Abstract—In this paper we have implemented a mobile personalized sports video customization system for mobile users by using a novel approach. The system is based on the B/S architecture, with this architecture, the whole system can be divided into two parts: a friendly client browser on smart phones and an intelligent multimedia analysis server. For the client browser, a friend UI is designed for mobile users to customize their favorite video clips. For the multimedia server, with the analysis of video content as well as user preference, semantics-specific video clips can be tailored to the particular users to effectively improve their visual enjoyment. The whole approach of the multimedia analysis is implemented in a two steps. First, sports video content is automatically indexed by detailed textual descriptions, which are derived from latent semantic analysis of web-casting text, based on the semantic annotation, a video database is constructed. And then video content and user preference are collaborated into event importance computation mechanism, after that the server can tailor the optimal set of video clips to the particular user. Particularly, we adopt basketball and football games as our initial sports genres because they not only widely adopted study bed but also globally popular sports, which possess great values in both research and application. Both quantitative and qualitative experiments validate the effectiveness of our system.

Keywords—sports video; mobile phone; personalized customization; semantic annotation

I. INTRODUCTION

With the rapid development of program production technology, an explosive proliferation of sports video is made available on broadcast and Internet. However, the amount of multimedia content is increasing so rapidly that people may not have enough time to browse all favorite video content. What’s more, the traditional sports programs are made by studio professionals without considering the diversity of audiences, hence cannot meet people's various demands towards particular players or events. For example, a basketball fans may only interest in the slam duck or three point shot in NBA matches. While an England fans may favorite shots concerning Beckham than that of other players.

In such condition, the ability of effectively retrieving the most desirable content by personal preference from the lengthy and voluminous sports video program is of great importance. To realize sports video personalized customization, the source video need to be annotated in a more refined scale and higher semantic level. The annotation should not only be broad to cover the general events in the video, but also be sufficiently deep to cover the semantics of the events, e.g. the events type, the events happening time, the name of the joined players, the teams which the players belong to, etc.

Moreover, with the advance of 3G technology, more and more people, especially sports fans, favorite in using multimedia-enable mobile devices to watch sports video, because they can freely enjoy their favorite matches at anytime in anyplace. The mobile video service may have tremendous commercial potentials. In view of this extensive prospect, we have implemented a mobile personalized sports video customization system for customs.

Related work [1], [2], [3] is focused on video analysis and customization. Previous sports video analysis approaches can be classified into two categories: using low-level features extracted from the video itself only and using external sources. For the content based method, various low-level features, such as audio, visual and textual features are utilized for video semantics extraction [4]. Tan et al. [5] used low-level information available directly from MPEG compressed video, combined with domain knowledge of basketball, to identify certain events and classify video into wide-angle and close-up shots. Xiong et al. [6] developed a unified framework to extract highlights from three sports: baseball, golf and soccer by detecting some of the common audio events that are directly indicative of highlights. Nepal et al. [7] presented temporal models with audio/visual features to detect goal segments. All these approaches were based on low-level features extracted from the video itself, which are difficult to extract detailed semantics from the video sources due to the semantic gap between low-level features and high-level concepts.

In order to get ideal event detection accuracy and high level semantic description, text-facilitated sports video analysis has been validated as an effective method under current state of the art. There are two external text sources were used for sports video annotation: close caption and web text. Closed caption was employed to identify semantic event segments in American football games [8]. Since closed caption is generated from speech to text directly, its colloquial words and poorly-defined structure will cause the difficulty for automatic text parsing. The rich information and high level semantics of web
texts have made it to be a trend to leverage these information sources for supporting event detection and content annotation. An approach was proposed by Xu and Chua [9] to utilize match report and game log obtained from web to assist event detection in soccer videos. Although this approach can get an encouraging performance, match report and video events were matched based on the global structures, it cannot achieve live event detection.

In this paper, we exploit web-casting texts to help detect events and extract semantics from sports games. In our previous work, we developed personalized sports video customization system for mobile devices based on web-casting text and broadcast video [10]. In this paper, we adopt a novel approach to perform automatic text analysis and provide personalized video customization service to mobile users by collaborating content and semantic analysis. The main contribution of our work can be summarized as follows: 1. propose a general approach, which based on unsupervised clustering method, instead of pre-defined keywords to automatically detect event from web-casting text; 2. provide a general and robust event importance computation scheme for the selected video segment groups by making full use of the video groups’ content and its semantic information; 3. a friendly, elegant and easy to use client user interface is designed to facilitate mobile customers enjoying their sports video customization trip.

