A Mid-level Representation Framework for Semantic Sports Video Analysis

Ling-Yu Duan[1], Min Xu [1], Tat-Seng Chua[2], Qi Tian[1], Chang-Sheng Xu[1]
[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Email: {lingyu, xumin, tian, xucs}@i2r.a-star.edu.sg
[2]Department of Computer Science, School of Computing,
National University of Singapore, Kent Ridge, Singapore 117543
Email: chuats@comp.nus.edu.sg

ABSTRACT
Sports video has been widely studied due to its tremendous commercial potentials. Despite encouraging results from various specific sports games, it is almost impossible to extend a system for a new sports game because they usually employ different sets of low-level features appropriate for the specific games and closely coupled with the use of game specific rules to detect events or highlights. There is a lack of internal representation and structure to be generic and applicable for many different sports. In this paper, we present a generic mid-level representation framework for semantic sports video analysis. The mid-level representation layer is introduced between the low-level audio-visual processing and high-level semantic analysis. It allows us to separate sports specific knowledge and rules from the low-level and mid-level feature extraction. This makes sports video analysis more efficient, effective, and less ad-hoc for various types of sports. To achieve robustness of the low-level feature analysis, a non-parametric clustering, mean shift procedure, has been successfully applied to both color and motion analysis. The proposed framework has been tested for five field-ball type sports covering duration of about 8 hours. Experiments have shown its robust performance in semantic analysis and event detection. We believe that the proposed mid-level representation framework can be used for event detection, highlight extraction, summarization and personalization of many types of sports video.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *abstracting methods, indexing methods.*

I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement – *Feature representation.*

General Terms
Algorithms, design, experimentation.

Keywords
Mid-level representation, semantics, events, sports video.

1. INTRODUCTION
The extensive amount of multimedia information available necessitates the development of content-based video indexing and retrieval techniques. Since humans tend to use high-level semantic concepts when querying and browsing multimedia databases, it is critical to develop techniques for semantic video analysis. Despite the significant progress in automated feature-based and structure-based indexing and retrieval, current users' expectations still far exceed the functionalities of today's computing systems,. The solutions currently available have one major drawback, viz. the generic low-level content metadata available from automated processing deals only with representing perceived content, but not their semantics. Thus, current research effort is geared towards modeling and extracting media-intrinsic as well as media-extrinsic semantics [1].

As an important video domain, sports video has been widely studied due to its tremendous commercial potentials [2-17]. The content of a video is intrinsically multimodal, since its creator uses visual, auditory, and textural channels to convey meaning. Many researchers have studied the respective roles of visual [4, 7, 9, 10, 11, 13, 14, 16, 17], auditory [2, 3, 15], and textural [5, 8] modalities in the sports video analysis.

Recently, the integrated use of different information sources is an emerging trend in video indexing research [2, 6, 8, 12]. Nepal *et al.* [12] employed heuristic rules to combine crowd cheer (auditory), score display (textural), and change in motion direction (visual) for detecting 'Goal' segment in basketball videos. Han *et al.* [6] used a maximum entropy method to integrate image, audio, and speech cues to detect and classify highlights from baseball video.

Despite numerous efforts in semantic sports video analysis, it is hard to develop a generic approach to sports video analysis. Currently most works focus on specific sports games in order to investigate the roles of different information sources or statistical learning algorithms in structure analysis and semantics extraction. Although it is possible to achieve promising results on limited dataset by adopting an advanced learning approach or strong domain rules, it is hard to extend the approach developed for one kind of sports game to another, and even to the same kind of game but for different matches. The main challenge lies in the amount of variation in low-level visual and auditory features, and game-specific rules.

In this paper, we present a generic mid-level representation framework for semantic sports video analysis. As shown in Figure

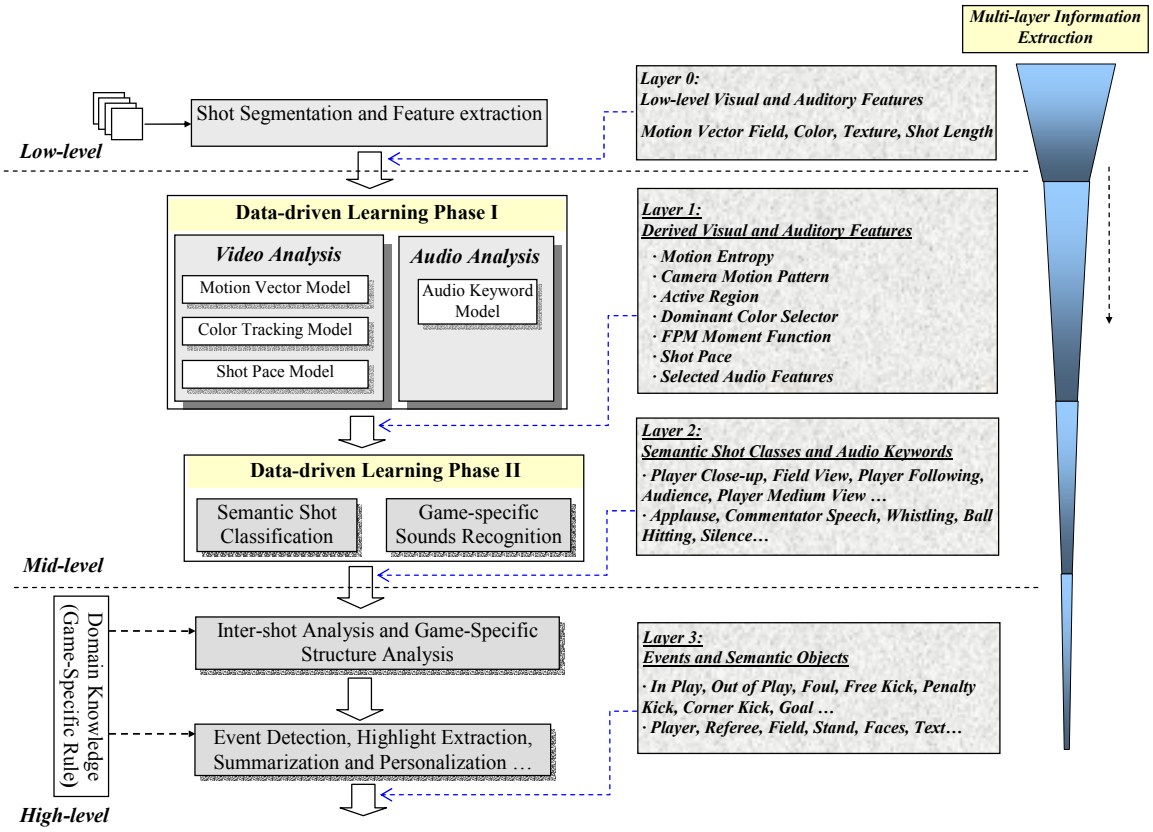

Figure 1. A mid-level representation framework for semantic sports video analysis

1, we establish a so-called mid-level representation layer between low-level audio-visual processing and high-level semantic analysis. This framework partitions sports video analysis into 4 layers. Layer 0 is to complete low-level audio-visual feature extraction and partition a video sequence into shots. Layer 1 and Layer 2 form the mid-level representation layers. They deal mainly with the engineering of knowledge in a data-driven learning way, with emphasis on the robust representation of prime concepts of semantic importance to the tasks among a sports sub-genre. Since this representation layer is constructed by generic learning approaches, it is feasible to make them reusable. Once we are equipped with robust and reusable mid-level representations, it is straightforward to design customized solutions in an elegant and flexible way, by combining with game-specific rules at Layer 3. We employ the semantic shot classes and audio keywords to represent the sports video production rules and game-specific rules, thus producing a high-level semantic analysis scheme at Layer 3 to detect events. From the perspective of machine learning scheme, the introduction of mid-level representations divides the learning procedure [6, 11] into several phases. We argue such decomposed learning phases should bring about more flexibility and robustness in the design of complex media processing systems.

