A Unified Framework for Semantic Shot Classification in Sports Video

Ling-Yu Duan, Min Xu, Qi Tian, *Senior Member, IEEE*, Chang-Sheng Xu, *Senior Member, IEEE*, and Jesse S. Jin, *Member, IEEE*

***Abstract*—The extensive amount of multimedia information available necessitates content-based video indexing and retrieval methods. Since humans tend to use high-level semantic concepts when querying and browsing multimedia databases, there is an increasing need for semantic video indexing and analysis. For this purpose, we present a unified framework for semantic shot classification in sports video, which has been widely studied due to tremendous commercial potentials. Unlike most existing approaches, which focus on clustering by aggregating shots or key-frames with similar low-level features, the proposed scheme employs supervised learning to perform a top-down video shot classification. Moreover, the supervised learning procedure is constructed on the basis of effective mid-level representations instead of exhaustive low-level features. This framework consists of three main steps: 1) identify video shot classes for each sport; 2) develop a common set of motion, color, shot length-related mid-level representations; and 3) supervised learning of the given sports video shots. It is observed that for each sport we can predefine a small number of semantic shot classes, about 5–10, which covers 90%–95% of broadcast sports video. We employ nonparametric feature space analysis to map low-level features to mid-level semantic video shot attributes such as dominant object (a player) motion, camera motion patterns, and court shape, etc. Based on the fusion of those mid-level shot attributes, we classify video shots into the predefined shot classes, each of which has clear semantic meanings. With this framework we have achieved good classification accuracy of 85%–95% on the game videos of five typical ball type sports (i.e., tennis, basketball, volleyball, soccer, and table tennis) with over 5500 shots of about 8 h. With correctly classified sports video shots, further structural and temporal analysis, such as event detection, highlight extraction, video skimming, and table of content, will be greatly facilitated.**

***Index Terms*—Semantic gap, shot representation, shot similarity, video classification, video databases indexing.**

Manuscript received June 30, 2003; revised November 28, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy K. Shih.

L.-Y. Duan is with the Institute for Infocomm Research, Singapore 119613 and also with the School of Design, Communication and Information Technology, University of Newcastle, NSW 2308, Australia (e-mail: lingyu@i2r.a-star.edu.sg).

M. Xu was with the Institute for Infocomm Research, Singapore 119613. She is now with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: MXu@ntu.edu.sg).

Q. Tian and C.-S. Xu are with the Institute for Infocomm Research, Singapore 119613 (e-mail: tian@i2r.a-star.edu.sg; xucs@i2r.a-star.edu.sg).

J. S. Jin is with the School of Design, Communication and Information Technology, University of Newcastle, NSW 2308, Australia (e-mail: Jesse.Jin@newcastle.edu.au).

Digital Object Identifier 10.1109/TMM.2005.858395

I. Introduction

THE ever-increasing amount of multimedia information is becoming inaccessible because of the lack of human resources to perform the time-consuming task of annotating it. The major goal of multimedia research is directed toward providing information for pervasive access and use [1]. To achieve this, it is critical to develop technologies to find the points of interest from the media chunks. However, current user expectations still far exceed the intelligence of today's computing systems, despite the significant progress in automated feature-based and structure-based indexing and retrieval techniques. The solutions currently available have one major drawback, *viz.* generic low-level content metadata available from automated processing deals only with representing perceived content, but not its semantics. Thus, more and more research effort is now geared toward modeling and extracting media-intrinsic, as well as media-extrinsic, semantics [27].

As an important video document, sports video has been widely studied due to tremendous commercial potentials [3]–[10], [12], [13]. Despite numerous research 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 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 order to adequately and flexibly identify and interpret meaning, we have to bridge the semantic gap between the richness of user semantics and the simplicity of available low-level perceptual visual and auditory feature (e.g., colors and motion in images, pitch and spectral shape of general sound, rhythm and harmonics in music). To address this issue, we have to develop various high-level semantic features and concepts such as "$Player$", "$Court/playing\ field$", "$Sporting\ event$", etc. These descriptions are meant to be clear to humans and we attempt to automate feature detection. To model real-world situations, the TREC video retrieval evaluation (TRECVID) [20] was proposed as an open, metrics-based evaluation to promote progress in content-based retrieval from digital video, where high-level feature extraction is one of three main tasks.

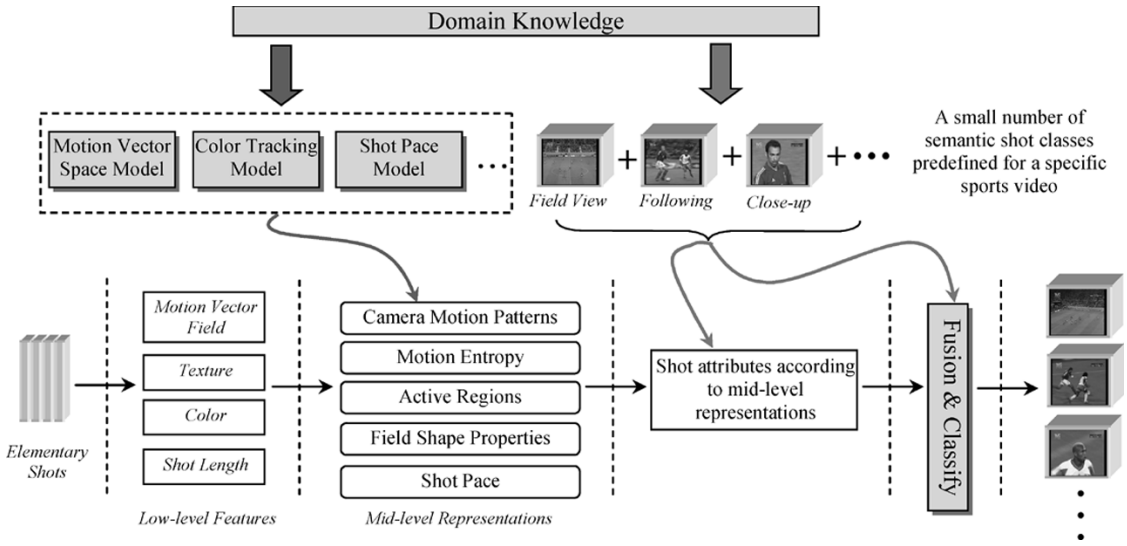


Fig. 1. Unified framework for semantic shot classification.

In this paper, we will present an effective high-level semantic concept, namely "semantic shot classes", which occurs frequently in broadcast sports video databases. Sport is governed by a set of rules or customs and often undertaken upon an open and level area with appropriate identifying colors. Most sports photograph exhibits various views from multiple angles and composition. It is the combination of shots that convey a message and helps viewers reconstruct activities as much as eyewitness observers would see them. Unlike movies stories, TV sports program exhibits limited and compact field production techniques since most athletic events take place in an area having specific dimensions. Moreover, most games consist of repeated actions accompanied by score of a competitive event. Therefore, it is feasible to identify a small number of shot classes, which cover a large percentage (greater than 90%) of a sports video. As the main goal of sports photograph is to follow sports actions, we can use domain knowledge to endue each shot class with some basic semantic meanings. According to extensive experimental observations, we have predefined a small number of shot classes for five field-ball type sports, i.e., tennis, basketball, volleyball, soccer, and table tennis. For example, we have identified eight shot classes for soccer video, i.e., $Replay$, $Field\ View$, $Player\ Following$, $Goal\ View$, $Player\ Close-up$, $Player\ Medium\ View$, $Audience$, and $Setting\ Bird\ View$. It is straightforward to find their corresponding semantic meanings, e.g., ($Player\ Close-up \rightarrow$ out of play), ($Field\ View$, $Player\ Following \rightarrow$ in play), ($Player\ Medium\ View \rightarrow$ Free Kick, Penalty Kick, Corner Kick, or Game Start), ($Goal\ View \rightarrow$ Corner Kick), ($Setting\ Bird\ View \rightarrow$ Game to be started), ($Replay \rightarrow$ Highlight, Foul). As a high-level semantic concept, the "*semantic shot classes*" is expected to act as an effective link between low-level video processing and high-level video content analysis.

Humans perceive semantic shot categories so accurately and with so little apparent effort. Yet it remains a difficult task how to create computer methods for automatic video shot class detection. The main challenge lies in the amount of variation in visual appearance. For example, playing fields/courts vary in size, shape, and coloring. An object (e.g., a player, a goalmouth, a net)'s orientation and distance from the camera affects its appearance. A more general difficulty is that visual information is ambiguous. Geometric ambiguity exists since the three dimensions of the world are projected onto two in the image. Moreover, video shots contain a large amount of image frames, each of which may carry important information. Computer power and memory limits us to use video information to its fullest extent. Fortunately, the physical world imposes constraints on the appearance of objects and the spatial relationship between an interesting object and its surroundings. As discussed above, sports video domain exhibits many constraints of interest to facilitate content analysis. In order to build a robust and flexible semantic shot classification system, we have to choose a set of effective representations within sports domain constraints and computing constraints.

In this paper, we will propose a unified framework for semantic shot classification in sports video, with an emphasis on knowledge representation and acquisition. As illustrated in Fig. 1, this framework consists of four stages: low-level feature extraction, mid-level representation, shot attributes production, and classification. Firstly, we derive low-level features (e.g., motion vectors field, texture map, dc images in compressed domain, and pixel-wise images in uncompressed domain) from video data. Secondly, we exploit nonparametric feature space analysis methods to perform the mapping from low-level features to mid-level representations such as camera motion patterns, action regions, field shape properties, etc. Thirdly, we use available mid-level representations to construct a feature vector for shot attributes' numerical description. Finally, we classify a shot into one of the predefined shot categories, in which various supervised learning algorithms can be used, such as decision trees, neural networks, support vector machines (SVMs), etc. [49] Note that the supervised learning procedure is constructed on the basis of effective mid-level representations rather than relies on blind training of large amounts of high-dimensional data.

A key issue in the design of complex media processing systems is the engineering of knowledge. These systems need to store information about the environment and objects of interest, in such a manner that an operational recognition scheme can be enacted [2]. We do not seek to match models (the models are often constructed by directly applying machine learning to low-level perceptual features) onto perceived information from the media data; rather the proposed framework is discriminatory. We attempt to do the least work necessary to discriminate objects, motion and production rules of interest from those items of no relevance. As a result, we do not have exhaustive feature extractors of low-level features; only those of semantic importance to the task, are included, namely, those intra-shot mid-level representations.

The use of machine learning is the means of acquiring task specific knowledge [4]. Alternatives to this approach include the use of human constructed knowledge [3]. The proposed framework takes advantage of the combination of machine learning and human constructed knowledge. As illustrated in Fig. 1, we use nonparametric clustering approaches to analyze motion vector space and perform adaptive field color tracking to generate motion and color related representations. On the other hand, the predefined semantic shot classes and the selection of appropriate mid-level representations are motivated by human constructed knowledge in sports domain.

The rest of this paper is organized as follows. Section II is a review of existing literatures devoted to content-based sports video analysis. In Section III, we introduce the predefined shot classes and their semantic meanings. In Section IV, we briefly discuss a feature space analysis method as the basis for our mid-level representations. In Section V, VI, and VII, we propose motion vector field model (MVFM), color tracking model (CTM), and shot pace model (SPM) to develop a common set of mid-level representations among field-ball type sports games. Section VIII explains the design of our semantic shot classifier. The experimental results are presented in Section IX. Finally, we conclude this paper in Section X.

II. Previous Work

The content of a video is intrinsically multimodal, since a content creator uses visual, auditory, and textural channels to convey meaning. In this section, we will review the state-of-the-art in sports video analysis according to various information modalities. To distinguish our work from other related works, we will discuss the roles of different statistical learning algorithms in structure analysis and semantics extraction.

