A Unified Framework for Semantic Shot Representation of Sports Video

Xiaofeng Tong[1], Qingshan Liu[1], Lingyu Duan[2], Hanqing Lu[1], Changsheng Xu[2], Qi Tian[2]

[1]National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, POBox 2728, Beijing, China 100080
Email: {xftong, qsliu, luhq}@nlpr.ia.ac.cn

[2]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Email: {lingyu, xucs, tian}@i2r.a-star.edu.sg

ABSTRACT
The development of mid-level shot description helps to bridge the gap between low-level feature and high-level semantics in video indexing and analysis. In this paper, we present a unified framework for semantic shot representation in field-ball sports genres, in which a video shot is characterized via three essential properties, namely, camera shot size, subject in a scene and video production technology. The three properties clearly represent the primary factors of a shot, and provide a unified viewpoint of semantic shot definition. Based on this framework, we design an effective architecture for semantic shot management comprising three main components as: 1) flexible shot clustering and retrieval by adjusting the weights of three properties according to different requirements; 2) semantics based video temporal segmentation for further event recognition; and 3) comprehensive sports video semantics analysis. Extensive experiments on soccer, basketball and tennis demonstrate the effectiveness and validity of this framework.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: clustering, search process. H.3.1 [**Content Analysis and Indexing**]: abstracting methods, indexing methods.

General Terms
Algorithms, design, experimentation.

Keywords
Mid-level shot representation, framework, video retrieval, and sports video.

1. INTRODUCTION
The enormous increase of multimedia information necessitates the development of content-based video indexing, retrieval and data management techniques [7,8]. In previous work, shot based analysis was a popular method for video indexing and retrieval, which was often performed by key frame matching with low-level feature similarity measurement. The direct mapping from low-level features to high-level semantics has been proved to be less effective or efficient in video retrieval. Although some strategies, such as region, concepts and objects relationship information, and statistical modeling, were used [9], the semantic gap still existed. There is a far distance between current technology and users' requirement. Presently semantic annotation for shots combining with video genre knowledge is a trend for video analysis [10, 11].

As a popular video genre, sports video attracts much attention for its wide viewer-ship and tremendous commercial potential. The three-level framework has been proved to be promising for sports video analysis [1]. At the low layer, low-level features, such as color, shape, motion, and texture, are directly extracted from raw video data. Semantics events are located at high layer. The mid-level descriptors including visual semantic shots and audio keywords play a critical role as bridging the gap between low-level features and high-level semantics. A semantic shot can be regarded as a basic shot adhering some clear meaning, which is an indication for event detection or context perception. Some researchers have applied semantic shots to semantic analysis, such as highlight extraction and event detection [2, 3, 4, 5, 6]. However, most of them defined the shot types according to domain-specific and application-driven information. The definition lacked a general viewpoint and unified explanation that could cover several sports genres, so it resulted in restriction of the ability for generic and comprehensive semantics analysis.

In this paper, we propose a unified semantic shot representation framework for sports video. A shot is described by three properties: camera shot size, subject in the scene, and video production technology, which can clearly and fully describe the composition of a shot as viewed from sports scenarios. This representation model can not only cover almost all the semantic shot types defined in existing work, but also can generate some refined useful types. Based on this framework, we develop an effective semantic shot management and analysis architecture, i.e., 1) shot clustering and retrieval based on adaptively adjusting the similarities of the three properties; 2) semantics based video temporal segmentation as considering the correspondence between play/break and shot types; and 3) comprehensive sports video semantics analysis. This framework can be applied in most field-ball sports, such as soccer, basketball, tennis, volleyball, table tennis, and so on.

MIR'05, November 10–11, 2005, Singapore.

The contributions of this work can be summarized as follows:

(1) We propose a unified shot representation model for field-ball sports programs. This model cannot only cover all the defined semantic shot types in the previous work, but also it can generate some new shot types for further analysis.

(2) We develop an effective semantic shot management and analysis architecture, such as clustering and retrieval; play-unit based video temporal segmentation, and semantic analysis.

The rest of this paper is organized as follows: the proposed framework is presented in Section 2. Semantic shot definition is described in Section 3. The detection algorithms of semantic shots are introduced in Section 4. Applications and experimental results are given in Section 5. Finally, we conclude our work in Section 6.

2. FRAMEWORK

The structure of our proposed model is illustrated in Figure 1. A shot in broadcast sports video is represented by three properties: camera shot size, subject in the scene and video production technology. A shot described by the meaningful factors is called a semantic shot. It can be defined by the following formula:

$$a\ semantic\ shot = \{shot\ size, subject, prod\text{-}tech.\}$$

Usually, the shot size set includes long shot, medium shot and close-up in sports programs. The subject set concerns the important object in scenes. Video production technology considers the post-production methods and editing rules for broadcast programs.