We summarize the main contributions of this work as follows:

- An effective mid-level representation framework for semantic sports video analysis.
- A set of novel mid-level features extraction approaches.
- A general temporal model for inter-shot analysis and game-specific structure analysis.
- The use of mid-level framework to perform accurate event detection in tennis video and soccer video.

The rest of this paper is organized as follows. In Section 2, we present the mid-level representation framework. In Section 3 we explain the audio-visual mid-level representations. In Section 4, we introduce the high-level analysis scheme based on the mid-level representations. To evaluate the effectiveness of this framework, we propose two event detection applications in tennis video and soccer video in Section 5. Finally, we conclude this paper in Section 6.

2. FRAMEWORK

Unlike movie stories, TV sports program exhibits limited and compact field production techniques since most athletic events take place in a confined area with specific dimensions. A game consists of recurrent actions accompanied by score of a competitive event. In order to make the story more visually interesting, most sports photograph exhibits various views from multiple angles and composition. The sports photographers tend to follow exciting actions with a tight shot rather than a wide shot. The wide shot provides a full view of the actions to reestablish a sense of the setting. Hence the shot classification and further intra- or inter-shot analysis is an effective approach towards high-level sports video content analysis.

Below we explain this mid-level representation framework by the multi-layer information extraction flow.

2.1 Low-level Processing

This level is to complete a common first step in video analysis, namely, shot segmentation and low-level feature extraction. Various shot-boundary detection algorithms have been proposed [18, 19, 34]. The low-level audio-visual features can be easily extracted from video data in (un)compressed domain. Video shots contain a large amount of important information. However, limitations in computer power prevent us from using video information to its fullest extent. We have to perform pre-processing to reduce the dimensionality of the input video data. At this level, we take the simplest way to reduce the dimensionality, i.e. discarding a subset of the original input data. These low-level features include motion vector field, texture map, DC images [34], sub-sampled image frames, shot length, etc.

2.2 Mid-level Representation

As shown in Figure 1, two data-driven learning phases are carried out to produce a set of audio-visual mid-level representations.

In Phase I, we employ non-parametric clustering techniques to develop a motion vector model and a color tracking model. These models are used to extract five important visual descriptors, i.e. motion entropy, camera motion pattern, active region, dominant color selector, field color probability map (FPM) moment function. We also develop a shot pace model to describe the production rule's effect on the shot length. In addition, we employ a single-layer SVM classifier to select good frequency-domain and time-domain audio features to classify game-specific sounds.

In Phase II, we use the available mid-level representations from Phase I to classify each shot into one of the predefined shot categories, e.g. Player Close-up, Field View, etc. in which we explore the use of various supervised learning algorithms. The supervised learning procedure is constructed on the basis of effective mid-level representations at Layer 1 rather than relies on the blind training of large amount of high-dimensional data. Moreover, we make use of selected audio features to train SVM classifiers to generate game-specific audio keywords, e.g. applause, whistling, etc.

The principle motivation for introducing Phase I is to seek a set of clever pre-processing approaches, which robustly and flexibly generates more appropriate features for the generalization performance. For practical solutions, data pre-processing is often one of the most important stages in the development of the solution, and the choice of pre-processing steps can often have a significant effect on the generalization performance [20]. If we can perform sufficiently clever pre-processing then the remaining operations become trivial.

The choice of shot classes in Phase II is based on the nature of sports video that we are studying. First, sports video shots can be summarized in a small number of shot classes with clear semantic meanings, and the semantic shot class transition patterns (at the inter-shot level) might occur in an event. Second, game-specific sounds are significant information to assist sports event detection (at the intra-shot level).

2.3 High-level Analysis

Sports domain knowledge plays an important rule in the high-level analysis. Domain knowledge consists of two aspects: production rule and game-specific rule. We use semantic shot classes to construct a general temporal model for representing sports production rules. However, different sports games exhibit much variation. Thus we employ game-specific rules to perform high-level semantic analysis for a concrete sports game.

The high-level analysis scheme can be described as follows. First, we employ this temporal model to coarsely identify event and non-event segments. Second, we make use of semantic linkages between shot classes and game-specific rules to perform inter-shot and structure analysis for locating game-specific event segments. Third, we combine rich user semantics and strong game-specific rules to detect rich events and semantic objects within related event segments.

3. MID-LEVEL REPRESENTATIONS

In this section, we present a set of novel audio-visual mid-level representations. In terms of data-driven learning approaches and semantic hints, this set of mid-level representations is different from the basic low-level features. In Section 3.1, we briefly introduce mean shift procedure widely used in the motion vector model and color tracking model. In Section 3.2, 3.3, and 3.4, we explain motion vector model, color tracking model, and shot pace model. In Section 3.5, we discuss the semantic shot classification. In Section 3.6, we briefly introduce the audio keywords.

3.1 Mean Shift Procedure

The histogram is the oldest and most widely used density estimator. However, the discontinuity of histograms causes extreme difficulty if derivates of the estimates are required. This weakness together with the inefficient use of data makes it necessary to explore alternatives to histograms particularly when the density estimate is an intermediate component [21]. Apart from histogram, the kernel estimator is elegant and of wide applicability. The mean shift procedure is derived from the kernel estimator. The repeated movement of data points to the sample means is called the *mean shift procedure* [22].

The mean shift vector always points towards the direction of the maximum increase in the density. In [22], Cheng have shown that mean shift is a mode-seeking process on a surface constructed with a "shadow" kernel and studied the convergence for mean shift iteration. Since efficient mean shift computation requires efficient range searching, Comaniciu *et al.* [23] proposed a computational module based on the mean shift procedure, and successfully applied it to two low-level vision tasks: discontinuity preserving filtering and image segmentation.

3.2 Motion Vector Model

We propose a cone-shaped motion vector space (MVS) to represent motion vectors. The MVS space provides a visualized representation of the motion vector fields (MVF), which provides us with a visual aid to understand and analyze the motion characteristics. We then map the analysis of the MVS space to the problem of feature space analysis. With the nonparametric clustering approach *Mean Shift Procedure*, we have come up with 5 major descriptors: entropy, pan, tilt, diagonal, active region.

3.2.1 Cone-shaped MVS Space

Most users tend to think of color in the same way that they perceive it – in terms of hue, purity, and brightness. So scientist came up with what they call *perceptual* color spaces. Similarly

we would like to think about motion vectors field in the way that is natural to non-technical people. According to motion vector characteristics and HSV parameter ranges, we propose a cone-shaped MVS space to represent motion vectors. Figure 2 (a) illustrates the cone-shaped MVS space. Figure 2 (b) shows the polar coordinates *Angle* (the radial coordinate) and *Magnitude* (the angular coordinate) of the motion vector.

We convert a motion vector into MVS space as follows:

$$Hue = Angle \quad , \quad 0 \le Angle < 360$$

$$Saturation = \begin{cases} \dfrac{255 * Magnitude}{Mag_{th}}, & if\ Magnitude < Mag_{th} \\ 255 \quad , & if\ Magnitude \ge Mag_{th} \end{cases}$$

$$Brightness = \begin{cases} \dfrac{255 * Texture}{Tex_{th}} & , if\ Texture < Txt_{th} \\ 255 & , if\ Texture \ge Txt_{th} \end{cases}$$

where Mag_{th}, Tex_{th} are normalizing thresholds. Texture measure is obtained by computing the high frequency energy generated by the variance of wavelet coefficients in the high frequency bands, or by computing AC energy from AC DCT coefficients.

