Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2010, Article ID 836357, 20 pages
doi:10.1155/2010/836357

Research Article

Personalized Sports Video Customization Using Content and Context Analysis

Chao Liang,[1,2] Changsheng Xu,[1,2] and Hanqing Lu[1,2]

[1] *National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[2] *China-Singapore Institute of Digital Media, 119615, Singapore*

Correspondence should be addressed to Chao Liang, liangchao827@gmail.com

Received 2 September 2009; Revised 11 December 2009; Accepted 26 January 2010

Academic Editor: Jungong Han

We present an integrated framework on personalized sports video customization, which addresses three research issues: semantic video annotation, personalized video retrieval and summarization, and system adaptation. Sports video annotation serves as the foundation of the video customization system. To acquire detailed description of video content, external web text is adopted to align with the related sports video according to their semantic correspondence. Based on the derived semantic annotation, a user-participant multiconstraint 0/1 Knapsack model is designed to model the personalized video customization, which can unify both video retrieval and summarization with different fusion parameters. As a measure to make the system adaptive to the particular user, a social network based system adaptation algorithm is proposed to learn latent user preference implicitly. Both quantitative and qualitative experiments conducted on twelve broadcast basketball and football videos validate the effectiveness of the proposed method.

1. Introduction

The proliferation of advanced program production technology and multiple TV broadcast channels have contributed to an amazing growth of sports video content and its increasing popularity among the public. However, such increasing availability has not yet been accompanied by an improvement in its accessibility, which means that audiences can do nothing but passively watch the whole match edited by studio professionals once they choose it. Since interesting segments usually account for a small portion of the whole match; such passive watching mode not only impairs audiences' viewing experience but also wastes their time and money. To solve this problem, the ability to provide personalized video content in accordance to the features of individual viewers is of great importance.

Intuitively, viewers difference first comes through in their diverse preference towards semantic content where particular players and events are appearing in the video. For example, a Beckham's fan may mainly focus his attention on this football star than any other players, while an NBA's audience may prefer to watch the slam dunk than any other events. To meet these requirements, the source video has to be analyzed in a more refined scale and higher semantic level. More precisely, the video analysis should not merely tag some salient events with simple concepts, for example, shots in basketball or fouls in football, but annotate various events with detailed semantic description, including the involved player(s), event type(s), and result consequence.

Besides video content personalization, viewers individuality also reflects in their diverse customization modes and environments. Here, video customization mode regulates the concrete selection criteria of video segments. For viewers focusing on the particular player or events, segments that are semantically consistent to viewers' interest are better, while for the viewer interested in a game's global situation, segments that can best capture the main body of the match are more preferable. As for the customization environment, it denotes multiple constraints during the practical usage, for example, memory capacity, electrical quantity and transmission bandwidth, and so forth. All these physical conditions differ from person to person, and directly affect the final customization result.

Video personalization not only lies in customizing pointed content, but also embodies in system adaptation to the particular viewer. It is an important function for an intelligent system that can be made easier to use as the user continues to use it. User preference learning is an effective measure to tackle such problem. By analyzing the explicit or implicit feedbacks given by the user, an intelligent system can automatically infer the latent user preference and adaptively adjust its structure or/and parameters for future more convenient usage.

In this paper, we aim to propose a sports video personalization framework, through which users can enjoy refined video segments containing their favorite semantics from the lengthy sports match at anytime in anyplace. Both subjective content preference and objective environment constrains will be well balanced so that the optimal visual experience can be brought to the particular viewer. Particularly, we adopt basketball and football games as our initial sports genres because they are not only widely adopted study bed but also globally popular sports, which possess great values in both research and application. Moreover, since the proposed framework is generic, we believe that our approach can be easly extended to other sports domains.

The rest of the paper is organized as follows. Related work on sports video analysis, retrieval, summarization, and user preference learning are reviewed in Section 2. The problem formulation and proposed framework are described in Section 3. The technical details of sports video annotation, personalized video customization, and system adaptation are presented in Sections 4, 5, and 6, respectively. Experimental results are reported in Section 7. We conclude the paper with future work in Section 8.

2. Related Work

Extensive research efforts have been devoted to sports video analysis and application due to their wide viewership and enormous commercial potential. In this section, we give a brief review of related work on sports video annotation, retrieval, summarization, and user preference learning.

2.1. Sports Video Analysis and Annotation. Sports video analysis and annotation aim at the detection and recognition of semantic content. Most of the previous work is based on the audio [1–3], visual [4, 5], and textual [6, 7] features directly extracted from video content itself. The basic idea of such methods is to utilize heuristic rules [8] or machine learning algorithms [9, 10] to infer semantic events from various low-level [11] or mid-level [12] features. Since sports video is an integration of various information modalities, algorithms based on multimodal fusion are more likely to achieve robust and accurate event detection result than single-modality methods. For example, audiovisual features were successfully used to detect events in basketball [13], soccer [14], cricket [15], and tennis [16]; audiovisual-textual features were also collaborated in analyzing baseball [17], cricket [18], and golf [19] matches.

Nevertheless, due to the semantic gap between low-level features and high-level semantics, the content-based methods, no matter using single- or multimodality, can only annotate certain salient events with simple concepts, which cannot meet viewers' personalized appetites for specific players and events. In order to obtain more abundant high-level semantics, external textual information is introduced to facilitate video annotation and has achieved encouraging results. Babaguchi et al. [20] proposed a multimodal strategy using closed caption for event detection and video indexing. Xu and Chua [21] raised an integrative approach to align text events with match phase information to detect multiple events in soccer video. Xu et al. [22] used web broadcasting text from sports websites to detect event semantics and achieved inspiring results.

2.2. Sports Video Retrieval and Summarization. As for the direct applications of sports video analysis, retrieval and summarization represent two typical customization modes. Video retrieval can be regarded as a point query, where users focus on the particular person or event, while video summarization can be considered as a plane query, where users are more concerned about the entire situation of the match.

For methods using only low-level features, slow-motion replay portions and various highlight segments are competent candidates for video summarization. In [8], Ekin et al. summarized soccer video by classifying shot types and detecting slow-motion replay. In [23, 24], the whole sports video was divided into a sequence of play/break segments and highlights were assigned to the play segments for complete sports video summarization. For methods using textual features, video customization can effectively incorporate user preference and operate on the semantic level. Fleischman and Roy [25] proposed an unsupervised sports video retrieval approach by pairing repeated temporal visual patterns with associated closed caption text. Babaguchi et al. [26] proposed a personalized video retrieval and summarization framework based on the rich semantic description obtained from closed caption recognition.

Considering various environment limitations in the practical usage, such as network traffic and device memory, resource-constraint video customization also received wide attention from both academy [27, 28] and industry [29].

2.3. System Adaptation. Adaptation is an important mechanism of enabling an information system to provide personalized service to the particular user. Through learning user preference, a personalization system can adaptively adjust its structure and/or parameters to provide more pointed service.

In previous work, relevance feedback is a representative approach of explicit user preference learning. The main idea is using human-computer interaction to directly guide system adaptation so that it can provide focused service to the particular user. Zhang et al. [30] designed a relevance feedback strategy to retrieve suitable sports video clips to meet user request. By computing both the semantic and visual consistency of selected video segments, users'

personalized preference can be properly quantified and satisfied. Amir et al. [31] proposed a mutual relevance feedback method for multimodal query formulation in video retrieval. Based on the relevant shots marked by the user, the system can automatically identify useful search terms to refine the retrieval result.

Due to the time-consuming interactions in explicit feedback, implicit feedback is utilized to make the system adaptive to the user with less interruption. User profile analysis is a typical method of implicit user preference learning. Syeda-Mahmood and Ponceleon [32] adopted a hidden Markov model to predict users' internal states from their history browsing behaviors, and then generated a specified video preview for the particular viewer. Zimmerman et al. [33] used two kinds of implicit recommenders, the Bayesian classifier and the C4.5 decision tree, to learn users' preference from their viewing histories and fused multiple recommendations with a neural network to generate user favorite TV shows.

3. Problem Formulation and Framework

Video annotation serves as the foundation of personalized video customization. It takes the responsibility of connecting low-level audiovisual segments with high-level semantics. Compared with content-based annotation, approaches utilizing external textual information can provide more detailed semantic description of video content. Currently, two types of external textual sources, closed caption and web-casting text, are used for semantic video analysis. For the closed caption, although it holds the inherent advantage of video-text synchronization, it faces the challenge of accurate information extraction from irregular and variable spoken language, while for the web-casting text, the reverse applies, which means it has well-defined syntax structure but lacks of direct video-text correspondence. Most related work [22, 30] adopted the timestamp as a key link to connect these two media. However, these methods are usually confined to the lower timestamp recognition accuracy due to the noisy broadcast video. Moreover, the availability and concrete styles of timestamp are always decided by the program producer, which further limit the expansibility of such timestamp-based approaches.

With the detailed semantic annotation, two research challenges need to be addressed for personalized video customization: first, the incorporation of high-level semantics in video content selection, second, the balance between user preference and environment constraints. If we consider the customization problem from the optimization point of view, the first challenge defines an objective function to evaluate the semantic importance of video segments, while the second challenge models a constraint optimization problem on the basis of the above objective function by adding environment constrains. Most previous work focused on only one aspect of the above two problems, either semantic content selection [25, 26] or resource-constrained application [28, 29]; all of which lacked an integrated consideration of the above two challenges.

To practice the concept of "human-centered multimedia" [34], it is the responsibility of the information system rather than the user to adapt itself to provide more convenient service. On the side of the user, an ideal personalization system should learn his/her preference as accurately and unconsciously as possible. However, these two requirements are usually incompatible in the practical implementation. Specifically, explicit feedback [30, 31] gives accurate user preference but requires additional interaction, while implicit inference [32, 33] needs less operations but may result in incomplete learning.

In this paper, we will address the challenges existing in the previous approaches for semantic annotation, personalized customization, and system adaptation. Solutions toward the above problems in these three fields constitute an integrated system to provide personalized sports video customization. Compared with previous work in the related fields, the main contributions of our approach are summarized as follows.

(1) We propose an integrated framework to provide personalized sports video customization. With a comprehensive strategy on semantic annotation, personalized customization, and preference learning, users can conveniently customize their interested video segments concerning specific players or events.

(2) We propose a novel sports video annotation approach, where video content and web-casting text are aligned by their semantic correspondence along the temporal sequences. Since semantics is an intrinsic existence in multimedia, it is more generic and robust for cross-media analysis.

(3) We propose a user-participant multiconstraint 0/1 Knapsack model for personalized video customization, where user content preference and environment limitations can be well balanced and satisfied.

(4) We propose a social network based system adaptation algorithm to propagate local user interaction information along the video semantics network, from which complete user preference can be implicitly inferred and learned.

Figure 1 illustrates the framework of our proposed approach. First, a hierarchical semantic-matching method is employed to generate detailed video annotation in an off-line manner. Then, a user-participant content customization algorithm is proposed to provide real-time video content customization under resource-constraint environment. Meanwhile, to facilitate the above customization process, a concept network is built to capture the latent user preference for effective system adaptation.

