the content attention via spatial, temporal, facial and global motion. Frames #78 and #6697 are globally static, so they have a decreased visual ARs. Frame #5850 is zooming in but object is stable, therefore, the visual AR

Fig. 6. Keyframe extraction for different event scenarios: (a) caught stealing with replay, (b) double with replay, and (c) home run with replay.

is also low. Frame #606 and frame #4753 both are the mid-distance views with local motion and the camera is horizontally panning. Moreover, the face is clear and near the center of the frame, they both get a high attention score. However, the frame #6697 is a close-up view resulting in a higher face attention but the static scene decreases AR. The frame #1106 has a high attention due to the rapid running of the batter who is located in the center region of the frame.

C. Typical Performance

Although the topics of video content mining have been intensively investigated, there is no perfect solution due to the difficulty of measuring the viewers' actual perceptual attention. The definition of interesting frame or the so-called keyframe is actually user-dependent. Generally speaking, the critical event is always composed of several video shots, and each shot captured by the video producer implies a certain semantic meaning of what he/she wants to convey. Therefore, there should be at least one frame selected as a keyframe of each video shot. Fig. 6 shows the results of the single keyframe extraction for various event scenarios and the corresponding calculated ARs. Three events are illustrated in this figure including caught stealing with replay, double with replay, and home run with replay. When an event contains a substantial amount of excitement, different angle shots and slow-motion replays will be followed as shown in Fig. 6(a). In our observations, there are usually four to six keyframes chosen for each highlighted event along with some additional keyframes for replay.

Unfortunately, some corresponding ARs are miscalculated due to the underestimated motion attention model, such as frames #4944 where duplicate choice (with frame #4949) resulting from the fast zoom-in as shown in Fig. 6(a), frame #1599 whose blurred frame due to the fast panning as shown in Fig. 6(b), and frame #4900 where duplicate choice (with frame #4942) resulting from the clustered object motion in Fig. 6(c).

D. Subjective Evaluation of the Re-Ranking Performance

During the re-ranking experiments, six subjective reviewers were asked to briefly review the testing video sequence in advance. Then, based on their personal subjective opinions, an as-

TABLE IV
EVALUATION ON KEYFRAME RANKING (%)

No.	HRP	RP	NE	RD	HRD	Avg. R_{score}
I	30.72	35.21	24.47	9.62	0.00	1.18
II	20.09	26.50	45.08	8.34	0.00	0.79
III	24.98	34.89	29.18	10.01	0.95	0.97
Avg.	**25.26**	**32.30**	**32.91**	**9.32**	**0.32**	**0.98**

TABLE V
EVALUATION ON KEYFRAME RE-RANKING (%)

No.	HRP	RP	NE	RD	HRD	Avg. R_{score}
I	45.48	23.86	27.09	3.57	0.00	1.57
II	28.14	34.58	32.08	5.20	0.00	1.14
III	40.87	31.15	25.62	2.68	0.00	1.36
Avg.	**38.16**	**29.86**	**28.26**	**3.81**	**0.00**	**1.35**

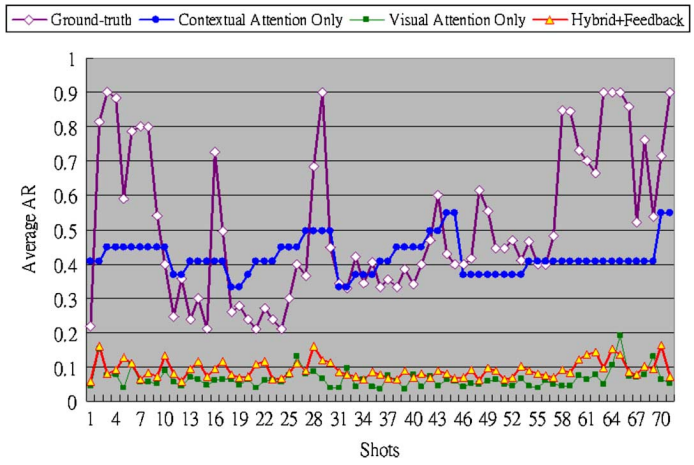

Fig. 7. Performance of attention ranking with visual attention information only, and combining the visual and the contextual attention models comparing to the ground-truth.

sessment was given for each selected keyframe, i.e., highly representative (HRP), representative (RP), neutral (NE), redundant (RD), and highly redundant (HRD). In order to obtain a quantitative evaluation, we quantified each reviewer's assessments for each selected keyframe into scores $3, 1, 0, -1$, and -3, corresponding to HRP, RP, NE, RD, and HRD, respectively. The precise definitions of the assessments are as follows.

