

# A Survey of Content-Aware Video Analysis for Sports

Huang-Chia Shih, *Member, IEEE*

**Abstract**—Sports data analysis is becoming increasingly large scale, diversified, and shared, but difficulty persists in rapidly accessing the most crucial information. Previous surveys have focused on the methodologies of sports video analysis from the spatiotemporal viewpoint instead of a content-based viewpoint, and few of these studies have considered semantics. This paper develops a deeper interpretation of content-aware sports video analysis by examining the insight offered by research into the structure of content under different scenarios. On the basis of this insight, we provide an overview of the themes particularly relevant to the research on content-aware systems for broadcast sports. Specifically, we focus on the video content analysis techniques applied in sportscasts over the past decade from the perspectives of fundamentals and general review, a content hierarchical model, and trends and challenges. Content-aware analysis methods are discussed with respect to object-, event-, and context-oriented groups. In each group, the gap between sensation and content excitement must be bridged using proper strategies. In this regard, a content-aware approach is required to determine user demands. Finally, this paper summarizes the future trends and challenges for sports video analysis. We believe that our findings can advance the field of research on content-aware video analysis for broadcast sports.

**Index Terms**—Action recognition, content-aware system, content-based video analysis, event detection, semantic analysis, sports video analysis, survey.

## I. INTRODUCTION

RESEARCH interest in sports content analysis has substantially increased in recent decades, because of the rapid growth of video transmission over the Internet and the demand for digital broadcasting applications. The massive commercial appeal of sports programs has become a dominant focus in the field of entertainment. Research on big data analytics has attracted much attention to machine learning and artificial intelligence techniques. Accordingly, content analysis of sports media data has garnered attention from various studies in the last decade. Sports data analysis is becoming large scale, diversified, and shared. The most pressing problem currently is how to access the most important information in a short time. Because of the massive demand for sports video broadcasting, many enterprises such as Bloomberg, SAP, and

Vizart employ sports content analytics. Content analysis with big data has used become a major emerging industry. In offline service, historical records can be used to analyze video content through machine learning. In online service, discovered latent knowledge can be for real-time tactic recommendation. In recent years, many books have contributed to the content analysis of statistics for baseball [1]–[3] and basketball [4]–[6]. At present, several sports intelligence systems and content analytics have been developed and applied.

- 1) Panasonic incorporated SAP SE to develop a video-based sports analytics and tracking system. Match Insights is an analytics prototype SAP developed with the German Football Association for the soccer World Cup 2014 in Brazil [7].
- 2) Vizrt, short for visualization (in) real time, is a Norwegian company that provides content production, management, and distribution tools for the digital media industry. Its products include applications for creating real-time 3D graphics and maps, visualizing sports analyses, managing media assets, and obtaining single workflow solutions for the digital broadcast industry. Vizrt has a customer base in more than 100 countries and approximately 600 employees distributed at 40 offices worldwide [8], [9].
- 3) PITCHf/x data set is a public resource presented by MLBAM [10] and Sportvision [11]. Brooks Baseball [12] makes systematic changes to this data set to improve its quality and usability. They manually review the Pitch Info using several parameters of each pitch's trajectory and validate the parameters against several other sources such as video evidence (e.g., pitcher grip and catcher signs) and direct communication with on-field personnel (e.g., pitching coaches, catchers, and the pitchers themselves). The default values of the trajectory data are slightly altered to align them more closely with the real values.
- 4) Sportradar [13], a Swiss company, focuses on collecting and analyzing data related to sports results by collaborating with bookmakers, national soccer associations, and international soccer associations. Their operating activities include the collection, processing, monitoring, and commercialization of sports data, which result in a diverse portfolio of sports-related live data and digital content.

## A. Surveys and Taxonomies

Since 2000, sports video analysis has continually drawn research attention, leading to a rapid increase in published

Manuscript received November 6, 2016; revised January 5, 2017 and January 15, 2017; accepted January 17, 2017. Date of publication January 19, 2017; date of current version May 4, 2018. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST105-2221-E-155-066 and in part by the Innovation Center for Big Data and Digital Convergence under Grant 228196. This paper was recommended by Associate Editor J. M. Martinez.