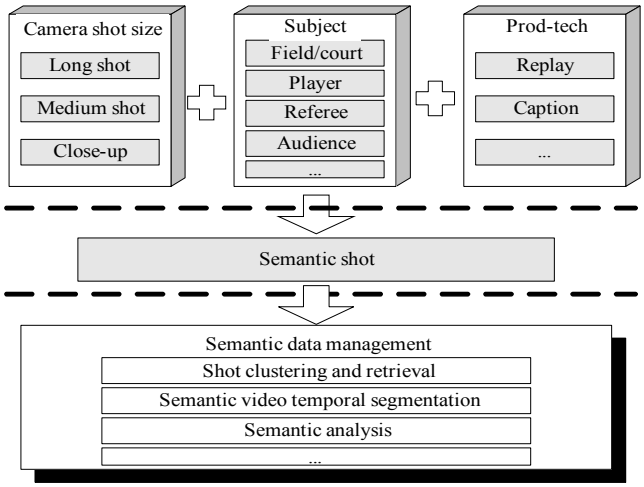


Figure 1. Structure of proposed framework

Based on this representation framework, we can apply semantic shots to semantic data management, such as shot clustering and retrieval; event based video temporal segmentation, and semantic analysis.

3. SEMANTIC SHOT DEFINITION

Three types of camera shot size are usually used in sports videos: long shot, medium shot and close-up. A long shot captures the global view of a scene. It is usually photographed with a wide-angle lens. A medium shot has less view coverage than a long shot. It is zoomed in to a specific part of a view. A close-up shot gives the details of a smaller part of a subject or view. Generally, an above-waist view of a person is captured in a close-up once the event of an excited offense or a foul happens.

Different subjects are concerned at different scenarios and in different sports videos. The appearance of subjects is usually associated with certain semantic meanings. For example~~s~~, in soccer video, a player close-up often appears after a shoot, an attack or a foul; a referee close-up is often presented with the occurrence of a severe foul; an excited audience view is usually an indicator of an interesting shoot, etc. Thus, further subject identification is useful for subtle event detection.

Broadcasting technologies (especially post-editing rules) are often utilized to make the program more attractive, which includes camera movement, replay scene, superimposed caption, the composition of shots, etc. We denote them as video production technology. For example, camera pan or tilting are often used to track player. Camera switching is applied to capture a scene from different angles. As a significant video editing technique, a replay usually highlights an interesting or important segment once or several times with a slow motion pattern. The scoreboard is superimposed on the screen when the score is changed. This is a caption technique. Composition of different shots and camera motion are utilized to express scenarios.

The value sets of the three factors may be different in different sports genre. For almost all field-ball sports programs, the value sets of camera shot size are the same: long shot, medium shot and close-up. Similarly, the broadcast video technologies of field-ball sports genres are the same. They include replay, caption, motion pattern, and so on. But subjects are related to the specific sports genre. For example, in soccer video, usually seven types of subjects are concerned: field, player, referee, goalkeeper, audience, goalmouth and goal-net. In basketball, court, player, referee, audience and basket are mostly interested. And in tennis, court, player, referee, and audience are particular attended. The value sets of shot size and production technology are shown in Table 1; and subjects in soccer, basketball and tennis are listed in Table 2.

Table 1. Value sets of shot size and prod-tech.

factor	soccer / basketball / tennis
shot size	long, medium, close-up
prod-tech.	replay, caption, camera-motion

Table 2. Value sets of concerned subjects

sports	value set of subject
soccer	player, referee, coach, field, goalmouth, audience, goal-net
basketball	player, referee, coach, court, basket, audience
tennis	player, referee, court, audience

With the framework and value sets, we can define some semantic shots for different sports genres. In this work, we take soccer, basketball and tennis as examples to evaluate our proposed framework. The labeling semantic shots in soccer, basketball and tennis are listed in Table 3 respectively. As an example, the shot types in soccer video are displayed in Figure 2.