A. State-of-the-Art

1) Visual Modality Based Techniques: Visual features are widely used in image/video indexing and retrieval [38], [39]. Various techniques have been developed based on color, texture, motion, shape, or a combination of them. An automatic soccer parsing system [8] was proposed to classify a sequence of frames into various play categories based on *a priori* model comprising line mark recognition, motion detection, and ball detection, etc. Camera motion was employed to annotate basketball videos [6]. In [9], the authors tried to classify frames into three kinds of views (global, zoom-in and close-up) and segment plays/breaks from the labels sequence. In [3], [28], simple color-based approaches were proposed to select tennis court clips from a raw tennis video. Besides frame-level works, shot-level parsing techniques were proposed [5], [23]. In [5], histograms were used to represent motion and color features of a shot for aggregating shots with similar low-level visual features. They tried to explain each cluster's semantic meanings. Domain knowledge was employed in [23] to perform a top-down classification of semantic video shots. Semantic shot categories were considered as a kind of mid-level representation to facilitate high-level analysis [24].

2) Auditory Modality Based Techniques: The auditory channel also provides strong clues for the presence of semantic events in video documents. Early work was done by [33] in trying to prioritize regions within a talk. In [32], audio analysis was primarily employed as an alternative tool for sports parsing. Their goal was to detect football touchdowns by spotting the key words "touchdown" or "fumble" and detecting "cheers". In [10], the authors tried to detect excited announcers' speech and ball hits from noisy and complex audio signals for extracting highlights in baseball videos. Hierarchial SVMs were employed in [21], [22] to train game-specific sound (e.g., "*Applause*", "*Whistling*", etc.) recognizers to detect events in tennis video. It was assumed that those sounds are closely related to interesting events with the help of specific sports game rules.

3) Textural Modality Based Techniques: Text in images and videos is one important source of high-level semantics. If these text occurrences could be detected, segmented, and recognized automatically, they would be useful for indexing and retrieval. With video OCR methods [35]–[37] the visual overlaid text object can be converted into a textural format, though the quality of the results varies. Reference [34] employed caption text detection and recognition to identify events in baseball videos. Closed caption (CC) text is another textural information source in broadcast video. It is a symbolic transcript of the speech part of the auditory stream, which is embedded in video signals as the textural stream. A method [29] was proposed to seek for time spans in which events are likely to take place through keywords extracted from the CC stream, and to index visual shots.

4) Multimodality Based Techniques: The integrated use of different information sources is a trend in video indexing. With the enhancements of content findings and more information available, video indexing results improve when a multimodal approach is followed. Reference [7] exploited heuristic rules to combine crowd cheer (auditory), score display (textural), and change in motion direction (visual) for detecting "Goal" segments in basketball videos. In [32], the authors used the line-marks and goal-posts to verify the results obtained by audio analysis in detecting football touchdowns. In [34], a video OCR method was heuristically combined with camera view recognition to discover the semantic events. An integration scheme of visual and auditory modalities was proposed to detect events in tennis video [21]. Beyond heuristic rules, hidden Markov models (HMMs) are frequently used as a statistical method for multimodal analysis, since it is not only capable to integrate multimodal features, but is also capable to incorporate sequential features [40]. A maximum entropy

method was used in [4] to integrate image, audio, and speech clues to detect and classify highlights from baseball videos. Recently, a probabilistic framework of multijects and multinets [41] was proposed to model the inter-conceptual relationships and integrate content elements by using a Bayesian Belief Network. A survey of multimodal video indexing methods is available in [26].

B. Statistical Learning in Video Structuring and Semantics Representation

Domain knowledge is useful for sports video indexing and retrieval [3], [6], [7], [21] in terms of easy implementation and computational efficiency. However, explicitly setting heuristic rules may not be easy in some cases, especially when the inference is based on a larger number of cues and complex context information. In recent years, more and more research efforts have focused on the roles of various statistical learning algorithms in video structuring and semantics modeling [4], [30], [31], [41]. HMMs were employed in [30] to model the structure *play/break* in broadcast soccer video. The maximum entropy principle was chosen to statistically model baseball highlights [4]. In [31], a Dynamic Bayesian Networks (DBNs) framework was proposed to represent temporal structure in video toward highlight extraction or violence detection. DBNs generalize HMMs by allowing the state space to be represented in factored form, instead of as a single discrete random variable. These works have explored the significant role of DBNs in the probabilistic representation of video structure and semantic concept, with an emphasis on modeling sequential data.

The modeling of sequential data incorporates video context and overcomes some disadvantages inherent to deterministic algorithms. It is different from key-frame based scene analysis and interpretation [19], [44], [45]. In [19], key-frames were used to classify sports video shots. They devised a frame-level visual feature to represent three kinds of shots: $playing\ field, player$, and $audience$. Unfortunately, dynamic characteristics within a shot were not sufficiently represented, which is required by fine classification of semantic sports video shots. The purpose of key-frames or salient images mainly lies in browsing [42], [44] and summarization [46] rather than semantic representation.

In this paper, we will present a new approach, semantic shot classification, for structuring sports video. Different from previous works [5], our shot classification system relies on the predefined shot categories obtained by manual observation of a specific sports video. The predefinition of shot categories have two advantages: 1) high classification accuracy and 2) clear semantic linkages between shot classes and potential events. A successful application of our semantic shot classification is an audio-visual integration scheme for detecting events in tennis videos [21]. In terms of statistical learning, our work is also different from DBN based methods [4], [30], [31]. We emphasize the construction of effective mid-level representations by combining human constructed knowledge and unsupervised learning algorithms. For each shot, we construct a feature vector by averaging measurements within a shot. SVMs are then employed to train classifiers. Unlike DBNs, the shot-level average operator is very simple and does not indicate any probability related to hidden states. However, promising results have been extensively achieved. As our focus is to construct effective mid-level representations, the promising result does not indicate whether DBNs is better or not in terms of semantic shot representation. In [24], semantic shot classes and audio keywords have been incorporated into a mid-level representation framework for semantic sports video analysis, where DBNs may model the labels sequences of shot classes and audio keywords to detect events.

III. Semantic Shot Classes

Camera shots are conventionally divided into three main categories: *long shot* (L.S.), *mid shot* (M.S.), and *close-up* (C.U.). The terms are usually used in connection with the human figure. A L.S. usually includes a subject's feet; a M.S. usually extends below the waist; and a C.U. does not include the hands. These three basic shots relate broadly to three different degrees of concentration. The shots can be divided into sub-categories such as "*Big close-up* (B.C.U.)", "*Medium close-up* (M.C.U.)", "*Medium long shot* (M.L.S.)", or "*Extreme long shot* (E.L.S.)" [47].

In this section, we will describe the predefined shot categories through manual observation of field-ball sports videos and formulate the problem of semantic shot classification.

A. Predefined Semantic Shot Categories

To get the best coverage of the game, photographers usually set up more than two cameras so that multiple angles can then be intercut during editing. All of the action, no matter how many cameras you use, must be strung together without any jump cuts. We introduce a general temporal model [24] and concretize semantic shot categories for a specific sport.

According to the focal distance and the main subject, we summarize the shots in eight classes (U_{1-6}, P_{1-2}) as shown in Fig. 2. By using these classes, we generally partition a sports video shot sequence into two logical segments, namely, *in play segment* (IPS) and *out of play segment* (OPS). IPS and OPS occur in successive turns. For a field-ball game, an IPS corresponds to the video segment when a ball is within the field boundaries and play has not been stopped by the referee. An OPS corresponds to the video segment when a ball is outside the field boundaries or play has been stopped by the referee. An IPS or OPS may comprise more than one shot. It is straightforward to derive the concept of "play/break" [9], [30] with IPS and OPS.

Fig. 3 illustrates the concretized shot categories for five typical field-ball games. We name a shot category C_i as follows:

$$C_i = \langle G_i, L_i \rangle,\ G_i \in \{U_j, P_k | j = 1, 2, \ldots, 6; k = 1, 2\}$$

where L_i denotes a linguistic description about the subject, and G_i denotes the class labels as listed in Fig. 2.

With the help of linguistic descriptions, we can derive semantic meanings by using domain knowledge. For example, ($\langle \mathrm{U}_4, \mathrm{Replay} \rangle \rightarrow$ Shot, Foul, Out of bound), ($\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle, \langle \mathrm{P}_1, \mathrm{Field\ View} \rangle$ and $\langle \mathrm{P}_2, \mathrm{Player\ Following} \rangle$) $\rightarrow$ in play, ($\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle \rightarrow$ out of play), ($\langle \mathrm{P}_1, \mathrm{Full\ Court\ Advance} \rangle \rightarrow$ Fast break, Drive), ($\langle \mathrm{P}_1, \mathrm{Penalty\ View} \rangle \rightarrow$ Throws induced by the

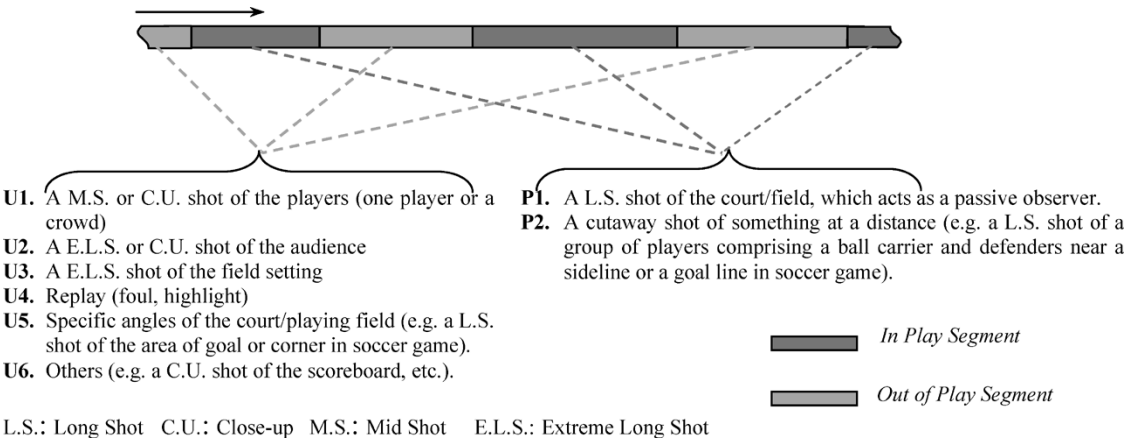


Fig. 2. General temporal model in broadcast sports video of field-ball type.

Fig. 3. Predefined shot classes for tennis, soccer, basketball, volleyball, and table tennis, along with the percentage of each class.

opponent's foul), ($\langle U_5, \text{Half Court View}\rangle \rightarrow$ To serve), ($\langle U_6, \text{Players \& Coach}\rangle \rightarrow$ Conversion between players and/or a coach during the pause requested by a coach), ($\langle U_2, \text{Audience}\rangle \rightarrow$ the response from crowds to actions) and ($\langle U_5, \text{Goal View}\rangle \rightarrow$ Corner Kick). Moreover, $\langle U_1, \text{Player Medium View}\rangle$ has specific meanings in different

games, i.e., "to serve" in tennis, "Free Kick, Penalty Kick, Corner Kick, or Game Start" in soccer, "jump ball" in basketball, etc.

B. Problem Formulation

Assume that each shot is represented by a vector $\vec{x} = \langle x_1, x_2, x_3, \ldots, x_n \rangle$, where $x_1, \ldots, x_n$ are the values of attributes $X_1, \ldots, X_n$. The problem of shot classification is to develop an algorithm which will assign any shot, represented by a vector $\vec{x}$, to one of predefined semantic shot categories, which we shall denote by C_k, $k = 1, 2 \ldots, M$. It is supposed that we are provided with a large number of sample video shots, which has already been classified by a human. Clearly, this is a multiclass classification problem.