The rest of this paper is organized as follows. Section II describes the proposed system framework. Section III describes the technical details of the web-casting text analysis, sports video analysis and video database construct. After that, the friendly, elegant client end user interface and personalized video customization mechanism are introduced in Section IV and V, respectively. Experimental results are shown in Section VI and Section VII concludes this paper.

II. THE PROPOSED SYSTEM FRAMEWORK

Based on the idea illustrated above, we propose a B/S framework for the proposed personalized video customization system, as shown in Fig.1. It consists of two major components: the multimedia server and the mobile client. For the multimedia server, the sports videos and its corresponding web-casting text are integrated to detect and annotate the events in the original videos. First, the web-casting texts are automatically captured from the web pages. We use regular expression to parse the play-by-play region of the web pages in which the web-casting text lies. In the text, the text events are detected by the keywords which are generated based on a Latent Semantic Analysis (LSA) method [11], and the semantics are also been extracted. Then we align the text events with the video to detect the video event clips and annotate them with the semantics from the text. Finally, the video event clips are segmented from the original video and indexed in a sports video database. The mobile client end provide a elaborate designed user interface to the users and allows them to submit their personalized preference request to customize their favorite sports video clips. Once the server receives the customization request, it will searches through the database and tailor the optimal set of video clips to the users.

III. SPORTS VIDEO ANALYSIS AND SEMANTIC ANNOTATION

A. Web-casting Text Analysis

In our system, high-level semantics play a very important role, it is the basis of our personalized sports video customization service. In our previous work [12], we used pre-defined keywords to match related words in the web-casting texts to detect text events, which is less general to different sports domains. In order to provide a more general approach to analysis the web-casting text, we employ an unsupervised approach to fulfill the task. In this approach, we firstly extract event descriptions from the web-casting texts and then use Latent Semantic Analysis (LSA) and K-means approach to analysis and cluster the event descriptions into different categories, after that, a rule-based part of speech tagger is utilized to tag the nouns and verbs in the groups. In each cluster, these nouns and verbs are ranked by their “term frequency-inverse document frequency” (tf-idf) weights [13]. The top ranked words are selected as the keywords for each cluster. Based on the keywords, we can directly detect the text events by searching the records which contain the keywords. Once the text events detected, the time-tag and the description of the detected event is used for text/video alignment. For more technical details, reader can reference [11].

B. Sports Video Analysis

The basic idea of web-casting text based video analysis is to detect time-tag, which indicates the event occurring moment during the game, from text events and map them into the video stream. So, as the events in the web-casting text have been detected, we firstly obtain the time-tag from the text events and recognize the game time in the video. Then we use them to detect the exact frame, which corresponds to the event time in the video stream, as the reference frame. Finally, a finite state machine based approach [12] is used to model event structure and detect the event boundary.

1) Game Time Recognition: In most sports game videos such as football and basketball, a digital clock is overlaid on the video to indicate the game lapsed time. As the clock digits
change periodically, we can propose an approach by using this important temporal pattern to detect the game time. Firstly, we segment the static overlaid region using a static region detection approach. Then the edge feature’s changing is employed to describe the digits’ temporal pattern and locate the digits area. After digit position location, the digit templates are automatically captured from 0 to 9 in the clock digit region as binary images. At last, those digit templates are used to recognize the game time by template matching in the clock digit position. For reader can reference our previous work [12] for more technical details.

2) Event Boundary Detection: After detecting the event moment in the video, we need further detect the event boundary to identify the accurate event clips. In our observation, we found that due to the common production rule in sports video broadcasting, different broadcast sports videos usually have similar structure for an event. When an event occurs, the broadcast director generates the related broadcast video segment by smoothly concatenating the video feeds from different cameras and inserting the replays. A typical event structure in broadcast video can be represented as a temporal pattern.

   a) Event with replay: far view shot, close-up shots, replay, close-up shots, far view shot.
   b) Event without replay: far view shot, close-up shots/medium view shot, far view shot.