4. Sports Video Annotation

Sports video annotation is responsible for the generation of detailed semantic description of video content. Recent work [22, 35] mainly focused on the timestamp-based video-text alignment, where timestamp's availability and recognition are two main bottlenecks restricting the application of

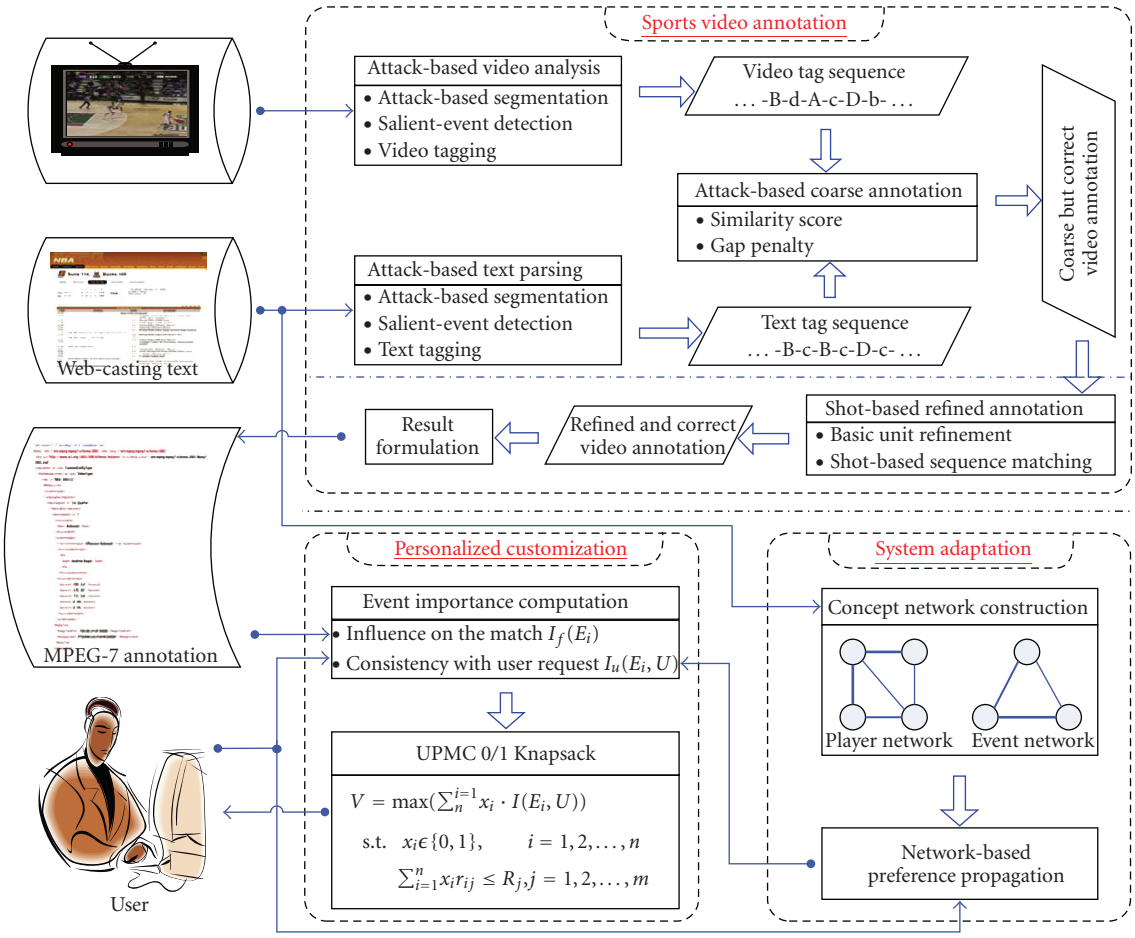


Figure 1: System framework of personalized sports video customization.

such method. To conquer these drawbacks, we propose a semantics-matching approach [36] to align the sports video and web-casting text based on their semantic correspondence. Since semantics is ubiquitous in various media and not susceptible to the postediting and transmitting, the proposed method is believed to be more generic to the cross-media analysis.

4.1. Attack-Based Coarse Alignment. To perform semantic matching between video and text, the first thing is to find a common abstract unit that can be reliably extracted from both media. Once such a unit is identified, the semantic bridge connecting the paired media can be effectively built. According to the sports video features, attack-based analysis method [37] provides a suited choice. Since the notion of attack exists in most opponent sports and has very clear semantics, a change in the attack side is a natural segmentation criterion for both game video and text.

4.1.1. Video Tagging. Video tagging module aims to generate a semantic tag sequence where each tag corresponds to an attack segment in the match. Here, attack is defined as a complete attempt of a team (player) in an opponent sport to score a point. It refers to a macroscopic process rather than a specific event; hence it includes not only obvious offensive events like shot or goal in an attack attempt, but also contains other nonoffensive events like foul or return pass during that process.

According to the above definition, consistent moving segment from one side of the court to the other usually corresponds to a complete attack unit. However, due to the fierce competition in the sports match, segmentation result is likely to be affected by a mass of blurring motion. To overcome this difficulty, we first smooth the horizontal camera motion [38] by the field zone information [21] so that blurring motion clips without obvious position change can be filtered out. Then, attack-based video segmentation is obtained with a sequence of boundaries at the start point of each remaining motion segment. A realistic example of the above process is illustrated in Figure 2, where blue solid line represents field zone information (1.5 denotes left field, −1.5 right field and 0 mid field) and green dash-dot line represents horizontal camera motion (1 denotes leftward moving, −1 rightward moving).

After attack-based video segmentation, a group of mid-level binary features including shot type transition (only long and nonlong shots are considered in our method) (*ST*) [39],

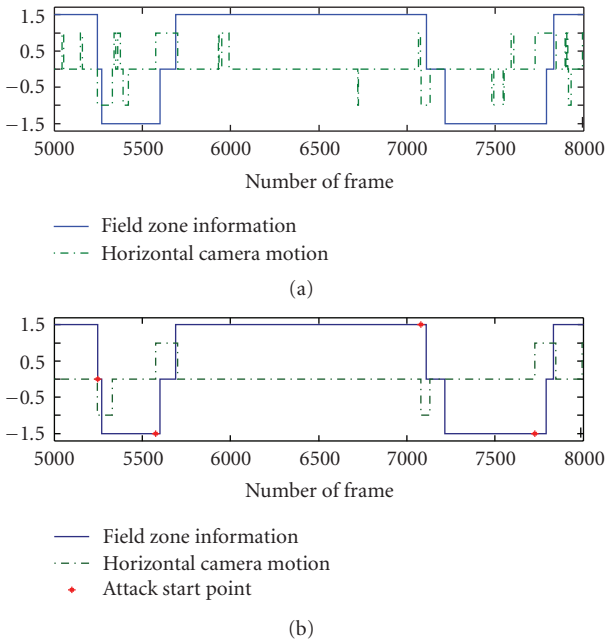


Figure 2: Attack-based sports video segmentation. (a) Initial field zone and horizontal camera motion information. (b) Smoothed horizontal camera motion and three attack segments (#1: 5210–5613; #2: 5613–7120; #3: 7120–7748).

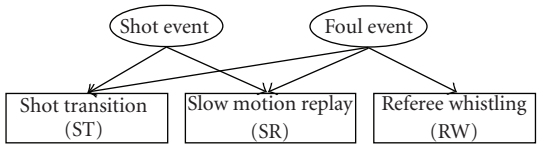


Figure 3: Naive Bayesian network for shot and foul events detection.

slow-motion replay (SR) [35], and referee whistling (RW) [13] are extracted from each attack segment and a heuristic Bayesian network (see Figure 3) is trained to detect shot and foul events as follows:

$$S^* = \underset{S}{\operatorname{argmax}} P(S \mid ST, SR, RW),$$
$$F^* = \underset{F}{\operatorname{argmax}} P(F \mid ST, SR, RW), \quad (1)$$

where S and F are binary variables corresponding to the existing states of shot and foul events in an attack segment and S^* and F^* are their inferred states. Finally, the detected attack direction and semantic events are further encoded by the combination of their binary status (see Table 1), and hence the whole video tagging process can be represented in a concise form as follows:

$$X^* = X(D, S^*, F^*) = \underset{X(D,S,F)}{\operatorname{argmax}} P(S, F \mid ST, SR, RW), \quad (2)$$

where X^* represents the final video tag encoded from the combination of attack direction (D) and the inferred semantic events (S^* and F^*).

4.1.2. Text Tagging. The utilization of textual information significantly facilitates video content analysis. Compared with caption text overlaid on the image [6] or encoded in the video [20], web-casting text [22] has following obvious advantages. (1) It is available in many famous sports websites such as ESPN (http://sports.espn.go.com/) and BBC (http://news.bbc.co.uk/sport2/hi/football/teams/) and can be timely accessed during or after the game. (2) It is organized in a well-defined structure that can be easily parsed. (3) It contains rich match description that are difficult to be obtained solely from content analysis (see Figure 4). Therefore, we utilize web-casting text in our method to facilitate semantic video annotation.

Similar to the content-based video tagging process, web-casting text analysis also aims to generate a semantic tag sequence where each tag corresponds to an complete attack attempt in the game text. Since web-casting text is tagged by sports professionals and provided by famous websites, its use of words and syntax structure are standard and fixed. Therefore, semantic events can be reliably detected by keyword-based searching. Table 2 lists the selected events and their related keywords used in our method; all of which are typically interesting events in the match and can be flexibly expanded according to users' preference.

With text event detection, attack-based text segmentation can be formulated as clustering adjacent text events into individual groups so that each group corresponds to a complete attack attempt in the match. By convention of composing the web-casting text, adjacent records with consistent attack directions always belong to the same attack process. Hence, we can cluster event records by analyzing the current attack side (CAS) in each text event. In opponent sports, the CAS is an important binary feature indicating which team is in the offense state when current text event is happening. Both one side's offense event and the other side's defense event correspond to the same CAS value. It can be reliably extracted from a text record by analyzing the player membership and event attack attribute (listed in Table 2). For example, given a text event saying "Michael Redd makes layup", we can know that it represents an offense event from the keyword "layup" and the CAS is the Bucks, the team that Michael Redd belongs to. Once all CAS features are extracted from text events, segment boundaries can be easily identified by sequentially comparing the adjacent two CAS features. To better understand the above process, the detailed algorithm flow and an example for CAS detection are given in Figure 5.

When using the above text segmentation method, one implementation detail needs to be addressed. As can be seen from Table 2, except rebound event, all other events having double attack attributes are related to the foul event. Since these events can happen in both offense and defense sides of an attack, their attack attributes are always difficult to be identified from a single text record. To solve this problem, we utilize their previous event's CAS feature and current event's related team to codetermine current event's attack attribute. Specifically, if current event is performed by a player belonging to the offense team in the previous event, its attack attribute is the offense, otherwise the defense. Such a rule is based on the assumption that the attack

Table 1: Code book used for video tagging.