The semantic shot classification essentially comprises two typical problems: 1) how to represent complex patterns and 2) how to establish decision boundaries excluding spurious (unstable) patterns. From the pattern recognition point of view, the first is treated as a kind of pre-processing operation, while the latter is solved by various learning algorithms such as Decision Tree, Neural Networks, SVMs, and naïve Bayesian classifier (NBC).

In many practical applications, the choice of pre-processing is the most significant factor in determining the performance of the final classification system. Preprocessing may take the form of a linear transformation of the input data. More complex preprocessing may involve reduction of the dimensionality of the input data. Another important way, in which classifier performance can be improved, sometimes dramatically, is through the incorporation of prior knowledge, which refers to relevant information which might be used to develop a solution and which is additional to that provided by the training data. Prior knowledge can either be incorporated into the classifier structure itself or into the preprocessing and post-processing stage [53]. This supports the design principle of our proposed framework; that is, we emphasize knowledge representation and acquisition by combining machine learning algorithms with human constructed knowledge in sports video domain. Instead of exhaustive training on high-dimensional data, nonparametric techniques are employed to construct visual mid-level representations of interest to the classification task. The supervised learning is then based on mid-level features.

IV. Mid-Level Representations

As shown in Fig. 1, we introduce a mid-level representation layer to bridge the gap between low-level visual features and high-level semantic concepts *semantic shot classes*. This layer differentiates our work from existing concept detection work [20] in two aspects. Firstly, our semantic concept model does not work directly on low-level features. We extract a set of semantic features by using nonparametric techniques. These features are compact and reasonably contribute to semantic shot classification. Secondly, although the shot classifiers have fused different shot attributes, this procedure is different from semantic context learning by Multinet [41], discriminate model fusion (DMF), or Ontology-based Boosing [20]. Our mid-level representations are lower than semantic concepts. The proposed concept is inherent to field-ball sports video. Compared with 18 concepts in TRECVID'03 [20] (e.g., outdoor, people, sports event, etc.), the concept of *semantic shot classes* is less generic but effective in sports video domain.

A. Low-Level Feature Versus Mode Seeking

Low-level feature extraction is the first step to semantic concept modeling. Low-level feature representation is fundamental to statistical learning for building concept. The histogram is the simplest and the most often used feature representation. It is often employed in combination with the Euclidean distance as a measure of dissimilarity, providing undemanding yet efficient retrieval method [5], [43]. However, the histogram representation does not match human perception very well and lacks discriminatory power in retrieval of large image and video databases. This weakness together with inefficient use of the data makes it necessary to use alternatives to histograms. Many studies have discovered that, when viewing the global color content, human visual system eliminates fine details and average colors within small areas. Hence, on the global level, humans perceive images only as a combination of few most prominent colors. These findings motivate us to address the issue of video representation from the viewpoint of dominant features in the context of groups of frames. Dominant features detection can be mapped to modes seeking in feature space. In our proposed framework, we employ the mean shift procedure to complete spatio-temporal features (i.e., motion and color) mode seeking. The mean shift procedure is derived by the density gradient estimation. Note that the discontinuity of histograms causes extreme difficulty if derivates of the estimates are required.

Assume that $X_1, \ldots, X_n$ is the given multivariate data set in the d-dimensional Euclidean space R^d. The multivariate kernel density estimator with kernel K and window width h is defined as

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left\{\frac{1}{n}(x - X_i)\right\}. \tag{1}$$

The term

$$M_h(x) \equiv \frac{1}{n_x} \sum_{X_i \in S_h(x)} [X_i - x] = \frac{1}{n_x} \sum_{X_i \in S_h(x)} X_i - x \tag{2}$$

is called the sample mean at $x \in X$. The repeated movement of data points to the sample means is called the *mean shift procedure* [54], [55].

The mean shift vector always points toward the direction of the maximum increase in the density. In [54], 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. Reference [16] developed a computational module of mean shift procedure, and successfully applied it to two low-level vision tasks: discontinuity preserving filtering and image segmentation.

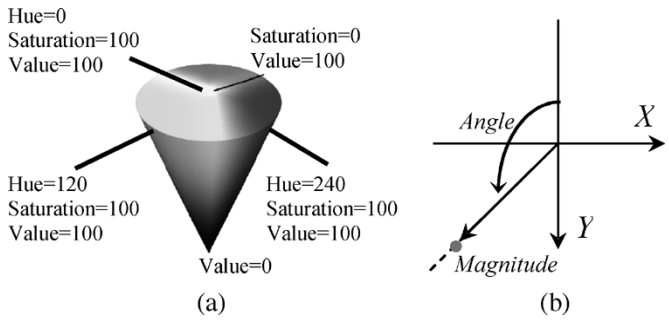


Fig. 4. (a) Cone-shaped MVS space. (b) Polar coordinates of a motion vector.

V. Motion Vector Field Model (MVFM)

Motion analysis plays a significant role in the interpretation of video shot content, particularly for sports video [5], [6], [23]. Most motion estimation approaches rely on a parametric model and dense motion field estimation [14], [15]. However, it is hard to achieve reliable model estimation in large amounts of video data. This failure is due to the violation of parametric assumption in the presence of large object motion and bad estimation of optical flow fields in low-textured regions. Particularly in sports video, rich camera movements, frequent large object motion and low textured regions (e.g., playing field in soccer) jointly cause the arbitrarily structured MVF.

Our proposed MVFM is to characterize MVF from the nonparametric clustering point of view. MVFM treats MVF characterization as the problem of feature space analysis. *Clustering* is an effective way to learn the structure of multidimensional patterns from a set of unlabeled samples. Arbitrarily structured feature space can be analyzed only by nonparametric methods as these methods do not come with embedded assumptions. Among numerous nonparametric clustering methods we choose the kernel density based clustering approach of mean shift procedure. MVFM has provided five descriptors: entropy, pan, tilt, diagonal, and active region.

In this section, we discuss MVFM from three aspects: motion display, MVF filtering, and descriptors.

A. Motion Display

According to motion vector characteristics and HSV parameters range, we propose a cone-shaped MVS space to represent a motion vector. The MVS visualizes the MVF and provides us with a visual aid to understand, analyze, and compare different kinds of motion characteristics. Hue represents motion direction, $saturation$ represents motion intensity, and $value$ represents confidence. Fig. 4(a) illustrates the MVS space. Fig. 4(b) shows the polar coordinates $Angle$ (the radial coordinate) and *Magnitude* (the angular coordinate) of a motion vector. The conversion of a motion vector to MVS is done as

$$\begin{aligned} Hue &= Angle, \quad 0 \le Angle < 360 \\ Saturation &= \begin{cases} \frac{255*Magnitude}{Mag_{\text{th}}}, & \text{if } Magnitude < Mag_{\text{th}} \\ 255, & \text{if } Magnitude \ge Mag_{\text{th}} \end{cases} \\ Value &= \begin{cases} \frac{255*Texture}{Tex_{\text{th}}}, & \text{if } Texture < Txt_{\text{th}} \\ 255, & \text{if } Texture \ge Txt_{\text{th}} \end{cases} \end{aligned} \tag{3}$$

where Mag_{th}, Tex_{th} are normalizing thresholds. $Texture$ measure is obtained by computing high-frequency energy according to the variance of wavelet coefficients in the high-frequency bands, or by computing ac energy from DCT accoefficients. The confidence measure $value$ relates to the intuition that a high-textured region should produce a "good" motion vector. Spatial confidence or temporal confidence measures [25] may be developed to perform confidence processing. Fig. 5 shows some examples of motion display in MVS.

B. MVF Filtering

In MVS space, the MVF is a two-dimensional lattice of three-dimensional vectors (Hue, $Saturation$, and $Value$). The space of the lattice is known as the spatial domain, while the color is represented in the range domain. In order to consider the spatial consistency of magnitude and direction, we have to concatenate the location and range vectors in the joint spatial-range domain of five dimensions. Our MVFM employs the mean shift algorithm to perform MVF filtering in the joint domain. A series of MVF filtering results are shown in Fig. 5. After filtering, we have decomposed an MVF into homogeneous colored tiles. The recognition of a motion pattern is to capture the spatio-range composition knowledge of colored tiles for each predefined motion pattern from training samples [50]. Currently, our MVFM employs only simple measures to characterize an MVF. However they suffice for shot classification. Readers are referred to [50] for more details on the learning based nonparametric motion characterization.

The MVF filtering is expected to remove noise and preserve salient information by using the local structure in feature space. The mean shift procedure has excellent discontinuity preserving smoothing performance and simple control parameter with clear physical meaning (i.e., the kernel bandwidths determine various spatial and range resolutions for analysis). We thus propose the filtering scheme as follows.

Stage 1): The mean shift procedure is employed to smooth the MVF. The kernel (window) is moved in the direction of the maximum increase in the joint density gradient. The joint domain kernel is defined as the product of two radically symmetric kernels and the Euclidean metric is employed

$$K_{h_s,h_r} = \frac{C}{h_s^2 h_r^3} k\left(\left\|\frac{\mathrm{X}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathrm{X}^r}{h_r}\right\|^2\right) \tag{4}$$

where X^s and X^r are the spatial part and range part, respectively, $k(x)$ is the normal kernel used in both two domains, h^s and h^r are the kernel bandwidths, and C is the normalization constant.

Stage 2): The outcome of the mean shift filtering is fed through a watershed algorithm [17], yielding the delineation of the clusters in the joint domain. Small spatial regions are easy to eliminate though post-processing.

Stage 3): We heuristically select significant clusters for analyzing MVF characteristics.

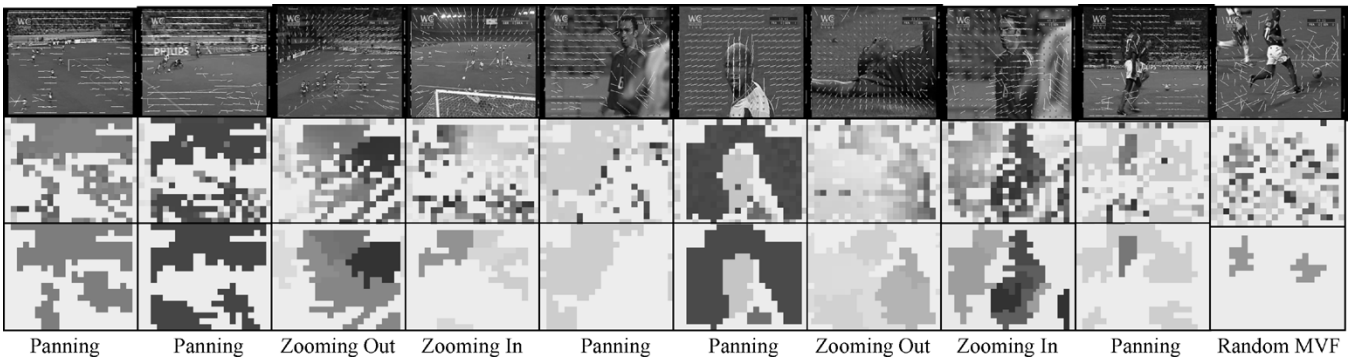


Fig. 5. Examples of mean shift-based MVF representation. First row lists frames overlapped with a MVF; second row displays the MVF in MVS space; third row lists the representations obtained by MVF filtering.