Hence, we can model these temporal transition patterns in sports video, and use it to detect the event start/end boundaries effectively. The technique details can be referred to [12].

C. Video Database Construct

After the accuracy event clip generated, we segment it from the original video and use the related semantics of the text events to annotate it. For each event clip, we extract text semantics, such as event time, event type, joined player, player's team, from the corresponding text event, and combine with the event start/end frame which obtained from the event boundary detection process, to annotate it. The text semantics are organized in following structure:

```xml
<EventTime><CurrentScore><EventType><JoinedPlayer>
<Player'sTeam><EventStartFrame-EventEndFrame>
```

Based on these annotations, we index all the video clips and build a video database. This video database contains event clips of the given soccer and basketball games. It is used for personalized retrieval. In our system, the personalized retrieval can be conducted based on the time, the event type, the joined player and the team according to the user's interests.

IV. CLIENT END USER INTERFACE DESIGN

A good user interface can facilitate finishing the task at hand and also without drawing any unnecessary attention of the user. In our system, a friendly graphic user interface (a browser) is designed on the client end to make the user’s interaction as simple and efficient as possible. By using the browser, mobile users can easily visit the website and submit their personalized customization requirements, including favorite matches, teams, event types, players and allowed viewing time. Meanwhile, the device's power capacity and memory size are all automatically submitted to the server. Once the server receive the user's request, the multimedia server will search through the whole database and find the optimal set of video segments. Fig. 2 illustrates a realistic example which is conducted on a Nokia N95 cell phone, throw this reader can better understand the operation flow which we introduced above.

V. PERSONALIZED CUSTOMIZATION AND SUMMARIZATION

Video customization and summarization is the key function of our system. In the client end of our system, the mobile user only need to visit the web-sit and submit their customization request, the multimedia server will provide as much relevant video content as possible to meet user's preference, automatically. However, due to the constrains of network and client end device capacity, the video content transmitted to client end is limited. In order to generate a optimal customization result for different users, we need to design an approach to rank the event clips in the database and consider the user's preference, the video content and its semantic significance synthetically, and well balance them automatically. In this paper we provide a general and robust event importance computation mechanism for the selected event clips by collaborating content and semantic analysis synthetically.

A. Event Significance Computation

Video customization result is a video abstract which not only includes the important event scenes of the original video streams but also should consider the user's preference or interests. In order to well balance both of them, we design an approach to collaborate the low-level features of event clips, the event occurrence time, the event intensity and the user's profile synthetically at the process of event significance computation.
1) Low-level Feature: There are various low-level features, such as audio, visual and textual features can be utilized for video content analysis. Due to the characteristic of computationally cheaper and good correlations with semantics of audio features, our approach uses audio features to compute the event’s importance. Based our observation, we found that when exciting events occur, the crowd’s cheer and commentator’s speech usually become louder, faster, and higher (in pitch) and less pauses occur. So, we calculate pitch value for each event clip to determining its interest degree.

To calculate pitch value of specified event clip, we applied the subharmonic-to-harmonic ratio-based pitch determination from Sun [14] for its reliability. Based on the pitch values, 

\[ S_d(E_i) = \frac{N_{hp}}{N_i} \alpha \]

where \( N_{hp} \) is the number of frames containing high-pitch speech in a event clip, \( N_i \) is the number of frames in a event clip, and \( \alpha (0 \leq \alpha \leq 1) \) is the adjustable parameter to consider how much the audio feature affects the significance of events.

2) Event Occurrence Time: In the basketball and football match, the score events which are close to the final stage of the game largely affect the game’s outcome, especially when the two teams tie or have slight score difference. Thus, such events are usually more attractive to users and of great significance. Then, 

\[ S_t(E_i) = 1 - \frac{N_j}{N} \beta \]

where \( N_j \) denotes the total event number of rank \( R_j \), the significance degree of rank \( R_j \) is defined as:

\[ S_j(R_j) = \frac{R_j - 1}{3} \]

where \( N \) denotes the total event number, \( i \) denotes the index number of event \( E_i \), and \( \beta (0 \leq \beta \leq 1) \) is the adjustable parameter to consider how much the occurrence time affects the significance of events.