Semantics	Attack direction		Semantic events	
Tag	Rightward	Leftward	Shot event	Foul event
"a"	true	false	false	false
"b"	true	false	true	false
"c"	true	false	false	true
"d"	true	false	true	true
"A"	false	true	false	false
"B"	false	true	true	false
"C"	false	true	false	true
"D"	false	true	true	true

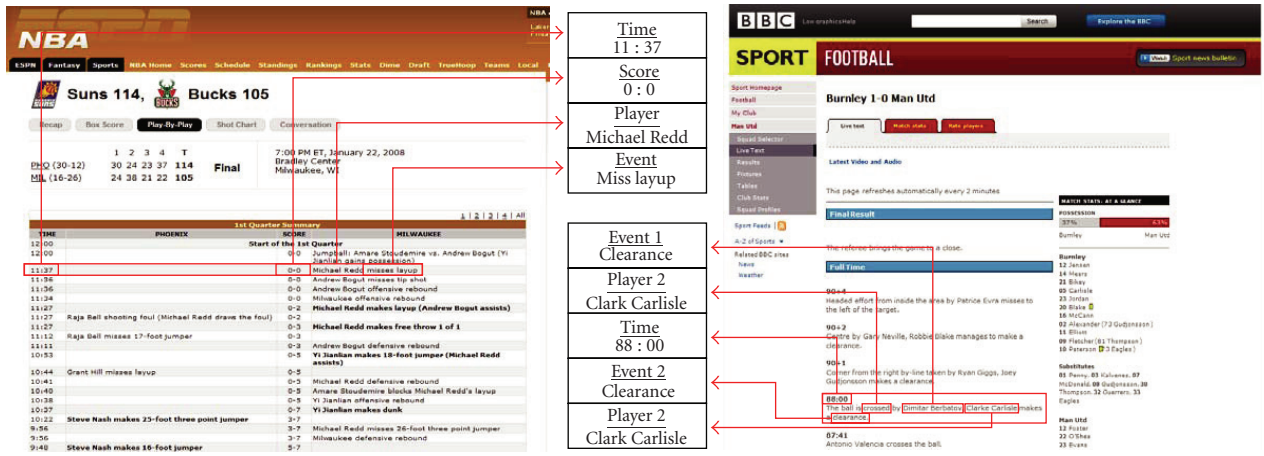


Figure 4: Web-casting text for (a) basketball and (b) football matches.

side is less likely to change in two adjacent events when the latter is a foul event (If not so, there must be an event in between causing the attack side change, which is contradictory with the assumption that two events are adjacent.). Our experiments also prove the rationality of such assumption.

With detected attack side in the match, the related attack direction can be easily inferred by comparing the attack segment ratios between two teams in the text and two directions in the video. Then, similar encoding process can be conducted to generate text semantic tag sequence.

Although video and text tagging share similar processing steps and output, we must stress that these two sequences are intrinsically different. For the former, each tag is a random variable and the finally derived video sequence is composed of the most likely semantic tags given an observation of mid-level features extracted from video attack segments. In contrast, the text sequence is a constant sequence with each tag being identified in a determinate way. Such a difference provides convenience to the tag similarity measurement in the following subsection.

4.1.3. Attack-Based Sequence Matching. The output of video analysis is a tag sequence with accurate attack boundaries (in terms of video shot) but inaccurate semantic tags (due to the semantic gap), while the output of text parsing is another tag sequence with accurate semantic tag (from keywords matching) but without video boundary information. Therefore, an intuitive way to annotate sports video on the level of attack is to align the accurate text tag sequence with its related video counterpart. Considering the semantic tag sequence, we propose a semantic-based Needleman-Wunsch algorithm [36] to match the probability video sequence with the constant text sequences. Compared with other algorithms, the proposed algorithm searches for the optimal alignment based on the global semantic correspondence, hence is more robust to local errors in the inaccurate video tag sequence.

The standard Needleman-Wunsch algorithm [40] is intrinsically a dynamic programming algorithm that initially aims to find the best protein or nucleotide sequence matching in bioinformatics. It is implemented in a multistage decision process, where the forward computation calculates the optimal local matching scores under various matching modes and the backward tracking identifies the global optimal sequence alignment. Consider a matching problem between two sequences, $\mathbf{v}$ and $\mathbf{t}$, which has m and n elements, respectively. In the forward computation, a score matrix $\mathtt{M}$ with m rows and n columns is first allocated, where each row corresponds to an element in $\mathbf{v}$ and each column to

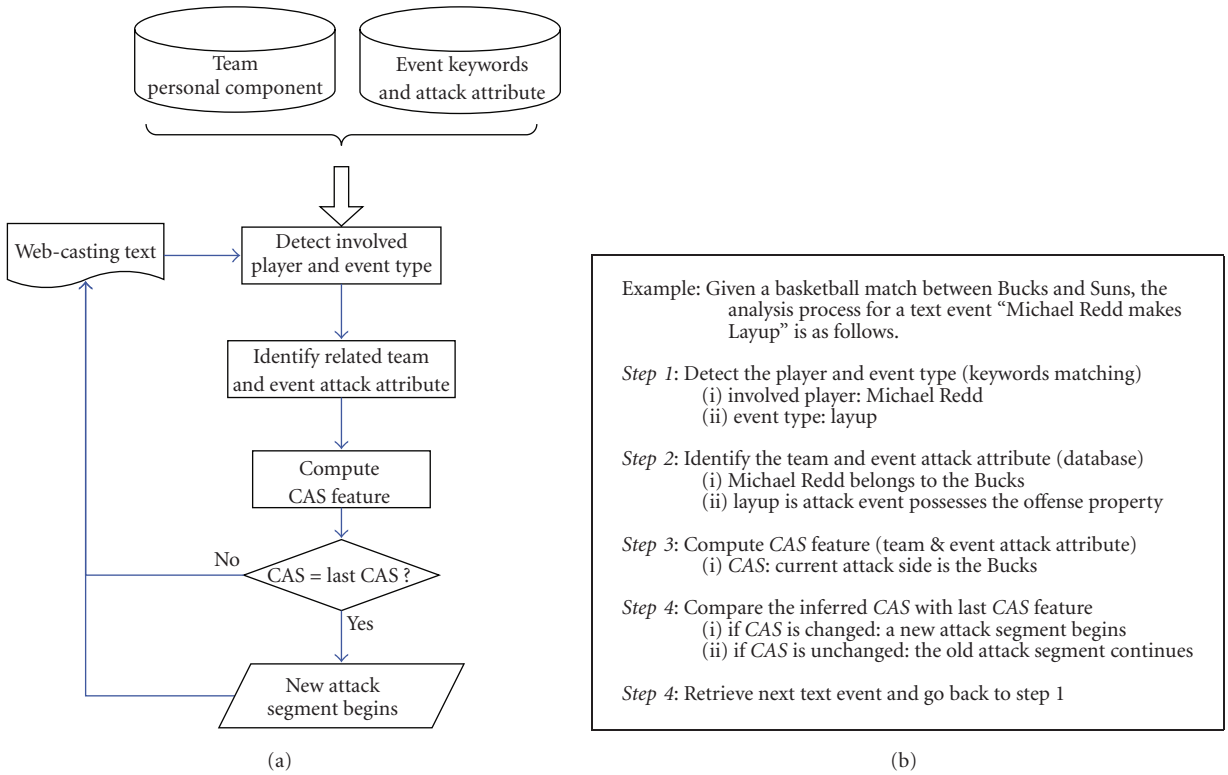


Figure 5: Attack-based text segmentation (a) algorithm flow and (b) an example.

Table 2: Event keywords and attack attributes.

Domains	Events	Related keywords	Attack attributes
Basketball	shot	makes layup, hook, jumper, free throw	offense
	miss	miss	offense
	block	blocks	defense
	stolen	steal, lose ball, bad pass	defense
	foul	foul	offense/defense
	rebound	offensive rebound, defensive rebound	offense/defense
Football	shot	shot, goal, head, scored	offense
	free kick	free kick	offense
	corner	corner	offense
	foul	foul	offense/defense
	card	red card, yellow card	offense/defense
	offside	offside	offside

an element in $\mathbf{t}$. The (i, j)th entry of the matrix $\mathtt{M}$ records a local optimal alignment score between subsequence $\mathbf{v}_1 \sim \mathbf{v}_i$ and $\mathbf{t}_1 \sim \mathbf{t}_j$. The forward computation is implemented in an iterative manner, where the value of $\mathtt{M}_{i,j}$ is decided by one of its three adjacent predecessors as follows:

$$\mathtt{M}_{i,j} = \max\left\{\mathtt{M}_{i,j-1} + \mathrm{Pg},\ \mathtt{M}_{i-1,j} + \mathrm{Pg},\ s_{i,j} + \mathtt{M}_{i-1,j-1}\right\}, \quad (3)$$

where $s_{i,j}$ is the similarity score between the ith video tag ($\mathbf{v}_i$) and the jth text tag ($\mathbf{t}_j$), Pg denotes the gap penalty given to an empty matching. In (3), the first two items correspond to the rightward (gap-element) and downward (element-gap) matching, where element in one sequence cannot find its counterpart in the other sequence; thus a gap matching generated. The third item in (3) denotes a diagonal (element-element) matching where the similarity between element $\mathbf{v}_i$ and $\mathbf{t}_j$ is computed. By comparing the final alignment scores generated from three directions, the highest score is assigned to $\mathtt{M}_{i,j}$ and the related direction is stored in the corresponding position of an equal-sized $(m \times n)$ backtracking matrix $\mathtt{B}$. As the whole score matrix is allocated, the related backtracking matrix is also filled. Starting from the bottom right of the backtracking matrix, the global optimal alignment path

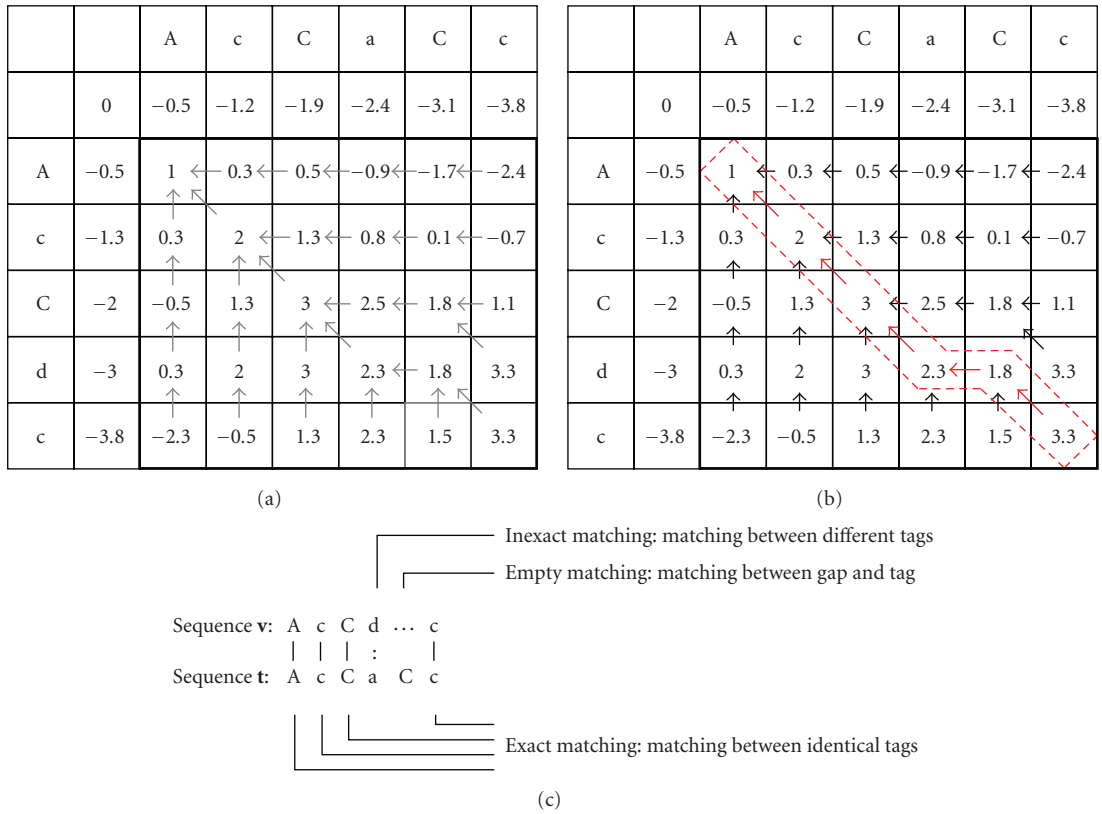


Figure 6: (a) Forward filling of score matrix and backtracking matrix, (b) backward tracing along best local alignment directories, and (c) sequence alignment result.