Table 3. Some examples of semantic shots in soccer

Sports genre	Semantic shots
soccer	field-view, goal-view, goal-net, audience, other long view; player medium still, player medium motion, other medium view; player close-up, referee close-up, goalkeeper close-up, coach close-up, other close-up; replay, caption; others
basketball	long court-view, basket-view, audience, other long view; player medium, other medium view; player close-up, referee close-up, coach close-up, other close-up; replay, caption; others
tennis	court-view, audience, other long view; player medium, other medium view; player close-up, other close-up; replay, caption; others

Figure 2. Some semantic shots in soccer

A semantic shot is denoted by combination of the three properties with a three-dimension vector:

$$ss_i = <s_i, b_i, p_i>,$$

where ss_i is the i^{th} semantic shot, and s_i, b_i and p_i are its shot size, subject type and production technology respectively. Through the combination of these variables, we can generate the meaningful semantic shots to cover almost all the previous work. Certainly, domain-specific knowledge should be considered in defining semantic shots. Some examples of semantic shots and their definition under this framework in soccer video are listed in Table 3, where the symbol "*" means that the value is arbitrary in value domain.

Table 4. Some examples of semantic shots in soccer

No	Semantic shots	value of <si, oi, bi>
1	field-view	<long, field, still/motion>
2	audience	<long, audience, still/motion >
3	goal-net	<long, goal-net, still>
4	goal-view	<long, goalmouth, still>
5	medium-view	<medium, field, still/motion>
6	player close-up	<close-up, player, still/motion>
7	Referee	<close-up, referee, still/motion>
8	Goalkeeper	<close-up, goalkeeper, still/motion>
9	replay	<*, *, replay>
⋮	⋮	⋮

We can see that these semantic shots cover almost all the defined shots in previous work of soccer video analysis, i.e., replay, field-view, player following, goal view, player close, player medium view, audience and setting bird view [2]. In our framework, a replay can be represented by <*, *, replay>; a field-view = <long shot, field, still/motion>; a player following = <medium shot, player, motion>; a goal view = <long shot, goalmouth, still>; a player close-up = <close-up, player, still/motion>; a player medium view = <medium shot, player, still>; an audience view = <long shot, audience, still/motion>. In the same way, the other shot types in basketball, tennis, volleyball, table tennis can also be defined under this framework.

Moreover, in addition to the existing defined shots in previous work, some new semantic shots can also be generated. For example, more close-ups can be defined through subject identification. Different subjects usually appear at different scenes, and correspondingly different subjects indicate different scenarios. For example, in a shoot event, player close-up views are often displayed; but in a foul event, the referee close-up view probably comes forth besides player close-up. Once we define the "goalmouth" and "goal-net", we can deduce the corresponding semantic shots, which are useful for event analysis.

4. SEMANTIC SHOT DETECTION

4.1 Camera Shot Size Detection

In sports video, three major types of camera shot sizes are frequently used: long shot, medium shot and close-up. Generally, a long shot gives a full view of the scene's subject at a distance. It is photographed by a wide-angle lens and used to establish a setting, reveal the location, develop a mood, set the environment, or follow action. It provides lots of visual information, but does not focus on the detail, which results in small scale of objects within it. A medium shot, captured by a medium camera size, is a zoom-in view of a specific part of the play in sports video. It gives a complete view of the subject and assumes that the viewer already has an understanding of the setting and that they recognize the subject's location. Similar to the long shot, the medium shot can be used to connect scenes and to show

interactions. It is often used to re-establish the setting after a series of close-ups. In sports video, it is often used to follow players who compete intensely. Close-up shots are tight shots of the subject that focus the viewer's attention. They can show the details, and are excellent way of showing emotions and reactions. A close-up shot aims at a focal person who is the leading actor of current event or a person related to present scenario, so it often shows above waist view. Different subjects will be displayed in close-ups at different scenes. Different shot size is used in different case, so the shot size also indicates certain semantics. Generally speaking, the reliable shot size should be detected through object segmentation and scale estimation of the object. But accurate object segmentation is a classical difficult problem in both still image and moving image sequence. We thus roughly estimate the shot size by combination of objects scale and view appearance in this work. The flowchart of the camera shot size detection algorithm is illustrated in Figure 3. The corresponding procedure is described as follows:

(1) If the ratio of playing field (FR) is high, we can directly estimate the object size via in-field object extraction. The procedure consists of dominant color detection, playfield extraction and object segmentation [12]. The color of playfield can be taken as the dominant color of a video. An accumulated color histogram or a Gaussian Mixture Model (GMM) is often used to extract the dominant color. Non-playfield pixel segmentation and connect-component analysis are used to extract object within the playfield. Small object size indicates a field-view; larger size corresponds to a medium view; a close-up is declared by the largest object size derived from special head size detection.

(2) Otherwise, if most parts of a view are non-field, the size of an object is estimated through texture measurement in the view. The contrast of the gray-level co-occurrence matrix (C_GLCM) is used for texture feature [12]. A GLCM describes how frequently two pixels with some gray levels appear in a certain spatial local window separated by a certain distance and orientation.

Six example images in a soccer video are given in Figure 4.

