Feature Article

An Interactive Attention-Ranking System for Video Search

Huang-Chia Shih and Chung-Lin Huang
National Tsing Hua University, Taiwan

Jenq-Neng Hwang
University of Washington, Seattle

The authors present an attention-ranking algorithm that measures user interest level for video frames in a way that is similar to how Google's PageRank algorithm determines the importance of Web sites.

Effectively measuring users' attention when they observe images or videos can be an important factor in the success of multimedia applications, especially when it comes to multimedia search and retrieval systems. Numerous researchers have focused on ways to determine user preference without the need for any explicit user input. In particular, modeling visual attention[1] can be an effective method for understanding video content through scene analysis, summarization, and indexing. With these considerations in mind, we have developed an automatic interpretation and context-extraction system that can be used to monitor events in the domain of sports videos.[2]

The question of how to provide ways for users to access only useful video content (resulting in a reduction in transmission cost) has led to key research focusing on this area. However, existing prototype systems don't provide a sufficient analysis of the video content, so they can't present the information that the user really wants, such as a brief overview or a representative summary of an entire video. Researchers responded to this issue by developing relevance-feedback mechanisms to allow users to interactively choose the preferred content to be retrieved. On the basis of retrieval results and user feedback about those results, feature weights can be adjusted automatically and filters changed appropriately. This method has proven to be a good way to determine relevance.

One of the most popular examples of this technique is PageRank,[3-5] an iterative algorithm used by Google to determine the relevance and importance of a Web page through analysis of inbound links. PageRank relies on connectivity-based ranking, with the main idea being that if there are lots of sites linked to a particular page, then that page must have the most relevant content. The system we present in this article operates according to the same principles, but focuses mainly on how interested people are in the video content they see. Our system relies on a content-driven attention-ranking strategy that enables clients to browse a video according to their preference. To make the system as effective as PageRank, we treat the video content as a Web page and use attention rank to indicate the content's importance value.

Attention ranking of multimedia data is critical when different models have to be constructed from different attention characteristics. This article describes a well-defined attention-ranking mechanism for measuring the importance of objects, frames, and events in various content-driven applications. It also describes a novel technique for understanding and modeling contextual knowledge for sports videos. Lastly, it describes a reranking strategy for updating user interest and accommodating user preference.

System framework

Video abstraction and summarization has been discussed at great length elsewhere, but it remains impossible to fully meet user expectations by providing them with a representative summary of a desired video. Thus a ranking system must be developed to analyze multimedia data and interpret its content significance, either via visual cues or through contextual information. We believe the way to meet this need is by developing a system around several strategies. First, a viewer attention model, derived from user feedback selections, can help determine the combination of extracted features that will best reflect user interest. Second, an adaptive learning strategy can help identify user interest by quantitatively analyzing chosen keyframes. Finally, an effective reranking mechanism can help update the attention models of the chosen keyframes.

1070-986X/09/$26.00 © 2009 IEEE Published by the IEEE Computer Society

As shown in Figure 1, our proposed framework for sports videos consists of three components: low-level feature extraction, attention modeling, and attention ranking and reranking. First, a group of feature maps and interpreted context details are extracted by a low-level, feature-extraction component. Second, visual modeling and attention modeling estimate the viewer's attention. Finally, the attention-rank-prediction component computes the highlight distribution of the ongoing game status on the basis of the attention-rank values, then determines a group of keyframes. On the basis of what keyframes are selected by the users, the system modifies the weights of the feature models in the attention-ranking phase. Reranking the score of the interest level in the game situation allows for more reliable results to be offered to users. To develop an extensible framework that can be applied to different sports videos with only small modifications, we developed several options for each component.

For low-level feature extraction in attention-driven applications to work effectively, it's necessary to deal with video object extraction, meaning objects that are semantically consistent with human perception must be segmented. To do so, we developed an object-segmentation method[6] that detects any change in the stationary background of a video and relies on heuristics that are based on human perception. This technique uses edge detection to extract the shape information of moving objects in spite of the fact that they may suffer from noise due to changes in lighting. We also developed a video object segmentation algorithm for cases where the background moves.[7] In addition, we developed a method for context extraction and interpretation to help viewers note events and highlights.[2]

