Personalized Retrieval of Sports Video

Yifan Zhang[1], Xiaoyu Zhang[1], Changsheng Xu[2], Hanqing Lu[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100080, China
{yfzhang, xyzhang, luhq}@nlpr.ia.ac.cn,

[2]Institute for Infocomm Research, Singapore 119613
xucs@i2r.a-star.edu.sg

ABSTRACT

There has been a growing demand for effective access to video information from media archives in recent years. Personalized video retrieval is one of the most challenging issues and has spurred a significant interest in many research communities. In this paper, a novel approach is proposed to achieve personalized retrieval of sports video, which includes two research tasks: semantic annotation of sports video and acquisition of user's preference. For semantic annotation, a multi-modal framework is employed to detect sports event and index the sports video content. Web-casting text, as external information, is utilized to detect semantic events in sport videos. The semantic concepts and keywords included in the web-casting text are extracted to annotate and index the sport event segments automatically. For user's preference acquisition, relevance feedback is applied to model user's preference and non-preference, and re-ranking is used to refine the results. First, the user is asked to label some video segments as desirable and undesirable. Then, we use these labels to infer the user's interesting points (e.g. the player, the event type, the team, etc.) by analysis of text keywords; the low-level video features are also adopted as a supplementary to reflect the user's preference. The overall new rank of the results is the combination of the user's high-level and low-level preference. Experiments conducted on real-world soccer game videos show that the proposed method has an encouraging performance.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstract methods, indexing methods*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval –*Retrieval model, Relevance feedback*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Semantic annotation, video indexing, video retrieval

1. INTRODUCTION

The need for content-based access to video information from database is growing due to the significant improvement in the video processing technology and availability of large storage systems. The variety of video services via new media channels such as online TV and mobile device has brought new consumer demand to personalized video retrieval. Sports video has wide view-ship and tremendous commercial potentials. The traditional one-to-many broadcast mode can not meet audience's various appetites. People may not be satisfied to view the whole game video or highlight collections generated by studio professionals. Their individual interesting points possibly focus on some specific players or certain type of events such as goal or free-kick in soccer and slam duck or shot in basketball. The ability of effective access to the most interesting contents from lengthy and voluminous sports video programs is required by the audience. The existing search techniques mainly rely on the textual metadata such as video title, author, date etc., which can not fulfill personalized demands. Hence, the source sports videos have to be annotated and indexed in a more particular scale and higher semantic level (e.g. which event happens, how the event is developed, and who is involved in the event, etc.) and the retrieval mechanism should be able to acquire and reflect users' preferences. In some video sharing and searching websites, video segments are labeled with content descriptions by users which would be labor-intensive and not always reliable. How to automatic annotate and index sports video content is a challenging problem. Furthermore, in common circumstances, users can search some specific events or players' video segments by keywords matching with the annotation labels. However, it sometimes can not achieve user's expectation and reflect user's preference. For example, one wants to collect Kaka's highlights in AC Milan's host red-black striped jersey; a soccer coach needs to view the far-view shots of free-kick to observe the players' positions. These kind features of video segments probably can not be described in their labels. It is essential to obtain visual similarities within text-based results for better retrieval precision. Therefore, it necessitates a new effective annotation and retrieval framework for personalized demands.

The fusion of visual and text retrieval methodology has been utilized in many content-based multimedia retrieval systems and proved to be effective. Donald *et al.* [1] reported on many fusion methods used in video shot retrieval. In [2], the search topics were first classified into several classes and query-class dependent weights were used for fusing results in a hierarchical mixture of expert framework. Hsu *et al.* [3] proposed a video shot re-ranking search via information bottleneck principle, which finds the optimal clustering that preserves the maximal mutual information between the search relevance and visual features in the initial text search results. These frameworks can not acquire and learn the user's searching preference. In addition, they did not propose a robust way to achieve automatic annotation and indexing of the video contents.

MIR'07, September 28–29, 2007, Augsburg, Bavaria, Germany.

In this paper, we present a framework and methodology for personalized retrieval in broadcast soccer video database. Web-casting text is utilized to detect semantic events in original video streams. The semantic concepts and keywords included in web-casting text are used to annotate and index the event segments automatically. After that, text-based retrieval in the video database can be conducted according to user's individual preference (e.g. the event type, the player's name, the team which the player belongs to, etc.). Based on the initial text research result, relevance feedback is applied to acquire user's preference. The feedbacks which can perform their detailed needs and interests are employed to learn a user preference model which concerns both the semantic labels' coherence and visual features' similarity. The overall new rank of the results is the combination of the user's high-level and low-level preferences.

The rest of the paper is organized as follows. Section 2 introduces the related work. The framework is described in Section 3. The technical details of semantic video annotation and acquisition of user's preference are presented in Section 4 and 5 respectively. Section 6 shows the whole mechanism of our personalized sports video retrieval system. Experimental results are reported in Section 7. We conclude the paper in Section 8.

2. RELATED WORK

There are two research tasks in our framework: semantic annotation of sports video and acquisition of user's preference for personalized retrieval, which are both challenging issues in information retrieval domain and have captured the attention of researchers in recent years.

2.1 Semantic video annotation

The existing sports video annotation approaches can be classified into two categories: using video content only and using external sources. Most of the previous annotation work using video content only is based on audio, visual and textual features directly extracted from video content itself. Rui *et al*. [4] detected the commentator's speech and ball-hit sounds from noisy and complex audio signals for extracting highlight of baseball videos. Visual features were used in [5] to analyze and summarize soccer videos. The framework included some video processing approaches such as dominant color region detection, referee detection and penalty-box detection. Textual features such as caption text overlaid on the video were utilized to annotate in soccer videos [6].

However, single modality based approaches which only use single stream are not able to fully characterize the contents in sports video. Hence, the integration of multi-modal analysis is developed to improve the robustness and accuracy. Audio/visual features were utilized for event detection in tennis [7], soccer [8] and basketball [9]; and audio/visual/textual features were utilized in baseball [10], basketball [11].