1) HRP: Highly representative, meaning right at the key moment.
2) RP: representative, meaning that viewers probably were gazing at the same region, and that this frame could be helpful for understanding the content.
3) NE: neutral, meaning a standard frame, might imply the particular information, but barely intuitive, nothing special.
4) RD: redundant, can be discarded, not relevant to the content.
5) HRD: highly redundant, without information, hard to understand or recognize.

The ranking results for all video sequences are shown in Table IV, where the equal weightings of attention models are assumed. The re-ranking performance is shown in Table V. The average ranking/re-ranking scores are listed in the right-most column of Tables IV and V. As shown in Table IV, the average R_{score} for all of the testing sequences from 0.98 to 1.35, shows an improvement of the HRP rate by about 12.9%, and the RD rate is decreased to 5.51%. These promising results show that the proposed attention model and the attention re-ranking is quite consistent with human perception. Unlike the computer centric approach where the user is unable to provide feedback, the proposed interactive approach allows users to make their decision together with the system. The proposed method greatly enables higher precision, content driven information mining.

E. Objective Evaluation

To evaluate the objective attention ranking performance, the average AR of all the frames in that shot is utilized. For comparison, we also show the ground-truth collected from subjective reviewers as the diamond mark in Fig. 7. The combined attention ranking (i.e., $\text{visual} + \text{contextual}$) is compared to the performance of using only individual attention. The contextual attention curve is shown by the blue line with solid circle mark. The steady flat intervals in Fig. 7 indicate that contextual information does not change. Since the context within SCB will usually change immediately following the occurrence of a new event, which usually consists of one or more than one shots. This is reflected by several dynamic changes of contextual attention (in blue) are made between events as highlighted by the grond-truth peaks (in purple).

Nevertheless, it is still hard to know the exact scores of the viewer excitement. We thus attempt to estimate a rising and falling distribution that can be close to ground-truth. One can see that the combined attention ranking works well especially when the part of the visual information is not evident enough. For example, a defensive event occurs in shot #27, which has a low visual AR, suffers from a small object in the distance. However the contextual attention value is relatively high. One may also notice that the flat contextual values from shot #53 to shot #68 due to the replay of the same event by slow-motion in different angles. We can observe that the combined approach creates ARs with their peaks relatively more consistent with the ground-truth ARs subjectively created by reviewers, when compared with using only visual or contextual attention model alone.

To be more specific, the correlations among three resulting AR curves with the ground-truth AR curve are computed to justify the claimed improved performance. The correlation values are 0.32, 0.41, and 0.45 for applying visual attention information only, contextual information only, and combined mechanism with user feedback respectively. Overall, it shows that the use of the combined attention model reflects better with the human excitement score than that of using the other methods.

V. CONCLUSION

In this paper, we proposed a content-driven attention ranking strategy which enables client users to effectively browse the videos according to their preference. A systematic object-based attention modeling was adopted to avoid the problems of unpredictable noises resulting from clutter and useless background

noise. The proposed system not only could more accurately reflect the human perceptual characteristics but it also effectively discriminated the video contents that might attract the viewers' attention. With the examples of baseball videos, a novel and well-defined algorithm for modeling the contextual description of the superimposed scoreboard images was introduced. The contextual attention model is successfully integrated with the visual attention models in this paper to deal with the attention ranking and re-ranking. Finally, the relevance feedback strategy is efficiently incorporated to update the resulting ARs by taking into account the user feedback preference. The system has a good potential to be applied to different sports domains with the superimposed caption box embedded in the video frames.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments, which were crucial in improving this paper.

REFERENCES

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

[2] 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.

[3] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neur. Netw.*, vol. 19, pp. 1395–1407, 2006.

[4] 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.

[5] W.-H. Cheng, C.-W. Wang, and J.-L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.

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

[7] D. Zhong and S. F. Chang, "Spatio-temporal video search using the object-based video representation," in *Proc. IEEE ICIP'97*, Santa Barbara, CA, Oct. 1997.

[8] M. R. 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.

[9] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[10] R. Zhang and Z. Zhang, "BALAS: Empirical Bayesian learning in the relevance feedback of image retrieval," *Image Vis. Comput.*, vol. 24, no. 3, pp. 211–223, Mar. 2006.