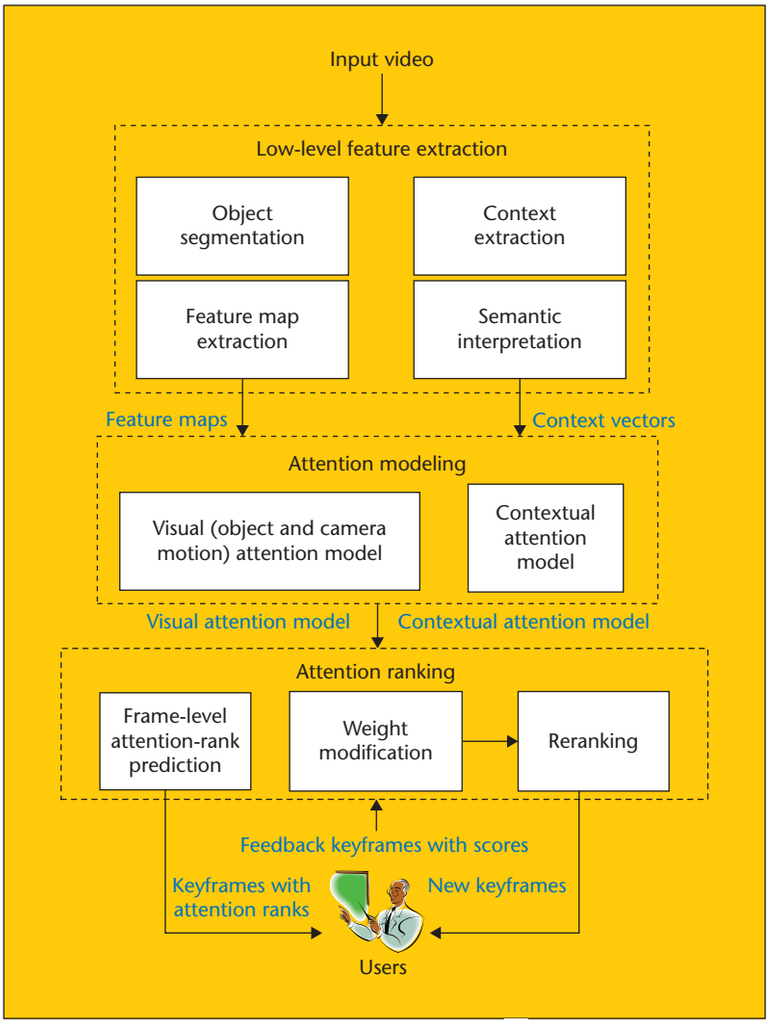

Figure 1. The architecture of the system framework.

Attention modeling

Attention modeling combines several representative feature models into a single salient map and then locates the regions to which the user is paying attention. Normally, most researchers use a frame-based attention model to analyze the visual contrast around a frame. However, it's possible for the contrast in the background to be higher than in the foreground. As a result, a lot of noise can be introduced in the attention computation. Therefore, our proposed attention-modeling system, which we call an *object attention model*, modifies the visual attention model using an object-based approach. In addition, we take into account the characteristics of the camera motion and the contextual description because they both provide numerous meaningful clues about the level of interest in the content. Figure 2 shows a conceptual diagram of our attention-modeling scheme. We measure the frame-level attention rank (that is, *intra-AR*) quantitatively by using an attention rank function defined as the weighted sum of the visual attention rank, AR_v, and the contextual attention rank, AR_c.

Visual attention rank

Most frame-based visual attention models are sensitive to background clutter. Therefore, instead of using a frame-based visual attention model, we adopted the object-based attention model to provide more accurate information to viewers. On the basis of their characteristics,

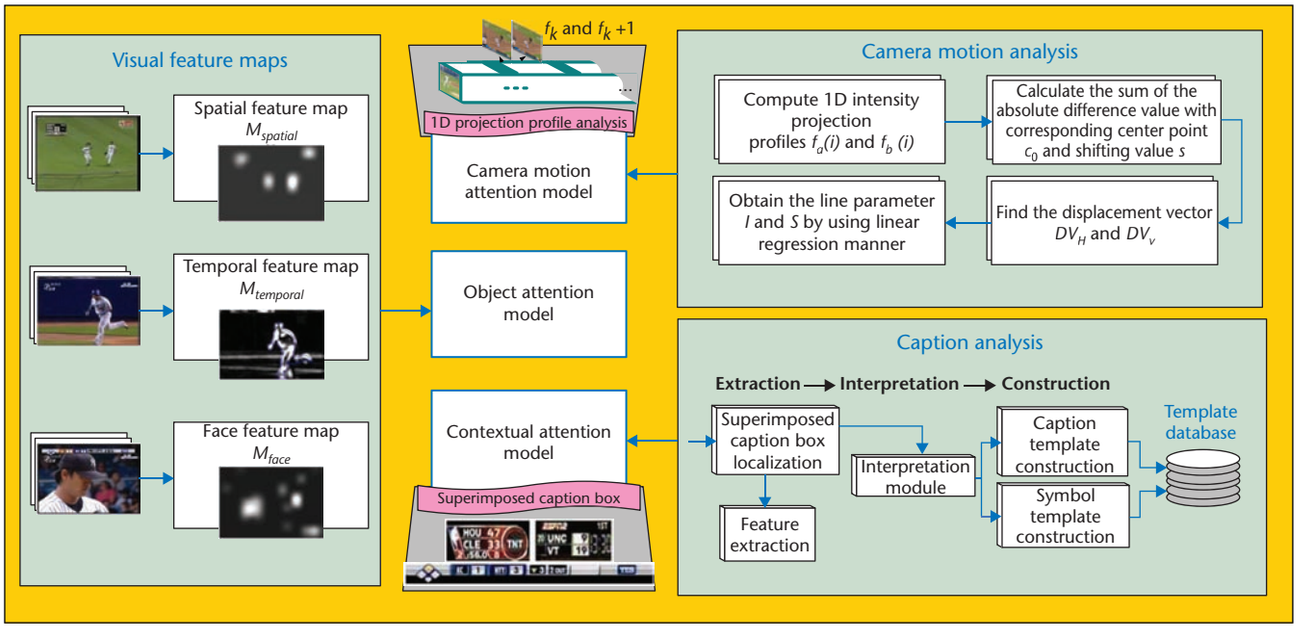

Figure 2. The object attention model is based on three types of visual features. The contextual attention model is obtained by the context description of the superimposed caption box, which changes with the game's status. The camera motion attention model is computed for each consecutive pair of frames.

the extracted feature maps can be classified into three types: spatial, temporal, and facial.