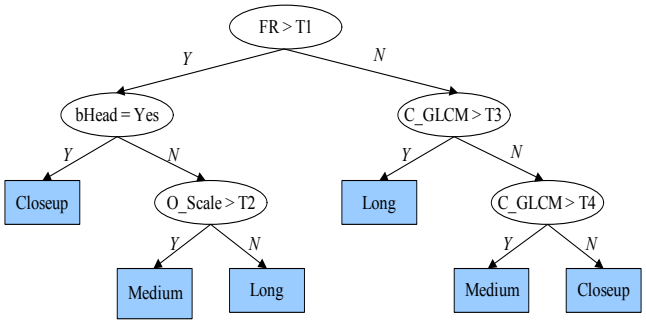


Figure 3. Decision tree for shot size detection

long | medium in field | close-up | long | medium out field | close-up

Figure 4. Examples of shot sizes in soccer video

4.2 Subject Identification

The subject in a shot is expressed by objects or interaction between objects and background. Actually, there is a certain relationship between a subject and the camera shot size. A long shot often emphasizes a background and the whole scene. Thus, the background is considered as the subject of a long shot. A medium shot shows interaction between objects. A close-up focuses on a person to show emotions and reactions, and we can discriminate different subjects (usually persons) through their color cues because of their big scale and enough color information. The concerned subjects in soccer, basketball and tennis are listed before. For the three sports genres, the subject identification methods are similar. In this section, we categorize the subjects into four classes and perform different identification procedure respectively.

(1) Playfield detection. Generally, we call the playfield of soccer as "field", and the playfield in basketball and tennis as "court". Dominant color and morphological filtering are usually used to extract the playfield region.

(2) Close-up person identification. In soccer, five classes of person are often displayed in close-ups: "a player" (two teams), "a referee" and "a goalkeeper" (two teams), the close-up subject class number $k = 5$. In basketball, "a player" (two teams) and "a referee" are often shown (class number $k = 3$), and in tennis, only two players frequently appear in close-ups ($k = 2$). Sometimes, it is hard to discriminate the two players in tennis because of their similar sportswear's color. But affirmatively, we can distinguish the player close-up from other close-ups. The above subjects are often shown in close-ups, which result in a number of samples for feature characterization. We firstly construct their above-waist color models and then use the models to identify the subject type. The color models are built up as follows: (a) Robust face detection is carried out to get the close-up views with the playfield as background. For these close-ups views, k classes' subjects usually appear. Their percentages of the restrained close-ups in the game of Brazil vs. England in 2002 World Cup are given in Figure 5. (b) A spatial relationship constraint within a face-body region [13] is used to locate the body region (above-waist). For each body region, mean shift based color characterization is performed to extract the color modes [14]. (c) k-means clustering is applied to construct the color models of these classes of subject. The clusters are sorted in a descendent order according to the size of cluster. For soccer, the first two big clusters correspond to two teams' players. The third one is "referee". The last two are goalkeepers. For basketball, the first two classes are players and the last one is referee. And for tennis, the two color models correspond to two teams' player. (d) After clustering, each subject's color is modeled as a GMM. With the GMM, we can perform pixel classification and post processing (morphological operations and region connection), and decide which subject appears.

player-A 38.1% | *player-B* 47.8% | *referee* 11.0% | *goalkeeper-A* 0.90% | *goalkeeper-B* 2.2%

Figure 5. Five classes of close-ups and their proportions

(3) Audience view detection. Here, the so-called "audience" is the audience view in a long shot. The reason has two folds: (a) Only "audience" in long shots are meaningful for sports video processing. (b) The "audience" in medium or close-up is difficult to be identified because of sparse samples for model construction. Statistical texture feature is enough for describing "audience" views.

(4) Other game-specific subject identification. In addition to the above general subjects, some game-specific subjects are also useful. In soccer [12, 13, 14], there are "goalmouth" and goal-net", and in basketball, there is a "basket". The "goalmouth" is a key indication for shoot events. In [15], a down-sample block segmentation and vertical line search were used to search the goalmouth under several experiential constraints. Actually, a simply playfield segmentation and field line slant angle estimation are enough for goalmouth view detection as viewed from event analysis, although the position cannot be accurately located. A "goal-net" view is captured by a camera placed at the back of the net. It is defined as the scene with the playfield background and uniform texture characteristics, which can be easily represented by edge histograms. In basketball, the "basket" view shows the scene that focuses on the penalty area. These views are often appear when free throw or offense at the penalty area. Similar to goalmouth detection in soccer, the detection of "basket" views can be performed through playfield extraction and the edge line slant angle estimation. Some subjects as discussed above are illustrated in Figure 6.