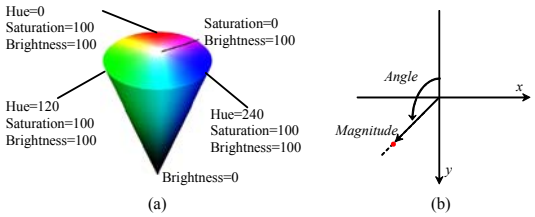

Figure 2. The cone-shaped MVS space and the polar coordinates of the motion vector

We can interpret the physical meanings of the MVS space. *Hue* represents direction, *saturation* represents intensity, and *brightness* represents confidence. The use of *brightness* to represent confidence is based on the intuition that a high texture region should produce a "good" motion vector. Apart from texture confidence measure, we may develop spatial confidence and temporal confidence measures [24] to perform confidence processing. Figure 3 shows some illustrative representations of MVF in MVS space.

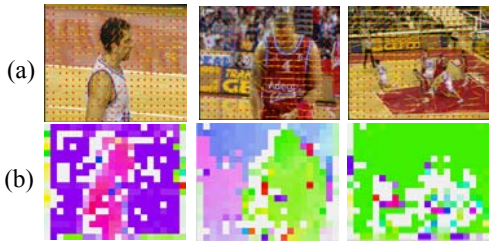

Figure 3. Examples of the MVF representations in MVS space: (a) original images overlapped by the MVF; (b) the representations in MVS space

3.2.2 Analysis of the MVS Space

The analysis of the MVS space is essentially the problem of feature space analysis. Most of previous research focused on the robust recovery of parametric motion models [25, 26]. In spite of the promising results and significant theoretical properties, it is hard to directly apply these algorithms to estimate camera motion and local motion from large amounts of video data. The failure is due to the violation of parametric assumption in real-world problems. *Arbitrary structured feature spaces can be analyzed only by nonparametric methods since these methods do not come with embedded assumptions*.

The MVF representation in MVS space is a two-dimensional lattice of three-dimensional vectors (Hue, Saturation, and Brightness). The space of the lattice is known as the *spatial* domain, while the color information is represented in the *range* domain. In order to consider the spatial consistency of motion magnitude and direction, we concatenate on the location and range vectors in the joint spatial-range domain of dimension five. Therefore, we think decomposition of the MVF into homogeneous tiles, is important for efficient motion analysis.

Since the Mean Shift Procedure's mode-seeking has excellent discontinuity preserving smoothing performance and simple control parameter with clear physical meaning (the kernel bandwidths determine various spatial and range resolutions of analysis), we propose the analysis scheme as follows:

Mean-shift based MVS Space Analysis Scheme

1. Run the mean shift procedure to smooth the motion vector field through moving the kernel (window) in the direction of the maximum increase in the joint density gradient. The joint domain kernel is defined as the product of two radially symmetric kernels and the Euclidean metric is employed: $K_{h_s,h_r} = \frac{C}{h_s^2 h_r^3} k\left(\left\|\frac{\mathrm{x}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathrm{x}^r}{h_r}\right\|^2\right)$, where x^s and x^r are the spatial part and range part respectively; $K(x)$ is the normal kernel used in both domains; h_s and h_r are the kernel bandwidths; and C is the normalization constant.
2. The outcome of the mean shift filtering is fed through a watershed algorithm [27], yielding the delineation of the clusters in the joint domain. The watershed drowning parameter is used to group together the clusters that are closer than h_r in the range domain. Small spatial regions are easy to eliminate through post-processing.
3. We heuristically select the significant clusters for analyzing the characteristics of the global motion and local motion.

3.2.3 Entropy of the Motion Vector Field

Assume that we have a set of clusters $\{C_1, C_2, \cdots, C_m\}$ in the joint domain with associated spatial regions $\{R_1, R_2, \cdots, R_m\}$. The entropy of the motion vector field is:

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i \quad , P_i = R_i \Big/ \sum_{j=1}^{m} R_j$$

where P_i denotes the probability of the cluster C_i.

The entropy H, which indicates the motion vector field's randomness or unpredictability, is derived from the analysis in MVS space. We can use the entropy to facilitate video shot classification. For example, close-up shots usually exhibit higher entropy since the foreground's active motion and the camera's following action result in the uncertainty of MVF. Wide-angle shots usually exhibit lower entropy since the dominant pan and tilt lead to more uniform distribution of MVF.

3.2.4 Camera Motion Patterns

Instead of trying to precisely and robustly recover the parametric model through solving an over-determined linear system, we introduce an angle quantization scheme to analyze the motion characteristics in the joint spatial-range domain.

The angle quantizer is formulated as:

$$r_k = [\ (-1)^{k+1}\alpha + \lfloor k/2 \rfloor \cdot 90^\circ, (-1)^k \alpha + \lfloor (k+1)/2 \rfloor \cdot 90^\circ\ ), k = 0 \cdots 7$$

$$where\ \ \alpha = 15^\circ.$$

Let $G = \{C_i\}_{i=1\ldots m}$ denote the homogeneous clusters by the mean shift filtering procedure, where $C_i = < R_i, \overline{Mag}_i, \overline{Ang}_i >$; R_i is the number of motion vectors associated with cluster C_i; $\overline{Mag}_i$ is the average magnitude of C_i; and $\overline{Ang}_i$ is the average angle of C_i. Let P_i denote the $\overline{Ang}_i$ quantization level of the i^{th} cluster by using the angle quantizer, then we can compute the camera motion rate in the three directions respectively as follows:

$$Pan = \sum_{i=1}^{m} \left((-1)^{\frac{P_i}{4}} \cdot \overline{Mag}_i \cdot R_i \right) \Big/ \sum_{i=1}^{m} R_i, P_i \in \{0,4\}$$

$$Tilt = \sum_{i=1}^{m} \left((-1)^{\left(\frac{P_i - 2}{4} - 1\right)} \cdot \overline{Mag}_i \cdot R_i \right) \Big/ \sum_{i=1}^{m} R_i, P_i \in \{2,6\}$$

$$Diagonal = \sum_{i=1}^{m} \left(\overline{Mag}_i \cdot R_i \right) \Big/ \sum_{i=1}^{m} R_i, P_i \in \{1,3,5,7\}$$

To eliminate the outlier effects, we eliminate those clusters with fewer motion vectors below a suitable threshold.

3.2.5 Active Region

In terms of content-based video analysis and indexing, a refined object boundary is not essential. It is well documented that user attentions tend to cluster around places with high gradients of change in the luminance distribution [33]. Moreover, motion plays an important role in focusing of attention within perceptions. Therefore we combine the analysis scheme of MVS space and the rules of sports video to develop the concept of *active region*. This can be easily understood as a strategy of paying different attention to different regions of the images at different times. Figure 4 shows some examples of active regions.

Figure 4. Examples of active regions from eight different sports video sequences: basketball, soccer, volleyball, tennis, golf, racing, table tennis, and gym.

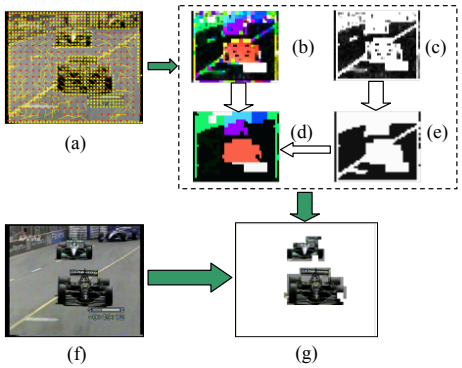

Figure 5. An illustrative procedure of extracting active regions. (a) Image frames overlapped with its associated motion vectors field (yellow dots denote blocks with high texture, arrows denote motion vectors), (b) the representation in MVS space, (c) the normalized texture map, (d) the representations after mean shift smoothing and watershed transform, (e) the texture map after mean shift smoothing and watershed transform, (f) the original image frame, and (g) the active regions according to (d) and (e).