Nevertheless, both single-modality and multi-modality approaches heavily reply on audio/visual features directly extracted from the video itself. Due to the semantic gap between low-level features and high-level events as well as dynamic structures of different sports games, it is difficult to use these approaches to address following challenges: (1) ideal event detection accuracy; (2) extraction of high level semantics. Therefore, we propose to seek available external sources for help.

There are two external sources: close caption and web text, which are both text sources. Closed caption, which is a manually tagged transcript from speech to text and encoded into video signals, can be employed to indentify semantic event segments in sports video [12, 13]. Since closed caption is generated from speech to text directly, it has some redundant or irrelevant information to the games.

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 in sports videos. An approach was proposed in [14, 15] to utilize match report and game log obtained from web to assist event detection in soccer video. Xu *et al*. [16] used web-casting texts from common sports websites to detect and identify semantic events and achieved ideal accuracy.

2.2 Acquisition of user's preference

In multimedia retrieval, acquiring the user's preference is of great importance. If we know the user's preference, we are able to arrange the retrieval results according to this preference. However, in most cases, it is not easy to acquire the user's preference exactly; sometimes, even the user himself will find it hard to describe his preference very clearly. During the retrieval process, what we know is the submitted query, which is far from enough to reveal the user's real preference.

In the everyday use of the popular multimedia search engines, such as Google, Yahoo, Baidu, etc., the user submits the query using text, and then the search engine returns the relevant images or videos. This text-based retrieval [17] is widely used because of its convenience (inputting text is much more convenient than submitting a sample image or video), and its accuracy (the high-level semantic concept of text is more reliable than the low-level features). The problem with the text-based retrieval is that when the user is not satisfied with the returned results, he has to rack his brains to modify the current query; and if he cannot find a query which can better describe his preference, no improvement will be done to the results. Therefore in text-based retrieval, the system passively accepts the user's preference by the gradually refined query.

Relevance feedback [18] is a powerful approach to get the user's preference in multimedia retrieval [19, 20, 21]. It is first applied to content-based image retrieval (CBIR) [22] to bridge the gap between high-level semantic concepts and low-level image features. The main idea of relevance feedback is using human-computer interaction to "inform" the system of the user's preference and improve the performance of the retrieval system. During relevance feedback, the user gives his feedback by labeling some results as relevant or irrelevant, and then the system learns from the labels and updates itself. With relevance feedback, if the result is not satisfactory, no refinement work of the query is required from the user. The only thing he should do is to give simple judgment (relevant or irrelevant) of some results, which is much easier than figuring out a better query. The system will then use the labeled results as instruction from the user, and acquire his preference gradually. However, traditional relevance feedback confines only to low-level features. In other words, it uses only low-level features of the labeled results to infer the user's preference, which is in fact not effective enough. If annotation of the images or videos is available, the high-level semantic concept of the labeled results is also very informative to acquire the user's preference.

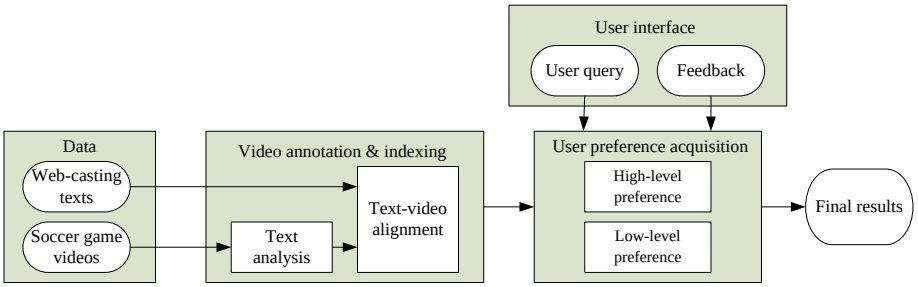

Figure 1. Framework of personalized sports video retrieval

3. FRAMEWORK

The framework (see Figure 1) of the proposed approach contains two major components: semantic annotation and acquisition of user's preference which are both aiming to personalized retrieval.

For semantic annotation task, it is divided into two modules. (1) web-casting text analysis. Due to the different presentation styles of web-casting texts, we cannot use prior knowledge to design a general heuristic approach for analysis. Hence, we employ Latent Semantic Analysis (LSA) [23] to cluster the texts and extract keywords in an unsupervised fashion. Then we detect events in the texts and formulate them with the corresponding keywords. (2) text/video alignment. Web-casting text lacks of time boundary information. In other word, it only contains the time point at which the event occurs. Accordingly, automatic synchronization of detected text events to the sports video is important. The events are firstly mapped into the original video stream by their time tag in the web-casting text. Then a statistical method based on Hidden Markov Model (HMM) is used to determine the exact event boundary. As a result, the event segments can be generated and annotated by the keywords for further retrieval.

For acquisition of user's preference, we use relevance feedback to refine the results. Different from the traditional relevance feedback, we combine the merits of both high-level semantic based feedback and low-level feature based feedback to solve the problem in personalized sports video retrieval. We analyze both the annotated high-level semantic label and the low-level video features, and model the user's high-level preference and low-level preference separately. Then, we combine the two parts of preference together and form the total preference of the user.

4. SEMANTIC VIDEO ANNOTATION

Most existing annotation systems based on audio and visual features extracted from the video are only effective in detecting events which have distinct characteristics and annotating them with simple semantic concepts. In order to achieve personalized retrieval, it is needed to extract high-level semantics (e.g. event type, how the event develops, who is involved, etc.) and use them to annotate and index the video segments. Hence, it motivates us to seek external information resources for help.

4.1 Web-casting text analysis

Web-casting text is a description of the sport game progress, which is available on many websites [24, 25]. An example of a web-casting text is given in Figure 2. Each item of description corresponds to an event, giving information on the time, the player, event type, and an optional remark. Exciting or important events which is essential to the course of the game will be recorded.

Figure 2. Web-casting text from ESPN.