can be finally obtained through tracing back along the stored local matching directory. A graphical demonstration of above algorithm flow is illustrated in Figure 6, where numbers and direction in the thick dark black box in Figure 6(a) constitute the score and backtracking matrixes, respectively, and red arrows in Figure 6(b) indicate the final backtracking directory corresponding to the optimal sequences alignment. Three different matching types are obtained in Figure 6(c), where the mark "|" denotes the exact matching between identical elements and mark ":" denotes the inexact matching between different elements. As for the empty matching, no mark is presented.

From above discussion, the forward computation of score matrix serves as the core part of the Needleman-Wunsch algorithm. It affects the final alignment result by deciding the local alignment directions. In the field of bioinformatics, the sequence elements usually correspond to certain chemical substances, and thus the matching scores are given based on their chemical structures and properties. However, in the particular application of sports video and game text alignment, since each sequence element is a character tag representing the combination of high-level features and events, the alignment scores should be marked according to their semantic similarities.

Because the video tag is intrinsically a discrete random variable with its value ($\mathbf{v}_i$) corresponding to the largest probability in the distribution generated by Bayesian network, the semantic similarity between video tag $\mathbf{v}_i$ and text tag $\mathbf{t}_j$ can be indirectly measured by the conditional probability difference when the video tag variable taking the values of $\mathbf{v}_i$ and $\mathbf{t}_j$, given an observation of the mid-level features. In our approach, we formulate the similarity measure between video and text tags as follows:

$$s_{i,j} = \begin{cases} -\infty & \text{inconsistent directions,} \\ \dfrac{P\left(X = \mathbf{t}_j \mid ST, SR, RW\right)}{P(X = \mathbf{v}_i \mid ST, SR, RW)} & \\ = \dfrac{P\left(X = \mathbf{t}_j \mid ST, SR, RW\right)}{\max P(X \mid ST, SR, RW)} & \text{consistent directions} \end{cases} \tag{4}$$

where X is the video tag variable and ST, SR, RW represent audiovisual features described in Section 4.1.1. Since the attack direction can be reliably identified from both video and text, tags corresponding to inconsistent attack directions are not allowed to be aligned in the approach. For the direction consistent condition, the more proximal the inference probabilities between tags $\mathbf{v}_i$ and $\mathbf{t}_j$ are, the more likely $\mathbf{t}_j$ can be used to replace $\mathbf{v}_i$ to annotate the related video segment, in other words, the more likely video segment tagged as $\mathbf{v}_i$ can be aligned with text group tagged as $\mathbf{t}_j$. Moreover, since $\mathbf{v}_i$ corresponds to the maximum condition probability in

the video tagging process, the tags similarity $s_{i,j}$ can never be larger than 1, which can only be reached in the exact matching.

As for the gap penalty, it is defined as a function of the semantic events contained in the attack segment:

$$\mathrm{Pg} = [-0.5 - 0.25 \times (S^* + F^*)] \times \theta, \tag{5}$$

where S^* and F^* represent the most probable existing state of shot and foul events generated by (1) and θ is the affine gap cost defined as 1 for the first gap and 0.95 for others in our following experiments. According to above equation, the more events contained in an attack, the less likely it cannot find a corresponding text tag, hence the severer punishment will be given to its empty alignment, and vise versa. Table 3 describes the proposed semantics-based Needleman-Wunsch algorithm for the tag sequence alignment between broadcast sports video and web-casting text.

If we consider a timestamp as a tag, previous timestamp-based methods can be regarded as a special case of our approach. However, two important differences exist. First, tags in our method represent high-level semantics rather than low-level visual features, which is an intrinsic and generic link across multimedia. Second, global structure rather than individual equivalence is utilized to align video and text sequences, which greatly improve the method's robustness against local errors. Therefore, the proposed semantics-matching approach is considered to be more effective for the generic cross-media analysis.

4.2. Shot-Based Refined Alignment. The output of the attack-based alignment is a coarse but accurate annotation result, where the basic unit corresponds to an attack segment rather than a specific event. To obtain a more elaborate event detection result, which locating the semantic event in a specific video shot, we repeat previous sequence matching algorithm on the scale of each attack to generate shot-level video annotation. With the variance of matching granularity, the basic sequence units change from the attack segment into a shot in the video and an event record in the text. Since long shots are always adopted to depict the global situation when the match is in play, they are the only shot type used in refined alignment.

Although the shot-based refined alignment lacks direct semantic correspondence between sports video and web text, it indeed generates more elaborate annotation on the level of shot. Moreover, our experiments also show that such semantic weakness during the refinement process based on the coarse alignment result is not significant and hence totally acceptable.

Final video annotation result is stored in the standard MPEG-7 XML format for efficient retrieval and management. As shown in Figure 7, the XML file is organized as a hierarchical structure in accordance with its related sports match, where the whole match includes several quarters and each quarter contains a group of events. For each event in the match, the XML file records the detailed information including its involved player(s), event type and moment, current score and the corresponding video segment.

5. Personalized Video Customization

Personalized video customization aims to tailor proper video content to the particular user. Different from generic highlight presentation work [23, 24] where interesting video clips are fully decided by video content analysis, the proposed customization scheme is featured in the cooperation of video semantics and user preference in video content selection process. Moreover, considering the conflict between the mass video content and limited user environment, a balanced customization strategy is also addressed to maximize audience's viewing enjoyment under various context constraints.

5.1. Event Importance Computation. To evaluate the quality of selected video segments, event importance computation is indispensable. Both event influence on the match and its relevance to user request are taken into account. For influence computation, event rank and occurrence time are two main factors [41]. For relevance measurement, the semantic consistency on involved players and event types are considered.

(1) Event Rank. The rank of event is directly determined by its influence to the game state. Considering the difference between basketball and football match, two representative rank criteria are adopted to evaluate the event importance in these two sports types. Since shot event is common in basketball matches, the change of game leader and score gap are used as the indicator for event importance evaluation. While for the football match, influential events are confined in limited types and thus only shot and card events are ranked. With the rank list given in Table 4, the rank-based event importance $I_{\mathrm{ra}}(E_i)$ $(0 \le I_{\mathrm{ra}} \le 1)$ is defined as follows:

$$I_{\mathrm{ra}}(E_i) = 1 - \frac{R_i - 1}{3} \cdot \alpha, \tag{6}$$

where R_i $(0 \le R_i \le 4)$ represents the rank level of event E_i and α $(0 \le \alpha \le 1)$ denotes the adjustable parameter to control the effects of rank difference on event importance computation.

(2) Event Occurrence Time. In the sports match, events occurring at the end of the match are usually critical to both sides because little time is left to change the final result. In our approach, the event occurrence time based importance $I_t(E_i)$ $(0 \le I_t \le 1)$ is defined as follows:

$$I_t(E_i) = 1 - \frac{N - i}{N} \cdot \beta, \tag{7}$$

where N is the total event number, i is the index of E_i, and β $(0 \le \beta \le 1)$ is the adjustable parameter to control the effects of event occurrence time on event importance computation.

(3) Event Relevance. Different from the type rank and occurrence time that are decided by event itself, the event relevance reflects the semantic consistency between event content and user's preference. This item plays an important role in our

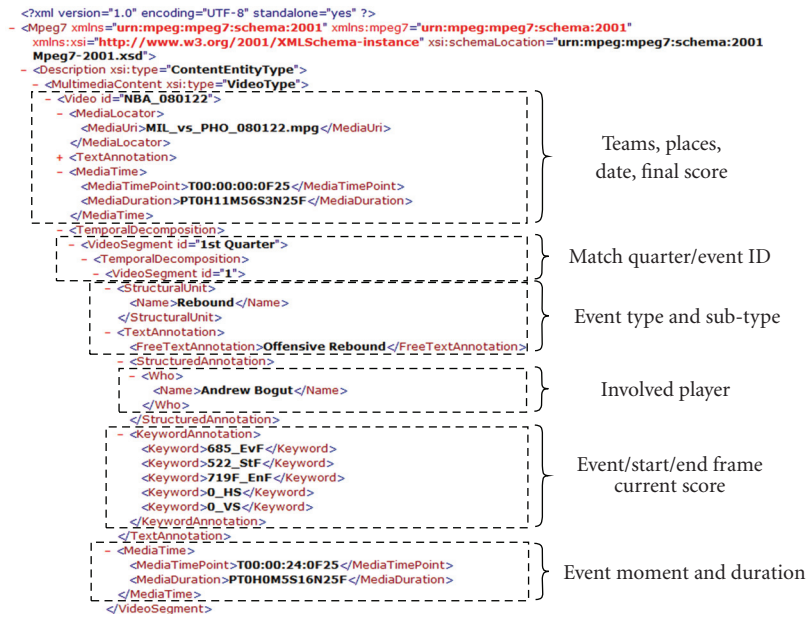


Figure 7: An example of MPEG-7 sports video description XML file.

Table 3: Sequence matching algorithm for video-text alignment.