The extraction of active regions is done in three steps as follows: a) We use the mean shift procedure to smooth the motion vector field, and then use the watershed algorithm to delineate the clusters in the joint domain; b) We heuristically select the seed region from the center to the periphery according to the region's texture and shape features; c) We investigate other homogeneous regions belonging to the same cluster as the seed region. For those regions also satisfy the shape requirement, we will consider them as active regions together with the seed regions. Figure 5 illustrates the procedure of extracting active regions.

3.3 Color Tracking Model

To represent a stable perception of color over varying lighting conditions, we propose an adaptive color tracking model consisting of two components: *Dominant Color Selector* (DCS) and *Field Color Probability Map Tracker* (FPMT).

3.3.1 Dominant Color Selector (DCS)

A dominant color is a color that is most characteristics of the scene. It usually determines the presence, appearance, and spatial relationships of objects (e.g. playing field/court, stand, players, etc.) of semantic importance to sports scene understanding. With the DCS component, it is straightforward to select a field/court region for further shape analysis. A unique advantage of the DCS component is the capability of representing multi-modal field/court colors.

In [4], we used mixture models for interpreting the concept of 'dominant', where a dominant color selection procedure is treated as the density estimation of color pixels within a video sequence. This modeling work was based on assumption that the uniform field regions are dominant in the image frames. In [28], we further proposed a nonparametric color characterization model based on mean shift procedure. With this model, it is easy to select dominant colors in an elegant and flexible way. Figure 6 illustrates the functional modules. Herein we exploit the mean shift's mode seeking function to realize the spatio-temporal clustering. For more details, please refer to [28].

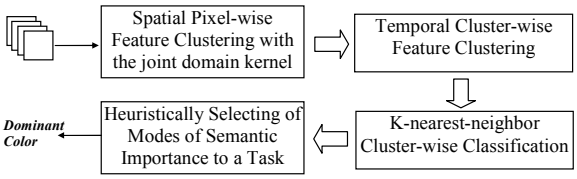

Figure 6. Functional modules of dominant color selector

3.3.2 Field Color Probability Map Tracker (FPMT)

We may use the field color probability map to represent a stable perception of color over varying lighting conditions. The FPMT component is designed to perform the tracking of field/court's pose variation according to the appearance geometric moment functions of its FPM, which can be used to generate various descriptors, e.g. Area, Orientation, Elongation, etc.

The FPM is built upon the ratio histogram between the model and the image. The definition of FPM uses the histogram backprojection method [30]. We perform adaptive FPM tracking as follows:

1. Choose the initial image patch to compute the model histogram and set the initial mean shift search window. This initial patch is selected from the first frame of a given major shot. The initial mean location is computed based on the zero- and first-order moment of FPM within the initial part.
2. The Mean Shift procedure is applied to seek the mode of the ratio histogram for the next frame within the shot. According to the mode seeking results and search window width, update the model histogram and compute the FPM with the above formulation. The new mean location is computed by the zero- and first-order moment of FPM.
3. Center the search window at the new mean location and adjust the window size according to the zero-order moment of this new FPM, so that the initial seed patch can grow to encompass the playing field/court after several iterations. Go to Step 2.

We make use of the geometric moment functions of the FPM to compute shape information such as total area, coordinates of the centroid, and orientation. Geometrical moments are defined with basis set $\{x^p y^q\}$. The $(p+q)^{th}$ order two-dimensional geometric moments $m_{pq} = \iint_{\zeta} x^p y^q f(x,y)dxdy, \quad p,q = 0,1,2\ldots$, ζ is the region of the pixel space in which the density function $f(x,y)$ is defined. The shape characteristics of FPM are represented by:

$$Position: \ x_0 = m_{10}/m_{00}, \quad y_0 = m_{01}/m_{00}$$
$$Area: A = m_{00}$$
$$Orientation: \ \theta = arctan(b,(a-c))/2$$
$$Length \ of \ major \ axis: l_1 = \sqrt{(a+c+\sqrt{b^2+(a-c)^2})/2}$$
$$Length \ of \ minor \ axis: l_2 = \sqrt{(a+c-\sqrt{b^2+(a-c)^2})/2}$$
$$Elongation: E = (l_1 - l_2)/(l_1 + l_2)$$
$$a = m_{20}/m_{00} - x_0^2, \ b = 2(m_{11}/m_{00} - x_0 y_0), \ c = m_{02}/m_{00} - y_0^2$$

Together with the Energy descriptor proposed in [4], we have 4 color-based descriptors: Area, Orientation, Elongation, Energy. These descriptors can be used to present camera perspective, view coverage, and field poses.

3.4 Shot Pace Model

Through extensive experimental study, we found that there is a distinguishable shot length difference between major shot classes (court/field view vs. close-up) in team-based sports videos, e.g., basketball, soccer, volleyball. The reasons are twofold: First, a game played by two teams of more than 3 players greatly relies on the cooperation of teammates to make an offense and defense, which leads to relatively looser structure compared with tennis and table tennis. Second, a photographer tends to use a wide shot (court/field view shot) to follow actions and use a close-up shot to track a player or a gathering of people, which makes the wide shots longer than the neighbor close-up shots.

Thus we use a *sliding window* to examine the m successive shot lengths. We introduce the normalized shot length measure SLP to represent the shot length-related pace characteristics, namely, the rate of the current shot's length to the maximum within a symmetric sliding window.

3.5 Semantic Shot Classification

So far we have discussed a set of effective representations of motion, color, and shot pace. In this section, we briefly introduce the design of semantic shot classifier. Figure 7 lists the predefined shot classes. Please refer to [29] for more details.

Figure 7. Predefined shot classes for tennis, soccer, basketball, volleyball, and table tennis, along with their percentage.

We exploit available visual mid-level representations derived from Phase I to construct the feature vector the shot attributes' numerical description and train the semantic shot classifier in accordance with the predefined shot categories as listed in Section 2.3.2. Currently the feature vector consists of 9 features, i.e. $< \overline{Entropy}, \overline{Pan}, \overline{Tilt}, \overline{Diagonal}, SLP, \overline{Elongation}, \overline{Orientation}, \overline{Energy}, \overline{Area} >$. We apply the 'average' operator to the series of feature values so that we get a 9 dimensional feature for representing a shot's attributes. In fact the 'average' operator leads to a big reduction in the dimensionality of the input space within a shot. Experiments showed that the averages of these 9 features at the shot level are enough for semantic shot classification.

We employ the C-Support Vector classification (Binary Case) to perform semantic shot classification. Given the training vectors $x_i \in R^n, i = 1,\ldots,l$ in two classes and a vector $y \in R^l$ such that $y_i \in \{1,-1\}$, C-SVC solves the following primal problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i, \qquad y_i(w^T \Phi(x_i) + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0, i = 1,\ldots,l.$$

We train all datasets only with the RBF kernel:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}.$$

Hence, two parameters C and γ are considered. We conducted a 10-fold cross validation on the whole training data to select these parameters. Two thirds are chosen as training data and one-third as test data. Tables 1-5 give the test data results. We could achieve an accuracy of around 85~95% over more than 5500 shots.

3.6 Audio Keyword

There are some significant game-specific sounds such as applause, whistling that have strong relationships to the action of players, referees, commentators, and audience in broadcast sports videos. These actions can be heuristically mapped to interesting events according to specific sports game rules.