To do spatial feature mapping, our system uses the contrast sensitive function, which is one of the most prevalent techniques for image compression, watermarking, and adaptive transmission.[8,9] We can derive the contrast sensitive function by measuring the contrast threshold for different spatial frequencies. For example, when an athlete plays an important role in a sports event, the video producer might use freeze frames so viewers can see him or her more clearly. A static scene like this can provide a lot of important information. In our system, three feature maps are chosen according to a saliency-based attention model.[10] These spatial feature maps are derived from intensity, color contrast, and the orientation of the chosen keyframes.

Object motion implies semantic information in a video sequence, whether the camera is stationary or moving. Generally speaking, a higher motion energy expresses a more exciting situation. Techniques for estimating motion vectors usually come at a high cost because of the huge amount of computation needed. To solve this problem in representing temporal attention, we use motion activity instead of motion vectors. Consequently, we compute the motion activity (MA) with $W \times H$ macroblocks for each object. We classify macroblocks into one of three types: foreground, neighboring, and background. The MA of the background region is rarely noticed by the viewer because it provides little information. On the other hand, the MA surrounding the object boundary implies the object's moving behavior and the displacement denotes the energy of the movement. Finally, the MA within the object region reflects the information regarding the texture of the corresponding object. Thus, we use different weights, based on the type of macroblock, to assess the contributions made to temporal attention.

The appearance of faces within a frame plays an important role in determining the amount of viewer attention. For instance, video producers will continuously track the face of a person involved in an interesting activity. Even though state-of-the-art face-detection techniques[11] can be used to achieve this task, they demand fairly large computation resources. Consequently, we use a skin color feature map instead a face-detection scheme. However, the detected skin region sometimes also includes the arms of a human object.

One of the main tasks of a photographer is to present the game status as completely as possible within a short time interval. Therefore, many cinematic models, such as field of view and focal length, can help video producers deliver the most critical content to viewers.[12,13] We use a slice-based motion characterization method to extract the camera motion model.[14] However, we replace the time-consuming calculation of 2D elements with two 1D vectors by

projecting the luminance values in vertical and horizontal directions.

The procedures for analyzing 1D camera motion are as follows:

- Assume that the frame size is $W \times H$ pixels. Let f_x be a 1D horizontal projection profile and f_y denote the projection profile in the vertical direction.
- Divide the projection profile into small slices, each one with width N.
- For each pair of consecutive frames, such as 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 the absolute difference (SAD) value with a shifted value sv for the horizontal and vertical projection profiles.
- Find the displacement vector DV corresponding to the minimum SAD values for the horizontal and vertical fields (DV_H, DV_v) for each slice with center pixel n_0 by means of the following formulation:

$$DV(n_0) = \arg\min_{sv}\{SAD(n_0,sv)\}, \quad \forall n_0 > 0$$

- Use linear regression to acquire two displacement curves for fitting DV_H and DV_v.

On the basis of the displacement curve distribution, we can obtain the curve parameters, such as intercept I and slope S, for the horizontal (I_H, S_H) and the vertical (I_v, S_v).

In this article, we define the camera motion attention models in terms of magnifier function (M_{cm}) and switch function (SW_{cm}) as functions of the displacement vector (that is, DV_H, DV_V) and estimated line parameters (that is, I_H, I_V), respectively. The values of $M_{cm}(f_k)$ and $SW_{cm}(f_k)$ are within the intervals of [1, 2] and [0, 1] respectively. We can emphasize or degrade the visual attention model of frame f_k through the values of $M_{cm}(f_k)$ and $SW_{cm}(f_k)$. The higher the $M_{cm}(f_k)$ and $SW_{cm}(f_k)$ values, the higher the possibility that viewers may pay attention to those frames.

At the frame level, we define object-based visual attention as the average contribution from each video object by integrating the feature maps of objects involved in the frame and weighted by a Gaussian template centered in the frame. If there is no object involved, then we simply take the camera motion into account. Let $M_v^m(f_i)$ indicate the visual attention model of frame f_i with feature map m. We can formulate this process as the average attention contributed by all objects involved, weighted by a Gaussian weighting function.[15] The camera motion attention model adjusts the visual attention, frame by frame, and is emphasized or degraded by the value M_{cm}, which is powered by the switch function SW_{cm}.

Contextual attention rank

For sports videos, researchers have focused on several specific domains to for contextual ranking.[16,17] For example, the superimposed caption box (SCB) contains details about a game's status. If the context of the SCB is updated, it indicates a change in the game status. In this article, we propose a new method that not only automatically segments the SCB from the video frame, but also identifies and interprets what semantics are contained in it.[2]

We combine color-based dynamics and temporal consistency to locate the SCB from a group of frames. We compute the color distance between two consecutive frames, f_k and f_{k+1}. We denote the hue-saturation-intensity color distance for pixel i in f_k and the corresponding pixel in f_{k+1} as DC_i. Then we define the motion activity, MA_i, for every pixel i of each frame f_k using block-based motion estimation. We classify pixel i with a threshold algorithm based on histograms $hist(DC)$ and $hist(MA)$ to determine whether pixel i belongs to the SCB. We then use a fine-tuning process to find the best thresholds. By analyzing $hist(DC)$ and $hist(MA)$, we use thresholds θ_{DC} and θ_{MA} to obtain the potential SCB mask M_{scb} in the kth frame. Because of transparency, the extracted region of the SCB might not be contiguous, so we apply a morphology operation to merge all the SCB pixels and obtain a contiguous SCB image region.