Output	Alignment sequence $\mathbf{a} : \mathbf{a}_1 \sim \mathbf{a}_{\max(m,n)}$
Input	Video tag sequence $\mathbf{v} : \mathbf{v}_1 \sim \mathbf{v}_m$ and text tag sequence $\mathbf{t} : \mathbf{t}_1 \sim \mathbf{t}_n$
Step 1	Forward computation
Step 1.1	Allocate a score matrix $\mathtt{M}$ and a direction matrix $\mathtt{B}$.
	Both matrixes are with m rows and n columns.
Step 1.2	Fill all entries in $\mathtt{M}$ and $\mathtt{B}$ from the up left to bottom right corner.
	(i) For each $\mathtt{M}_{i,j}$, its value is calculated by (3), where $s_{i,j}$ and Pg are computed by (4) and (5).
	(ii) For each $\mathtt{B}_{i,j}$, its value is the optimal matching direction indicated by $\mathtt{M}_{i,j}$ in (3).
Step 2	Backward tracing
Step 2.1	Allocate an alignment sequence a with $\max(m, n)$ elements.
Step 2.2	Fill the entries of $\mathbf{a}$ in a reverse direction by tracing back along $\mathtt{B}$.
	(i) For each $\mathbf{a}_k$, its value is equal to $\mathtt{B}_{i,j}$, where (i, j) corresponds to the coordinate of the kth backtracking step in $\mathtt{B}$.
Step 3	Result explanation
Step 3.1	Three possible values (directions) may appear in $\mathbf{a}_k$, which are
	(i) diagonal, corresponding to the "$\mathbf{v}_i - \mathbf{t}_j$" alignment,
	(ii) downward, corresponding to the "empty-$\mathbf{t}_j$" alignment,
	(iii) rightward, corresponding to the "$\mathbf{v}_i$-empty" alignment.

Table 4: Event importance rank list.

Rank Level	Event description	
	Basketball	Football
1	Score events and their inductive foul events that can change the game leader	Score (goal events)
2	Score events and their inductive foul events that retain the game leader but change the scoring gap	Offense events and red card event
3	Offense events that failed to score a goal and common foul events	Yellow card event
4	All other events that are not in the rank 1 to 3	All other events

Remark: in above description, score event refers to attack event resulting in a score, while offense event refers to attack event without resulting in a score.

personalized customization because it effectively introduces the user's opinion in event importance computation. By increasing the importance score of semantically related events, the system can finally tailor the personalized video content to the particular user. With semantic video annotation, the event relevance based importance $I_{\text{re}}(E_i)$ $(0 \le I_{\text{re}} \le 1)$ can be defined as:

$$I_{\text{re}}(E_i) = \gamma^{\text{Dist}_{\text{player}}(E_i,U)} \cdot (1-\gamma)^{\text{Dist}_{\text{event}}(E_i,U)}, \tag{8}$$

where function $\text{Dist}_{\text{player}}$ and $\text{Dist}_{\text{event}}$ measure the semantic consistence between event E_i and user request U on the subjects of involved players and event types, and adjustable parameter γ $(0 \le \gamma \le 1)$ denotes the user preference between the above two subjects.

(4) Preference Learning. With user preference learning, system can provide more appropriate video content to the particular users as they continue to use it. This function is realized by adaptively adjusting the concept weight in accordance with user's input. The calculation of the preference learning-based event importance $I_p(E_i)$ $(0 \le I_p \le 1)$ will be discussed in Section 6.

Based on the above analysis, the event importance can be calculated as:

$$\begin{aligned} I(E_i,U) &= \lambda \cdot I_f(E_i) + (1-\lambda) \cdot I_u(E_i,U) \\ &= \lambda \cdot I_{\text{ra}}(E_i) \cdot I_t(E_i) + (1-\lambda) \cdot I_{\text{re}}(E_i,U) \cdot I_p(E_i), \end{aligned} \tag{9}$$

where λ $(0 \le \lambda \le 1)$ is the fusion parameter distributing the weights on event influence on the match $I_f(E_i)$ and its semantic consistency to the user request $I_u(E_i)$.

5.2. User-Participant MultiConstraint 0/1 Knapsack Model. With the above event importance evaluation scheme, user preference can be effectively incorporated into the video content selection process. However, compared with mass suited video data, user's available viewing conditions, for example, device memory and watching time, are usually not limitless. Hence, how to provide optimal video content under resource-constraint environment is of great practical importance and has been studied in [28, 42]. Merialdo et al. [42] raised a 0/1 Knapsack Problem to model the viewing-time-limited TV program personalization. However, due to the lack of semantic analysis of video content, only category interest rather than content preference was considered in their video personalization system. Wei et al. [28] proposed a Multichoice Multidimension Knapsack strategy to maximize the gross information under multiple environment constraints. Since their method requires to include every video segments (on various abstraction levels) to the final summarization, it is not appropriate for the target-specific video customization, where only video segments containing particular semantics are needed. Motivated by the optimization model used in previous work, we formulated our personalized video customization as a user-participant multiconstraint (UPMC) 0/1 Knapsack problem:

$$\begin{aligned} V &= \max\left(\sum_{i=1}^{n} x_i \cdot I(E_i,U)\right), \\ &\text{s.t.} \quad x_i \in \{0,1\}, \quad i = 1,2,\ldots,n \\ &\sum_{i=1}^{n} x_i r_{ij} \le R_j, \quad j = 1,2,\ldots,m, \end{aligned} \tag{10}$$

where $I(E_i,U)$ represents the integrative importance value of event E_i under specified user request U, r_{ij} be the jth resource consumption of the ith event, R_j be the client-side resource bound of the jth resource, x_i denotes the existence of the ith event in the selected optimal set, and n and m represent the number of events and resource types.

As can be seen from (9) and (10), the proposed video customization strategy adopts a constraint optimization model to well balance the user content preference against multiple resource limitations, from which only video segments possessing higher importance values can be selected into the final customization result. With different fusion parameters, two typical video customization modes, retrieval and summarization, can be treated in an unified manner and seamlessly switched to each other. Specifically, In the case of λ approximating to 0, event importance is mainly decided by user preference, thus only semantically consistent video segments can be assigned higher importance scores and finally presented to the user. While in the case of λ approximating to 1, event importance is largely up to the match itself, hence the selected events can reflect the global situation of the match. Moreover, with unbiased configuration of the fusion parameter ($\lambda = 0.5$), a new customization mode, the personalized video summarization can be realized. In this situation, users can query a video abstract about the specified player or event.

6. System Adaptation

Since user preference is relatively stable in a period of time, customization system with adaptation function is expected to respond to the particular user with more appropriate results but less required interactions. To achieve this goal, complete user preference should be learned as implicitly as possible. However, the completeness usually conflicts with the implicitness and hence previous work can hardly achieve an optimal balance between these two indexes. To conquer this difficulty, we propose a social network based approach to identify latent user preference without additional interactions. By building the social network of video semantics, user preference towards specific concepts can be effectively propagated along the network edge so that their latent attitudes toward other unspecific concepts can be implicitly inferred.

6.1. Concept Social Network. We borrow the idea of social network analysis (SNA) [43] to depict the concept relationship in sports video. A social network is a social

structure made of individuals called "nodes", which are connected by one or more specific types of interdependency. With a graphical representation of individuals' relationship, SNA can utilize the related property and theory in the graph theory to discover hidden structures/properties that cannot be directly perceived or measured [44]. According to the semantic sports video annotation, two parallel social networks for the player and event entities are defined as follows:

$$G_p = \left\langle V_p, E_p, W_p \right\rangle, \qquad G_e = \langle V_e, E_e, W_e \rangle, \tag{11}$$

where $V_p = \{v_{p_1}, v_{p_2}, \ldots, v_{p_n}\}$ represents the set of players appearing in the match, $E_p = \{e_{p_{i,j}} \mid e_{p_{i,j}} = 0 \text{ or } 1\}$ is a binary matrix indicating the relationship existence between players i and j, and $W_p = \{w_{p_{i,j}}\}$ denotes the strength of the relationship between players i and j. Similar explanations can be obtained from event network G_e.

To build above weighted concept network, the quantization of concepts' relationship is critical. In the scenario of sports matches, the relation between concepts is developed when they appear in the similar match context. Here, the match context refers to the event type when we consider the player network and refers to the player when we consider the event network, in other words, the player and event are mutual match context of each other. Therefore, the more often two players appear in the same events, the closer relationship is built between them in the player network, and we can quantify this relationship as the number of cooccurrence in the same match context between two players. Similar conclusion can be also drawn from the event network.

With the obtained semantic sports video annotation, a match can be viewed as a bipartite graph in Figure 8(a), where the square nodes denote events, the circle nodes denote players and the edge between a pair of event and player nodes represents their cooccurrence in the same text event. For a match with m events and n players, the above bipartite graph can be represented as a matrix $\mathtt{A} = [a_{ij}]_{m\times n}$, where the entry a_{ij} represent the cooccurrence times of the ith event and jth player in the same text events. The jth column, $\mathbf{a}_j^\top = \{a_{1j}, a_{2j}, \ldots a_{mj}\}$, of matrix $\mathtt{A}$ represents the cooccurrence times between the jth player and other m events, and the ith row, $\bar{\mathbf{a}}_i = \{\bar{a}_{i1}, \bar{a}_{i2}, \ldots, \bar{a}_{in}\}$ of matrix $\mathtt{A}$ represents the cooccurrence times between the ith event and other n players. Based on the event-player bipartite graph, we can build the concept social networks as follows:

$$\begin{aligned}
&\mathtt{E} = \mathtt{A}\mathtt{A}^\top = \left[E_{ij}\right]_{m\times m},\\
&\quad \text{where } E_{ij} = \begin{cases} \sum_{k=1}^{n} \bar{a}_{ik}\bar{a}_{jk} = \bar{\mathbf{a}}_i\bar{\mathbf{a}}_j^\top, & \text{when } i \neq j,\\ 0, & \text{when } i = j, \end{cases}\\
&\mathtt{P} = \mathtt{A}^\top\mathtt{A} = \left[P_{ij}\right]_{n\times n},\\
&\quad \text{where } P_{ij} = \begin{cases} \sum_{k=1}^{m} a_{ki}a_{kj} = \mathbf{a}_i^\top\mathbf{a}_j, & \text{when } i \neq j,\\ 0, & \text{when } i = j, \end{cases}
\end{aligned} \tag{12}$$

where E and P correspond to the event and player networks, respectively.

In the example of Figure 8, the generated social network between events and players are illustrated in Figures 8(b) and 8(c). The thicker an edge is, the more similar of two concepts in their match context. Take the edge between e_1 and e_3 in Figure 8(b) as an example, it is the thickest one among all three edges, meaning these two events usually occur with the same players (p_3 and p_4) in the match. Hence, for users who like event e_1, they may be also interested in e_3 because they are potentially fond of p_3 and p_4. This analysis can be applied to the player network where its match context refers to the cooccurring events.

6.2. User Preference Learning. With the help of social network analysis, complete user preference can be effectively inferred from the finite customization process. For the convenience of the following discussion, we introduce a concept-weight pair, $\langle C_k, W_k \rangle$, to describe the user preference degree W_k of the kth concept (C_k) in the match. In addition, due to the symmetric processing and similar conclusion of the event and player concepts, we will not differentiate these two kinds of concepts unless it is necessary.

Initially, the weights of all concepts are set to 0. When the kth concept is used as a keyword in video customization, it is regarded as an obvious interesting concept to the user. Hence, the new weight W_k^{new} of the kth concept whose last weight is W_k^{old} is given by the following equation:

$$W_k^{\text{new}} = \phi + W_k^{\text{old}}, \tag{13}$$

where $\phi > 0$ is a constraint increment of concept preference.