[11] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," *Knowl. Eng. Rev.*, vol. 18, no. 2, pp. 95–145, Jun. 2003.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, The Pagerank Citation Ranking: Bringing Order to the Web Stanford Univ., Palo Alto, CA, Stanford Digital Library Technologies Working Paper, 1999-0120, 1998.

[13] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proc. 7th Int. World Wide Web Conf.*, New York, 1998, pp. 107–117.

[14] M. R. Henzinger, "Hyperlink analysis for the web," *IEEE Internet Comput.*, vol. 5, no. 1, pp. 45–50, Jan.–Feb. 2001.

[15] L. Itti and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," in *Proc. SPIE Human Vision and Electronic Imaging IV*, San Jose, CA, Jan. 1999.

[16] C. Kim and J.-N. Hwang, "Fast and robust moving object segmentation in video sequences," in *Proc. IEEE ICIP'99*, Kobe, Japan, Oct. 1999.

[17] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 122–129, Feb. 2002.

[18] C. Kim and J.-N. Hwang, "Video object extraction for object-oriented applications," *J. VLSI Signal Process.—Syst. for Signal, Image, and Video Technol.*, vol. 29, no. (1/2), pp. 7–22, Aug. 2001.

[19] H. C. Shih and C. L. Huang, "Semantics interpretation of superimposed captions in sports videos," in *Proc. IEEE-MMSP07'*, Chania, Crete, Greece, Oct. 1–3, 2007.

[20] A. Rosenfeld, R. Hummel, and S. Zucker, "Scene labeling by relaxation operations," *IEEE Trans. Syst., Man, Cybern.*, vol. 6, no. 6, pp. 420–433, Jun. 1976.

[21] H. C. Shih, C. L. Huang, and J.-N. Hwang, "Video attention ranking using visual and contextual attention model for content-based sports videos mining," in *Proc. IEEE-MMSP07'*, Chania, Crete, Greece, Oct. 1–3, 2007.

[22] T. N. Cornsweet, *Visual Perception.*. New York: Academic, 1970.

[23] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, and H. J. Zhang, "A probabilistic semantic model for image annotation and multi-modal image retrieval," *Multimedia Syst. J.*, vol. 12, no. 1, pp. 27–33, Aug. 2006.

[24] A. D. Doulamis and N. D. Doulamis, "Optimal content-based video decomposition for interactive video navigation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 757–775, Jun. 2004.

[25] K.-H. Choi and J.-N. Hwang, "Automatic creation of a talking head from a video sequence," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 628–637, Aug. 2005.

[26] Z. Zhang, R. K. Srihari, and A. Rao, "Face detection and its applications in intelligent and focused image retrieval," in *Proc. IEEE-ICTAI'99*, 1999, pp. 121–128.

[27] D. Arijon, *Grammar of the Film Language.*. Los Angeles, CA: Silman-James, 1991.

[28] F. Beaver, *Dictionary of Film Terms.*. New York: Twayne, 1994.

[29] G. Davenport, T. A. Smith, and N. Princever, "Cinematic primitives for multimedia," *IEEE Comput. Graph. Applic.*, vol. 11, no. 4, pp. 121–133, 2002.

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

[31] M. K. Kim, E. Kim, D. Shim, S. L. Jang, and G. Kim, "An efficient global motion characterization methods for image processing application," *IEEE Trans. Consumer Electron.*, vol. 43, no. 4, pp. 1010–1018, Nov. 1997.

[32] W.-N. Lie and S.-H. Shia, "Combining caption and visual features for semantic event classification of baseball video," in *Proc. IEEE ICME'05*, Jul. 6–8, 2005.

[33] D. Zhang, R. K. Rajendran, and S.-F. Chang, "General and domain-specific techniques for detecting and recognizing superimposed text in video," in *Proc. IEEE ICIP'02*, Sep. 22–25, 2002.

[34] S.-H. Sung and W.-S. Chun, "Knowledge-based numeric open caption recognition for live sportscast," in *Proc. IEEE ICPR'02*, Aug. 11–15, 2002.

[35] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted American football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, Aug. 2004.

[36] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An automated content based video search system using visual cues," in *Proc. ACM Multimedia'97*, Seattle, WA, 1997, pp. 313–324.

[37] M. Naphade and T. Huang, "A probabilistic framework for semantic indexing and retrieval in video," in *Proc. IEEE-ICME'00*, New York, 2000, pp. 475–478.

[38] 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.

[39] N. D. Doulamis, A. D. Doulamis, and S. D. Kollias, "A neural network approach to interative content-based retrieval of video databases," in *Proc. IEEE-ICIP'99*, Vancouver, BC, Canada, 1999, pp. 116–120.