Our proposed system transforms a semantic interpretation problem into an object-labeling problem. We locate the text embedded in the SCB first, then classify the text into annotative objects or digital objects. Finally, we employ relaxation labeling[18] to deal with semantic interpretation according to the following principles:

- The digital and the annotative objects correspond if the in-between distance is near enough.
- The digital objects and their associated annotative objects are always on the same horizontal line or vertical line.

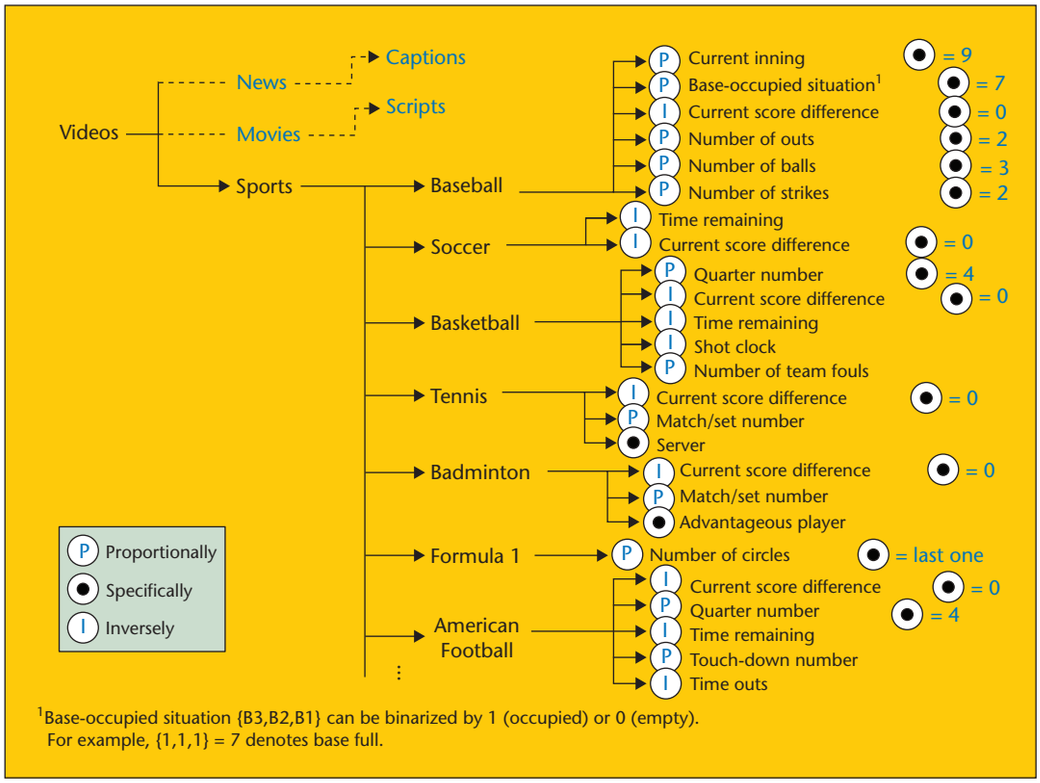

Figure 3. Contextual attention modeling is a domain-specific problem. Generally speaking, the context within the superimposed caption box can be categorized into three types, which are proportionally, specifically, or inversely based on the relationship between its semantic meaning and human excitement.

- Two digital objects may sandwich the same annotative object, but two independent digital objects may not correspond to the same annotative object.

Clearly, contextual information is domain-specific and is an important cue for providing semantic knowledge in many content-analysis applications. In our method, shown in Figure 3, we attempt to approximate viewer attention behavior in different contexts. For instance, the less time remaining in a game the more exciting the situation becomes if the score difference is relatively small. Also, the smaller the score difference, the more viewers will pay attention. We divide contextual information into three classes that are based on the relationship between the value of the context and the possibility of the viewer being interested, proportionally, specifically, and inversely. After classifying the contextual description, we use contextual matrices to model human excitement characteristics.

Let's consider a context vector C consists of v classes, identified by a label l that represents the semantic meaning from the SCB. Each element of context vector $C = [c^1, c^2, \ldots, c^n, \ldots, c^v]$ 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 three sets $\{c_p, c_i, c_s\}$ in accordance with the relevance types $\{r_p, r_i, r_s\}$, which are indicated proportionally, inversely, or specifically. We then use a relevance function $R(.)$ to translate the label of the context to reflect the real-world human attention. For example, $R(c^n) = r_i$ denotes that the label of context c^n is inversely proportional to the viewer's excitement. In other words, a specific number of contexts or a certain ratio of paired contexts appearing indicates that an exciting event held the viewer's attention. A proportional or an inversely proportional context might simultaneously belong to a specific relevance type r_s.

We construct the contextual attention model from a set of contextual matrices $\{Mc_1, Mc_2, \ldots, Mc_m, \ldots, Mc_k\}$, where each contextual matrix Mc_m is defined in terms of a group of composite relationships from context vector $C = \{c^1, c^2, \ldots, c^m\}$, $1 \le m \le v \le$. More specifically,

$$M_{cm} = T(c^1, c^2, \ldots, c^m | c'_p, c'_i, c'_s) = T[(c'_p)]/T[(c'_p)] \times T[(c'_s, c_{\text{key}})]$$

where T is an observation function denoting the level of interest from three sets of classified

contexts $\{c'_p, c'_i, c'_s\}$ that are a subset of C. Note that c_{key} is the key value of context c'_s, which supports additional interest when the context reaches the corresponding value c_{key}.