C. Entropy of MVF

Assume 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 MVF is

$$H = -\sum_{i=1}^{m} P_i \log_2 P_i, \quad P_i = \frac{R_i}{\sum_{j=1}^{m} R_j} \tag{5}$$

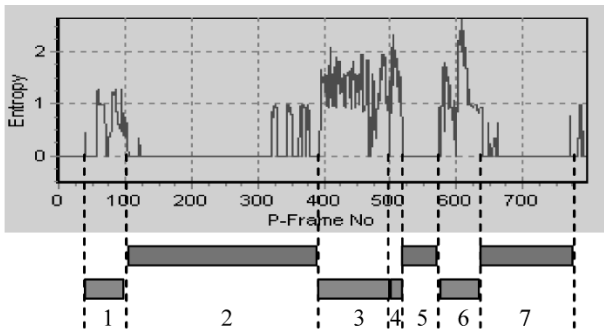
where P_i denotes the probability of the cluster C_i. The entropy H is a measure of the randomness or unpredictability of MVF.

A close-up shot usually features high entropy as player induced or camera induced movements may increase the MVF's uncertainty. Within a wide-angle shot, the entropy is often low as dominant camera movements (e.g., Pan, Tilt) tend to produce a uniform MVF. However, the Zooming may yield high entropy even within a wide-angle shot due to the diversity of motion vectors' direction and magnitude (see Fig. 5). The entropy measure thus contributes to video shot classification. For example, tennis video shots can be roughly classified into $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ and $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ through comparing the entropy with a suitable threshold. However, the entropy alone is insufficient to robust shot classification on extensive sports videos with more complex motion and structure (e.g., soccer and basketball, etc.). Thus, a coarse yet effective camera motion pattern analysis algorithm is needed, which relies on the clusters themselves, not on their probability only.

Fig. 6 illustrates an entropy curve computed from a series of P-Frames in an MPEG compressed tennis video. Clearly, $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ shots exhibit higher entropy than $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shots. Within the second and seventh $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shots, we observe large slopes, which are caused by the camera's following action in the presence of rapid and extended exchanges of a player's position. Within the sixth $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ shot, the peak indicates a Zooming.

D. Camera Motion Patterns

Our MVFM employs a quantization scheme to roughly estimate camera motion patterns. Based on preliminary results in [23] and extensive observations, we summarize three major motion factors of interest to shot classification and event detection: 1) the direction and duration of Pan and Tilt; 2) the strength of local motion caused by a foreground object; and 3) the variation of Pan, Tilt, and local motion within a shot.



<U$_1$, Player Close-up> Shots: 1, 3, 4, and 6.
<P$_1$, Court View> Shots: 2, 5, and 7.

Fig. 6. Entropy curve computed from the MVFs extracted from an MPEG compressed tennis video sequence.

We consider five typical patterns: pan left (PL), pan right (PR), tilt up (TU), tilt down (TD), and diagonal (DL). The pattern DL is associated with camera zooming and rotating. DL practically indicates the erratic MVF in case of a camera's fast following action or a foreground object's motion in a B.C.U. shot. Currently we simply use DL to represent complex motion characteristics in such cases.

Let $G = \{C_i\}_{i=1\ldots m}$, $C_i = \langle R_i, \overline{Mag}_i, \overline{Ang}_i \rangle$ denotes the clusters obtained by the MVF filtering, where the number of motion vectors associated with cluster C_i is R_i, $\overline{Mag}_i$ is the average magnitude of C_i, $\overline{Ang}_i$ 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 as follows:

$$r_k = \left[(-1)^{k+1}\alpha + \left\lfloor \frac{k}{2} \right\rfloor \cdot 90^\circ, (-1)^k \alpha + \left\lfloor \frac{(k+1)}{2} \right\rfloor \cdot 90^\circ \right)$$
$$k = 0 \cdots 7, \quad \alpha = 15^\circ. \tag{6}$$

Three camera motion rates are computed as shown in (7), at the bottom of the next page.

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

Fig. 7 illustrates the Pan, Tilt, Diagonal rates computed from a series of P-Frames in an MPEG compressed volleyball video. Within the second $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shot, there is no prominent motion as the camera is still and awaits game start.

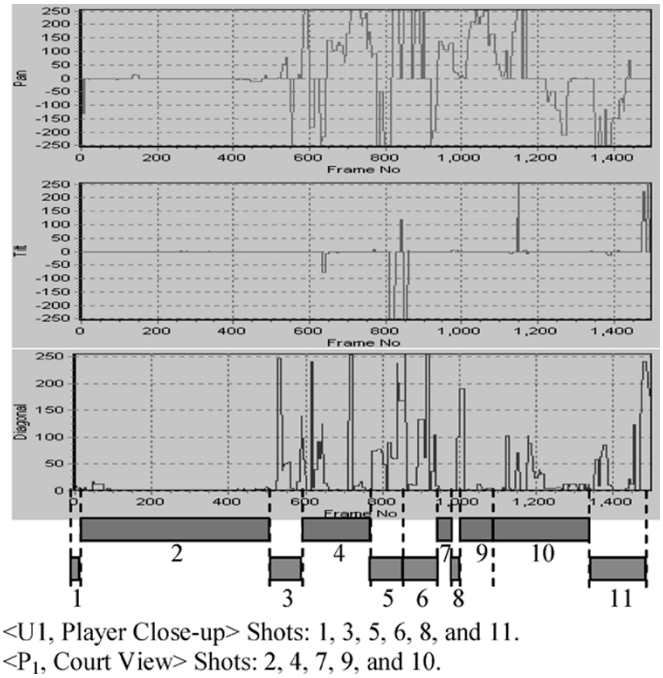


Fig. 7. Camera motion rates computed from the MVFs extracted from an MPEG compressed volleyball video sequence.

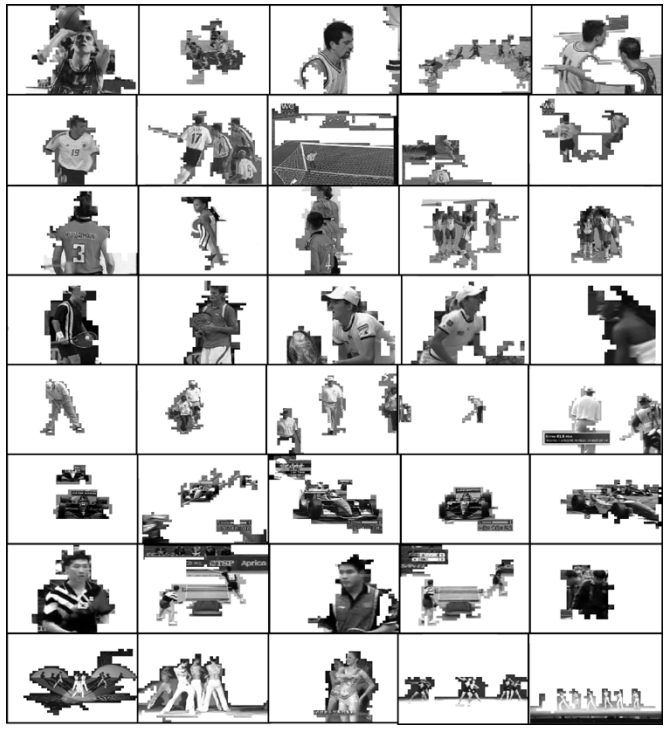


Fig. 8. Examples of active regions extracted from eight different sports video sequences.

The serve and embrace actions cause strong DL patterns within the third and fifth $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ shots, respectively. The fourth $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shot has such a transition pattern $(\mathrm{DL}, \mathrm{PL}) \rightarrow \mathrm{PR\ (persistent)} \rightarrow (\mathrm{PL}, \mathrm{DL})$, which corresponds to pass, hit, defense, and failed pancake. The sixth to ninth shots involve a $\langle \mathrm{U}_4, \mathrm{Replay} \rangle$ shot. The seventh and ninth $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shots feature persistent PR (offense), which are separated by the eighth $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ shot (split block) with a prominent DL. The tenth $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shot is another serve-receive-pass-spike procedure. Different from the fourth shot, the tenth $\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$ shot starts with PR, which means side out. We can notice a segment between PR and PL has zero pan rate but large diagonal rate as an assist triggers the following action in the diagonal direction.

E. 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 [57]. Moreover, motion plays an important role in focusing of attention within perceptions. Therefore we may combine the MVS filtering scheme 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. Fig. 8 shows some examples of active regions. Normally an active region corresponds to foreground objects such as a player or a gathering of players. In a general context, the property of an active region can be developed for different tasks, e.g., team classification, referee detection, etc. Compared with entropy and camera motion patterns, the active region is more semantic, but otherwise more domain constraints are introduced.

Basically the extraction of an active region is done in three stages: 1) we perform MVF filtering; 2) we heuristically select the seed region from the center to the periphery based on the region's texture and shape features; 3) we investigate other homogeneous regions belonging to the same cluster as the seed region. For those regions satisfying shape requirements, we will consider them as active regions together with the seed region. At the second step, we consider two heuristic rules: 1) despite various camera movements, a foreground object is always located

$$
\begin{aligned}
Pan &= \sum_{i=1}^{m} \left((-1)\frac{P_i}{4} \cdot \overline{Mag}_i \cdot R_i \right) / \sum_{i=1}^{m} R_i, P_i \in \{0,4\} \\
Tilt &= \sum_{i=1}^{m} \left((-1)\left(\frac{P_i - 2}{4} - 1 \right) \cdot \overline{Mag}_i \cdot R_i \right) / \sum_{i=1}^{m} R_i, P_i \in \{2,6\} \\
Diagonal &= \sum_{i=1}^{m} (\overline{Mag}_i \cdot R_i) / \sum_{i=1}^{m} R_i, P_i \in \{1,3,5,7\}.
\end{aligned} \tag{7}
$$

around the frame center and 2) a foreground object usually features much higher texture than the court/playing field. Undoubtedly, lots of heuristic rules have been incorporated. Moreover, it is assumed that MVF filtering can delineate semantic objects. These limits motivate us to develop an autonomous algorithm of active regions extraction in future work. Readers are referred to [60] for the illustration of active region extraction.

We want to mention that our semantic shot classification does not use any active region's feature currently. The active region is considered as a natural continuation of MVS space analysis.

VI. Color Tracking Model

Color is very useful in locating and recognizing objects that occur in artificial environment. Sports is governed by a set of rules or customs and often undertaken upon an open and level area with appropriate identifying colors. An adaptive color characterization plays a significant role in interpreting "*What/Where*" from broadcast sports video.

The term "*Tracking*" differentiates our CTM from traditional color histogram. Our CTM employs sports video structure constraints (i.e., a limited set of camera views) and the spatial constraints (i.e., a uniform field setting). The temporal "*tracking*" of distinguishing colors is performed to capture semantic concepts (e.g., "$Playing\ Field$", "$Player\ Clothing$", "$Audience\ Crowd$", etc.).

CTM consists of two components: *dominant color selector* (DCS) and *field color probability map tracker* (FPMT). DCS component combines the frame-based spatial features clustering and the temporal features clustering to seek modes for representing those colors of semantic importance to content analysis. The spatial-temporal mode seeking enables DCS to represent multimodal court/field colors in a nonparametric way. FPMT component performs pose tracking of a court/playing field within a shot. FPMT outputs a series of field color probability maps (FPM). We employ geometric moment functions of the FPM to generate various descriptors for representing view coverage and court/field pose.

A. DCS

In this paper, we define a dominant color as "a particular kind of color that is most characteristic of a sports video and usually determines the presence, appearance, and spatial relationships of objects (e.g., a playing field/court, a stand, a player, etc.) of semantic importance to sports scene understanding". There are two challenges to an autonomous and robust dominant color selector: Firstly, we have to design good algorithms capable of perceiving a stable perception of color over varying lighting conditions; secondly, we are not provided with any 'class label'. Color space conversion can be used to solve the first challenge to some extent. The latter is said to be the incomplete data problem. There are some training methods to solve it, i.e., maximum likelihood, EM and stochastic sequential estimation [53].