field in soccer | goal-net in soccer | basket in basketball | court in tennis | audience

Figure 6. Another subjects

4.3 Video Production Technology Detection

A replay scene may consist of several kinds of shots. If a shot belongs to a replay scene, it is called a replay shot in this paper. It is hard to discriminate a replay shot from a live shot without any contextual information. In this work, we combine an unsupervised logo detection method with a replay scene context recognition procedure to identify replay scenes. Firstly, we mine the logo template and detect all logos with template matching. Thus, a video is divided into segments with taking logos as boundaries. Then, for each segment, we examine its motion and shot transition context and use an SVM classifier to discriminate replay scenes. This method can accurately locate replay boundaries and robustly identify replay context [16].

Caption indicates content or compensatory information of scene in videos. We only consider manual-label caption that is superimposed upon the original video stream via posterior video edit. It contains the semantic description for current video content. Text in caption is treated as a special texture aligned by vertical strokes. The gradients of local neighbors in text region are greater and more uniform than those in other regions. Caption region is detected by local-accumulated gradient [17], which consists of gradient computation, run-length smoothing, morphological open operation, region segmentation and region verification. In practice, only the bottom of the screen needs to be examined in sports video.

As referring to motion in field-ball sports videos, the global motion feature is usually considered. It characterizes a dynamic scene, which indicates the status of a play. We use a simple block-matching method to extract the motion vectors and only consider pan and tilt motion patterns of the camera.

5. APPLICATIONS AND EXPERIMENTAL RESULTS

The framework not only supplies a general viewpoint for shot definition, but also serves for an effective semantic shot management. With this management architecture, we can perform some interesting applications, such as shot clustering and retrieval, semantic based temporal video segmentation, and typical semantics analysis. The three-factor definition model is a semantic representation for a shot. So it is more flexible, effective and proper for semantic shot operation.

5.1 Shot Clustering and Retrieval

5.1.1 Shot Retrieval Strategy

In previous work, the shot clustering or retrieval task was carried out on key frames, and the low-level features, such as color, texture, shape, etc. were usually used to measure the similarity between key frames. Thus, a so-called semantic gap exists in traditional image clustering and retrieval, which results in obstacle for semantic information indexing. It should be noted that the most contribution of the framework is the three-factor' representation manner for the application of shot retrieval. This provides a novel and effective data structure.

We parse a shot into three key and semantic properties. Correspondingly, the problem of shot clustering and retrieval can be performed on the three-factor structure. The similarity of two semantic shots is measured through evaluation on the pairs of the three factors.

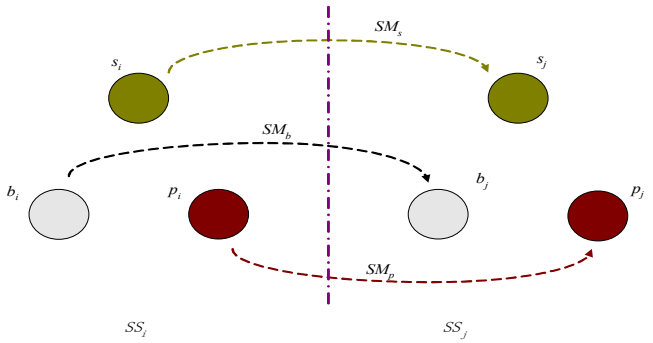


Figure 7. Illustration for semantic shot similarity measurement.

As illustrated in Figure 7, the similarity between two semantic shots is measured by combination of the three-factor's similarity measurement. The similarity between semantic shots ss_i and ss_j is defined as below:

$$Sim(ss_i, ss_j) = \omega_s SM_s(s_i, s_i) + \omega_b SM_b(b_i, b_j) + \omega_p SM_p(p_i, p_j)$$
$$\omega_s + \omega_b + \omega_p = 1$$

where SM_s, SM_b and SM_p are similarity measurements for shot size, subject in scene, and video production technology, respectively; and w_s, w_b, and w_p are their weights. s_i, s_j, b_i, b_j, p_i, p_j are shot components of shot ss_i and ss_j.

Compared to the previous work, the proposed method has the following advantages:

(1) For the framework itself, the three factors of a shot have certain semantics. Therefore, the task of shot clustering and retrieval can meet the demands for semantics or content-based video indexing and retrieval.

(2) The similarities of the three factors are individual measured, which leads to flexibility in semantic shot clustering and retrieval. Thus, the similarity can be respectively defined according to the domain-specific knowledge and measurement strategy.

(3) Each similarity item is weighted by a coefficient, which facilitates customized shot clustering according to different tasks by adjusting the weights.