Each contextual matrix reflects the level of interest through a different contextual combination. Assuming there are k' contextual matrices obtained from the composite contexts with a corresponding weight term $\{\omega_i \mid i = 1 \sim k'\}$, the mixed contextual model based on these k' contextual matrices is $M_{CAM} = \{(\omega_1, Mc_1), (\omega_2, Mc_2), \ldots, (\omega_{k'}, Mc_{k'})\}$. We adopt the first norm exponential kernel for integrating the contextual information provided by M_{CAM}, and we derive the contextual attention rank, AR_c from M_{CAM}: $AR_c(f_i) = p(M_{CAM} \mid C) = Mc_x/Mc_y \times Mc_z$, where

$$M_{c_x} = \prod_{c^x \in c'_i} \exp\{-\omega_x[(c^x)]\},$$

$$M_{c_y} = \prod_{c^y \in c'_p} \exp\{-\omega_y[(c^y)]\}, \quad \text{and}$$

$$M_{c_z} = \prod_{c^z \in c'_s} \exp\{-\omega_z|c_{\text{key}} - c^z|\}$$

We only need an approximate distribution of a viewer's excitement through contextual annotation with the context vectors. The reason we use the first norm exponential function in this article is because the exponential distribution not only is easy to compute and modify, but also is one of the most popular models in signal processing. Empirically, the exponential function also provides the best performance when dealing with the human attention model.

We form the contextual attention model on the basis of the context vector $C = [c^1, c^2, \ldots, c^v]$, which includes v classes. In terms of a similarity measure, the contextual information of different keyframes impacts viewer attention in different ways. As a result, we must define a similarity measure to account for different games and different context vectors. We use a cosine correlation measurement to compute the contextual similarity between f_l and f_k with the corresponding context vectors C_l and C_k, as follows:

$$d(M_{CAM}(f_l), M_{CAM}(f_k)) = \text{Cosine}(C_l, C_k)$$

Interactive attention ranking

There is no perfect solution to the problem of measuring viewer attention because the definition of an interesting keyframe strictly depends on each user. However, when a video shot contains a high number of attention-ranked keyframes, we can assume that it's worth noting. Each shot captured by a camera operator implies a certain intended meaning. Therefore, we assume that there should be at least one most-representative frame selected as the keyframe in each shot. Let $AR_{intra}(f_i)$ indicate the frame-level attention rank (*intra-AR*) of frame f_i. We can measure this rank quantitatively by using a function that is defined as the weighted sum of the visual attention rank, AR_v, and the contextual attention rank, AR_c, with a bias factor β. We derive AR_v by transforming visual attention models to a human attention rank.[15] Specifically, we can acquire the visual attention rank of frame f_k by the weighted sum of M different types of visual attention models with weights $\lambda_{m|m=1\sim M}$, which are initialized with equal probabilities. Finally, we normalize the rank score by the sum of the visual attention models of all frames f' belonging to the same shot S_i.

On the other hand, if we let $AR_c(f_k)$ indicate the contextual attention rank of frame k, which is directly affected by context vector C, then we can obtain contextual attention rank, which corresponds to M_{CAM} for frame k, by combining the classified contextual metrics: $AR_c(f_k) = Mc_x(c_i)/Mc_y(c_p) \times Mc_z(c_s)$. By using the attention rank discriminant function, we can determine the keyframes from each shot S_j of the video sequence by selecting the best P_k frames of the highest attention ranks, with the sum of the attention ranks of these P_k frames being larger than N percent of the total attention rank T_{AR}.

Let $KF^* = \{K_j \mid 1 \le j \le J\}$ be the keyframes that are the best keyframes chosen within the video sequence (with J shots), where $K_j = \{k_j \mid 1 \le j \le P_k\}$ is the keyframe collection in video shot S_j selected by the discriminant function

$$KF^* = \bigcup_{j=1}^{J} K_j = \min_{P_k}\left\{\bigcup_{i=1}^{P_k} f_i\right\}$$

which satisfies the condition

$$\sum_{i=1}^{P} AR_{\text{intra}}(f_i) > (T_{AR} \times N\%)$$

where

$$AR_{\text{intra}}(f_i) \ge AR_{\text{intra}}(f_j),$$
$$\forall j \in S_j, \quad j \ne i \quad \text{and} \quad 1 \le j \le J$$

We can rank the results of the keyframe extraction by the score of the attention rank discriminant function. However, the initial weights might not be good enough to provide the correct content for the user. Thus, user feedback is necessary. In addition to selecting the keyframes on the basis of the calculated attention rank discriminant function, we can further modify the attention rank value to guarantee that the results match user preference.

Weights modification

Suppose that the system receives the N_{RT} feedback keyframes with scores R_{score}, then $KF^*_{FB} = \{f^*_r, R^r_{score} | 1 \le r \le N_{rt}\}$. It's critical to predict what attention features users prefer, so we introduce two factors:

- If the values of the attention model M^m_v of the keyframes chosen by the users are quite different from each other, then M^m_v is not a good indicator.
- If all the attention rank values of the keyframes chosen by the user are persistently high, then the proposed attention model calculation is a good indicator about user interest.