Mixture models are a natural choice for interpreting the "dominant" concept. In [23], we employed Gaussian mixture models (GMM) to estimate the density of region-based color values given a training video sequence. An on-line K-means approximation of an exact EM algorithm was used to train the GMM. We order the Gaussian distributions based on the mixing parameters and variances. The most likely dominant color distributions remain on the top and less probable distributions are eventually replaced by new distributions. This approach was used to model the field colors in soccer videos.

However, GMM based modeling work has one major limitation. In [23], we have assumed that the uniform field region is dominant in image frames. Motion was used to remove distracting shots followed by selecting candidate field regions. The qualified regions were finally used for GMM training. Although this procedure is well tuned for field modeling in soccer video, it is not generic enough for various dominant colors (e.g., players' clothing color, stand color) characterization of other sports.

Our DCM consists of four stages as follows [58].

Stage 1): Spatial Feature Clustering: We employ the joint domain kernel

$$K_{h_s,h_r} = \frac{C}{h_s^2 h_r^3} k\left(\left\|\frac{\mathrm{X}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathrm{X}^r}{h_r}\right\|^2\right) \tag{8}$$

to perform mean shift clustering of color pixels within each image frame, where X^s and X^r are the spatial part and range part, respectively, h_s and h_r are the kernel bandwidths, C is the normalization constant. According to the clustering results, each image frame F_j can be represented by

$$F_j = \{A_i\}_{i=1,\ldots,m}, \quad A_i = \langle o_i, \overline{c}_i, r_i \rangle \tag{9}$$

where o_i denotes the pixels or motion vectors associated with the cluster A_i, $\overline{c}_i$ is the average color of o_i, r_i is the normalized cluster size of A_i, $\sum_{i=1}^{m} r_i = 1$, $0 \le r_i \le 1$.

Stage 2): Temporal Feature Clustering: We employ the joint domain kernel (8) to perform temporal mean shift clustering of spatial modes

$$\begin{aligned} S &= \{V_{ij}\}_{j=1,\ldots,k;i=1,\ldots,m(j)} \\ V_{ij} &= \langle \overline{c}_{ij}, r_{ij} \rangle \end{aligned} \tag{10}$$

obtained from the spatial feature clustering on a series of image frames $\{F_j\}_{j=1,\ldots,k}$, where $m(j)$ denotes the number of clusters in F_j, V_{ij} denotes the i^{th} spatial cluster in the j^{th} frame. Differently the range part is the mode feature and the spatial part is the mode percentage. A proper normalization is employed to compensate their different natures. Suppose the outcome of temporal feature clustering contains M clusters $\{C_\ell | \ell = 1, \ldots, M\}$, where C_ℓ contains N_ℓ feature points. We have $\sum_{\ell=1}^{M} N_\ell = N$, N is the total number of feature points. $N = \sum_{j=1,\ldots,k} m(j)$. Finally, we get spatio-temporal modes as follows:

$$\mathrm{MODE} = \{Mode_\ell\}_{\ell=1,\ldots,M}, \quad Mode_\ell = \langle O_\ell, Y_\ell, N_\ell \rangle \tag{11}$$

where O_ℓ, Y_ℓ, and N_ℓ denotes the feature points, the cluster center, and the number of feature points in C_ℓ, respectively.

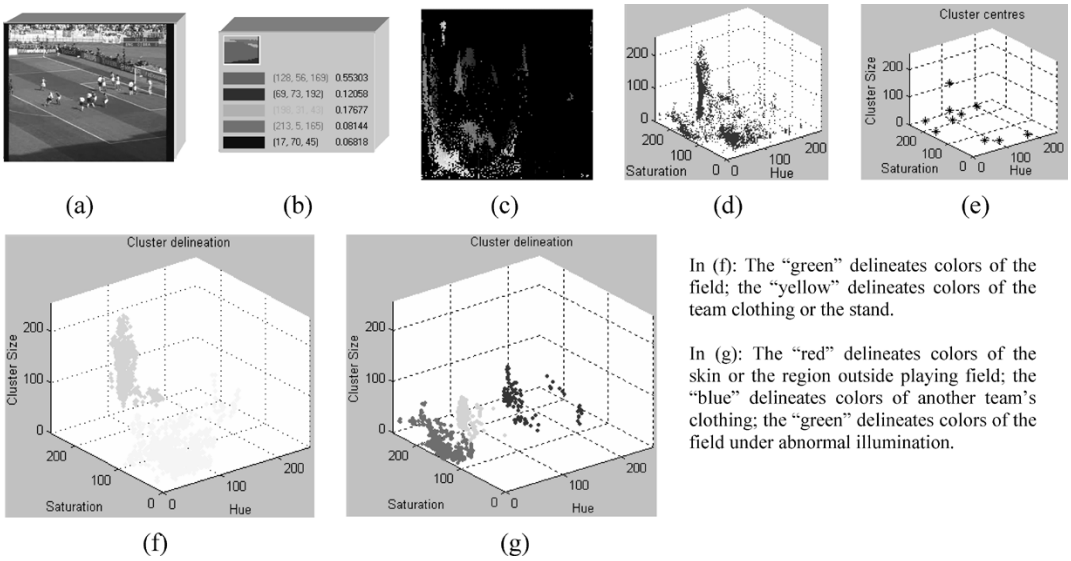


Fig. 9. Example of DCS in a soccer video sequence. (a) Training data comprising a series of continuous video frames; (b) data preprocessing step; (c) 2-D visualization of data points after preprocessing ("horizontal axis" is Hue, "vertical axis" is Saturation; (d) 3-D visualization of data distribution; (e) the clusters centers after mean shift filtering; (f) dominant modes with large ratios; and (g) dominant modes with small ratios but of semantic importance.

Stage 3): KNN classification: The KNN rule is employed to classify A_i into one of $\{Mode_\ell\}_{\ell=1,\ldots,M}$. Thus we may map a frame F_j into a set of modes, namely

$$\text{MODE}(F_j) = \{\{m_i\}_{i=1,\ldots,n} | m_i \in \{Mode_\ell\}_{\ell=1,\ldots,M}, 0 \le n \le M\}. \quad (12)$$

We draw a hyper sphere around the point of color $\langle \overline{c}_i, r_i \rangle$, which encompasses K points irrespective of their class labels. Suppose that this sphere, of volume V, contains K_k points from class C_k. Then we approximate the class-conditional densities in the form $p(\overline{c}_i, r_i|C_k) = K_k/(N_k V)$, and the unconditional density in the form $p(\overline{c}_i, r_i) = K/NV$, the priors $P(C_k) = N_k/N$. According to Bayes' theorem, we have $p(C_k|\overline{c}_i, r_i) = (p(\overline{c}_i, r_i|C_k)P(C_k))/p(\overline{c}_i, r_i) = K_k/K$. Thus, to minimize the probability of misclassification $\langle \overline{c}_i, r_i \rangle$ should be assigned to the class C_k for which the ratio K_k/K is largest.

Stage 4): Heuristically Selecting of Modes of Semantic Importance to a Task: It is feasible to exploit *a priori* knowledge to heuristically select those color modes associated with an interesting concept. Selected modes are used to determine the presence of interesting colors by using KNN rule. Cluster size is useful for selecting color modes as it directly corresponds to the concept of "dominant". The "size" factor is in two ways: 1) the size of clusters from spatial features clustering and 2) the size of clusters from temporal feature clustering. Moreover, domain knowledge can be introduced to heuristically select modes even if their associated cluster size is not "dominant" enough. Fig. 9 illustrates the production of a DCM model.

Fig. 10. Examples of FPM. The top row lists original frames; the bottom row visualizes the resulting FPMs by mapping the probability values from [0, I] to [0, 255].

B. Field Color Probability Map (FPMT)

DCS is to adaptively characterize colors in the context of a structured video. Sometimes there is a need to track the appearance variation of special regions within a shot. Particularly sports action usually occurs on a confined "ground" (i.e., playing field/court region). Once we can track it, geometric moment functions can be employed to represent its appearance change. Thus we propose an FPMT model to perform field tracking.

The FPMT works on an FPM. The definition of FPM is similar to the histogram backprojection method [18]. Given an image with color distribution $f(x, y)$, and $h(c)$ denotes the histogram function that maps a color c to a histogram bin. Let I denote an image histogram, and M a target model histogram. The FPM can be written as

$$\text{FPM}(x, y) = R_{h(f_{x,y})}, \quad R_j = \min\left(\frac{M_j}{I_j}, 1\right). \quad (13)$$

Fig. 10 gives some FPM examples of soccer and tennis video.

Our proposed FPMT consists of three stages, as follows.

Stage 1): Detecting color modes associated with a playing field /Court: DCS is employed to decide

whether there exist any dominant color modes to track within a shot. If yes, we select a small number of frames within this shot to compute the model histogram. An initial patch is selected within the first frame. The initial mean location is computed by the zero- and first-order moment of FPM within the initial patch.

Stage 2): Tracking the mode of ratio histogram: For the next frame, we compute the ratio histogram by (13). The mean shift procedure is used to seek the ratio histogram's mode. According to resulting modes and search window, we update the modal histogram and compute new FPM. New mean location is computed.

Stage 3): Adjusting search window size: We center the search window at the new mean location and adjust the search window size according to the zero-order moment of the new FPM. Go to *Stage 2*.

In this way, the initial seed patch can grow to encompass the playing field/court after several iterations.

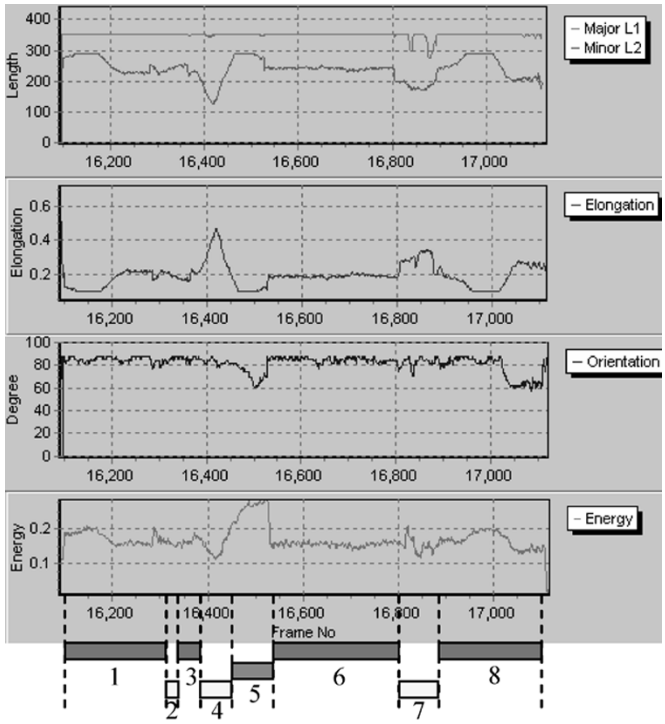
C. Geometrical Moment Representation

Geometrical moment functions is employed to represent the FPM's shape.