From the texts, the players' and teams' name can be identified by Name Entity Recognition (NER) automatically. We also need to design an approach to extract the event type keywords from the corpus. Based on our observation, discarding of the different players' and teams' name, the descriptions of same type of events in the web-casting text have the similar sentence structure and word usage. Figure 3 gives several descriptions of the event "shot" in the game of Chelsea v.s. Valencia in 06-07 season European Champion League. Based on the same structure and textual feature, we can therefore cluster these descriptions into one category. In this category, it is easy to find that the word "shot" has the highest occurrence frequency and explicit semantic meaning among all the verbs and nouns. It can be selected as the representative keyword of this category.

9:13 Shot by Salomon Kalou (Chelsea) right-footed from left channel, hit bar.

11:02 Shot by David Villa (Valencia) left-footed from left side of penalty area, missed left.

21:38 Shot by Del Horno (Valencia) left-footed from left channel, save (caught) by Petr Cech (Chelsea).

45:00 Shot by Didier Drogba (Chelsea) right-footed from right side of penalty area, over the bar.

Figure 3. Example descriptions of web-casting text.

Inspired from this, we decide to employ an unsupervised approach to first cluster the descriptions into different categories corresponding to certain types of events and then extract keywords from the descriptions in each category for event detection. Here we use Latent Semantic analysis (LSA) to cluster events into different categories. LSA is a technique in natural language processing, in particular in vectorial semantics. It is assumed that there are underlying or latent structures in word usage corresponding to se-

mantic meanings of documents. LSA uses a term-document matrix which describes the occurrences of terms in documents; it is a sparse matrix whose rows correspond to documents and whose columns correspond to terms. The element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up-weighted to reflect their relative importance. The term-document matrix is then analyzed by singular value decomposition (SVD) to derive the particular latent semantic structure model. Finally, the cosine distance between vectors is computed to measure the document similarity in the reduced dimensional space.

Before applying LSA, we firstly pre-process the web-casting text corpus by a standard stop-word list. NER is then utilized to recognize the players' and teams' names. These are filtered out from the corpus and built a players' and teams' name list for retrieval. After the pre-process, we build the term-document matrix $A_{t\times d}$ by regarding each description as one document. The SVD is computed by decomposing the matrix $A_{t\times d}$ into the product of three matrices:

$$A_{t\times d} = T_{t\times n} \times S_{n\times n} \times D^T_{d\times n} \quad (1)$$

where *t* is the number of terms, *d* is the number of descriptions, *n* is the rank of A, *T* and *D* are orthonormal and *S* is diagonal. Only the first largest k (k < n) singular values and the corresponding columns from the *T* and *D* matrices are used to give the estimate matrix $\tilde{A}_{t\times d}$.

$$\tilde{A}_{t\times d} = T_{t\times k} \times S_{k\times k} \times D^T_{d\times k} \quad (2)$$

Each column in this reduced model is each document's textual feature vector $\vec{v}_i$. Cosine distance is used to measure the similarity between two documents:

$$S(\vec{v}_i, \vec{v}_j) = \frac{1}{k}\sum_{n=1}^{k} \left\| \vec{v}_i^{\,n} \vec{v}_j^{\,n} \right\| \quad (3)$$

where $\vec{v}_i^{\,n}$ and $\vec{v}_j^{\,n}$ are the *nth* component of $\vec{v}_i$ and $\vec{v}_j$ respectively. The document-to-document textual similarity matrix is calculated by:

$$\tilde{A}^T_{t\times d} \times \tilde{A}_{t\times d} \quad (4)$$

K-means approach is employed for clustering in our case. As it is conducted in an unsupervised fashion, we need to determine the optimal category number *N*. Thus, an evaluation function *F* is determined as follows:

$$F = \sum_{i=1}^{N} \sum_{\vec{v}\in C_i} S(\vec{v}, \vec{m}_i) \quad (5)$$

$$\vec{m}_i = \frac{1}{n_i} \sum_{\vec{v}\in C_i} \vec{v}$$

where C_i is the *ith* category, $\vec{m}_i$ is the center of C_i , n_i is the number of documents in C_i , *N* is the category number to determine. The value of *N* iterates in an empirical range (e.g. 1 ~ 20), as prior knowledge tells us there will be no more than 20 categories in the whole corpus. The optimal category number is determined while *F* reaches the maximum.

In each category, a rule-based part of speech tagger is utilized to tag the words of descriptions and recognize nouns and verbs. Finally, we rank the nouns and verbs by their *tf–idf* weights， where *tf* is the word occurrence frequency in each group and *idf* is the inverse word occurrence frequency in the entire document corpus. Several top rank words are set as the keywords of each category.

After proper keywords are extracted, the events in the web-casting text can be detected by searching the description which contains the keywords and analyzing context information in the description. The time-tag of the detected event is used for text/video alignment.

4.2 Text-Video Alignment

As the events in the web-casting text have been detected, we need to map them into the video stream. From the text, we obtain the time-tag which indicates the event occurring moment during the game. The game time in the video is also recognized. Then we use the time-tag to detect the exact frame which corresponds to the event time as the anchor frame. Finally, the event boundary is detected by a HMM-based approach.

4.2.1 Anchor frame detection

In broadcast sports video, the starting points of game time and video time are not synchronized due to non-game scenes such as player introduction, ceremony, half-time break. Hence, we need to recognize the exact game time in the video. In most broadcast sports video, a digital clock is overlaid on the video to indicate the game lapsed time. As the clock digits change periodically, we can locate and recognize the digits by using this important temporal pattern. We first use the static region detection method to segment the static clock overlaid region. Then the edge feature's changing is employed to describe the digits' temporal pattern and locate the digits area. The template of the 0~9 digits are captured after the SECOND digit area is located. At last, the game time is recognized by template matching:

$$Digit(i) = \arg\max_i \{ \sum_{(x,y)\in R} V(x,y) \odot T_i(x,y) \} \quad (6)$$

$$i = 0, 1, 2, \ldots, 9, 10$$

where $T_i(x,y)$ is the image pixel value in position *(x,y)* for the template of digit number *i*, $V(x,y)$ is the image pixel value in position *(x,y)* for the digit number to be recognized, *R* is the region of digit number, $\odot$ is EQV operator, i =10 corresponds to the region without any digit. On every frame, we recognize the game time, and therefore build a time-frame index. The anchor frame of a specific event can be determined by searching the time-frame index.