5.1.2 Similarity Measurement

In this work, we use a simple binary measurement for each factor's similarity. If two instances of a factor are the same, the similarity is one, otherwise zero. The three weights for each item are equally set to be 1/3. That is, we regard the three factors equally. For example, the similarity matrix of "shot size" is listed below.

Table 5. Similarity matrix for "shot size"

	long	**medium**	**close-up**
long	1	0	0
medium	0	1	0
close-up	0	0	1

The similarity measurements for the other two factors are alike. In fact, the similarity and weight for each factor can be customized for different applications.

5.1.3 Experiments

The interface of semantic shot retrieval is displayed in Figure 8. It consists of several parts: play windows for query and retrieval shots (up-left); list of retrieval results (up-right); list of original all shots (bottom-left), and operation panel (bottom-right). The similarity of each shot is also given. The retrieval results can be ranked by their similarities. The higher similarity between a shot and the query data, the former position it settles.

A semantic gap usually exists between low-level features and high-level semantics. The retrieval with low-level feature usually results in pseudo resemblance i.e., similar in low-level feature, but total difference in semantic level.

Shot retrieval results under this framework can be ranked with an ascending sort of similarities. The most similar data are listed at the former, and the shots with much distance are shown as the last. A user can adjust the weights of the three factors according to different requirement and concerned item.

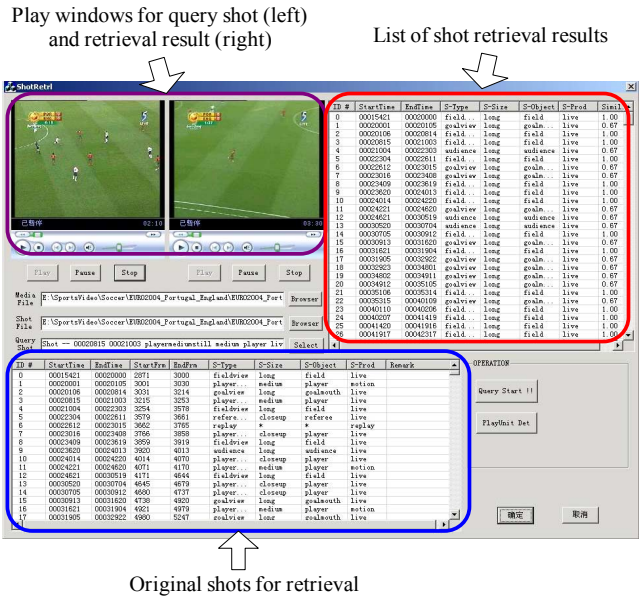


Figure 8. Interface of semantic shot retrieval

5.2 Semantic Video Temporal Segmentation

We can perform video temporal segmentation based on semantic shot as considering the correspondence between semantics and shot types. We define a semantic temporal segment notion "play-unit" as view from sports analysis. A play-unit is composed by consecutive play scenes and break scenes induced by the previous play (for examples, replay, close-ups). Intuitively, we can category the above shot types into two basic classes: play or break. These attributes can be used in play-unit segmentation.

A play-unit can be taken as an individual segment that usually contains key indication cues of semantics. Actually, the task of event detection contains both approximate video temporal segmentation and semantic analysis. But a prior semantic temporal segmentation can further facilitate video analysis. Suitable temporal play segmentation should: 1) reflect the appropriate inherent structure of video data; 2) be consistent with human understanding for sports; 3) facilitate video indexing and retrieval [18]. Previous work focused on play/break structure analysis, but actually the play/break itself has no much meaning for semantics analysis.

Based on semantic shots and play-units, we can create a hierarchical structure of video temporal segmentation (See Figure 9). At the low level, there are shot factors mentioned above. Semantic shots are generated at the middle level. With some criterion, we can group some consecutive semantic shots into a play-unit. The hierarchical structure can be served for video browsing.

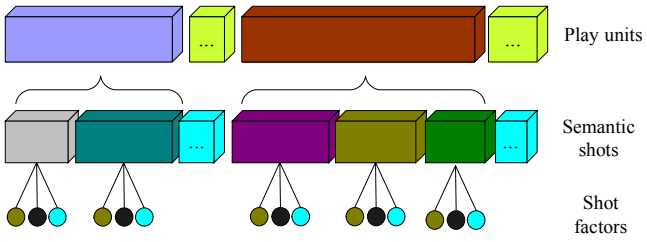


Figure 9. Play-unit video temporal segmentation

In soccer domain, a video is first parsed into shots, and the shots are classified under the proposed framework. A long field-view and medium field-view shots correspond to play process, and the others belong to break. A play-unit starts with a long field-view shot, and ends with a break shot induced by the play. Successive play shots are merged into one, but this strategy does not applied for consecutive break shots. The detailed criterion is listed in the following:

1) Check two adjacent long field-view shots with break shots between them.