We use all the attention models of the frames of KF^*_{FB} to compute the variance vector $\Phi = \{\sigma_1, \sigma_2, ..., \sigma_m, ..., \sigma_M\}$, and normalize these variances using the Gaussian normalization method to create $\{\sigma'_m\}$. Thus we can update the attention model weights under tth iteration:

$$\lambda^t_m = \lambda^{t-1}_m + \tau_m$$

where τ_m is a scaling factor defined as

$$\tau_m = \tau_{\max} \times \left(\frac{1}{\sigma'_m}\right) = \max_{R_{score}}\left\{\forall f_i \in KF^*_{FB}\right\} \times \left(\frac{1}{\sigma'_m}\right)$$

The max function denotes the confidence of the feedback score by the users for each selected keyframe, and R_{score}: {+3, +1, 0, −1, −3} indicates highly representative, representative, neutral, redundant, and 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 τ_{max} and σ'_m will not be zero. Finally, we normalize the attention model weights using the following equation:

$$\lambda^{t'}_m = \lambda^t_m / \sum_{m=1}^{M} \lambda^t_m$$

The reranking mechanism

We have attempted to extend the prediction of user interests to the event level. The ranking procedure, which is based on user feedback, is affected by the relevant keyframes located in the same event, the so-called inter-attention rank, AR_{inter}. The calculation of the *inter-AR* for frame f_i in event E_i summarizes the user interest in the game status in three ways:

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

With these observations in mind, we built the reranking model in *inter-AR* as follows:

$$AR_{inter}(f_i) = e * \sum_{f' \in (KF^*_{FB}, E_i)} \frac{R_{score}(f') \times AR(f')}{1 + d(M_{CAM}(f_i), M_{CAM}(f'_k))} + (1-e) * R_{SM}$$

where R_{SM} is similar to the Random Surfing Model in Google's PageRank algorithm.

We assume that $R_{SM} = 1/N_E$, where N_E is the frame number in the event E_i, and the keyframe with change in context is identified as the event boundary. The notation e is a damping factor that is empirically set to be $e = 0.85$, and KF^*_{FB} denotes the set of user-selected keyframes in E_i. These keyframes should have a higher attention rank and help promote the attention rank 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, we collect several keyframes. Thus, we can updated the new attention model by adding *inter-AR*:

$$AR^t(f_i) = AR^{t-1}(f_i) + AR_{\text{inter}}(f_i).$$

and $\forall f_i \in E_i$.

The reranking system gathers the relevant keyframes determined by the attention rank discriminant function. Then users can retrieve the video sequence that is based on the new ranking score (AR^t). Only a few interactions are required to find the most interesting video

clips and events. Efficient attention models can help the system speed up the retrieving procedures, not only the frame-level retrieval, but also the shot-level and event-level retrievals.

Simulation results

We collected four game videos and divided them into eight video sequences, for a total of about 47,960 frames in 690 shots. The video streams were AVI format with a digitization rate of 10 frames per second and a resolution of 352 × 240 pixels in 24-bit color. We chose complicated videos of baseball games captured by a DVD recorder from Major League Baseball broadcasts in 2006. Figure 4 shows the results of the single keyframe extraction for various event scenarios, and the corresponding calculated attention ranks. This figure illustrates five event types: ground out, walk, pickoff attempt, caught stealing with replay, and fly ball.

Figure 4. Keyframe extraction for different event scenarios: (a) ground out, (b) walk, (c) pickoff attempt, (d) caught stealing and replay, and (e) fly ball.

Table 1. Evaluation of keyframe ranking.

Video sequence number	Highly representative (%)	Representative (%)	Neutral (%)	Redundant (%)	Highly redundant (%)	Average R_{score} (%)
I	33.75	40.00	18.75	8.75	0.00	1.33
II	30.14	39.73	21.92	8.22	0.00	1.22
III	28.57	26.19	33.33	11.90	0.00	1.00
IV	18.06	22.22	55.56	4.17	0.00	0.72
V	22.12	30.77	34.62	12.50	0.00	0.85
VI	18.18	42.86	36.36	2.60	0.00	0.95
VII	30.48	43.81	18.10	5.71	1.90	1.24
VIII	19.48	25.97	40.26	14.29	0.00	0.70
Average	25.10	33.94	32.36	8.52	0.24	1.00

Table 2. Evaluation of keyframe reranking.

Video sequence number	Highly representative (%)	Representative (%)	Neutral (%)	Redundant (%)	Highly redundant (%)	Average R_{score} (%)
I	43.59	29.49	24.36	2.56	0.00	1.58
II	42.86	22.86	30.00	4.29	0.00	1.47
III	50.00	19.23	26.92	3.85	0.00	1.65
IV	26.87	35.82	29.85	7.46	0.00	1.09
V	29.41	33.33	34.31	2.94	0.00	1.19
VI	31.25	35.00	32.50	1.25	0.00	1.28
VII	43.75	31.25	21.25	5.00	0.00	1.26
VIII	39.19	28.38	31.08	1.35	0.00	1.45
Average	38.37	29.42	28.78	3.59	0.00	1.37

During the reranking experiments, we asked six people to review the testing video sequence in advance and assess each selected keyframe as highly representative, representative, neutral, redundant, and highly redundant, which we quantified as 3, 1, 0, −1, and −3. Table 1 shows the ranking results and the weighting of the attention models for all video sequences. Table 2 shows the reranking performance. The average ranking and reranking scores are listed in the right columns of Tables 1 and 2. As shown in Table 2, the average R_{score} for all of the testing sequences shows an improvement of about 13 percent, and a decrease in redundancy of 3.59 percent. These results indicate that our attention model and reranking strategy are consistent with human perception. Unlike the computer-centered approach where the user is unable to provide feedback, our interactive approach allows users to make decisions together with the system, enabling higher-precision, content-driven data mining.