As for other concepts that are not specified by the user, we further divide them into two classes, which are concepts that are connected with the kth concept in the social network and those are not. For the former, we distribute the weight increment ϕ to these latent concepts according to their link strength to the kth concept as follows:

$$W_i^{\text{new}} = \phi \cdot \frac{L_{ki}}{\sum_l L_{kl}} + W_i^{\text{old}}, \tag{14}$$

where L_{kl} represents the weight of the edge between the kth and lth concepts in the social network. For the rest concepts, which are neither specified by the user nor connected with the specified concept, their preference weights are decreased as follows:

$$W_j^{\text{new}} = \eta \cdot W_j^{\text{old}}, \tag{15}$$

where $1 > \eta > 0$ is a constant damping factor of concept preference.

With equations (13)–(15), user preference to all concepts in the match are identified. To avoid the weight divergence, a normalization process is adopted:

$$W_k = \frac{W_k^{\text{new}}}{\max\left\{W_k^{\text{new}}\right\}}, \tag{16}$$

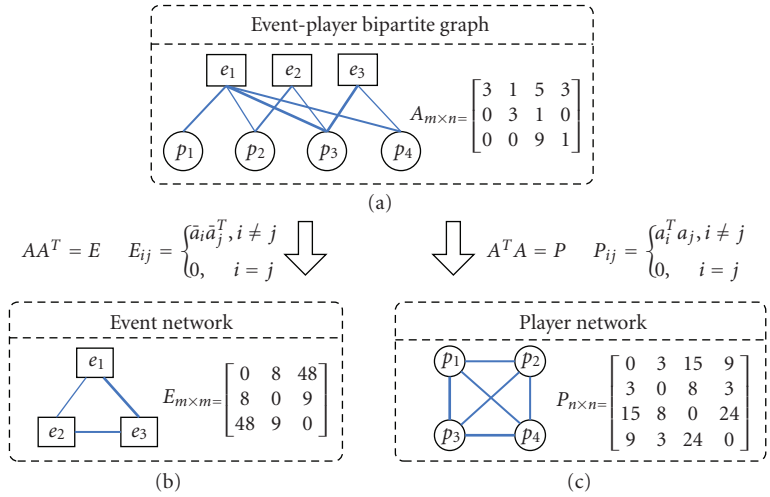


Figure 8: Graphical example of player and event network construction.

so that the normalized concept weight is not larger than 1. The corresponding preference learning based event importance $I_p(E_i)$ $(0 \le I_p \le 1)$ can be computed as:

$$I_p(E_i) = \frac{1}{2} \cdot \left(W_{\text{player}}(E_i) + W_{\text{event}}(E_i)\right), \tag{17}$$

where $W_{\text{player}}(E_i)$ and $W_{\text{event}}(E_i)$ correspond to the normalized preference weights of the player and event concepts in event E_i.

7. Experimental Result

In order to evaluate the proposed method, we conduct our experiment on more than 1000-minute broadcast sports video, including 2 NBA 2005-2006 and 4 NBA 2007-2008 basketball matches, 2 Euro-Cup 2004 and 4 World Cup 2006 football matches. The corresponding web-casting texts are obtained from the ESPN and BBC sports websites. In average, there are about 400 text events happening in one 100-minute basketball match and 50 text records in one 90-minute football match.

7.1. Sports Video Annotation. In this part, attack-based video and text segmentation, coarse and refined video annotation, and comparative experiment with timestamp-based method are reported to validate the effectiveness of the proposed semantic-matching algorithm.

To evaluate the obtained segmentation result, the "purity" index proposed in [45] is adopted in our experiment. Given a sequential data, a ground truth segmentation $\mathbf{S} = \{(s_1, \triangle t_1), \ldots, (s_g, \triangle t_g)\}$, and an automatic segmentation $\mathbf{S}^* = \{(s_1^*, \triangle t_1^*), \ldots, (s_g^*, \triangle t_g^*)\}$, the purity π is defined as:

$$\pi = \left(\sum_{i=1}^{g} \frac{\tau(s_i)}{T} \sum_{j=1}^{a} \frac{\tau^2\left(s_i, s_j^*\right)}{\tau^2(s_i)}\right) \cdot \left(\sum_{j=1}^{a} \frac{\tau\left(s_j^*\right)}{T} \sum_{i=1}^{g} \frac{\tau^2\left(s_i, s_j^*\right)}{\tau^2\left(s_j^*\right)}\right), \tag{18}$$

where $\tau(s_i, s_j^*)$ is the length of overlap between the scene segmentation s_i and s_j^*, $\tau(s_i)$ is the length of the scene s_i, and T is total length of all scenes. In each parenthesis, the first term is the fraction of recording for which a segment accounts, and the second term is a measure of how much a given segment is split into small fragments. The purity value ranges from 0 to 1. The larger the value is, the closer the result approaches the ground truth.

With the manually labeled ground truth segmentation, the attack-based video and text segmentation results are given in Figure 9. The consistent advantage of text segmentation result over that of sports video reflects the semantic intuitionism of the attack-based segmentation. The individual errors in the web text are mainly caused by the omission of some text records when the attack side is changed. As for the comparatively lower purity in video segmentation, large-scale back passings and inaccurate field zone information (especially in football matches) are primary reasons. However, despite some inaccuracies, the attack-based video and text segmentations in both basketball and football matches are generally feasible for the following semantic video annotation.

Leave-one-cross validation is adopted to evaluate the video annotation result. For each sports type, five matches are used to train the Bayesian network for salient event detection and the left one is used to test the annotation algorithm. With the attack-based segmentation, the coarse video-text alignment result is listed in Table 5, where the inference accuracy denotes the occupation of correctly detected tags in total video tags and alignment accuracy represents the occupation of all correctly matched tags in total video tags.

As shown in Table 5, the average inference accuracy is only about 77% for basketball matches and 86% for football matches, which reflects the negative effects of the semantic gap in content-based event detection. However, in contrast to the limited inference accuracy, the coarse alignment accuracy is still satisfactory (around 98% for basketball and 95% for

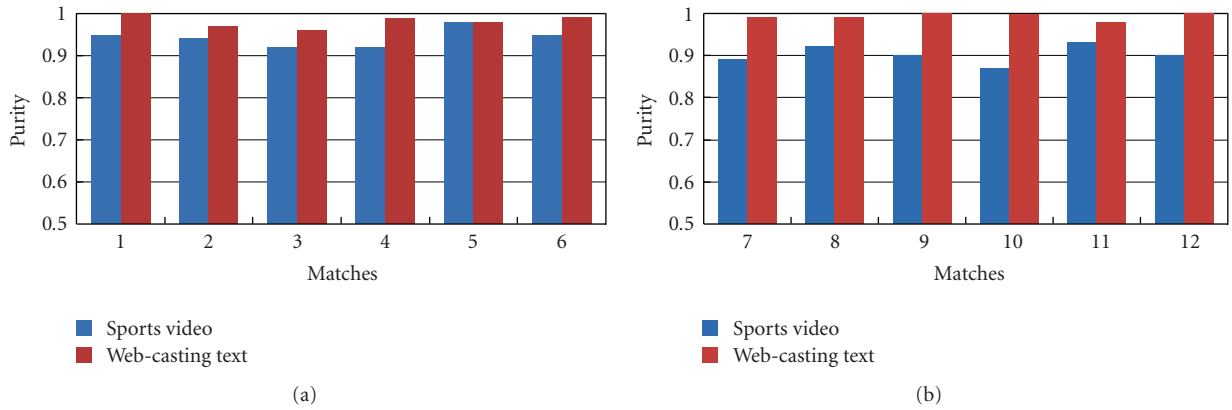


Figure 9: Attack-based sports video and web-casting text segmentation result in (a) basketball and (b) football matches. In both figures, the horizontal axes denote the match ID and the vertical axes represent the purity value of the segmentation result.

Table 5: Attack-based coarse alignment result.

No.	Sports type	Total video tags	Exact matching	Inexact matching	Inference accuracy	Alignment accuracy
1	basketball	192	143	44	74% (143/192)	97% (187/192)
2	basketball	196	152	42	78% (152/196)	99% (194/196)
3	basketball	188	133	49	71% (133/188)	97% (182/188)
4	basketball	176	144	26	75% (144/176)	97% (170/176)
5	basketball	187	150	33	80% (150/187)	98% (183/187)
6	basketball	185	140	38	76% (140/185)	96% (178/185)
7	football	55	45	8	82% (45/55)	96% (53/55)
8	football	50	43	6	86% (43/50)	98% (49/50)
9	football	53	44	5	83% (44/53)	93% (49/53)
10	football	53	47	3	88% (47/53)	95% (50/53)
11	football	51	45	4	88% (45/51)	96% (49/51)
12	football	45	39	2	87% (39/45)	91% (41/45)

football matches resp.). This result demonstrates the strong robustness of the proposed sequence matching algorithm and can be attributed to the semantics-based tags similarity measure and the global optimization strategy, where different tags can be effectively aligned with reference to the global matching status.

Attack-based video-text alignment generates a coarse but accurate video annotation, where the basic unit corresponds to an attack in the match. To obtain a more elaborate annotation, a refined alignment is carried out to locate each text event within a shot. Figure 10 gives the refined annotation result of 6 events in basketball and 7 events in football matches. Due to the lack of semantic relation, the shot-based alignment is not as accurate as the attack-based one. However, with the help of accurate coarse alignment, the following shot-based refinement can still annotate most events in an acceptable accuracy. As for the lower precision/recall rate of the shot event in basketball matches, it is mainly due to the irregular photography in free throw events where short shots rather than long shots were adopted by cameramen. Similar results also appear in the free kick and foul events in football matches, where the camera transition during those events is frequent and dynamic.

To further demonstrate the robustness of our proposed approach, a comparative experiment of our approach with timestamp-based method [35] is conducted on the evaluation data and their balanced F-measure results are shown in Figure 11. In the ideal condition, timestamp-based approach can achieve very high event detection accuracy if the timestamp can be correctly recognized. However, according to our experiments on both basketball and football matches, the above advantage is either not obvious (Figure 11(b)) or even in-existent (Figure 11(a)). This result can be explained from two aspects: First, timestamp cannot be always robustly located and recognized in the practical noisy broadcast video, which affects the event location precision of the timestamp-based method; Second, the basketball match has plenty of clock pauses, which make the timestamp-based method confused to locate the accurate event segment during that period. As for the lower detection accuracy of the proposed approach on football matches, it is mainly caused by the performance degradation when the sequence

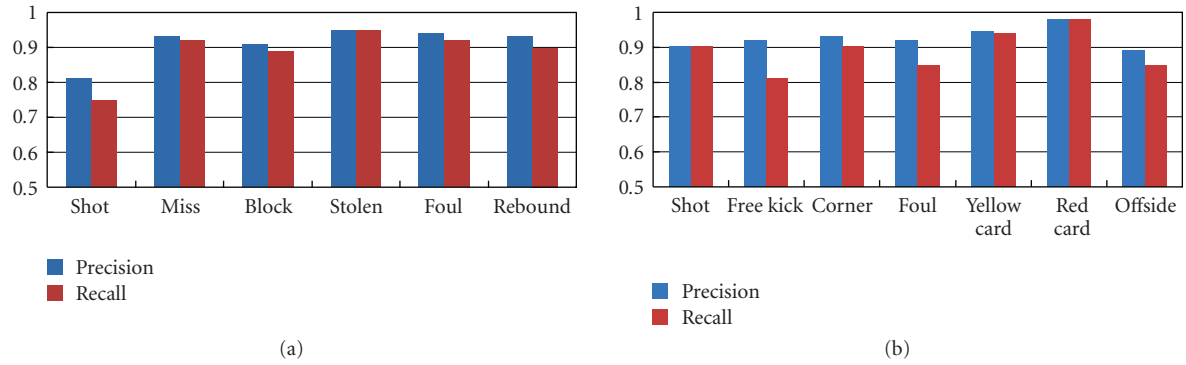


Figure 10: Refined event detection result on (a) basketball and (b) football matches.