We exploit representations of the audio signal in terms of time-domain and frequency-domain measurements to train the game-specific sound recognizers (i.e. Audio Keyword) using SVM. These measurements include zero-crossing rate (ZCR), spectral power (SP), mel-frequency cepstral coefficients (MFCC), linear prediction coefficient (LPC), short time energy (STE), and linear prediction cepstral coefficients (LPCC). Feature selection is important for discriminating audio signals. To select good features suitable for the classification of audio keywords, we make use of a single-layer SVM classifier to extensively evaluate the performance of a single feature in the classification [2, 3]. Finally we choose the most suitable features to construct different hierarchical SVM classifiers for different sports game videos. Currently, we have generated audio keywords for three typical sports game videos: tennis, soccer, and basketball. Table 6 shows the performance of audio keyword classifiers. For more details, please refer to [2, 3].

4. HIGH-LEVEL ANALYSIS

In this section, we present a high-level semantic analysis scheme to evaluate the functionalities of mid-level representations from the event detection viewpoint.

As shown in Figure 8, this scheme works in the top-down style.

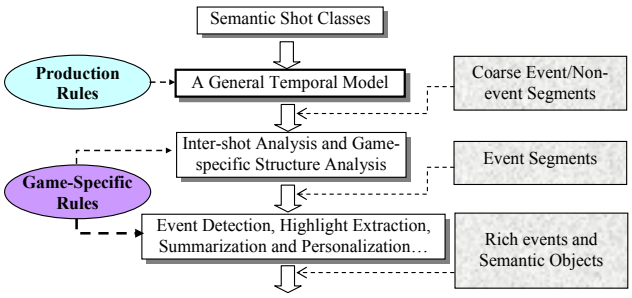

Figure 8. A high-level semantic analysis scheme and its relationship with domain knowledge

The semantic shot classes facilitate high-level analysis from three aspects: a) we could use the ordered sequence of shot classes as a reliable indicator of event/non-event, particularly in looser sports game video such as soccer videos; b) we could coarsely locate segments with a certain event according to shot classes' semantic linkage; c) we could incrementally perform intra-shot processing within selected potential segments of events to detect rich events and semantic objects.

Some heuristic rules have to be considered in this scheme since domain knowledge is nearly always advantageous to detect rich events with high accuracy [12]. However, it will incur the risk of making the solution less generic and automated.

It is feasible to seek supporting general models that might underlie the generation of the sequence of semantic shot classes among a sub-genre of sports games. Thus we model the sports domain knowledge in two aspects: production rules and game-specific rules. The first set of rules models the combination of shots from multiple angles determined by field production techniques. The latter set of rules is inherent to specific sports games. At present, we consider the class of ball games and manually construct a general temporal model to represent the production rules. However, it is difficult to abstract a model to represent the game-specific rules due to their variations. Thus, we individually craft game-specific rules for a concrete game.

4.1 A General Temporal Model

According to the focal distance of the shot and the main subject, we summarize the shots in 8 classes ($U_{1\text{-}6}$, $P_{1\text{-}2}$) as shown in Figure 9. With these 8 classes, we generally partition sports video shot sequences into two logical segments, namely, *in play segments* (IPS) and *out of play segments* (OPS). The IPS and OPS occur in successive turns. For the class of field-ball game, the IPS corresponds to the video segment when a ball is within the boundaries of the field and play has not been stopped by the referee. On the other hand, the OPS corresponds to the video segment when a ball is outside the boundaries of the field or play has been stopped by the referee. An IPS or OPS may comprise more than one shot.

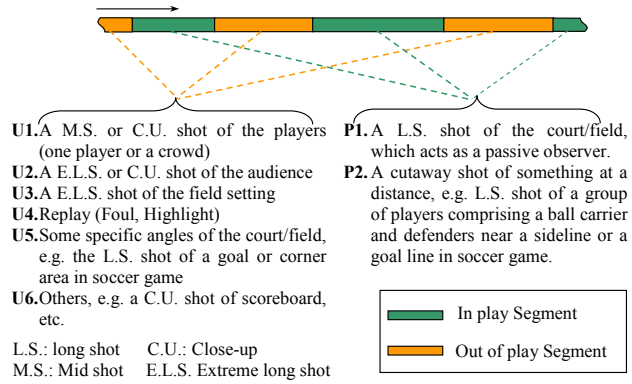

Figure 9. A general temporal model in broadcast field-ball game video.

It is straightforward to derive the concepts of 'play/break' in [13] with IPS and OPS. However, the label of 'play/break' in [13] is computed at the frame level with low-level features. The 'play/break' resolved by IPS and OPS is based on the semantic shot classes, thereby incorporating the context information at the shot level. Clearly, the IPS and OPS provide a general way to

solve the basic structure analysis problem of 'play/break' for the field-ball game video.

Figure 7 illustrates the concretized shot categories for five typical ball games. The shot category C_i is named as follows:

$$C_i =< G_i, L_i >,\ G_i \in \{U_i, P_j | i = 1,2,\ldots,6; j = 1,2\}$$

where L_i denotes a brief linguistic description related to the subject, G_i denotes the class label as listed in Figure 9.

4.2 An Event Hierarchy

We employ the general temporal model to identify any sequence transition pattern that might help to detect an event.

According to the characteristics of field-ball games, the transition of IPS and OPS is an indicator of a set of regular events. For example, in soccer video, a *Field View* shot is normally followed by a *Player Following* shot unless an event occurs, such as offside, foul, out of bounds cross a sideline or a goal line. The pairs of *Field View* and *Player Following* shot compose the IPS segment. The OPS segment consists of a series of shots between two consecutive IPS segments. During an OPS segment, we can check the occurrence of events, e.g. free kick, corner kick, throw-in, goal, etc. It is observed that the transition between IPS and OPS embodies the events' relationships between the cause and effect at the shot level in field-ball game videos.

As compared to soccer, tennis has a compact structure. In tennis video, the IPS segment normally comprises one *Court View* shot. The transition from IPS to OPS is caused by a set of regular tennis events, such as score, ace, fault, double fault, etc. Similarly, in volleyball, table tennis, and basketball video, the IPS comprises one shot, such as *Court View* shot, or *Full Court Advance* shot. It is game-specific regular events that lead to the transition between IPS and OPS.

Based on the above analysis, we present an event hierarchy in the field-ball game videos as illustrated in Figure 10. We classify events into three categories: structure events, regular events, and tactic events. A tactic event always occurs in the IPS segment. It can be detected by the spatial relationship inference of ball trajectory and the players, or by the pose analysis of the players. Regular events comprise two sets: Set I and Set II. An event of Set I leads to the transition from IPS to OPS. An event of Set II leads to the transition from OPS to IPS. In order to identify a regular event, we exploit the semantic shot classes and audio keyword to derive heuristic rules according to game-specific rules. Structure events depend on the scoring system of a specific game. We choose tennis and soccer video to concretize the event hierarchy as follows (please refer to the glossaries in [31, 32]):

Tennis events hierarchy:

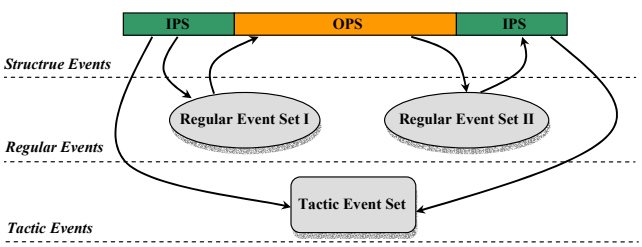

Figure 10. An event hierarchy posed by the temporal model

Structure events: in play, out of play, point, deuce, game, set.

Regular event Set I: serve, reserve, score, ace, fault, double fault.

Regular event Set II: to serve.

Tactic events: return, take the net, volley, rally, wrong-foot, back hand, forehand, drop shot, forcing shot, pass, retrieve, smash, winter.

Soccer events hierarchy:

Structure events: in play, out of play.