3) Event Intensity: In this paper, the game we focused on is played between two teams, say team A and team B, and the team’s goal is to get more scores than its opponent. The game situation can be divided into three states: “two-team tie”, “A team leads”, and “B team leads”. Motivated by this, we can further rank the event into four levels based on the event's influence to the game state:

- Rank 1: goal state change event.
- Rank 2: score event but not change the game state.
- Rank 3: common foul events or events closely related to score events.
- Rank 4: all other events that are not Rank 1 to 3.

Then, \( S_r(R_i) \), the significance degree of rank \( R_i \) is defined as:

\[ S_r(R_i) = \frac{R_i}{4} \]

As we can see from Fig. 3(a), the text event, which we get from web-casting text, corresponding to a time-tag is always composed by a number of events. For example, the text event “83:00 Outswinging corner taken from the left by-line by Florent Malouda, Ricardo Carvalho takes a shot. Save made by Jose Reina.” is composed by three events, i.e. “corner, shot, save”. For a text event, it is evident that the more events it contains the intensifier it is. Based on this observation, we construct a text events’ rank network to compute the text event’s intensity, once the text event’s intensity computed, all the events it contains can be given the same intensity as itself.

In order to build the text events’ rank network, we treat a match as a bipartite graph, as shown in Fig. 3(a). The square node denotes rank level (1~4), and the circular denotes the time of a text event. The edge between the \( i_{th} \) circular node and the \( j_{th} \) square node represents that one text event accurred at time \( i_{th} \) can be ranked to \( j_{th} \) level. For a match that consists of \( N \) text events and 4 ranks as defined before, we can express the status of events occurrence by a matrix \( A = [a_{ij}]_{4 \times N} \), where the element

\[ a_{ij} = N_j S_i(R_j) \]

where \( N_j \) denotes the total event number of rank \( R_j \) at the time \( i_{th} \), \( S_i(R_j) \) denotes the significance degree of rank \( R_j \).

Figure 3. A graphical example to show how to compute the text events’ importance.
The matrix presentation of Fig. 3(a) is shown in Fig. 3(b). More specifically, the $i$th column vector, $a_i = [a_{i1}, a_{i2}, a_{i3}, a_{i4}]^T$, of $A$ denotes the different levels of significance degree of $i$th text event. Based on this matrix, the intensity of $i$th text event $E_i$ can be identified by:

$$s_{t_i}(E_i) = \sum_{j=1}^{4} a_{ij}$$  \hspace{1cm} (5)

4) **User’s Preference**: For the user’s request plays a main role in our personalized customization system, we strongly introduce an approach to effectively model how the user’s preference affects the event importance evaluation. Through increasing the significance degree of preference matched events, our system can finally tailor the appropriate video clip to the user. Based on the video annotation result, $s_u(E_i)$, the user’s preference based significance degree of a event $E_i$, is calculated as:

$$s_u(E_i) = \frac{\chi}{\lambda}$$  \hspace{1cm} (6)

where $\chi$ is the coefficient to consider how much user’s preferences affect the significance of events, and $\chi=1$ means that the user’s preference has no effect on the significance of the events, $N_u$ denotes the number of user’s request, which is composed of sets of a keyword, such as a favorite team, player, and event, and $N_{ri}$ denotes how many keywords in the annotation of event $E_i$ match user’s request.

As a consequence, the significance degree of an event $E_i$ can be formulated as:

$$S(E_i) = \lambda \cdot S_M(E_i) + (1 - \lambda) \cdot S_u(E_i)$$  \hspace{1cm} (7)

where, $S_M(E_i)$ is the semantic consistency to the user $s_u(E_i)$ . With the fusion parameter $\lambda$, we can treat the two modes of mobile-based sports video service, i.e. personalized customization and general summarization in a unified manner. When $\lambda$ is approximating to 0, event importance is mainly decided by user’s preference, in this case, only semantically consistent events can be assigned higher significance degree, thus a personalized customization result is generated. While in the other case of $\lambda$ approximating to 1, event’s significance degree is largely depends on the event content itself, and the selected events are mainly reflect the global situation of the match, hence, a general summarization can generate in this case.

B. **Mobile-based Personalized Customization**

In our system, once the user submit their customization request at the client end, the multimedia sever will provide as much relevant video content as possible to meet user's Preference. While, due to the constrains of network and client end device capacity, the video content transmitted to client end is limited. In order to present as much desired video content as possible within various resources constraints, we raised an user-participant multi-constraint 0/1 knapsack problem to formulate the mobile-based personalized customization. The technique details can be referred to [10].