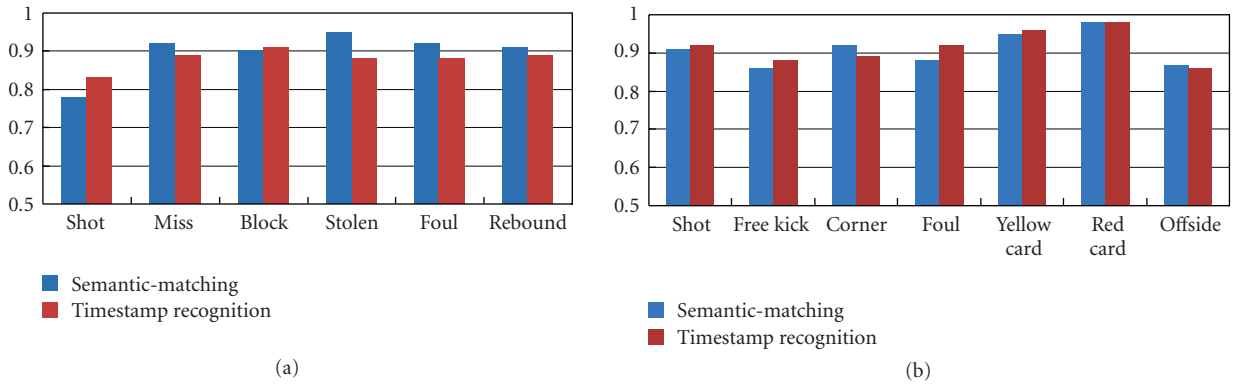
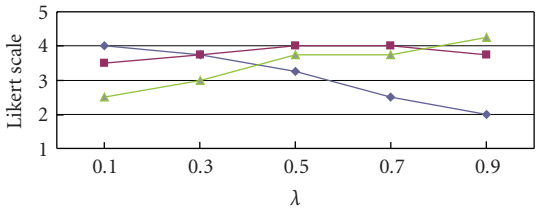


Figure 11: Comparative event detection experiments between semantics-matching and timestamp recognition methods on (a) basketball and (b) football matches.



Consistency	4	3.75	3.25	2.5	2
Conciseness	3.5	3.75	4	4	3.75
Converge	2.5	3	3.75	3.75	4.25

Figure 12: User evaluation of personalized video customization.

matching algorithm is applied to two sequences with obvious length difference (multiple long shots vs. single event) in the shot-level sequence matching.

7.2. Personalized Video Customization. Before we give the experimental results of personalized video customization, we first mention the setting of the adjustable parameters for event importance computation. As can be seen from (6) and (7), α and β are control coefficients to decide the effects of event rank and occurrence time on event importance computation. In our following experiments, both coefficients are set to 1, meaning that event rank and occurrence time fully affect the event semantic significance. In addition, the adjustable parameter γ in (8) is set to 0.5, denoting the equal weights for event type and involved player. As for the preference learning, the constant gain and damping is set to 0.1 and 0.99, respectively, implying the concept preference degree should be increased or decreased gradually. Based on the concept network, the Dist function used in (8) is computed as follows:

$$\mathrm{Dist}(E_i, U) = \begin{cases} 0, & \text{direct connection,} \\ 1 - \dfrac{L_{ki}}{\sum_l L_{kl}}, & \text{indirect connection,} \\ 1, & \text{no connection,} \end{cases} \tag{19}$$

where E_i and U represent the semantic event and user request, respectively, L_{ki} denotes the semantic context similarity between the kth and ith concepts in social network, where the kth concept is customized by U and the ith concept is contained in E_i. The "direct connection" refers to the case when E_i and U are consistent on the player or event type, and the "indirect connection" refers to the case when concept in E_i is not contained in U but indirectly connected with U in the concept network. As for the "no connection" case, neither direct nor indirect relationship can be found between E_i and U.

Table 6: Customized video segments under different fusion parameters.

λ	ID	Moment	Player	Event	Duration (s)
0.1	45	02 : 39	Steve Nash	three point	3.16
	33	06 : 41	Steve Nash	layup	3.08
	11	10 : 22	Steve Nash	jumper	3.40
0.5	54	01 : 10	Grant Hill	three point	3.65
	45	02 : 39	Steve Nash	three point	3.16
	33	06 : 41	Steve Nash	layup	3.08
0.9	54	01 : 10	Grant Hill	three point	3.65
	45	02 : 39	Steve Nash	three point	3.16
	50	01 : 54	Charlie Villanueva	jumper	3.10

Table 7: Learned user preference on player concepts.

Users	Player concept learning results (on 3 football and 2 basketball matches)				
	1st	2nd	3 rd	4th	AP
No. 1	Zidane (12, 3, 4)	Totti (14, 2, 1)	Henry (6, 12, 8)	Piero (10, 1, 3)	(0.83, 0.73)
No. 2	Messi (16, 1, 1)	Gonzalez (8, 6, 7)	Schweinsteiger (10, 4, 5)	Ballack (10, 2, 4)	(0.85, 0.67)
No. 3	Zidane (15, 1, 2)	Ronaldo (13, 2, 4)	Ronaldinho (12, 3, 3)	Carlos (8, 9, 8)	(0.86, 0.60)
No. 4	Redd (15, 2, 1)	Nash (14, 3, 2)	Yi (9, 6, 6)	Bogut (5, 7, 8)	(0.77, 0.75)
No. 5	Kobe (12, 3, 2)	Yao (12, 2, 1)	Gasol (9, 5, 7)	Odom (4, 8, 11)	(0.78, 0.70)

Due to the intrinsic subjectivity of personalized video customization, we carry out a user study to evaluate the performance of our customization system. Our study involves 12 volunteers, who are all first-time users of the customization system and have certain knowledge about basketball and football games. For each user, after watching 3 or 4 integrated matches, he/she is asked to use the system to customize their favorite video clips with various personalized preference, such as players, events and time, under different fusion parameters.

To validate the effectiveness of the proposed personalized customization algorithm, a specially designed questionnaire is handed out to the participants to get their feedbacks to the generated video content. Motivated by the work in [46], we define a new "3C" criterion in the questionnaire as follows:

(1) Consistency. Whether the generated video clip is consistent with user request on content semantics, such as involved players and event types.

(2) Conciseness. Whether the generated video clip capture the main body of the match without including irrelevant events.

(3) Coverage. Whether the generated video clip covers all important events happening in the match under current viewing time limit.

All above three indexes need to be answered on a five-grade Likert scale [47] where 1 denotes strongly reject, 2 reject, 3 marginally accept, 4 accept, and 5 strongly accept.

As can be seen from Figure 12, when the fusion parameter λ increases from 0.1 to 0.9, average user evaluation on result consistency gradually declines from 4 to 2. Meanwhile, the evaluation on result coverage behaves oppositely, rising from 2.5 to 4.25. This contrast reveals the important role that λ plays in balancing the game content and user preference in the video customization process. When λ is small, the algorithm will emphasize more on user's request, hence more semantically related events are selected. On the contrary, when λ is big, say equals to 0.9, the result is largely up to the match status and thus the globally interesting events are presented. Table 6 lists the customization results under different fusion parameters in response to the user request "selecting 10 seconds highlight about Steve Nash's shot event in the first quarter of NBA 2008 Suns vs. Bucks". As shown in the table, the selected segments are highly related to user request when λ equals to 0.1. As λ is growing from 0.5 to 0.9, the customized video clips gradually change to other more important shot events in the end the match. In those cases, user request is not strictly obeyed and followed, for example, the involved player.

7.3. System Adaptation. In this section, we investigate whether the proposed system can adaptively acquire user's

Table 8: Learned user preference on event concepts.

Users	Event concept learning results (on 3 football and 2 basketball matches)				
	1st	2nd	3rd	4th	AP
No. 6	Shot (10, 2, 1)	Free Kick (4, 4, 6)	Corner (8, 5, 5)	Red Card (2, 6, 7)	(0.57, 0.62)
No. 7	Goal (2, 6, 8)	Shot (5, 1, 2)	Offside (4, 5, 4)	Yellow Card (5, 3, 1)	(0.73, 0.81)
No. 8	Foul (5, 3, 4)	Corner (10, 1, 1)	Yellow Card (3, 5, 5)	Shot (1, 6, 7)	(0.73, 0.67)
No. 9	Shot (12, 1, 1)	Foul (10, 2, 2)	Stolen (5, 6, 5)	Rebound (9, 3, 4)	(0.92, 0.88)
No. 10	Shot (10, 2, 1)	Miss (5, 5, 6)	Foul (8, 1, 2)	Block (8, 3, 5)	(0.95, 0.82)

personalized preference from the customization process. For the convenience of result evaluation, a set of preselected concepts, including players and events, are given to users so that they can focus attention on these semantics in their following operations. Each user is assigned a sports match with 4 emphasis concepts, and is asked to use the proposed system for one or two hours, evoking the procedure of user preference leaning. After about 50 rounds of interaction, we sort the concepts in accordance with their preference weights, and list the ranks of those preselected concepts in the sorted sequence. To properly measure the system learning ability, we adopt the average precision (AP) index used in information retrieval to objectively depict the system's learning ability with user interaction.

Tables 7 and 8 show the experimental results of learned user preference on player and event concepts in football and basketball matches. In both tables, each row corresponds to a user preference learning result on one match, and each cell of its central part is filled with an emphasized concept name and a triple group, marked as (#F, #R1, #R2), where #F, #R1 and #R2 denote the concept queried frequency and its final rankings with and without network-based preference propagation. Similarly, the AP result corresponding to each match is a pair of numbers, marked as (#AP1, #AP2), representing the AP result of the customization system with and without concept network analysis. Take the up-left cell in Table 7 as an example, the user beforehand selected the most interesting player is "Zinedine". After 12 queries with that keyword, the derived weight rankings are 3 and 4, respectively, and the related AP results on the four preselected concepts are 0.83 and 0.73, respectively, corresponding to the learning algorithm with and without using concept network.

The experimental results indicate that 75% and 60% of the predecided player concept and 60% and 55% of those for event concepts were placed in the top-4 rank of the learned concept sequence with and without concept social network analysis (CSNA), respectively. Both of which reflect the effectiveness of the proposed user preference learning algorithm, especially the network-based preference propagation strategy. For the player concept learning in Table 7, CSNA acts more effectively on football matches than basketball matches, which is mainly due to the player context difference. Specifically, the selected football players are highly connected in attack events than that of basketball players (because the basketball players are usually involved in both offense and defense events in the match), thus it is more easily for the proposed CSNA algorithm to locate those football forwards than those in basketball matches. Similar explanations can be also applied to the event learning result in Table 8. In summary, the proposed user preference learning algorithm can effectively acquire user preference from their customization operations, and concept network analysis further improves the system performance especially when users appetite is fixed and specified.