Regular event Set I: foul, offside, shot, goal.

Regular event Set II: free kick, corner kick, penalty kick, throw-in, kickoff, goal kick.

Tactic events: dribbling, clear, passing, assist, possession, penetrate, volley, tackling, steal, shielding.

4.3 Event Detection

Event detection is one of the main tasks of the semantic analysis. The use of sports domain knowledge and machine learning are the means of detecting sports events. We employ the high-level analysis scheme (See Figure 8) to detect tennis and soccer events.

4.3.1 Key Design Considerations

Due to so many types of sports games and sports events, it is infeasible to present extensive event detection results for various sports games. Our aim is to demonstrate that the proposed mid-level representation framework contributes to more efficient, effective, and less ad-hoc event detection solutions.

The key issues considered are:

a) **The choice of representative sports games**. We consider tennis and soccer as two typical field-ball games. The tennis has compact structure, while the soccer video generally has looser structure.

b) **The choice of convincing events**. We mainly focus on the structure and regular events, plus some comparatively simple tactic events. The reasons are twofold: 1) structure and regular events are tightly related to mid-level representations in terms of shot transition patterns or sound transition patterns, while the detection of tactic events raises demanding requirements of low-level image processing techniques; 2) too detailed tactic event detection requires strong game-specific knowledge, resulting in less generic properties.

c) **The extensibility of event detection approaches**. The event hierarchy implies different relationships between events and domain knowledge. The structure and regular events produce a relatively loose relationship with game-specific knowledge, while the tactic events produce a tight relationship. This means different extensibility of the approaches. The structure and regular events detection approaches are easily extensible, while the tactic events detection approaches are less extensible.

4.3.2 Approaches

4.3.2.1 Detection of IPS and OPS

According to the temporal model (see Figure 9) and the semantic shot classes (see Figure 7), we can determine the IPS and OPS by checking the shot transition pattern.

For tennis, volleyball, and table tennis, there is only one class $< P_1, Court\ View >$ that belongs to the IPS segment. For basketball, there is only one class $< P_1, Full\ Court\ Advance >$ that belongs to the IPS segment. For soccer, however, there are two classes $< P_1, Field\ View >$ and $< P_2, Player\ Following >$ that belongs to the IPS segment, but they normally appear in pairs. Thus it is straightforward to locate the IPS and OPS by checking the appearance of these shot classes.

As illustrated in Figure 10, Regular Event Set I occurs in IPS, Regular Event Set II occurs in OPS, and Tactic Event Set occurs in IPS. So the located IPS and OPS also signify coarse event/non-event segments in addition to the *in play* and *out of play* events.

4.3.2.2 Detection of Regular Events

After locating IPS and OPS, we perform the detection of Regular Event Set I in the IPS and Regular Set II in the OPS. Since the OPS comprises much more shots than the IPS, the detection of Event Regular Set II needs to consider more inter-shot information in addition to intra-shot analysis. Figure 11 illustrates the multi-modal approach to detect regular events.

For Regular Event Set I, the coarse event segment is composed of the shots within IPS. For Regular Event Set II, the coarse event segment is composed of the shots within OPS. According to the shot classes' semantic linkage, we heuristically select shots containing interesting events to be detected via intra-shot analysis or signifying interesting events to be detected via inter-shot analysis. Once the event segments are found, we heuristically search the game-specific sounds to support events of interest. The video shots are not necessarily aligned with audio segments, we enlarge the audio search range by one or two shots. Finally, we design concrete decision rules for event inference. The decision rules are based on the semantic shot classes and audio keyword instead of low-level features.

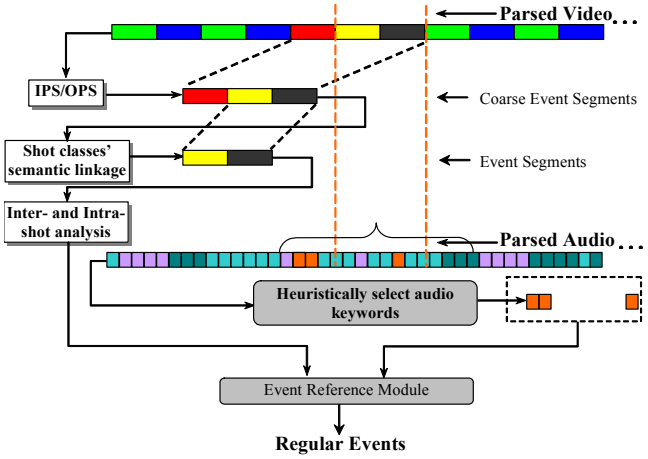

Figure 11. A multi-model approach to detect regular events

4.3.2.3 Detection of Structure Events

In play and *out of play* are two common structure events for field-ball games. Other structure events (e.g. *point*, *deuce*, *game*, and *set* in tennis) have to be derived by a regular event (e.g. *score* in tennis) and the scoring system of a specific game. Due to its loose structure, soccer has no structure events like in tennis.

4.3.2.4 Detection of Tactic Events

The tactic event detection approaches are flexible and dependent on strong game-specific knowledge. For example, we exploit a single player's trajectory to detect *take the net* and *rally* in tennis. It is advantageous to locate the event segments (refer to Figure 11) before applying careful event analysis methods.

5. EVENT DETECTION APPLICATIONS

In this section, we use event detection to illustrate the effectiveness of our mid-level representation framework, including the mid-level representations and the high-level analysis scheme.

5.1 Event Detection in Tennis Video

5.1.1 Detection of IPS and OPS

By detecting the shot class $< P_1, Court\ View >$, we locate the IPS and OPS segments.

5.1.2 Detection of Regular Events

Within the IPS segments, we employ the audio keywords *Hitting Ball* and *Applause* to compute the ball hitting times and the intervals between two ball hits, and check the applause sound around the end of court view shots. In this way, we are able to detect the regular events, which comprise *serve*, *reserve*, *score*, *ace*, *return* [2]. *Score* indicates the structure event *point*.

Now it is straightforward to detect fault and double fault.

1) Within one *point*, two *serve* $\rightarrow$ *fault*;

2) Within one *point*, two *serve*, no *return*, no *Applause* $\rightarrow$ *double fault*;

5.1.3 Detection of Structure Events

The IPS segments are temporally aggregated to form a *game*. This is because the distance between two *games* is much greater than the distance between two successive IPS segments. According to the rules, (0:4), (1:4), (2:4) are three normal cases within one *game*. If the total points are greater than 6, *deuce* (3:3) must occur in that game. A *set* is a group of *games*. In major tournaments, there are usually five *sets* in a men's match and three in a women's match.

5.1.4 Detection of Tactic Events

We employ the player tracking to detect *take the net* and *rally*.

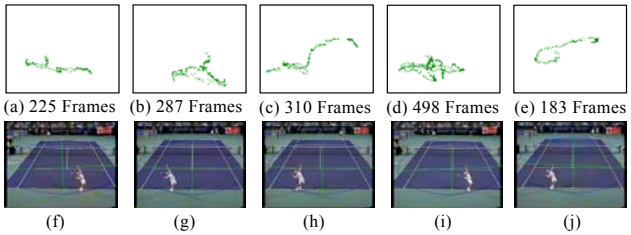

Figure 12. (a)-(e): the single player's trajectories by tracking the player within the <P_1, Court-View> shots, (f)-(j): the first frames within the < P_1, Court-View> shots.

As shown in Figure 12, continuous upward moving in (c) and (e), indicates *take the net*, the rapid and extended exchange of player's position in (a), (b) and (d) indicates *rally*.

We can use the player trajectory to improve the detection of *points*. Since the server alternates sides with each *point* in singles, we can determine *point* through judging the relative position between the server and the vertical minor principle axis of FPM, i.e. (f) → (g, h) → (i) → (j).