4.2.2 Event boundary detection

Based on the anchor frame, we will obtain a search range. Within the range, the exact event boundary will be detected. We use HMM to model the visual features' transition pattern to determine the boundary of the event portion in the video stream.

4.2.2.1 Visual feature extraction

As video is a continuous stream of multimedia information, we segment the whole video stream into shot as a suitable basic unit for content representation. The mean absolute difference algorithm of consecutive frames is utilized for shot boundary detection [16]. In broadcast sport video, the occurrence of events will cause shot transition. We classify the shots into three view type classes: far-view, medium-view, close-up. We want to use the transition pattern of the shot view type to character the feature of event in videos. In addition, replay is a special video effort to highlight the

event and extract audiences' attention. Thus, it is also a critical feature and visual cue to imply the ending of the event.

Shot classification is conducted by a majority voting of the frame view type. If one view type's frame is dominant in the shot, the shot is determined as this view type. Each frame is classified by color and edge features. Figure 4 illustrates examples of frames in far-view, medium-view and close-up.

(a) (b) (c)

Figure 4. Video frames (a) far-view frame (b) medium-view frame (c) close-up frame

Replay detection is relying on the flying logo matching method. In most broadcast sports videos, there are flying logo at the beginning and ending of a replay. (see Figure 5) We use template matching technique to detect the flying logos, and the shots between two successive logos are identified as replay shots. The detected replay/non-replay state of each shot is denoted by value 1 and 0 respectively and collected as a shot sequence.

Figure 5. Examples of flying logos in replay

4.2.2.2 Model fitting

After extracting visual features, we combine them into a vector for each shot. The shot feature sequences are used to train a HMM for each type of event. We search the event boundaries within a search range in the video stream by model fitting. The search range is empirically set to start from the first far-view shot before the anchor frame, and end at the first far-view shot after the anchor frame. The shot which contains the anchor frame is also called anchor shot.

Within the search range, the shots which are not much relevant to the event are regarded as noise and cannot be included in the event. We set all the possible partitions of shot sequences in the search range as candidate sequences and send them to the trained HMMs to calculate the probability scores. The partition which generates the highest probability score is selected as the detected event segments (see Eq.(7)) and the event boundaries can be obtained from the boundaries of the first and last shot of this event.

$$I = \arg\max_i P(S_i \mid Hmm) \tag{7}$$

where S_i is the *ith* possible partition of shot sequence in the search range. $P(S_i \mid Hmm)$ is the probability scores of S_i.

4.3 Summarization

Since the event segments from the original video data are generated, they each are annotated with the event type keywords and the player's and team's name in the corresponding description of web-casting text. A label which contains the information is recorded to tag with the video segment and an indexing is built in the video database. Hence, the text-based retrieval can be conducted based on the event type, the player's and the team's name according to the user's interests. However, the users sometimes can not describe their preference by text. It is therefore necessary to integrate the low-level visual features of videos together with text to modeling the users' preference.

5. USER'S PREFERENCE ACQUISITION

The label automatically tagged to each video segment is comprised of four items, which are player, player's team, opponent team and event type. We introduce the notation of *unit* to define the set of video segments in which all their four label items are exactly the same. Different units which share one or more identical items form a *cluster*. The relationship between video segment, unit and cluster is illustrated in Figure 6.

---

video segments: *s*1, *s*2,…, *s*13

---

*unit*1: (Beckham, England, France, goal) = {*s*1, *s*2, *s*3}

*unit*2: (Beckham, England, Argentina, foul) = {*s*4, *s*5}

*unit*3: (Beckham, England, Argentina, free kick) = {*s*6, *s*7, *s*8}

*unit*4: (Beckham, Real Madrid, Valencia, goal) = {*s*9, *s*10}

*unit*5: (Beckham, Manchester United, Arsenal, goal) = {*s*11}

*unit*6: (Zidane, France, Brazil, goal) = {*s*12, *s*13}

---

*cluster*1: (Beckham, *, *, goal) = {*unit*1, *unit*4, *unit*5}

*cluster*2: (Beckham, *, *, *) = {*unit*1, *unit*2, *unit*3, *unit*4, *unit*5}

*cluster*3: (*, *, *, goal) = {*unit*1, *unit*4, *unit*5, *unit*6}

Figure 6. Example of video segments, units and clusters.

5.1 Cluster Model

If the user's query is "unspecific", which means the query dose not cover all the items of the label, the results correspond to a cluster. For example, the query "Beckham" corresponds to cluster: (Beckham, *, *, *). Commonly in a huge video database, the cluster returned as results contains a large number of video segments. Therefore, it is important to acquire the user's preference and rank the video segments accordingly, so that the video segments in which the user is most interested can be arranged on top of the cluster.

We acquire the user's preference using relevance feedback. After the user has submitted the "unspecific" query, video segments of the corresponding cluster are randomly ranked, and the top-*t* of them are displayed on the first screen, where *t* is the number of video segments that can be displayed on one screen. If the results are not satisfactory, the user can label some video segments as desirable or undesirable according to his or her personal preference. Based on these labels, we re-rank all the video segments in the cluster.

For each video segment, its new rank after a round of relevance feedback is decided by two key factors, i.e. the user's preference on high-level semantic concept and low-level visual features. The former is essentially the rank of each unit in the cluster, while the latter can be seen as the rank of each video segment within the units. By combining the influence of both inter-unit and intra-unit

ranking, new results are achieved and returned to the user. The process of relevance feedback can be repeated until the user is satisfied with the results.