The author is with the Department of Electrical Engineering, Yuan Ze University, Taoyuan 32003, Taiwan (e-mail: hcshih@saturn.yzu.edu.tw).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSVT.2017.2655624

work and surveys in this field. In [14], a preliminary survey of sports video analysis was conducted to examine diverse research topics such as tactic summarization, highlight extraction, computer-assisted referral, and content insertion. Kokaram *et al.* [15] demonstrated the development trends in the topics of sports-related indexing and retrieval. They identified two broad classes of sports, court and table sports and field sports. However, this was not a favorable classification for content analysis. An alternative classification criterion, the time- or point-driven structure in the progress of a game, was then considered. Sports such as baseball and soccer are time driven because the high-excitement events occur sparsely and randomly. Conversely, point-driven sports consist of regular events and constructs within a domain-specific scenario. In these types of sports, such as tennis and snooker, highlight events yield particular points in the game structure. Different from the viewpoint of Kokaram *et al.* [15], Santiago *et al.* [16] focused on the application of team tracking techniques to sports games. They categorized the reviewed systems into two classes, intrusive and nonintrusive. In intrusive systems, tags or sensors are placed on the targets, whereas nonintrusive systems introduce no extra objects to the game environment, instead using vision as the main sensory source and employing image processing techniques to perform player and team tracking. Intrusive systems are typically not considered highly mature in providing high-level information. A domain-specific survey of soccer video analysis was presented in [17]. The authors reviewed studies on soccer video analysis at three semantic levels of interpretation: video summarization, provision of augmented information, and high-level analysis. In addition, the computational problems at these three semantic levels were discussed. First, the two main tasks of video summarization include extracting the most interesting parts of the video and omitting the less interesting parts. Studies on event detection and highlight extraction were reviewed from the viewpoints of features, models, and classifiers. Second, studies on provisions of augmented information for presenting the game status to viewers were discussed. Third, extension of the level of video summarization was proposed. Considering the interactions between objects, more complex analyses, such as team statistical analysis and real-time event analysis, can be conducted. Specifically, an overview of automatic event detection in soccer games was presented [18]; this overview involved simply classifying approaches into a hierarchical structure on the basis of their analysis levels (i.e., low, middle, and high). Similarly, Rehman and Saba [19] reviewed feature extraction for soccer video semantic analysis. In addition, audio and video feature extraction methods and their combination with textual methods have been investigated. The data sources, methodologies, detected events, and summarization applications used in event-oriented soccer video summarization have been compared. Table I shows a comparison of previous surveys.

In summary, previous surveys have focused on the methodologies of sports video analysis from the spatiotemporal viewpoint instead of a content-based viewpoint. Few of these studies considered semantics. The current study adopted a different framework, developing a proper and

TABLE I  
KEY ISSUES OF PREVIOUS SURVEYS

Works	Years	Sport Genres	Major subjects	Minor subjects	Goal	# of ref.
[14]	2003	Comprehensively	Event	Tactics analysis, highlight extraction, tracking, computer-assisted refereeing, content insertion	Sports Video Analysis	28
[15]	2006	Snooker, tennis, badminton, cricket, soccer, American football, baseball, sumo	Object, Event, Context	Feature extraction, event detection, video summarization, indexing, and retrieval	Sports content retrieval Court/table sports vs. field sports Bridging the semantic gap	54
[16]	2010	Football, handball, basketball, squash, skating, indoor soccer	Objects	Team tracking	Intrusive vs. nonintrusive systems Indoor vs. outdoor sports	43
[17]	2010	Soccer	Event	Video summarization, provision of augmented information, high-level analysis	Low-level visual features vs. high-level semantic analysis	89
[18]	2011	Soccer	Object, Event	Highlights, summarization, event detection	General models for soccer analysis.	28
[19]	2014	Soccer	Event	Feature extraction, highlight event detection, video summarization	Audio and video features, textual method integration	56

deeper interpretation of content-aware sports video analysis. Specifically, we present the insight offered by research into the structure of content under different scenarios. On the basis of this insight, we provide an overview of the themes particularly relevant to the research on content-aware systems for sports.