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

[41] B. Li, H. Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer," in *Proc. IEEE-ICASSP'03*, Hong Kong, China, Apr. 2003, pp. 169–172.

[42] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 243–255, Feb. 2005.

[43] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen, "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 240–251, Feb. 2008.

Huang-Chia Shih (M'08) received the B.Sc. degree (with the highest honors) in electronic engineering from the National Taipei University of Technology, Taipei, Taiwan, in 2000 and the M.S. and Ph.D. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2002 and 2008, respectively.

His research interests are content-based video summarization, video indexing and retrieval, object-based video representations, applications of statistical models in multimedia processing, and model-based human motion capturing and recognition. He has published more than 20 journal and conference papers in the areas of sports video analysis and human motion capturing and recognition. From September 2006 to April 2007, he served as a Visiting Scholar in the Department of Electrical Engineering, University of Washington, Seattle.

Dr. Shih has received several awards and prizes, including the Excellent Student award in the field of engineering from The Chinese Institute of Engineers in 2000 and awards from the China Youth Corps in 2000.

Jenq-Neng Hwang (F'01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from the Department of Electrical Engineering, University of Southern California, Los Angeles, in 1998.

After two years of obligatory military service, in 1985 he enrolled as a Research Assistant at the Signal and Image Processing Institute, Department of Electrical Engineering, University of Southern California. He was also a visiting student at Princeton University, Princeton, NJ, from 1987 to 1989. In the summer of 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, where he was promoted to Full Professor in 1999. He also served as the Associate Chair for Research and Development in the Electrical Engineering Department from 2003 to 2005. He has published more than 240 journal and conference papers and book chapters in the areas of image/video signal processing, computational neural networks, multimedia system integration, and networking.

Dr. Hwang served an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1992 to 1994, an Associate Editors for the IEEE TRANSACTIONS ON NEURAL NETWORKS from 1992 to 2000, and an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 1998 to 2006. He is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and an Editor for the *Journal of Information Science and Engineering*. He is also on the Editorial Board of the *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*. He served as the Secretary of the Neural Systems and Applications Committee of the IEEE Circuits and Systems Society from 1989 to 1991 and was a member of the Design and Implementation of Signal Processing Systems Technical Committee for the IEEE Signal Processing Society (SPS). He is a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE SPS. He served as the Chairman of the Neural Networks Signal Processing Technical Committee of the IEEE SPS from 1996 to 1998, and was the Society's representative to the IEEE Neural Network Council from 1996 to 2000. He was the conference Program Chair of 1994 IEEE Workshop on Neural Networks for Signal Processing held in Ermioni, Greece, in September 1994, was the General Co-Chair of the International Symposium on Artificial Neural Networks held in Hsinchu, Taiwan, in December 1995, and chaired the Tutorial Committee for the IEEE International Conference on Neural Networks (ICNN'96) held in Washington, DC, in June 1996. He is the Program Co-Chair of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle 1998. He also served as the Conference Chairs for the IASTED Signal and Image Processing Conference and IASTED Internet Multimedia Systems and Applications in 2006. He is the Special Session Co-Chair of ISCAS 2008 and Program Co-Chair of the International Computer Symposium (ICS) 2008 and International Symposium on Circuits and Systems (ISCAS) 2009. He received the 1995 IEEE SPS's Annual Best Paper Award (with S.-R. Lay and A. Lippman) in the area of Neural Networks for Signal Processing.

Chung-Lin Huang (SM'04) received the B.S. degree in nuclear engineering from National Tsing-Hua University, Hsinchu, Taiwan, in 1977, the M.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1979, and the Ph.D. degree in electrical engineering from the University of Florida, Gainesville, in 1987.

From 1987 to 1988, he was with the Unisys, Orange County, CA, as a Project Engineer. Since August 1988, he has been with the Electrical Engineering Department, National Tsing-Hua University, Hsinchu, where he is currently a Professor. His research interests are in the area of image processing, computer vision, and visual communication.

Dr. Huang received the Distinguish Research Award from the National Science Council of Taiwan in 1993 and 1994, the Best Paper Award from the ACCV, Osaka, Japan, in 1993, the Best Paper Award from the CVGIP Society, Taiwan, in 1996, and the Best Paper Award from the IEEE ISMIP Conference, Taipei, in 1997. In 2002, he received the Best Annual Paper Award from the *Journal of Information Science and Engineering*, Academia Sinica, Taipei.