Now we discuss the modeling of user's preference in detail. We use $S = \{s_1, s_2, \ldots, s_n\}$ to denote the whole dataset of video segments. $q = \{q_1, q_2, \ldots, q_l\}$ is the user's query, and $\{item_1, item_2, \ldots, item_m\} = Q \bigcup U$ is all the items of a label, where Q is the items appear in the query, and U is the rest unmentioned items (obviously, $l \le m$). For example, if a user's query is "Kaka' goal", then we set q = {Kaka', goal}, Q = {player, event type}, U = {team, opponent team}. When the user's query is "unspecific", i.e. $l < m$, q corresponds to a cluster $C(q_1, q_2, \ldots, q_l, *, \ldots, *) = \{u_1, u_2, \ldots, u_N\}$, where $u_i (1 \le i \le N)$ is the unit in C. If relevance feedback is performed, we use P (positive) to denote the video segments which the user labeled as desirable, while N (negative) is the undesirable video segments.

5.1.1 Inter-unit ranking

It is common that when a user gives an incomplete query, he or she may place different implicit emphasis on the rest unmentioned items. For example, a user querying "Beckham" may prefer the video segments of Beckham's goal or free kick, while regardless which specific game the video segments belong to. In this case, the item "event type" seems to be more important than other unmentioned items. Thus, different importance of each unmentioned item can well reveal the user's personal preference.

In this paper, we quantify the importance of each unmentioned item according to its coherence in the labeled video segments. Based on the video segments which are labeled desirable and undesirable, we calculate the diversity of unmentioned items separately. Here the diversity is defined as the number of different values in a certain item. Take Figure 6. as an example again, suppose the results returned randomly for the query "Beckham" is {*s*1, *s*4, *s*5, *s*7, *s*10, *s*11}, and the user label {*s*1, *s*7, *s*10, *s*11} as desirable video segments, then the desirable diversity of the items "player's team", "opponent team" and "event type" are 3, 3 and 2. The less diverse an item is, the more coherent it is, and thus the more important it is. Therefore, we calculate the weight of each item as follows:

$$\begin{cases} w_i^+ = \dfrac{1}{Z^+ \cdot div_i^+} \\ w_i^- = \dfrac{1}{Z^- \cdot div_i^-} \end{cases} (l+1 \le i \le m), \tag{8}$$

where div_i^+ and div_i^- are the diversity of $item_i$, and Z^+ and Z^- are used for normalization:

$$\sum_{i=l+1}^{m} w_i^+ = \sum_{i=l+1}^{m} w_i^- = 1. \tag{9}$$

For each unit $u_j \in C$, we compare its unmentioned items with all the video segments that have been labeled ($s_l \in P \bigcup N$), and define its score as:

$$\begin{cases} score^+(u_j) = \max\limits_{s_l \in P} \sum\limits_{i=l+1}^{m} w_i^+ \delta\left(u_j(item_i), s_l(item_i)\right) \\ score^-(u_j) = 1 - \max\limits_{s_l \in N} \sum\limits_{i=l+1}^{m} w_i^- \delta\left(u_j(item_i), s_l(item_i)\right) \end{cases}, \tag{10}$$

where

$$\delta(x,y) = \begin{cases} 1 & , x = y \\ 0 & , x \ne y \end{cases}$$

From the equations above, we can see that $score^+(u_j)$ depicts the similarity between unit u_j and the desirable units, while $score^-(u_j)$ depicts the dissimilarity between u_j and the undesirable units.

Put $score^+$ and $score^-$ together, and we can get the inter-unit rank of each $u_j \in C$ ($1 \le j \le N$):

$$r^{\text{inter}}(u_j) = \frac{1}{2}\left[score^+(u_j) + score^-(u_j) \right] \tag{11}$$

The inter-unit ranking focuses on the high-level semantic concept of the user's preference. The larger the inter-unit rank of a unit, the more desirable it is.

5.1.2 Intra-unit ranking

Beside the high-level inter-unit ranking which ranks the units in cluster C, we use low-level video features to refine the ranks of the video segments within the units, since several features, such as color histogram, texture, the proportion of shot view type in a segment, etc., can also reflect the user's preference from various aspects.

From the machine learning point of view, the process of re-rank can be considered as a classification problem which classifies the video segments into desirable or undesirable. The video segments which have been labeled by the user can be used as the training data, and we can use them to train a classifier in the feature space. Since many features can be extracted from the video segments, we train classifiers out of each feature separately, and then combine them according to the importance of each feature.

Suppose there are K features altogether, then we have K corresponding classifiers: $f_1, f_2, \ldots, f_K$. To make the classifiers' classification results comparable to each other, we normalize them to [0, 1]. As a result, for each video segment $s_j \in C$, its intra-unit rank with the kth feature ($1 \le k \le K$) is:

$$r_k^{\text{intra}}(s_j) = \frac{f_k(s_j) - minf_k}{maxf_k - minf_k}, \tag{12}$$

where

$$\begin{cases} maxf_k = \max\limits_{s_j \in C} f_k(s_j) \\ minf_k = \min\limits_{s_j \in C} f_k(s_j) \end{cases}$$

Similarly, we can get each classifier's importance by analyzing the coherence of the labeled video segments' feature, which can be revealed by variance. A small variance of a feature means that this feature can, to some extent, reflect the user's like or dislike, hence we put emphasis on this feature. For the kth feature, we calculate the variance in both desirable and undesirable video segments separately, denoted by var_k^+ and var_k^-, and then set the