Geometrical moments are defined with basis set $\{x^p y^q\}$. The $(p+q)^{\text{th}}$ order two-dimensional geometric moments are defined by $m_{pq} = \int\int_{\zeta} x^p y^q f(x,y)dxdy$, $p,q = 0,1,2\ldots$, where ζ is the region of the pixel space in which the density function $f(x,y)$ is defined. We use the zero-, first-, and second-order moments to represent the FPM shape as follows:

$$\begin{aligned}
&\text{Position} : (x_0, y_0) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right)\\
&\text{Area} : A = m_{00}\\
&\text{Orientation} : \vartheta = \frac{arctan\,(b,(a-c))}{2}\\
&\text{Length of major axis} : l_1 = \sqrt{\frac{\left(a+c+\sqrt{b^2+(a-c)^2}\right)}{2}}\\
&\text{Length of minor axis} : l_2 = \sqrt{\frac{\left(a+c-\sqrt{b^2+(a-c)^2}\right)}{2}}\\
&\text{Elongation} : E = \frac{(l_1-l_2)}{(l_1+l_2)},\\
&\text{where } a = \frac{m_{20}}{m_{00}} - x_0^2, \quad b = 2\left(\frac{m_{11}}{m_{00}} - x_0 y_0\right),\\
&\qquad c = \frac{m_{02}}{m_{00}} - y_0^2. \qquad (14)
\end{aligned}$$

An example is given in Fig. 11. The energy measure is obtained from the 1-D field-players interaction curve (FPIC) [23]. We try to interpret those curves. Within the fifth $\langle \mathrm{U}_1, \text{Player Close} - \text{up}\rangle$ shot, a large energy is due to FPIC's large variance in top-down player close-up views. Compared with the first and eighth $\langle \mathrm{P}_1, \text{Field View}\rangle$ shots, the shape measure of sixth $\langle \mathrm{P}_1, \text{Field View}\rangle$ shot are more stable. Within the first $\langle \mathrm{P}_1, \text{Field View}\rangle$ shot, to follow the "goal kick" action,



<P$_1$, Field View> Shots: 1, 3, 6, and 8.
<P$_2$, Player Following> Shots: 2, 4, and 7.
<U$_1$, Player Close-up> Shot: 5.

Fig. 11. Set of shape characteristics curves of eight successive soccer video shots computed by the FPMT.

camera movement leads to the transition of view coverage away from the goal area to the midfield, which increases elongation and decreases energy. Within the eighth $\langle \mathrm{P}_1, \text{Field View}\rangle$ shot, a prominent increase of elongation along with an orientation change is associated with the switch from the midfield passing in the sixth $\langle \mathrm{P}_1, \text{Field View}\rangle$ shot to the goal area offense in the eighth $\langle \mathrm{P}_1, \text{Field View}\rangle$ shot. Within the fourth and seventh $\langle \mathrm{P}_2, \text{Player Following}\rangle$ shots, horizontal views cause lower energy but a sharp increase of elongation.

VII. Shot Pace Model

SPM is motivated by the fact that there is a distinguishable and consistent shot length difference between major shot classes (i.e., $\langle \mathrm{P}_1, \text{Field View}\rangle$ versus $\langle \mathrm{U}_1, \text{Player Close} - \text{up}\rangle$) in team-based sports video (e.g., basketball, soccer, volleyball, etc.). It can be interpreted in two aspects: Firstly, a game played by two teams of more than three players mainly relies on the teammate cooperation to make an offense and defense, which leads to looser structure than tennis and table tennis; secondly, a photographer tends to use a wide shot ($\langle \mathrm{P}_1, \text{Court View}\rangle$ or $\langle \mathrm{P}_1, \text{Field View}\rangle$) to follow actions and use a close-up shot to track a player or a gathering of people, which makes a wide shot's length longer than its neighbor close-up shots. The shot rhythm for content identification was explored in unstructured movie data [11]. Our SPM focuses on the sports video domain.

We employ a sliding window to examine the lengths of m successive shots. Assume $S_i, i = 1,2,\ldots,N$ denote a series of successive shots. We may declare a wide shot

($\langle \mathrm{P}_1, \mathrm{Court\ View} \rangle$, or $\langle \mathrm{P}_1, \mathrm{Field\ View} \rangle$) if the shot length of S_i is maximum within a symmetric sliding window of size m and is also n times of the second largest maximum within this sliding window. This criterion is practically infeasible since it is difficult to select appropriate parameters of m and n.

Thus, our SPM takes a soft method. We calculate the rate $\overline{SLR}$ of each shot's length to the maximum shot length within a symmetric sliding window. This simple measure can be combined with other features discussed above to achieve various tasks including semantic shot classification.

VIII. The Design of Semantic Shot Classifer

In this section, we will construct a feature vector for shot attributes' numerical description and train the semantic shot classifiers in accordance with the predefined shot categories.

A. Feature Choice

The choice of distinguishing features is a critical design step and depends on the characteristics of the problem domain. Instead of exhaustively extracting high-dimensional low-level features, we employ the mid-level representations to construct a common set of features for the field-ball type sports video. Currently the feature vector consists of nine dimensions, i.e.,

$$\langle \overline{Entropy}, \overline{|Pan|}, \overline{|Tilt|}, \overline{Diagonal}, \overline{SLR}, \overline{Elongation}, \overline{Orientation}, \overline{Area}, \overline{Energy} \rangle \quad (15)$$

The first four dimensions are average motion measurements computed by MVFM within a shot. The last four dimensions are average color measurements computed by CTM within a shot. The fifth dimension is computed by SPM.

Color measurements are obtained by DCS. Equation (14) considers the dominant color modes associated with the playing field/court only. As those measurements are based on geometric moment functions, they do not involve any color ranges. They provide numerical descriptions of semantic concepts (i.e., camera view coverage and field poses under various views). For a new sports video, we sample a segment of around 2.5 min from the whole video to train the DCS (the DCS in Fig. 9 is constructed by training 5000 continuous image frames, $5000/(29.97 \times 60) = 2.78$ min). We employ the learned DCS to detect a playing field/court. If yes, we compute the measurements by geometric functions in (14); otherwise we set those four measurements to zero.

The computation of motion measurements is based on the MVFs of P-frames extracted in MPEG compressed domain. Texture measure is obtained by computing ac energy of ac DCT coefficients. The recovery of ac DCT coefficients can be achieved by using the manipulation method in [59].

MVFM and CTM automatically process video data and deliver a series of feature values for all frames within a shot. An "average" operator is applied to get a feature vector of nine dimensions for representing shot attributes. The "average" greatly reduces the dimensionality of the input space. The shot-level average measurements are shown to suffice for semantic shot classification.

Finally, we have to perform the scaling of feature values except $\overline{Entropy}$, $\overline{SLP}$, $\overline{Elongation}$, $\overline{Energy}$ as follows:

$$\begin{aligned} &\overline{|Pan|} : [0,255] \rightarrow [0,1]; \quad \overline{|Tilt|} : [0,255] \rightarrow [0,1]; \\ &\overline{Diagonal} : [0,255] \rightarrow [0,1]; \quad \overline{Orientation} : [0,90] \rightarrow [0,1]; \\ &\overline{Area} \rightarrow \frac{\overline{Area}}{Sample\ Frame\ Size}. \end{aligned} \quad (16)$$

B. Classifier Training

Two learning approaches are used, namely, NBC and SVMs. Through performance comparison, we evaluate the mid-level representations in terms of efficiency and noise insensitivity. Moreover, we notice the data unbalance (as listed in Fig. 3) and the limited availability of labeled data of minor classes such as $\langle \mathrm{U}_3, \mathrm{Setting\ Bird\ View} \rangle$, $\langle \mathrm{U}_6, \mathrm{Player\ \&\ Coach} \rangle$, and $\langle \mathrm{U}_2, \mathrm{Audience} \rangle$. Two ways are taken to solve it. Firstly, we combine those minor shot classes to form a new category for classification, and further perform classification within this category. Secondly, *cross-validation* is adopted for model selection instead of the *hold out* method [53].

1) Using a Naíve Bayesian Classifier: We have to estimate the *priori* probability $P(C_k)$ and the class-conditional densities $p(x_i|C_k)$. $P(C_k)$ is set to be the percentage of each class listed in Fig. 3. The densities $p(x_i|C_k)$ are estimated by histogram.

Given an origin x_0 and a bin width h, we define the histogram bins as the intervals $[x_0 + mh, x_0 + (m+1)h)$ for integers m. $\hat{f}(x) = 1/nh(no.\ of\ X_i\ in\ same\ bin\ as\ x)$ defines the histogram. We set the origin to zero. An appropriate bin width is determined by cross-validation. We select bin width parameters as: $h = [2^{-2}, 2^{-3}, 2^{-4}, \ldots, 2^{-9}]$. In theory, for each problem we have to try $8^9 = 134\,217\,728$ combinations. Practically, we let $\overline{Entropy}$, $\overline{|Pan|}$, $\overline{|Tilt|}$, $\overline{Diagonal}$ use the same bin width and $\overline{SLP}$, $\overline{Elongation}$, $\overline{Orientation}$, $\overline{Area}$, $\overline{Energy}$ the same bin width. Thus we have to try $8 \times 8 = 64$ combinations. A 10-fold cross-validation is conducted on the training data.

NBC is easy to construct simply by estimating $P(x_i|C_k)$ on training examples. Intuitively, this might be inaccurate, since the conditional independence assumption rarely holds true. Many empirical comparisons have shown that Naïve Bayes predicts just as well as C4.5 [48].

2) Using Support Vector Machines: C-Support Vector Classification (C-SVC) [49] (Binary Case) is employed. We train all datasets only with the radial basis function (RBF) kernel: $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$. Two parameters ($C$ and γ) are considered. We consider C and γ as the same for all binary problems. Different kernel parameters are used to estimate the accuracy

$$C = [2^{-2}, 2^{-1}, \ldots, 2^{10}], \gamma = [2^{-10}, 2^{-9}, \ldots, 2^{4}].$$

For each problem, we have to try $13 \times 15 = 195$ combinations. A tenfold cross validation is conducted on the training data.

SVMs were originally designed for binary classification. There are two types of approaches for multiclass SVM. One is

Fig. 12. Two sets of SDVE indicating the replay scenes.

by constructing several binary classifiers while the other is by directly considering all data in one optimization formulation, "one-against-all", "one-against-one", and DAGSVM [51]. We employ the "one-against-one" [52] approach in which we train a classifier for each possible pair of classes. For M classes, this results in $(M-1)M/2$ binary classifiers. In classification, we use a voting strategy: each binary classification is considered to be a voting where votes can be cast for all data points so that each point is designed to be in a class with the highest number of votes. As some works have shown "one-against-all" does not perform as well as "one-against-one", we do not consider it.

3) A Special Shot Class: Replay: One of the post production techniques is to insert replay scenes in broadcast sports video. The main purpose is to provide an especially significant or interesting event. A replay scene is a reliable indicator of sports highlights.

A replay scene usually consists of several shots. In our shot classification system, we think of all of the shots within a replay scene as a special shot class Replay. In general, it might be very difficult to classify scenes into either live or replay by means of slow motion analysis. We resort to the detection of a special digital video effect (SDVE) inserted at the beginning and the end of a replay scene. The overlapped "flying graphics" is a typical kind of SDVE, as shown in Fig. 12. To robustly represent a SDVE, we collect a set of SDVE video segments for training, and perform mode seeking to capture dominant colors that best describe the overall color appearance. Then we employ a sliding window technique and the earth mover distance (EMD) [61] to perform similarity matching over the video data.

This method is through feature space analysis on training segments. It does not rely on any robust shot segmentation, key frame extraction, or complex logo tracking. It is more flexible and generic. For more details, readers are referred to [62].

IX. Experimental Results

We collect video data from TV by using SONY digital video camera recorder. Snazzi III USB2 is used to transfer video data from tape to PC in MPEG-1(352 × 288, 2.00 Mbps, 25.00 fps, Audio 44.1 kHz, 16 bits, Stereo). The shot boundary is detected by MGI VideoWave 4.0 [56].