2) A play-unit begins with a long field-view shot.

2.1) If there is a replay between the two adjacent long field-views, go to 2.1.1); otherwise go to 2.2):

2.1.1) If the shot is not a close-up followed the replay, the play-unit ends with the replay; else go to 2.1.2).

2.1.2) If it is a player close-up, the play-unit ends with the close-up.

2.2) If there is a player close-up or a referee close-up followed by the last play shot, go to 2.2.1), otherwise go to 2.2.2).

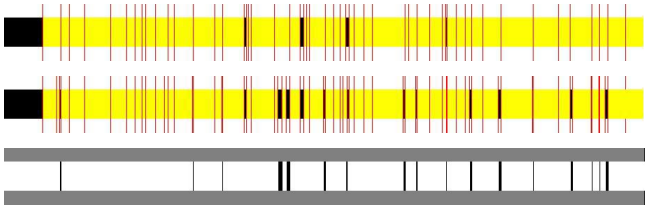
2.2.1) If the following shot is play close-up or referrer close-up, the play-unit ends with this close-up.

2.2.2) The play-unit ends with this break shot.

3) The segment containing the rest shots from end of the play-unit to the next long field-view shot is taken as a break-unit.

4) A play-unit with short long field-view is merged into the latter one.

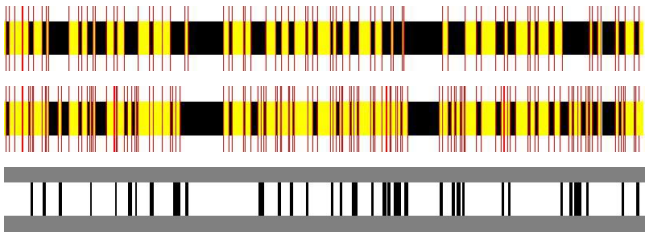
Figure 10 shows the play-unit segmentation results of the game between Portugal and England in EURO2004. The play- and break-units are listed and assigned by different color along a timeline. The ground truth and detection results are shown at the top and second lines in Figure 10. Play- and break-units are displayed as yellow and black bars respectively. A vertical red line is inserted to separate two adjacent units. The comparison result, i.e., difference between truth and detected data, is listed at the third line. The same part is noted as white, and the difference is represented by the black bar. The performances evaluated by precision and recall are given at the bottom.



Detect	Miss	False	Precision	Recall
40173	2036	107	0.9973	0.9518

Figure 10. Play- and break-units in a soccer clip. Top: ground truth; Middle: detection result; Bottom: comparison result.

The similar operation can be carried out for basketball and tennis, in which the domain-specific and video production rules are incorporated. In basketball and tennis, a long field/court view corresponds to play, and the other shot types belong to break. An experimental result of play-unit segmentation for a tennis clip is shown in Figure 11. The labeling method in this figure is the same as above. This 30 minutes length clip is cut from the game between Federer and Safin in Australia Open 2004.



Detect	Miss	False	Precision	Recall
18042	266	6879	0.7240	0.9801

Figure 11. Play- and break-units in a tennis clip. Top: ground truth; Middle: detection result; Bottom: comparison result.

5.3 Semantic Analysis

The semantic shot representation framework can also be used for semantic analysis and event detection, as described in previous work [1, 12, 16]. In [1], eight types of shots were defined in soccer video: replay, field-view, player following, goal view, player close, player medium view, audience, and setting bird view; seven types in basketball: replay, full court view, penalty view, player close, player medium view, audience and setting bird view; six types in tennis video: replay, court view, player close, audience, player medium view and setting long view. Based on a reasonable mid-level representation framework, we used these semantic shots, audio cues, and incorporated with domain knowledge to detected eight semantic events in soccer: foul or offside, free kick, penalty kick, corner kick, shot, goal, in play and out-of play; and eleven events: game, deuce, point, serve, reserve, return, ace, fault, double fault, take the net, and rally. In another work [19], we combined video and web broadcast text (WBT) to replay scene classification. WBT is a textual record of live commentary on sports game in Web. It is widely available from

web sites and provides more details with more refined granularity than match report. For example, the players' name, action, substitution, etc. can be got from WBT. Actually, visual information is important to multimedia data and can provide special information for semantic analysis. In [19], we used mid-level visual shot descriptors and WBT key words to classify soccer replay scenes into seven categories: 1) goal replay (GR), 2) shoot replay (SR), 3) attack replay (AR), 4) foul replay (FR), 5) offside replay (OR), 6) out of bound (OBR), and 7) others (OTR). The experiments were tested on seven full matches and an overall accuracy of 79.9% was achieved.