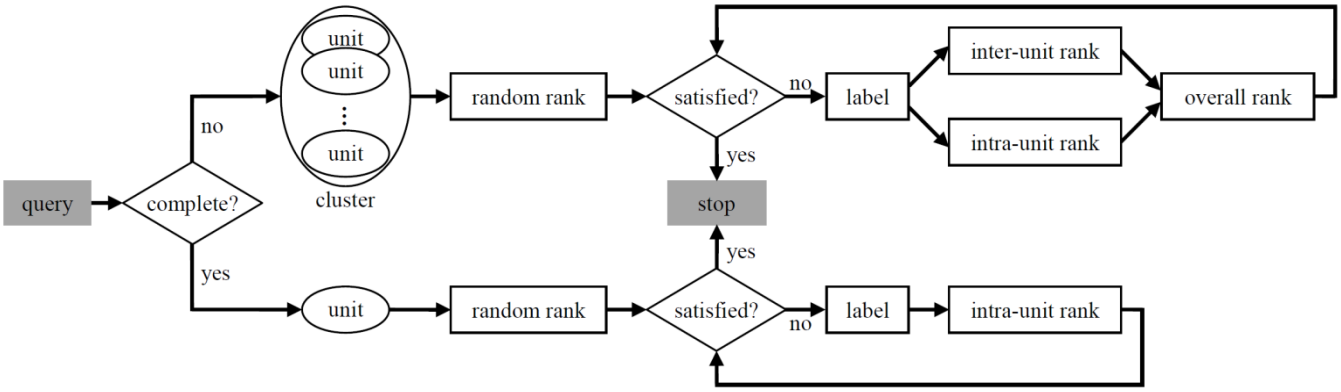

Figure 7. The flow chart of user's preference modeling.

weight of classifier f_k as:

$$w_k = \frac{1}{Y}\max\left(\frac{1}{var_k^+}, \frac{1}{var_k^-}\right), \tag{13}$$

where Y are used for normalization:

$$\sum_{k=1}^{K} w_k = 1. \tag{14}$$

Then, we calculate the intra-unit rank of each $s_j \in C$, which is the linear combination of the intra-unit ranks using each single feature with its corresponding weight:

$$r^{\text{intra}}(s_j) = \sum_{k=1}^{K} w_k r_k^{\text{intra}}(s_j) \tag{15}$$

The intra-unit ranking places emphasis on the low-level feature similarity between the unlabeled video segments and labeled ones, and can be seen as a refinement of the inter-unit ranking.

5.1.3 Overall ranking

Taking both inter-unit and intra-unit ranking into consideration, we calculate the new overall rank of each video segment.

As discussed above, the inter-unit ranking, which reflects the user's semantic preference, indicates the rank of units in the cluster, thus it plays a primary role in the overall ranking. In contrast, the intra-unit ranking is less influential and we use it as a supplementary to inter-unit ranking.

We calculate the overall rank for each $s_j \in C$ by combining inter-unit and intra-unit ranking as follows:

$$R(s_j) = 2^{r^{\text{inter}}(unit(s_j))} \cdot r^{\text{intra}}(s_j) \tag{16}$$

where $unit(s_j)$ is the unit which s_j belongs to.

The overall rank R quantifies the degree of the user's personal preference. A large R of a video segment represents a large probability that the user will be interested in it. Therefore we return the video segments with downward value of R to the user.

5.2 Unit Model

If the user's query is "specific", which means the query includes all the items of the label, the results correspond to a unit. For example, if the user submits the query "Beckham, England, France, goal", the unit: (Beckham, England, France, goal) is immediately returned.

For some small video databases in which a unit contains only a few video segments, the ranking of video segments is unnecessary. While for large video databases, relevance feedback is still indispensable.

Since in this case all the results are confined to the same unit, they are of the identical inter-unit rank. Hence only intra-unit ranking is necessary. We calculate the intra-unit rank for each $s_j \in C$ just as we did above in subsection 5.1.2, and then simply set it as the new overall rank:

$$R(s_j) = r^{intra}(s_j) \tag{17}$$

5.3 Summarization

After the new rank of video segments has been acquired, new results will be returned to the user. If the user is still not satisfied with the results, a new round of relevance feedback can be conducted to further improve the results.

The whole process of user's preference modeling with re-ranking is summarized in Figure 7.

Note that the re-ranking is totally based on the user's personal preference, thus different users may have different retrieval results, even if they have submitted the same query.

6. PERSONALIZED RETRIEVAL

In this section, we introduce the mechanism of our personalized sports video retrieval system.

In our sports video database, each video segment corresponds to a highlight in a game, with a series of text keywords (player, team, opponent team, event type) as its annotation. Every time a new game is to be added into the database, the video segments which correspond to the events in the web-casting text will be segmented and annotated by the text keywords. If the user is not satisfied with the automatic annotation of a certain video segment, he can manually modify it in a convenient way, and the new annotation will be adopted for the future use.

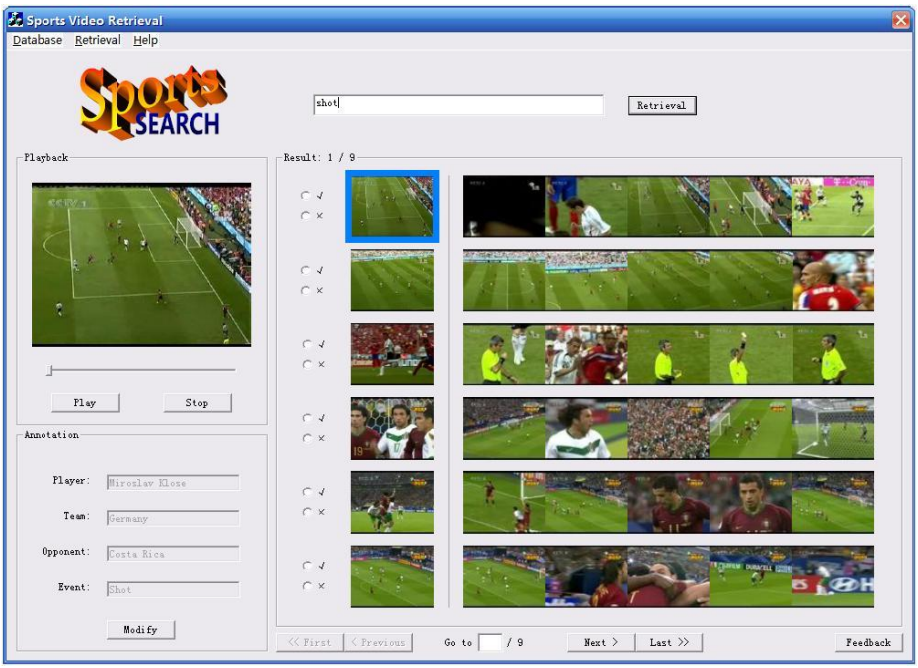

Figure 8. User interface of sports video retrieval system

The user interface (UI) of our sports video retrieval system is shown in Figure 8, which can be divided into three parts: the query region, the browsing region, and the playback region.