VI. **Experiment**

In this section, we show some experimental results. The experiments are conducted on five Euro-Cup 2009 football matches and five NBA 2009 basketball matches. The corresponding web-casting texts are collected from BBC for football and ESPN for basketball.

A. **Video Analysis and Semantic Annotation**

In web-casting text analysis module, after LSA, unsupervised clustering and rank process, the documents are divided into 9 clusters for both football and basketball. The top rank word and its corresponding semantic event of each cluster for both football and basketball are listed in Tables I. From the result we can see that these 9 clusters contain most semantic events in both the football and basketball game, we can use the generated keywords to represent the semantic events and annotate video clips.

We conduct experiment to detect event boundary in football and basketball videos. Boundary detection accuracy (BDA) [12] is used to evaluate the detected event boundaries in the testing video set. BDA is defined as follows:

$$BDA = \frac{T_{db}(\cap_{mb} \cup_{mb})}{T_{db} \cup_{mb}}$$  \hspace{1cm} (8)

where $T_{db}$ and $T_{mb}$ are the automatically detected event boundary and the manually labeled event boundary, respectively. The event boundary detection results for five football matches and five basketball matches are shown in Tables II. It is observed that the overall result of BDA is promising, which lays a solid foundation for the personalized video customization. As for the lower BDA scores of “free kick” in football matches and “substitution” in basketball matches, it is mainly due to these types of events in the video do not have distinguishable temporal patterns for event boundary modeling.

B. **Personalized Customization and Summarization**

The values of the parameters were set as $\alpha = 1, \beta = 1$ and $\chi = 10$ so that audio feature, event occurrence time and user preference can fully affect the significance of events. As for the parameter $\lambda$ in (7), we initially set it to 0.5 denoting the equal preference between event contents and user preference. It can also be changed according to user’s request.

In order to evaluate the performance of our system, we invited 20 volunteers to participate the evaluation. These volunteers are all sports fans, first-time users of the system and very familiar to the operations on smart phones. After watching the integral sports matches, the volunteers are asked to use the mobile phone to customize their favorite video clips with the
users are satisfied with the system. The score is based on following scale (5-very good, 4-good, 3-common, 2-bad, 1-very bad). Table III shows the result of the evaluation. The average scores on the three criterions are 4.25, 4.55 and 4.2, respectively. This demonstrates that most users are satisfied with the system.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Football Event</th>
<th>Basketball Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Goal</td>
<td>Shot</td>
</tr>
<tr>
<td>2</td>
<td>Shot</td>
<td>Dunk</td>
</tr>
<tr>
<td>3</td>
<td>Save</td>
<td>Foul</td>
</tr>
<tr>
<td>4</td>
<td>Foul</td>
<td>Block</td>
</tr>
<tr>
<td>5</td>
<td>Card</td>
<td>Throw</td>
</tr>
<tr>
<td>6</td>
<td>Corner</td>
<td>Layup</td>
</tr>
<tr>
<td>7</td>
<td>Offside</td>
<td>Jumper</td>
</tr>
<tr>
<td>8</td>
<td>Kick</td>
<td>Rebound</td>
</tr>
<tr>
<td>9</td>
<td>Cross</td>
<td>Enter</td>
</tr>
</tbody>
</table>

VIII. CONCLUSION

In this paper, we proposed a novel framework for mobile-based personalized sports video customization and summarization. In this framework, sports videos are automatically annotated by the web-casting texts. Based on the annotation, highlight event clips of the match are generated from the whole game videos and indexed in a video database. And with the help of a novel event importance computation mechanism, our system can provides timely multimedia service to mobile users and meet different users' personalized preference. Experiments conducted on basketball and football matches validated the effectiveness of our approach. Our future work will focus on extending the proposed approach to other sports to provide more abundant visual enjoyment to mobile users, and further improving the performance of video semantic annotation.

ACKNOWLEDGMENT

I would like to thank Dr. Hanqing Lu and Jian Cheng and Chao Liang, Yifan Zhang for their help when I was an intern student in Institute of Automation, Chinese Academy of Sciences, Beijing, China.

REFERENCES