5.1.5 Results

Table 7 summarizes the event detection results of tennis video. The total duration of tennis video used in our test is about 53 minutes. These tennis videos consist of *2002 Western & South Financial Group Masters between HEWITT vs. MOYA* (32 Minutes), and *2003 Australia Open Men's Singles Semifinal between FERREIRA vs. AGASSI* (21 Minutes). The recall and precision for structure event detection are perfect due to the canonical production rules and compact structure of tennis video. The recall and precision for regular event detection are relatively low because of the sensitivity of audio keyword to environmental noise. The recall and precision for the two simple tactic events are promising due to the precise event segment location and robust player tracking at the corresponding segments.

5.2 Event Detection in Soccer Video

5.2.1 Detection of IPS and OPS

We locate the IPS and OPS segments by detecting the pairs of $< P_1, Field\ View >$ and $< P_2, Player\ Following >$.

5.2.2 Detection of Structure Events

The only two structure events *in play* and *out of play* are derived by IPS and OPS segments.

5.2.3 Detection of Regular Event Set I

These events are detected by:

a) Checking that the audio keyword *Whistling* happens within the last one shot in the IPS segment. If *Whistling* is detected, then an event of *offside* or *foul* must occur in this IPS.

b) Checking that the audio keyword *Excited Commentator Speech* and *Excited Audience* occurs within the last $< P_1, Field\ View >$ shot in the IPS segment. If *Excited Commentator Speech* or *Excited Audience* are detected, and view coverage switches from the mid-field to the goal area (according to FPM's orientation), then a *shot* must occur at the end of this IPS.

c) Checking whether there are many $< U_1, Player\ Close\text{-}up >$ shots, persistent *Excited Commentator Speech* and *Excited Audience*, and long duration within the OPS segment. If yes, then a *goal* must occur at the end of the previous IPS followed by *kick off* at the end of this OPS.

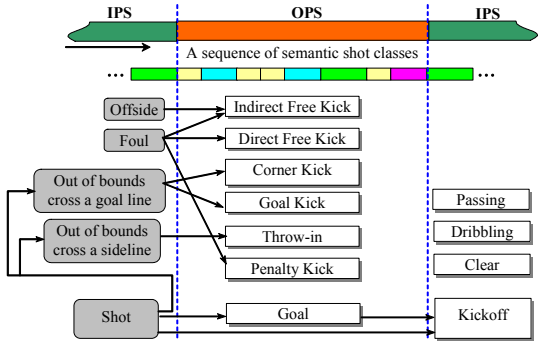

Figure 13. Events' relationship between cause and effect

5.2.4 Detection of Regular Event Set II

These events can be detected by:

a) Checking that the audio keyword *Whistling* occurs within the last two shots in the OPS segment. If *Whistling* is not detected and there is not $< U_5, Goal\ View >$ shot within the OPS segment, then a *throw in* must occur in the OPS.

b) Checking that the audio keyword *Whistling* happens within the last two shots in the OPS segment. If *Whistling* is detected and there are some $< U_5, Goal\ View >$ shots and some $< U_1, Player\ Medium\ View >$ shots within the OPS segment, then a *corner kick* must occur in the OPS.

c) If a *corner kick* cannot be decided by rule (b), then check whether there are some $< U_5, Goal\ View >$ shots, and long duration within the OPS segment. If yes then this is a *penalty kick*, else it is a *free kick*.

To help in understanding the above heuristic rules, Figure 13 illustrates the soccer events' relationships between the causes and effects.

5.2.5 Results

Table 8 summarizes the event detection results of soccer video. The ground truth comes from the official site of The 2002 FIFA World Cup. URL: http://fifaworldcup.yahoo.com/en/t/s/g.html. The recall and precision for the detection of *free kick* and *foul/offside* events are promising due to the lower error rate of *Whistling* keyword and the precise identification of the IPS and OPS segments. The precision for *goal* event is low because of the confusion from the loud environmental audience sounds. In practice, we can relax, constrain or insert some heuristic rules to favor higher recall or higher precision of a certain event. The average recall and precision of 75~85% for regular events shows that it is feasible to develop simple heuristic rules to distinguish among structure and regular events.

6. CONCLUSIONS

In this paper, we propose a mid-level representation framework to facilitate semantic sports video analysis. Compared with previous work in sports video domain, this framework exhibits unique features in the introduction of two mid-level representation layers. Under this framework, sports domain knowledge can be represented by the robust and reusable mid-level representations, which makes high-level video content analysis more efficient, effective, and less ad-hoc for various types of sports.

With this framework, we have accomplished the detection of rich events with strong semantic meanings in tennis and soccer videos. We believe that the proposed mid-level representation framework has provided an effective way to bridge the semantic gap between the richness of user semantics and the simplicity of available low-level perceptual visual and auditory features. It brings us a major step towards the identification and interpretation of meanings in sports video.

Clearly, this framework is open and extensible. Currently, we are working towards the development of robust mid-level feature

extraction approaches and the task-driven inter-shot analysis scheme.

7. REFERENCES

[1] C. Dorai, etc., "Media Semantics: Who Needs It and Why?", In *Proc. of ACM Multimedia* 2002, pp. 580-583, 2002.

[2] M. Xu, L.-Y. Duan, C.-S. Xu, Q. Tian, "A Fusion Scheme of Visual and Auditory Modalities for Event Detection in Sports Video," In *Proc. of ICASSP* 2003, pp. 189-192, 2003.

[3] M. Xu, N. C. Maddage, C.-S. Xu, M. Kankanhalli, Q. Tian, "Creating Audio Keywords For Event Detection in Soccer Video," In *Proc. of ICME* 2003, pp. 281-284, 2003.

[4] L.-Y. 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.

[5] D.Q. Zhang, S. -F. Chang, "Event Detection in Baseball Video Using Superimposed Caption Recognition," In *Proc. of ACM Multimedia 2002*, pp. 315-318, 2002.

[6] M. Han, W. Hua, W. Xu, and Y.H. Gong, "An integrated Baseball Digest System Using Maximum Entropy Method," In *Proc. of ACM Multimedia*, pp. 347-350, 2002.

[7] J. Assfalg, M. Bertini, C. Colombo, and A. D. Bimbo, "Semantic Annotation of Sports Videos," *IEEE Multimedia* 9(2): 52-60, 2002.

[8] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration," *IEEE Transactions on Multimedia* 4(1): 68-75, 2002.

[9] L.-Y. Duan, M. Xu, X.-D. Yu, and Q. Tian, "A Unified Framework for Semantic Shot Classification in Sports Videos", In *Proc. of ACM Multimedia* 2002, pp. 419-420, 2002.

[10] D. Zhong, S.-F. Chang, "Structure Analysis of Sports Video Using Domain Models," In *Proc. of ICME* 2001.

[11] C.W. Ngo, T.C. Pong, and H.J. Zhang, "On Clustering and Retrieval of Video Shots", In *Proc. of ACM Multimedia* 2001, pp. 51-60, 2001.

[12] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic Detection of Goal Segments in Basketball Videos", In *Proc. of ACM Multimedia* 2001.

[13] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, H. Sun, "Algorithms And Systems for Segmentation and Structure Analysis in Soccer Video," In *Proc. of ICME* 2001.

[14] Y.-P. Tan, D.D. Saur, S.R. Kulkarni, and P.J. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation," *IEEE Transactions on Circuits and Systems for Video Technology* 10(1): 133-146, 2000.

[15] Y. Rui, A. Gupta, A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," In *Proc. of ACM Multimedia*, pp. 105-115, 2000.

[16] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval," In *Proc. of IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 81-90, 1998.