In addition, we performed experiments to evaluate the sensitivity of the system. For the performance metric, we adopted the noninterpolated average attention rank, which corresponds to the mean value of all the frames in a shot. As illustrated in Figure 5, we compared hybrid attention ranking (visual and contextual) to the performance of individual attention ranking. And we collected an excitement score from the subjective reviewers as a benchmark. The example of the baseball videos shows that the context vector C consists of the current inning number, the on-base situation, the score of the home team, the score of the visiting team, the number of outs, the number of balls, and the number of strikes. In this experiment, we integrated the contextual attention with four contextual matrices that follow the composite relations from context vector C mentioned in the ''Contextual attention rank'' section.

We evaluated the system's performance with three correlation measurements for the

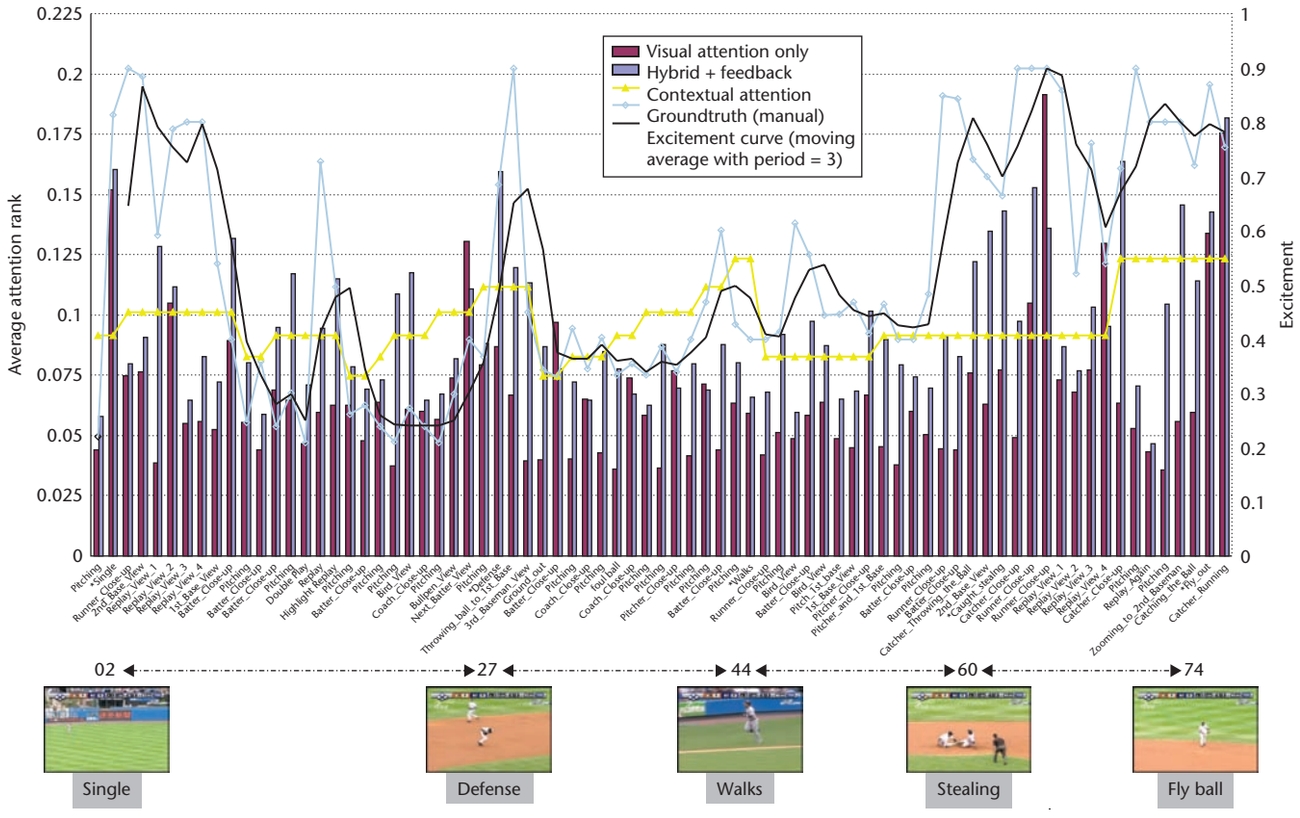

Figure 5. Performance of attention ranking with visual attention information only, and combining the visual and the contextual attention models for the 77 semantic shots, which are then compared to the excitement curve.

Table 3. Visual, contextual, and hybrid ranking in comparison to user excitement.