Generally, the more refined shot types are defined, the more semantic events can be mined from sports video. Semantics based video browsing or highlight indexing is the most frequent request for a common user.

6. CONCLUSIONS

In this paper, we proposed a uniform framework for semantic shot representation, and designed an effective architecture for semantic shot management from three aspects: 1) flexible shot clustering and retrieval; 2) semantics based video temporal segmentation; and 3) comprehensive sports video semantics analysis.

Compared with previous work, this framework provided a clear and general viewpoint for visual shot definition and representation, which facilitated semantic data management for sports video.

Base on this framework, we do semantic shot retrieval based on the three factors' individual similarity and their combination. The three-factor representation results in flexible and various objection of shot retrieval according to different requests. Semantics based video temporal segmentation and further semantic analysis can be also performed under this framework.

7. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 60475010 and 60121302) and International Joint Project between China and Singapore.

8. REFERENCES

[1] L. Duan, M. Xu, T. Chua, Q. Tian, C. Xu, "A mid-level representation framework for semantic sports video analysis", In *Proc. of ACM MM 2003*, pp. 33-44, 2002.

[2] L. Duan, M. Xu, and Q. Tian, "Semantic shot classification in sports video", In *Proc. of SPIE Storage and Retrieval for Media Database 2003*, pp. 300-313, 2003.

[3] R. Dahyot, N. Rea, and A. Kokaram, "Sports video shot segmentation and classification", In *Proc. of Visual Communications and Image Processing*, pp.8-11, 2003.

[4] J. Wang, E. Chng, and C. Xu, "Soccer replay detection using scene transition structure analysis", In *Proc. of ICASSP*, pp. 433-437, 2005.

[5] J. Assfalg, M. Bertini, C. Colombo, and A. bimbo" Semantic annotation of sports video", *IEEE Multimedia*, 9(2): 52-60, 2002.

[6] A. Ekin, A. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization", *IEEE Trans. Image Processing*, 12(7), pp. 796-807, 2003.

[7] A. Smeulders, M.Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years", *IEEE Trans. PAMI*., 22(12), pp.1349-1380, 2000.

[8] S. Smoliar, and H. Zhang, "Content-based video indexing and retrieval", *IEEE Multimedia*, 2(1), pp.63-75, 1994.

[9] M. Naphade, and T. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval", *IEEE Trans. Multimedia*, 3(1), pp. 141-151, 2001.

[10] I. Ide, K. Yamamoto, and H. Tanaka, "Automatic video indexing based on shot classification", In *Proc. of First Int. Conf. Advanced Multimedia Content Processing*, pp.87-102, 1999.

[11] C. Snoek, M. Worring, "Multimedia event based video indexing using time intervals", IEEE Trans.Multimedia, 2005 (in press).

[12] X. Tong, L. Duan, H. Lu, C. Xu, Q.Tian, and J. Jin, "A mid-level visual concept generation framework for sports analysis", In *Proc. of ICME 2005*.

[13] X. Tong, H. Lu, and Q. Liu, "An effective and fast soccer ball detection and tracking method", In *Proc. of ICPR 2004*, pp.795-798, 2004.

[14] X. Yu, C. Xu, H. Leong, Q. Tian and K. Wan, "Trajectory-based ball detection and tracking with applications to semantics analysis of broadcast soccer video", In *Proc. of ACM Multimedia 2003*, pp. 11-20, 2003.

[15] K. Wan, X. Yan, X. Yu, C. Xu, "Real-time goal-mouth detection in MPEG soccer video", In *Proc. of ACM Multimedia 2003,* pp. 311-314, 2003.

[16] X. Tong, H. Lu, Q. Liu, H. Jin, "Replay detection in broadcasting sports videos", In *Proc. of Intl' Conf. On Image and Graphics*, Hong Kong, pp. 337-340, Dec, 2004.

[17] X. Tong, Q. Liu, and H.Lu, "Semantic units based events detection in soccer video", In *Proc. of ICIP 2004*.

[18] L. Wang, M. Lew, and G.Xu, "Offense based temporal segmentation for event detection", In *Proc. of ACM Workshop on Multimedia Information Retrieval 2004*, pp. 259-266, 2004.

[19] J. Dai, L. Duan, X. Tong, C. Xu, Q. Tian, H. Lu, J. Jin, "Replay scene classification in soccer video using web broadcast text", In *Proc. of ICME*, Netherlands, July, 2005.