8. Conclusion

Video personalization is an important mechanism to provide particular viewers with their favorite content. Considering the diverse content preference and environment limitations, detailed video semantics should be fully mined and reasonably evaluated so that suited video segments can be collected for the specific user. This is a challenging task and its difficulty embodies in every submodules including video annotation, personalized customization and system adaptation.

In this paper, we presented an integral framework for personalized sports video customization. In the off-line annotation, a hierarchical video-text matching method is raised to align the multimedia information based on their semantic correspondence, which generates both refined and accurate video content description. In the on-line customization, a user-participant multiconstraint 0/1 Knapsack model is proposed to realize semantic content retrieval and summarization under resource-limited condition. To facilitate the above on-line customization process, a concept network based system adaptation algorithm is designed to implicitly infer the complete user preference.

The approach's complexity focuses on the semantic video annotation. Compared with the simple web-text analysis, intensive video processing, for example, camera motion estimation and replay shot detection, costs the majority of

computation resources. Moreover, since semantic-matching algorithm needs the integral video and text sequences as its inputs, the alignment result will not come out unless the whole match is over. For these reasons, video annotation is performed in an off-line manner in our approach. However, once the semantic annotation is obtained, video customization and system adaptation (concept weight adjustment) only involve some mathematical calculations, and thus can achieve real-time processing. Both quantitative and qualitative experiments conducted on more than 1000 minutes sports video validate the effectiveness of the proposed approach.

Another point needs to be addressed is the expansibility of the proposed framework. Since the on-line customization and adaptation are irrelevant with specific sports genres, the expansibility is mainly up to the video annotation. Although our work is implemented on the field sports, like football and basketball, we think the basic idea and framework of the semantic matching is general and can be easily adapted to other opponent sports. The evidences supporting our thought come from two sides: First, live casting text is now a standard service in famous sports websites, such as ESPN and BBC, which covers the majority of popular sports genres in our daily lives. Hence, there is no lack of textual descriptions for sports video annotation. Second, the notion of "attack" exists in most opponent sports and has very clear semantics, hence its robust detection in other sports types is also feasible (although may be not exactly the same as our current methods for football and basketball matches). Therefore, with some necessary modifications according to the specific environment, the basic framework of semantic matching can still be applied in other sports genres.

In the future, we plan to improve current work in both theoretic and application aspects. In semantic video annotation, Bayesian network (BN) is adopted to tag the semantic content of video segment and sequence matching algorithm is applied to align the video and text tag sequences. Such a method ignores the temporal dependence of neighboring semantic tags and thus may impair the final matching precision. To cover this shortage, more advanced sequential models, like hidden Markov model (HMM), or dynamic Bayesian network (DBN), can be employed to model the text facilitated video annotation, where video-text alignment can be treated as a hidden state inference problem and solved with Veterbi-like inference algorithm. On the other hand, video personalization in this paper mainly focuses on the content selection. However, as a practical system, problems such as video encoding and decoding in different communication channels and video displaying on various devices are also of great importance. To the end user, video content selection, data transmission, and terminal displaying constitute an integral solution for personalized video customization.

Acknowledgment

This work is supported by Natural Science Foundation of China (no. 90920303).

References

[1] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 105–115, Los Angeles, Calif, USA, 2000.

[2] M. Xu, N. C. Maddage, C. S. Xu, M. S. Kakanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Proceeding of the IEEE International Conference on Multimedia and Expo (ICME '03)*, vol. 2, pp. 281–284, Baltimore, Md, USA, 2003.

[3] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlight extraction from baseball, golf and soccer games in a United Framework," in *Proceeding of the IEEE International Conference on Acuoustics, Speech and Signal Processing (ICASSP '03)*, vol. 5, pp. 632–635, Hong Kong, April 2003.

[4] L.-Y. Duan, M. Xu, and Q. Tian, "Semantic shot classification in sports video," in *Storage and Retrieval for Media Database*, vol. 5021 of *Proceedings of SPIE*, pp. 300–313, January 2003.

[5] Y.-P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 133–146, 2000.

[6] D. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 315–318, Juan-les-Pins, France, December 2002.

[7] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 285–305, 2003.

[8] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.

[9] A. S. David and E. O. C. Eoel, "Event detection in Field sports video using audiovisual features and a support vector machine," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225–1233, 2005.

[10] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic analysis of soccer video using dynamic Bayesian network," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 749–760, 2006.

[11] L. Xing, Q. Ye, W. Zhang, Q. Huang, and H. Yu, "A scheme for racquet sports video analysis with the combination of audio-visual information," in *Visual Communications and Image Processing*, vol. 5960 of *Proceedings of SPIE*, pp. 259–267, 2005.

[12] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 33–44, Berkeley, Calif, USA, 2003.

[13] M. Xu, L. Duan, C. Xu, M. Kankanhalli, and Q. Tian, "Event detection in basketball video using multi-modalities," in *Proceeding of IEEE Pacific Rim Conference on Multimedia (PCM '03)*, vol. 3, pp. 1526–1530, Singapore, December 2003.

[14] K. Wan and C. Xu, "Efficient multimodal features for automatic soccer highlight generation," in *Proceedings of the International Conference on Pattern Recognition (ICPR '04)*, vol. 3, pp. 973–976, Cambridge, UK, August 2004.

[15] M. H. Kolekar and S. Sengupta, "A hierarchical framework for generic sports video classification," in *Proceedings of the 7th Asian Conference on Computer Vision (ACCV '06)*, vol. 3852

of *Lecture Notes in Computer Science*, pp. 633–642, Springer, Hyderabad, India, January 2006.

[16] M. Xu, L.-Y. Duan, C.-S. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 3, pp. 189–192, Hong Kong, 2003.

[17] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 347–350, Juan-les-Pins, France, December 2002.

[18] M. H. Kolekar and S. Sengupta, "Event-importance based customized and automatic cricket highlight generation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 1617–1620, July 2006.

[19] C. Jung and J. Kim, "Player information extraction for semantic annotation in golf videos," *IEEE Transactions on Broadcasting*, vol. 55, no. 1, pp. 79–83, 2009.

[20] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 68–75, 2002.

[21] H. Xu and T.-S. Chua, "The fusion of audio-visual features and external knowledge for event detection in team sports video," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '04)*, pp. 127–134, New York, NY, USA, October 2004.

[22] C. S. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and web-casting text," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MM '06)*, pp. 221–230, Santa Barbara, Calif, USA, October 2006.

[23] D. Tjondronegoro, Y.-P. P. Chen, and B. Pham, "Integrating highlights for more complete sports video summarization," *IEEE Multimedia*, vol. 11, no. 4, pp. 22–37, 2004.

[24] A. Ekin and A. M. Tekalp, "Generic play-break event detection summarization and hierarchical sports video analysis," in *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME '03)*, vol. 1, pp. 167–172, Baltimore, Md, USA, July 2003.

[25] M. Fleischman and D. Roy, "Situated models of meaning for sports video retrieval," in *Proceeding of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL '07)*, pp. 37–44, Rochester, NY, USA, April 2007.

[26] N. Babaguchi, K. Ohara, and T. Ogura, "Learning personal preference from viewer's operations for browsing and its application to baseball video retrieval and summarization," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 1016–1025, 2007.

[27] W. Gao, Q.-M. Huang, S. Q. Jiang, and P. Zhang, "Sports video summarization and adaptation for application in mobile communication," *Journal of Zhejiang University: Science A*, vol. 7, no. 5, pp. 819–829, 2006.

[28] Y. Wei, S. M. Bhandarkar, and K. Li, "Video personalization in resource-constrained multimedia environments," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 902–911, Augsburg, Germany, September 2007.

[29] B. L. Tseng, C. Y. Lin, and J. R. Smith, "Video summarization and personalization for pervasive mobile devices," in *Storage and Retrieval for Media Database*, vol. 4676 of *Proceedings of SPIE*, pp. 359–370, 2002.

[30] Y. Zhang, X. Zhang, C. Xu, and H. Lu, "Personalized retrieval of sports video," in *Proceedings of the International Workshop on Multimedia Information Retrieval (MIR '07)*, pp. 313–322, Augsburg, Germany, September 2007.

[31] A. Amir, M. Berg, and H. Permuter, "Mutual relevance feedback for multimodal query formulation in video retrieval," in *Proceedings of International Workshop on Multimedia Information Retrieval (MIR '05)*, pp. 17–24, Singapore, November 2005.

[32] T. Syeda-Mahmood and D. Ponceleon, "Learning video browsing behavior and its application in the generation of video previews," in *Proceedings of the 9th ACM International Conference on Multimedia (MM '01)*, pp. 119–128, Ottawa, Canada, September 2001.

[33] J. Zimmerman, K. Kurapati, A. L. Buczak, D. Schaffer, S. Gutta, and J. Martino, "TV personalization system: design of a TV show recommender engine and interface," in *Personalized Digital Television: Targeting Programs to Individual Viewers*, pp. 27–51, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[34] A. Jaimes, N. Sebe, and D. Gatica-Perez, "Human-centered computing: a multimedia perspective," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MM '06)*, pp. 855–864, Santa Barbara, Calif, USA, September 2006.

[35] C. Xu, J. J. Wang, H. Q. Lu, and Y. F. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 421–436, 2008.

[36] C. Liang, Y. Zhang, C. S. Xu, J. Q. Wang, and H. Q. Lu, "A hierarchical semanticsmatching approach for sports video annotation," in *Proceedings of the 10th Pacific Rim Conference on Multimedia (PCM '09)*, vol. 5879 of *Lecture Notes in Computer Science*, pp. 684–696, Bangkok, Thailand, December 2009.

[37] L. Wang, M. Lew, and G. Xu, "Offense based temporal segmentation for event detection in soccer video," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '04)*, pp. 259–266, New York, NY, USA, October 2004.

[38] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 497–501, 2000.

[39] C. Xu, Y. F. Zhang, G. Y. Zhu, Y. Rui, H. Q. Lu, and Q. M. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Transaction on Multimedia*, vol. 10, no. 7, pp. 1342–1355, 2008.

[40] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

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

[42] B. Merialdo, K. T. Lee, D. Luparello, and J. Roudaire, "Automatic construction of personalized TV News programs," in *Proceedings of the ACM International Multimedia Conference*, pp. 323–331, Orlando, Fla, USA, October 1999.

[43] J. Scott, *Social Network Analysis: A Handbook*, Sage, Newbury Park, Calif, USA, 1991.

[44] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[45] A. Vinciarelli and S. Favre, "Broadcast news story segmentation using social network analysis and hidden Markov models," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 261–264, Augsburg, Germany, September 2007.

[46] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *Proceedings of the ACM International Multimedia Conference and Exhibition*, pp. 489–498, Orlando, Fla, USA, October 1999.

[47] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.