Methods	Visual attention	Contextual attention	Hybrid and feedback
L_2-norm distance	0.465	0.406	0.384
Simple correlation	0.739	0.792	0.825
Cosine correlation	0.773	0.787	0.847

distribution similarities between the resulting attention rank and the excitement curve. We computed the square difference values of the gradient of normalized attention ranks in advance, then for each shot we applied L_2-norm distance, the simple correlation, and the cosine correlation. Table 3 shows that the hybrid attention ranking more effectively reflects the human level of interest for all videos, and that it matches well with the attention score of the excitement curve. Normally, hybrid attention ranking works well, especially when the part of the visual information is not sufficiently clear. Overall, we found that the resulting attention rank excitement score is reasonable for the actual occurrence of an exciting event, because the higher excitement score is consistent with the appearing highlights. However, unpredictable issues, such as personal viewer preferences or comments from the sportscaster, may affect the accuracy of attention ranking throughout the entire game.

Conclusions

In this article, we proposed a self-adaptive, content-driven mining mechanism that incorporates visual and contextual attention models. We believe a key strength of this system is its

reliance on systematic, object-based attention modeling to avoid the problems associated with unpredictable noise. In our tests, we found that the proposed system effectively determined video content that might attract viewer attention. We believe this proposed system has great potential for application to different sports domains, particularly videos that use a superimposed caption box. **MM**

Acknowledgment

We thank the anonymous reviewers for their valuable suggestions and comments, which were crucial in improving this article.

References

1. J.K. Tsotsos et al., ''Modeling Visual-Attention via Selective Tuning,'' *Artificial Intelligence,* vol. 78, nos. 1-2, 1995, pp. 507-545.
2. 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, 2008.
3. L. Page et al., ''The Pagerank Citation Ranking: Bringing Order to the Web,'' *Stanford Digital Library Technologies Working Paper, 1999–0120,* Stanford Univ., 1998.
4. S. Brin and L. Page, ''The Anatomy of a Large-Scale Hypertextual Web Search Engine,'' *Proc. 7th Int'l World Wide Web Conf.,* Elsevier, 1998, pp. 107-117.
5. M.R. Henzinger, ''Hyperlink Analysis for the Web,'' *IEEE Internet Computing,* vol. 5, no. 1, 2001, pp. 45-50.
6. C. Kim and J.-N. Hwang, ''Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications,'' *IEEE Trans. Circuits and Systems for Video Tech.,* vol. 12, no. 2, 2002, pp. 122-129.
7. C. Kim and J.-N. Hwang, ''Video Object Extraction for Object-Oriented Applications,'' *J. VLSI Signal Processing—Systems for Signal, Image, and Video Technology,* vol. 29, nos. 1-2, 2001, pp. 7-22.
8. T.N. Cornsweet, *Visual Perception,* Academic Press, 1970.
9. B.B. Huang and S.X. Tang, ''A Contrast-Sensitive Visible Watermarking Scheme,'' *IEEE MultiMedia,* vol. 13, no. 2, 2006, pp. 60-66.
10. L. Itti, C. Koch, and E. Niebur, ''A Model of Saliency-Based Visual Attention for Rapid Scene Analysis,'' *IEEE Trans. Pattern Analysis Machine Intelligence,* vol. 20, no. 11, 1998, pp. 1254-1259.
11. M.-H. Yang, D. J. Krieqman, and N Ahuja, ''Detecting Faces in Images: A Survey,'' *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 1, 2002, pp. 34-58.
12. D. Arijon, *Grammar of the Film Language.* Silman-James, 1991.
13. G. Davenport, T.A. Smith, and N. Princever, ''Cinematic Primitives for Multimedia,'' *IEEE Computer Graphics and Applications,* vol. 11, no. 4, 2002, pp. 121-133.
14. 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, 2005, pp. 449-459.
15. H.C. Shih, J.N. Hwang, and C.L. Huang ''Content-Based Video Attention Ranking Using Visual and Contextual Attention Model for Baseball Videos,'' *IEEE Trans. Multimedia,* vol. 11, no. 2, 2009, pp. 244-255.
16. W.-N. Lie and S.-H. Shia, ''Combining Caption and Visual Features for Semantic Event Classification of Baseball Video,'' *Proc. IEEE Int'l Conf. on Multimedia and Expo* (ICME), IEEE CS Press, 2005, pp. 1254-1257.
17. N. Babaguchi et al., ''Personalized Abstraction of Broadcasted American Football Video by Highlight Selection,'' *IEEE Trans. Multimedia,* vol. 6, no. 4, 2004, pp. 575-586.
18. A. Rosenfeld, R. Hummel, and S. Zucker, ''Scene Labeling by Relaxation Operations,'' *IEEE Trans. Systems, Man, and Cybernetics,* vol. 6, no. 6, 1976, pp. 420-433.

Huang-Chia Shih is a postdoctoral research fellow in electrical engineering at the National Tsing Hua University, Taiwan. His research interests include content-based video analysis, multimedia data mining, semantic computing, and model-based human-motion capturing and recognition. Shih has an PhD in electrical engineering from National Tsing Hua University. Contact him at hc.shih@gmail.com.

Chung-Li Huang is a professor in electrical engineering at the National Tsing Hua University, Taiwan. His research interests include image processing, computer vision, and visual communication. Huang has a PhD in electrical engineering from the University of Florida, Gainesville. Contact him at clhuang@ee.nthu.edu.tw.

Jenq-Neng Hwang is a professor in the Electrical Engineering Department at the University of Washington. His research interests include image and video signal processing, computational neural networks, multimedia system integration, and networking. Hwang has a PhD in electrical engineering from the University of Southern California. Contact him at hwang@u.washington.edu.