The query region is on the top of the UI, and it is for the user to input the query.

On the lower-right part of the UI is the browsing region. After the user has submitted a query, the system will return the relevant video segments and display them in this region. Every video segment is displayed using its first frame followed by five representative frames for the user to browse. In this region, the user can skim over the returned video segments, using the buttons "Previous" or "Next" to go to the previous or next page. If the user is not satisfied with the results, he or she can give the feedback by labeling some video segments as desirable or undesirable (√ or ×). After clicking the button "Feedback", the labels will be sent back to the system for user's preference modeling and results refinement.

The video segment selected by the user will be displayed in the playback region, which is on the lower-left part of the UI. The user can play, pause or stop the video segment using the buttons below the playback window. The keywords annotation (player, team, opponent team, event type) of the corresponding video segment is given in this region as well, which reveals the semantic concept of the video segment.

Using our personalized sports video retrieval system, the user does not need to give a very specific query. The system can acquire the user's preference by human-computer interaction, and refine the results accordingly.

7. EXPERIMENTS

The experiments are conducted on 5 World Cup 2006 soccer games, 3 European Championship 2004 soccer games and 3 European Champion League (2006-2007) soccer games. The videos are recorded from TV using Hauppauge PCI-150 TV capture card.

7.1 Semantic annotation

In video semantic annotation module, keyword generation experiment is conducted using web-casting texts from ESPN website[23]. After LSA and unsupervised clustering, the corpus of descriptions is divided into 10 categories. Here 10 is the optimal category number which is determined by the evaluation function in section 4.1. Then the nouns and verbs in each category are ranked by their *tf–idf* weights. The top 3 ranking words of each category are listed in Table 1. The last column in the table is the ground truth which is set as follows. We observe the clustering results. If the number of descriptions for one event type in one category is dominant (in our case, over 80% of all the descriptions), this event type will be set as ground truth for this category. It is used to evaluate whether the keywords generated by the system are representative for the categories and give explicit semantic meanings.

Table 1. Keywords generated from web-casting text

Category	Rank1	Rank2	Rank3	Event
1	goal	shot	target	**Goal**
2	shot	block	save	**Shot**
3	foul	kick	pass	**Foul**
4	card	foul	behavior	**Card**
5	corner	clearance	kick	**Corner**
6	offside	pass	foul	**Offside**
7	cross	block	shot	**Cross**
8	kick	foul	pass	**Free kick**
9	throw	defend	block	**Throw in**
10	substitution	tactic	block	**Substitution**

From the result we can see that these 10 categories are the most important semantic events in a soccer game and cover the main

content of the video. To make the approach more general, we ignore some low probability events such as penalty kick and handball or some unsporting behaviors like player conflicting. In fact, penalty kick belongs to the "goal" or "shot" event and handball belongs to "foul" event. These two events can both be retrieved by using the keywords of "goal" event or "foul" event respectively. It can be seen that the Top 2 ranking words in the table can be used as keywords to represent the semantic events and annotate video segments.

We conducted experiment for text/video alignment to detect event boundary in broadcast soccer videos. Boundary detection accuracy (BDA) is used to evaluate the detected event boundaries in the testing video set. BDA is defined as follows:

$$BDA = 1 - \frac{\alpha |t_{ms} - t_{ds}| + (1-\alpha)|t_{me} - t_{de}|}{\max\{(t_{de} - t_{ds}), (t_{me} - t_{ms})\}} \quad (18)$$

where t_{ds} and t_{de} are automatically detected start and end event boundaries respectively, t_{ms} and t_{me} are manually labeled start and end event boundaries respectively, α is a weight and set to 0.5 in our experiment.

Table 2. Soccer game event boundary detection

Event	BDA	Event	BDA
Goal	89.7%	shot	85.9%
foul	79.2%	card	77.3%
corner	84.8%	offside	79.0%
Cross	82.8%	Free-kick	67.2%
Throw in	81.4%	Substitution	60.5%

The event boundary detection result is shown in Table 2. The BDA scores of "free kick" and "substitution" are relatively lower than others. It is because they do not have distinguishable temporal patterns for event boundary modeling. The shot transition during a free kick event is very frequent and fast, while the substitution event always has loose structure and various temporal transition patterns.

To evaluate the keywords generated by our system, we use the top ranking one word of each category to search the corresponding events in web-casting texts. The result is shown in Table 3.

Table 3. Text event retrieval

Event	Precision/Recall	Event	Precision/Recall
goal	100%/100%	shot	83.2%/100%
foul	84.8%/98.5%	card	97.5%/95.6%
corner	100%/100%	offside	100%/100%
Cross	95.7%/97.9%	Free-kick	99.0%/98.5%
Throw in	100%/100%	Substitution	98.7%/99.1%

It can be seen that most events are well retrieved by their corresponding keywords. The penalty kick and handball events are also successfully retrieved in the results of "goal" and "foul" respectively. The precisions of "shot" and "foul" events' are low. This is because some "goal" events whose descriptions have the word "shot", thus these "goal" events are also retrieved by the keyword "shot". However, in common sense, users often do not revolt against the appearance of "goal" events in the results of "shot" query. The same situation is also occurred between "foul" and "card". Their descriptions share the same word "foul" too.