A. Replay Detection

Our proposed replay detection approach has been tested on four matches of soccer video from the 2002 FIFA World Cup. Table I lists the performance in terms of precision and recall. As illustrated in Fig. 13, the representation of "flying graphics" are trained from 54 replay scenes in SEN-FRA match. The duration

TABLE I
Performance on Replay Detection

Match	Total	Correct	Recall (%)	Precision (%)
GER-KOR (25/06/02)	67	65	97.0	94.2
GER-BRA (30/06/02)	33	30	90.9	85.7
SEN-TUR (22/06/02)	48	46	95.8	92.0
SEN-FRA (31/05/02)	54	52	96.3	89.7

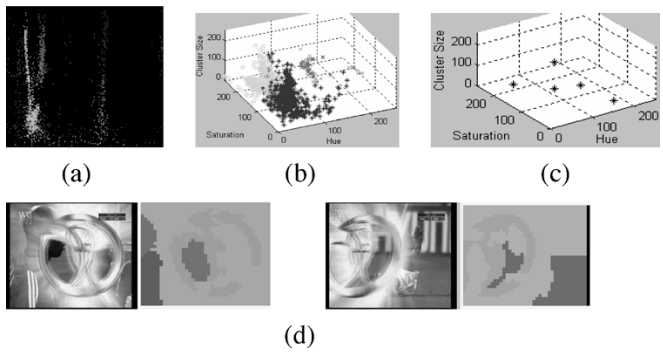


Fig. 13. Mode-based representations of flying graphics in the broadcast video of 2002 FIFA World Cup. (a) 2-D visualization of feature points from 54 replay scenes; (b) delineation of dominant color modes; (c) cluster centers after spatio-temporal mode seeking; (d) sample images along with spatial clustering results, jointly represented by the modes delineated in (b).

TABLE II
Experimental Result on Tennis Video by SVMs

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

of each SDVE segment is about 0.49 s on average. The similarity matching on four matches are all based on the trained SDVE model shown in Fig. 13(b). A promising performance, *Recall* 90%–97% and *Precision* 85%–95%, has been achieved.

B. Shot Classification in Tennis Video

The total length of tennis video is about 120 min (1350 shots) consisting of *2002 Western & South Financial Group Masters HEWITT vs. MOYA* (32 min), *2003 Australia Open Men's Singles Semifinal FERREIRA vs. AGASSI* (21 min), *2003 French Open Women's Final Henie vs. Kim* (17 mins), *2003 French Open Men's Final Costa vs. Ferrero* (50 min).

Table II and III summarize the performances on testing data by using SVMs and NBC. We choose two thirds for training and one-third for testing. A tenfold cross-validation is performed on the training data. This setting is also applicable to the following experiments. For SVMs, we choose $C = 2^3, \gamma = 2^{-2}$ with the best cross-validation rate of 94.7%. For NBC, we choose the

TABLE III
EXPERIMENTAL RESULT ON TENNIS VIDEO BY NBC

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

TABLE IV
EXPERIMENTAL RESULT ON SOCCER VIDEO BY SVMS

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

bin widths 2^{-6} (for motion measurements) and 2^{-8} (for color measurements) with the best cross-validation rate of 89.9%.

C. Shot Classification in Soccer Video

The total length of soccer video is about 200 min (2500 shots) consisting of *2002 FIFA World Cup FRA vs. SEN (May 31)* (80 min), *ENG vs. BRA (Jun.21)* (60 min), *GER vs. BRA (Jun 30)* (60 min).

Tables IV and V summarize the performances on testing data. We choose $C = 2^4, \gamma = 2^{-2}$ with the best cross-validation rate of 90.3% for SVMs. For NBC, we choose the bin widths 2^{-6} (for motion measurements) and 2^{-8} (for color measurements) with the best cross-validation rate of 84.8%.

D. Shot Classification in Basketball Video

The total length of basketball video is about 50 min (375 shots) consisting of *2003 NBA Detroit Pistons vs. Orlando Magic* (30 min), *New Jersey Nets vs. Milwaukee Bucks* (20 min).

Tables VI and VII summarize the performances on testing data. We choose $C = 2^3, \gamma = 2^{-2}$ with the best cross-validation rate of 96.2% for SVMs. For NBC, we choose the bin widths 2^{-4} (for motion measurements) and 2^{-7} (for color measurements) with the best cross-validation rate of 93.8%.

E. Shot Classification in Volleyball Video

The total length of volleyball video is about 65 min (550 shots) consisting of *2002 Volleyball Masters Cuba vs. Netherlands* (Jun. 5) (65 min).

Table VIII and IX summarize the performances on testing data. We choose $C = 2^3, \gamma = 2^{-1}$ with the best cross-validation rate of 93.2% for SVMs. For NBC, we choose the bin

TABLE V
EXPERIMENTAL RESULT ON SOCCER VIDEO BY NBC

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

TABLE VI
EXPERIMENTAL RESULT ON BASKETBALL VIDEO BY SVMS

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

TABLE VII
EXPERIMENTAL RESULT ON BASKETBALL VIDEO BY NBC

Shot Class	Total	Correct	Recall (%)	Precision (%)
<P_1, Full Court Advance>	30	28	93.3	90.3
<P_1, Penalty View>	12	12	100.0	85.7
<U_1, Player Close-up>	48	41	85.4	93.2
<U_2, Audience>	11	8	72.7	66.7
<U_1, Player Medium View>	6	5	83.3	83.3
<U_3, Setting Bird View>	3	3	100.0	100.0

widths 2^{-5} (for motion measurements) and 2^{-7} (for color measurements) with the best cross-validation rate of 89.7%.

F. Shot Classification in Table Tennis Video

The total length of table tennis video is about 70 min (810 shots) consisting of *2001 Men's Team Championship Tournament Semi Final OH Sang Eun vs. Liu Guozheng* (70 min).

Table X and XI summarize the performances on testing data. We choose $C = 2^2, \gamma = 2^{-2}$ with the best cross-validation rate of 92.6% for SVMs. For NBC, we choose the bin widths 2^{-4} (for motion measurements) and 2^{-6} (for color measurements) with the best cross-validation rate of 88.5%.

G. Discussion

Referring to Tables II–XI, we have come up with several points, as follows.

1) Overall, major shot classes exhibit higher accuracy than minor shot classes. In particular, we have achieved

TABLE VIII
EXPERIMENTAL RESULT ON VOLLEYBALL VIDEO BY SVMS

Shot Class	Total	Correct	Recall (%)	Precision (%)
<P_1, Court View>	67	63	94.0	90.0
<U_1, Player Close-up>	88	80	90.9	93.0
<U_2, Audience>	5	4	80.0	80.0
<U_5, Half Court View >	8	7	87.5	100.0
<U_6, Players & Coach>	3	3	100.0	100.0

TABLE IX
EXPERIMENTAL RESULT ON VOLLEYBALL VIDEO BY NBC

Shot Class	Total	Correct	Recall (%)	Precision (%)
<P_1, Court View>	67	63	91.0	89.7
<U_1, Player Close-up>	88	80	89.8	91.9
<U_2, Audience>	5	4	80.0	66.7
<U_5, Half Court View >	8	7	87.5	87.5
<U_6, Players & Coach>	3	3	100.0	100.0

TABLE X
EXPERIMENTAL RESULT ON TABLE TENNIS VIDEO BY SVMS

Shot Class	Total	Correct	Recall (%)	Precision (%)
<P_1, Court View>	86	81	94.2	89.0
<U_1, Player Close-up>	156	145	92.9	96.0
<U_2, Audience>	5	4	80.0	80.0
<U_5, Half Court View >	4	4	100.0	100.0
<U_6, Players & Coach>	3	3	100.0	100.0

TABLE XI
EXPERIMENTAL RESULT ON TABLE TENNIS VIDEO BY NBC

Shot Class	Total	Correct	Recall (%)	Precision (%)
<P_1, Court View>	86	78	90.7	81.3
<U_1, Player Close-up>	156	137	87.8	93.8
<U_2, Audience>	5	4	80.0	80.0
<U_3, Setting Long View>	4	4	100.0	100.0
<U_6, Players & Coach>	3	3	100.0	100.0

quite good accuracy (*recall* 93.3%–96.9% and *precision* 89.0–100.0%) for P_1 shots. This is due to the regularity of field-ball type sports video and the uniform color and motion attributes within P_1 shots. The promising performance on major shot classes is significant for sports video structuring and interesting event location.

2) The accuracy of $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ is comparatively lower (5% on average) than P_1 shots. When we try to use the feature vector (18) to distinguish $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$, it is implicitly assumed that this shot class lacks in a uniform and persistent MVF, and dominant field colors. However this assumption might be invalid when the player remains completely still, or a $\langle \mathrm{U}_2, \mathrm{Audience} \rangle$ shot is a close-up and features an erratic MVF. An alternative is to seek more effective representations of $\langle \mathrm{U}_1, \mathrm{Player\ Close-up} \rangle$ shot.

3) We have successfully accomplished the classification of minor but important shots with good performance, such as $\langle \mathrm{U}_5, \mathrm{Goal\ View} \rangle$, $\langle \mathrm{P}_1, \mathrm{Penalty\ View} \rangle$, $\langle \mathrm{U}_5, \mathrm{Half\ Court\ View} \rangle$. It benefits from the capability of FPM-based geometric moments to represent playing field/court appearance under various views.
4) By using NBC, we have achieved promising results comparable to that of SVM. This fact indicates our representations are effective and insensitive to noise. The incorporation of knowledge contributes to the classification after all.
5) As SPM model is involved in representing a shot, the accuracy of shot segmentation indirectly affects performance. Although our target is to construct a uniform feature vector, we can employ feature selection to favor high performance for a certain shot category. We will evaluate different features' roles in classifying different shot categories in future work.
6) Currently the complexity of our proposed system mainly lies in the construction of mid-level representations. The system works in four rounds. Compressed domain feature extraction, the construction of DCS model and the computation of motion measurements are completed at round 1–3, respectively. Each round works at more than real-time speed on the Pentium 1000-MHz PC. Classifier training is done within round 4. It is completed less than 10 s.

X. CONCLUSIONS

We have presented an effective high-level semantic concept of "semantic shot classes", which frequently occurs in broadcast sports video. In order to detect this concept, we have proposed a unified framework for semantic shot classification, with an emphasis on knowledge representation and acquisition. This framework relies on mid-level representations instead of exhaustive low-level features. Experiments have shown that an appropriate construction of mid-level representations can improve the accuracy and flexibility of shot classification.

The proposed mid-level representations can be extended to other sports video analysis work such as event detection, highlight extraction etc. The marriage of machine learning algorithms and human constructed knowledge proves to be effective for deriving mid-level representations. We have justified the proposed mid-level representations through the task of video shot classification. Our future work includes the evaluation of individual features for various tasks.

Our proposed unified framework has assumed that it is feasible to predefine a set of shot classes with a large coverage for a specific sports video. This assumption is valid for most field-ball type sports video with prominent structure constraints. However, there do exist large amounts of sports video with a loose structure such as golf, racing. An alternative is to apply a shot clustering approach to the mid-level representations.

REFERENCES

[1] Teleexperience: communicating compelling experience, in Keynote Speech: ACM Multimedia, Ottawa, ON, Canada, Sep. 30, 2001.

[2] T. Caelli and W. F. Bischof, *Machine Learning and Image Interpretation*. New York: Plenum, 1997, pp. 189–224.

[3] D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in *Proc. Int. Conf. Multimedia & Expo*, Tokyo, Japan, Aug. 22, 2001, pp. 920–923.

[4] M. Han, W. Hua, W. Xu, and Y. H. Gong, "An integrated baseball digest system using maximum entropy method," in *Proc. ACM Multimedia*, Juan-les-Pins, France, Dec. 1, 2002, pp. 347–350.