[17] Y.H. Gong, L.T. Sin, C. H. Chuan, H.J. Zhang, M. Sakauchi, "Automatic Parsing of TV Soccer Programs," In *Proc. of International Conference on Multimedia Computing and Systems*, pp.167-174, 1995.

[18] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved," *IEEE Transactions on Circuits and Systems for Video Technology* 12(2): 90-105, 2002.

[19] H. Zhang, A.Kankanhalli, and S.W. Smoliar, "Automatic Partitioning of Full-motion Video," *Multimedia System* 1(1): 10-28, 1993.

[20] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995, pp. 295-329.

[21] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986, pp.7-74.

[22] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE PAMI*, 17(8): 790-799, 1995.

[23] D. Comaniciu, P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE PAMI* 24(5): 1-18, 2002.

[24] R. Wang, H.J. Zhang, and Y.Q. Zhang, "A Confidence Measure Based Moving Object Extraction System Built for Compressed Domain," In *Proc. of ISCAS* 2000.

[25] M.J. Black and P. Anandan, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-smooth Flow

Table 1. Shot classification results of tennis video (test data)

Shot Class	Total	Correct	False Alarm	Recall (%)	Precision (%)
<P_1, Court View>	130	126	11	96.9%	92.0%
<U_1, Player Close-up>	176	157	14	89.2%	91.8%
<U_2, Audience>	46	39	6	84.8%	86.7%
<U_1, Player Medium View>	66	54	11	81.8%	83.1%
<U_3, Setting Long View>	12	10	2	83.3%	83.3%

Table 2. Shot classification results of soccer video (test data)

Shot Class	Total	Correct	False Alarm	Recall (%)	Precision (%)
<P_1, Field View>	261	249	25	95.4%	90.9%
<P_2, Player Following>	169	138	24	81.7%	85.2%
<U_5, Goal View>	34	31	1	91.2%	96.9%
<U_1, Player Close-up>	198	167	30	84.3%	84.8%
<U_2, Audience>	41	35	7	85.4%	83.3%
<U_1, Player Medium View>	23	18	1	78.3%	94.7%
<U_3, Setting Bird View>	9	8	1	88.9%	88.9%

Table 3. Shot classification results of basketball video (test data)

Shot Class	Total	Correct	False Alarm	Recall (%)	Precision (%)
<P_1, Full Court Advance>	30	28	0	93.3%	100.0%
<P_1, Penalty View>	12	12	2	100.0%	85.7%
<U_1, Player Close-up>	48	44	3	91.7%	93.6%
<U_2, Audience>	11	9	3	81.8%	75.0%
<U_1, Player Medium View>	6	5	1	83.3%	83.3%
<U_3, Setting Bird View>	3	3	0	100.0%	100.0%

Table 4. Shot classification results of volleyball video (test data)

Shot Class	Total	Correct	False Alarm	Recall (%)	Precision (%)
<P_1, Court View>	67	63	7	94.0%	90.0%
<U_1, Player Close-up>	88	80	6	90.9%	93.0%
<U_2, Audience>	5	4	1	80.0%	80.0%
<U_3, Half Court View>	8	7	0	87.5%	100.0%
<U_6, Players & Coach>	3	3	0	100.0%	100.0%

Table 5. Shot classification results of table tennis video (test data)

Shot Class	Total	Correct	False Alarm	Recall (%)	Precision (%)
<P_1, Court View>	86	81	10	94.2%	89.0%
<U_1, Player Close-up>	156	145	6	92.9%	96.0%
<U_2, Audience>	5	4	1	80.0%	80.0%
<U_3, Setting Long View>	4	4	0	100.0%	100.0%
<U_6, Players & Coach>	3	3	0	100.0%	100.0%

Fields," Computer Vision and Image Understanding 6(4): 348-365, 1995.
[26] J.M. Odebez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," Journal of Visual Communication and Image Representation 6(4): 348-365, 1995.
[27] L. Vincent, P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE PAMI* 24(5): 1-18, 2002.
[28] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu, "Nonparametric Color Characterization Using Mean Shift," to appear on *ACM Multimedia 2003.*
[29] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu, "A Unified Framework for Semantic Shot Classification in Sports Video," *Technical Report*, Institute for Infocomm Research, Jun 2003.
[30] M.J. Swain, D.H. Ballard, "Color Indexing," *International Journal of Computer Vision* 7(1): 11-32, 1991.
[31] http://www.hickoksports.com/glossary/gtennis.shtml
[32] http://www.firstbasesports.com/soccer_glossary.html
[33] V. Cantoni, S. Levialdi, V. Robert. Artificial Vision, *Academic Press*, 1997, pp. 1-52.
[34] B.L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Videos," *IEEE Transactions on Circuits and Systems for Video Technology* 5(6): 533-544, 1995.

Table 6. Performance of audio keyword classifiers

Sports	Audio Keywords	Potential Events	Error Rate (%)
Tennis (45 Mins)	Applause	Score	7.23
	Commentator Speech	At the end (or the beginning) of a point	11.29
	Silence	Within a point	7.62
	Hitting Ball	Serve, Ace or Return	1.1
Soccer (90 Mins)	Long-whistling	Start of free kick, penalty kick, or corner kick, Game start or end, offside	7.27
	Double-whistling	Foul	10.89
	Multi-whistling	Referee reminding	8.93
	Excited commentator speech	Goal or Shot	23.58
	Plain commentator speech	Normal	20.24
	Excited audience	Goal or Shot	26.37
	Plain audience	Normal	24.15
Basketball (30 Mins)	Whistling	Foul	0.55
	Ball hitting backboard or basket	Shot	0.82
	Excited commentator speech	Fast break, Drive or Score	21.56
	Plain commentator speech	Normal	20.91
	Excited audience	Fast break, Drive or Score	19.86
	Plain audience	Normal	16.29

Table 7: Event detection results in tennis video

	game	deuce	point	serve	reserve	return	ace	fault	double fault	take the net	rally
Total	13	4	88	120	32	280	17	39	7	8	19
Correct	13	4	88	120	27	271	14	32	5	7	16
Missed	0	0	0	0	1	9	1	2	2	1	2
False Alarm	0	0	0	0	2	13	2	4	1	2	4
Recall	100%	100%	100%	100%	84.4%	96.8%	82.4%	82.1%	71.4%	87.5%	84.2%
Precision	100%	100%	100%	100%	81.8%	95.4%	77.8%	78.0%	83.3%	77.8%	76.2%

Table 8: Event detection results in soccer video

Match	Performance	Foul or Offside	Free Kick	Penalty Kick	Corner Kick	Shot	Goal	In Play (Mins)	Out of Play (Mins)
GER-BRA (Jun. 30, 2002)	Total	41	41	0	16	21	2	50	40
	Correct	36	36	0	13	16	2	47	43
	Missed	2	2	0	1	4	0		
	False Alarm	4	4	0	3	2	2		
	Recall	87.9%	87.9%		81.3%	76.2%	100%		
	Precision	83.7%	83.7%		72.3%	74.2%	50.0%		
ENG-BRA (Jun. 21, 2002)	Total	46	46	0	7	15	3	53	37
	Correct	40	40	0	6	11	3	52	38
	Missed	4	4	0	0	4	0		
	False Alarm	5	5	1	1	3	2		
	Recall	87.0%	87.0%		85.7%	73.3%	100%		
	Precision	85.1%	85.1%		75.0%	78.6%	60.0%		
GER-KOR (Jun. 25, 2002)	Total	33	33	0	14	22	1	54	36
	Correct	27	27	0	12	16	1	52	38
	Missed	2	2	0	1	4	0		
	False Alarm	1	1	1	4	5	2		
	Recall	81.8%	81.8%		85.7%	72.7%	100%		
	Precision	84.4%	84.4%		70.6%	69.6%	33.3%		