7.2 Personalized retrieval

In order to illustrate the effectiveness of our personalized sports video retrieval system, we ask 20 students from the institution to evaluate it. These students are all first-time users of the system. Each student is asked to submit 5 queries he or she desires, and use relevance feedback to refine the results. For each query, the student is supposed to write down how many rounds of relevance feedback it takes before satisfactory results are received. The feedback rounds of 100 queries (by 20 students) are shown in Table 4. From the votes, we can see that for most queries, one or two rounds of relevance feedback is enough to get the preferable results, which means our system is effective and efficient in acquiring the user's preference.

Table 4. The feedback rounds of 100 queries

Rounds of relevance feedback	1	2	3	4	⩾5
Votes	65	27	6	2	0

The students are also asked to evaluate the personalized sports video retrieval system by giving a numerical score based on the following scale: 1-Strongly dissatisfied, 2-Dissatisfied, 3-Neutral, 4-Satisfied, 5-Strongly satisfied. The evaluation scores are tabulated in Table 5, which indicates that most first-time users are satisfied with our system.

Table 5. The evaluation of 20 students

Evaluation score	1	2	3	4	5
Votes	0	0	1	7	12

8. CONCLUSION

Personalized retrieval is a challenging task in sports video analysis. In this paper, we have presented a novel framework for video semantic annotation and personalized retrieval. The major contributions include: (1) an unsupervised approach is proposed to extract keywords and annotate video content automatically. Video event segments are also generated and indexed in the database. (2) A relevance feedback mechanism is utilized to acquire users' preference and refine the retrieval results. Text knowledge and low-level visual features are fused while user preference modeling. Our future work will focus on extending the proposed approach for other sports domains, further improving the performance of semantic annotation and integrating more effective features and knowledge such as audio information to enhance the current modeling mechanism

9. REFERENCES

[1] K. M. Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," In *Proc. of Computer Vision and Image Understanding*, pp.61-70, Singapore, 2005

[2] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," In *Proc. of ACM Multimedia* 2004, pp 548–555, New York, NY, Oct. 2004.

[3] W. H. Hsu, L. Kennedy, and S. -F. Chang, "Video search reranking via information bottleneck principle," In *Proc. of ACM Multimedia* 2006, pp.22-27, Santa Barbara, USA, Oct. 2006

[4] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs", In *Proc. of ACM Multimedia*, Los Angeles, CA, pp. 105-115, 2000.

[5] A. Ekin, and M. Tekalp, "Automatic soccer video analysis and summarization," In *Proc. of IS&T/SPIE03*, Santa Clara, CA, Jan. 2003.

[6] J. Assfalg, M. Bertini, C. Colombo, A. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," In *Proc. of Computer Vision and Image Understanding*, Vol. 92, pp. 285–305, November 2003.

[7] M. Xu, L. Duan, C. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video", In *Proc. of IEEE International Conference on Acoustics, Speech, & Signal Processing*, Hong Kong, China, Vol.3, pp.189-192, 2003.

[8] K. Wan, C. Xu, "Efficient multimodal features for automatic soccer highlight generation", In *Proc. of International Conference on Pattern Recognition*, Cambridge, UK, Vol.3, pp.973-976, 23-26 Aug. 2004.

[9] M. Xu, L. Duan, C. Xu, M.S. Kankanhalli, and Q. Tian, "Event detection in basketball video using multi-modalities", In *Proc. of IEEE Pacific Rim Conference on Multimedia*, Singapore, Vol.3, pp.1526-1530, 15-18 Dec, 2003.

[10] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method", In *Proc. of ACM Multimedia*, pp.347-350, 2002.

[11] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos, In *Proc. of ACM Multimedia*, Ottawa, Canada, pp.261-269, 2001.

[12] N. Babaguchi, Y. Kawai, T. Ogura, T. Kitahashi, "Personalized Abstraction of Broadcasted American Football Video by Highlight Selection," In *Proc. of IEEE Trans. Multimedia*, 6:575--586,2004.

[13] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," In *Proc. of IEEE Trans. on Multimedia*, Vol. 4, pp. 68–75, March 2002.

[14] H. Xu and T. Chua, "The fusion of audio-visual features and external knowledge for event detection in team sports video," In *Proc. of Workshop on Multimedia Information Retrieval*, Oct 2004.

[15] H. Xu and T. Chua, "Fusion of multiple asynchronous information sources for event detection in soccer video", In *Proc. of IEEE ICME'05*, Amsterdam, Netherlands, pp.1242-1245, 2005.

[16] C. S. Xu, etc., "Live sports event detection based on broadcast video and web-casting text," In *Proc. of ACM Multimedia 2006*, CA, USA.

[17] Sivic, J., Zisserman, A. Video Google: a text retrieval approach to object matching in videos. In *Proc. of the 9th IEEE International Conference on Computer Vision*, 2003, 1470-1477.

[18] Rui, Y., Huang, T.S., Ortega, M., Mehrotra., S. "Relevance feedback: a power tool for interactive content-based image retrieval," In *Proc. of IEEE Transactions on Circuits and Systems for Video Technology* 8, 1998, 644–655

[19] Yoshitaka A., Ichikawa, T. "A survey on content-based retrieval for multimedia databases," In *Proc. of IEEE Transactions on Knowledge and Data Engineering*, 1999.

[20] Aigrain, P., Zhang, H., Petkovic, D. "Content-based representation and retrieval of visual media: A state-of-the-art review," In *Proc. of Multimedia Tools and Applications*, 1996

[21] Aslandogan, Y. A., Yu, C. T. "Techniques and systems for image and video retrieval," In *Proc. of IEEE Transactions on Knowledge and Data Engineering*, 1999.

[22] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R. "Content-based image retrieval at the end of the early years," In *Proc. of IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 2000, 1349–1380

[23] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, vol. 41, iss. 6, pp.391-407. 1990.

[24] http://soccernet.espn.go.com/

[25] http://uk.eurosport.yahoo.com/fo/matchcast.html