[5] C. W. Ngo, T. C. Pong, and H. J. Zhang, "On clustering and retrieval of video shots," in *Proc. ACM Multimedia*, Ottawa, ON, Canada, Sep. 30, 2001, pp. 51–60.

[6] 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 Trans. Circuits Syst. Video Technol.*, vol. 10, no. 1, pp. 133–146, Feb. 2000.

[7] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos," in *Proc. ACM Multimedia*, Ottawa, ON, Canada, Sep. 30, 2001, pp. 261–269.

[8] Y. H. Gong, L. T. Sin, C. H. Chuan, H. J. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc. Int. Conf. Multimedia Computing and Systems*, Washington, DC, May 15, 1995, pp. 167–174.

[9] P. Xu *et al.*, "Algorithms and systems for segmentation and structure analysis in soccer video," in *Proc. Int. Conf. Multimedia & Expo*, Tokyo, Japan, Aug. 22, 2001, pp. 184–187.

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

[11] B. Adams, C. Dorai, and S. Venkatesh, "Finding the beat: an analysis of the rhythmic elements of motion pictures," *Int. J. Image and Graph.*, vol. 2, no. 2, pp. 215–245, 2002.

[12] K. A. Peker, R. Cabasson, and A. Divakaran, "Rapid generation of sports highlights using the MPEG-7 motion activity descriptor," in *Proc. SPIE Storage and Retrieval for Media Database*, vol. 4676, San Jose, CA, Jan. 2002, pp. 318–323.

[13] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. Int. Conf. Image Processing*, vol. 1, Barcelona, Spain, Sep. 14, 2003, pp. 5–8.

[14] M. J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise-smooth flow fields," *Comput. Vis. and Image Understanding*, vol. 6, no. 4, pp. 348–365, 1995.

[15] J. M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *J. Vis. Commun. and Image Repres.*, vol. 6, no. 4, pp. 348–365, 1995.

[16] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 1–18, 2002.

[17] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 6, pp. 583–589, 1991.

[18] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[19] J. Assfalg, M. Bertini, C. Colombo, and A. D. Bimbo, "Semantic annotation of sports videos," *IEEE Multimedia*, vol. 9, no. ., pp. 52–60, Junl 2002.

[20] Website.. [Online]http://www-nlpir.nist.gov/projects/tvpubs/papers/ibm.concept.detect.slides.pdf

[21] M. Xu, L.-Y. Duan, C.-S. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, Apr. 6, 2003, pp. 189–192.

[22] M. Xu, N. C. Maddage, C.-S. Xu, M. Kankanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Proc. Int. Conf. Multimedia & Expo*, Baltimore, MD, Jul. 6, 2003, pp. 281–284.

[23] L.-Y. Duan, M. Xu, and Q. Tian, "Semantic shot classification in sports video," in *Proc. SPIE Storage and Retrieval for Media Database*, vol. 5021, San Jose, CA, Jan. 2003, pp. 300–313.

[24] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proc. ACM Multimedia*, Berkeley, CA, Nov. 1, 2003, pp. 33–44.

[25] R. Wang, H. J. Zhang, and Y. Q. Zhang, "A confidence measure based moving object extraction system built for compressed domain," in *Proc. Int. Symp. Circuits and Systems*, Geneva, Switzerland, May 2000, pp. 21–24.

[26] C. M. Snoek and M. Worring, Multimodal video indexing: A review of the state-of-the-art, in ISIS Tech. Rep. Series, Univ. Amsterdam, Amsterdam, The Netherlands, vol. 2001-20, Dec. 2001.

[27] C. Dorai *et al.*, "Media semantics: who needs it and why?," in *Proc. ACM Multimedia*, Juan-les-Pins, France, Dec. 1, 2002, pp. 580–583.

[28] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Proc. IEEE Int. Workshop Content-Based Access of Image and Video Database*, Bombay, India, Jan. 1998, pp. 81–90.

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

[30] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, FL, May 13, 2002, pp. 4096–4099.

[31] A. Mittal, L.-F. Cheong, and T.-S. Leung, "Dynamic bayesian framework for extracting temporal structure in video," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, Kauai, HI, Dec. 11, 2001, pp. 110–115.

[32] Y.-L. Chang, W. Zeng, I. Kamel, and R. Alonso, "Integrated image and speech analysis for content-based video indexing," in *Proc. Int. Conf. Multimedia Computing and Systems*, Hiroshima, Japan, Jun. 17, 1996, pp. 306–313.

[33] B. Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *Proc. Int. Conf. Spoken Language Processing*, Yokohama, Japan, 1994, pp. 1931–1934.

[34] D. Q. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. ACM Multimedia*, Juan-les-Pins, France, Nov. 1, 2002, pp. 315–318.

[35] Y. Zhong, H.-J. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 385–392, Apr. 2000.

[36] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing," *Multimedia Syst. J.*, vol. 8, no. 1, pp. 69–81, 2000.

[37] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 147–156, Jan. 2000.

[38] M. Flickner *et al.*, "Query by image and video content: the QBIC system," *IEEE Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[39] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE Multimedia*, vol. 4, no. 3, pp. 12–20, Sep. 1997.

[40] A. A. Alantan, A. N. Akansu, and W. Wolf, "Multi-modal dialogue scene detecting using hidden markov models for content-based multimedia indexing," *Multimedia Tools and Applic.*, vol. 4, no. 2, pp. 137–151, 2001.

[41] M. R. Naphadem and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 141–151, Mar. 2001.

[42] H. J. Zhang, C. Y. Low, and S. W. Smoliar, "Video parsing and browsing using compressed data," *Multimedia Tools and Applic.*, vol. 1, no. 1, pp. 89–111, 1995.

[43] A. K. Jain, A. Vailaya, and X. Wei, "Query by video clip," *Multimedia Syst.*, vol. 7, pp. 369–384, 1999.

[44] B.-L. Yeo, "On fast microscopic browsing of MPEG-compressed video," *Multimedia Syst. J.*, vol. 7, no. 4, pp. 269–281, 1999.

[45] Y. Zhuang *et al.*, "Adaptive key frame extraction using unsupervised clustering," in *Proc. Int. Conf. Image Processing*, Chicago, IL, Oct. 4, 1998, pp. 866–870.

[46] Y. F. Ma, L. Lu, H. J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Multimedia*, Juan-les-Pins, France, Dec. 1, 2002, pp. 533–542.

[47] D. Cheshire, *The Complete Book of Video—Techniques, Subjects, Equipment*. London, U.K.: Dorling Kindersley, 1990.

[48] P. Langley, W. Iba, and K. Thomas, "An analysis of bayesian classifiers," in *Proc.10th Nat. Conf. Artificial Intelligence*, San Jose, CA, 1992, pp. 223–228.

[49] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[50] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu, "Mean shift based nonparametric motion characterization," in *Proc. Int. Conf. Image Processing*, Singapore, Oct. 24, 2004, pp. 1597–1600.

[51] J. C. Platt, N. Gristianini, and J. S. Taylor, "Large margin DAG's for multiclass classification," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 547–553.

[52] U. Krebel, "Pairwise classification and support vector machines," in *Advances on Kernel Methods—Support Vector Learning*. Cambridge, MA: MIT Press, 1999, pp. 255–268.

[53] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[54] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 8, pp. 790–799, Aug. 1995.
[55] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.
[56] Roxio.. [Online]http://www.roxio.com/en/products/videowave/
[57] V. Cantoni, S. Levialdi, and V. Robert, *Artificial Vision*. New York: Academic, 1997, pp. 1–52.
[58] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu, "Nonparametric color characterization using mean shift," in *Proc. ACM Multimedia*, Berkeley, CA, Nov. 1, 2003, pp. 243–246.
[59] S. F. Chang and D. G. Messerschmitt, "Manipulation and compositing of MC-DCT compressed video," *IEEE J. Select. Areas Commun.*, vol. 13, no. 1, pp. 1–11, Jan. 1995.
[60] "I2R Tech. Rep.," Inst. Infocomm Research, http://www1.i2r.a-star.edu.sg/~lingyu/report/shotclassification.pdf, Jun. 2003.
[61] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 59–66.
[62] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu, "Mean shift based video segment representation and applications to replay detection," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, May 17, 2004, pp. 709–712.

Ling-Yu Duan received the B.E. degree in applied physics from Dalian University of Technology (DUT), Dalian, China, in 1996, the M.S. degree in automation from University of Science and Technology (USTC), Hefei, China, in 1999, and the M.S. degree in computer science from the National University of Singapore (NUS), Singapore, in 2002. He is currently pursuing the Ph.D. degree in the School of Design, Communication, and Information Technology, University of Newcastle, NSW, Australia.

He is a Research Scientist at the Institute for Infocomm Research, Singapore. His current research interests include image/video processing, multimedia, computer vision and pattern recognition, and machine learning.

Min Xu received the B.E. degree in automation from the University of Science and Technology of China (USTC) in 2000, and the M.S. degree in computer science from the National University of Singapore (NUS) in 2003.

She is a Research Associate in the Center of Multimedia and Networking Technologies, School of Computer Engineering, Nanyang Technological University, Singapore. Her current research interests include audio/video processing and multimedia.

Qi Tian (S'83–M–86–SM'90) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree from the University of South Carolina, Columbia, all in electrical and computer engineering.

He is a Principal Scientist at the Institute for Infocomm Research, Singapore. His main research interests are image/video/audio analysis, multimedia content indexing and retrieval, computer vision, pattern recognition, and machine learning. He is also an Adjunct Professor at Beijing University, Beijing, China. He joined the Institute of System Science, National University of Singapore, in 1992 as a Research Staff, and was subsequently the Program Director for the Media Engineering Program at the Kent Ridge Digital Labs, then Laboratories for Information Technology, Singapore (2001–2002). In 1985, he was a post-doctorate Researcher at the University of California, San Diego. He has published over 110 papers in peer reviewed international journals and conferences. He has two U.S. patents issued and four pending.

Dr. Tian served as an Associate Editor for the IEEE TRANSACTION ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (1997–2003), and served as a Chair and a member of technical committees for several international conferences, including the IEEE Pacific-Rim Conference on Multimedia (PCM), the IEEE International Conference on Multimedia and Expo (ICME), and Multimedia Modeling (MMM).

Chang-Sheng Xu (M'97–SM'99) received the Ph.D. degree in electric engineering from Tsinghua University, Beijing, China, in 1996.

From 1996 to 1998, he was a Research Associate Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He joined the Institute for Infocomm Research (12R), Singapore, in March 1998. Currently, he is Head of the Media Analysis Lab of 12R. His research interests include multimedia content analysis, indexing and retrieval, digital watermarking, computer vision, and pattern recognition. He has published more than 100 papers in those areas.

Jesse S. Jin (M'98) was born in Beijing, China, in 1956. He received the B.E. degree from Shanghai Jiao Tong University, Beijing, China, the M.E. degree from China Textile University (CTU), China, and the Ph.D. degree from the University of Otago, Otago, New Zealand, in 1982, 1987, and 1992, respectively, all in computer science.

He held academic positions at the University of Otago, the University of New South Wales, NSW, Australia, and the University of Sydney, Sydney, Australia, and is the Chair Professor of Information Technology (IT) in the School of Design, Communication, and IT, University of Newcastle, NSW. He has published 175 articles, has one patent, and is in the process of filing three more patents. His research interests include multimedia, medical imaging, computer vision and pattern recognition.

Dr. Jin is a member of ACM, ASTS, and IS&T. He established a spin-off company which won the 1999 ATP Vice-Chancellor New Business Creation Award.