### B. Scope and Organization of This Paper

A video analysis system can be viewed as a content reasoning system. In this paper, we review the developments in sports video analysis, focusing on content-aware techniques that involve understanding and arranging the video content on the basis of intrinsic and semantic concepts. We focus on the video content analysis applied in sportscasts over the past decade from three aspects—fundamentals and general review, a content hierarchical model, and challenges and future directions.

- 1) We introduce the fundamentals of content analysis, namely, the concept of the content pyramid, sports genre classification, and the overall status of sports video analytics.
- 2) We review state-of-the-art studies conducted in this decade on the content hierarchical model (i.e., content pyramid). The information used to represent high-level semantic knowledge can be divided into three groups: object-, event-, and context-oriented groups.
- 3) We review the prominent challenges identified in the literature and suggest the promising directions for future research on sports video content analysis.

The remainder of this paper is organized as follows. In Section II, we discuss the fundamentals of content-aware video analysis for sports. A systematic and detailed review of state-of-the-art studies according to three content awareness levels is presented in Section III. In Section IV, we present the challenges and promising future directions regarding to the content-aware sports video analysis for expert users. A conclusion is drawn in Section V.

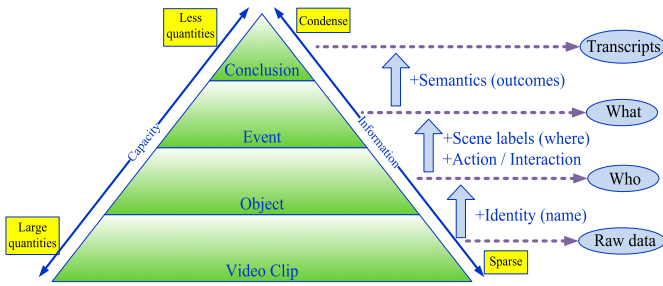


Fig. 1. Content Pyramid: a hierarchical content model.

## II. FUNDAMENTALS OF CONTENT-AWARE SPORTS VIDEO ANALYSIS

### A. Content Pyramid

In developing a content-aware retrieval system, the retrieval process must be designed according to users' intention. Several sports video analysis systems have focused on managing visual features in the spatial domain. For example, Han *et al.* [20] introduced a general framework for analyzing broadcast court-net sports videos from the viewpoints of the pixel, object, and scene levels. Using this framework, a group of predefined events at different levels can be identified in advance. According to the semantic significance, the video content was divided into four layers (Fig. 1): 1) video; 2) object; 3) action; and 4) conclusion layers [21], [22]. The volume of the layer denotes the quantity of the implied concepts. The compactness of information decays from the top down. The concept of the content pyramid is used to analyze context concerns for a video entity. Each layer of the content hierarchy represents various key components for characterizing the video content. The video layer consists of video clip frames, each of which consists of a video clip tag and raw video data. The object class layer consists of object frames, which represent the key objects in the video. In each object frame, an object tag and pointers link each key object to the corresponding video clips. An object that has a particular posture or movement or interacts with other objects yields an action or interaction tag. The event class layer consists of event frames, which represent the action of the key object. Actions combined with scene information construct an event tag. Each event frame consists of an event tag and an object tag, representing the related action or interaction among multiple objects. The top layer is the conclusion layer, which consists of conclusion frames representing the semantic summarization of the video sequence. Each conclusion frame consists of the event tags and corresponding results. A game summary is drafted according to transcripts and outcomes from events.

### B. Sports Genre Categorization

Recently, the proliferation and tremendous commercial potential of sports videos have strongly indicated the need for relevant applications. The technique of genre classification has become a common preliminary task for managing sports media. Fig. 2 shows a tree structure of sports genre classifications, presenting the reviewed papers by year of publication. The sports typically broadcasted on TV can be classified into

three categories: field, posture, and racket. As Fig. 2 shows, more than 80% of papers addressed baseball, basketball, soccer, and tennis. Fig. 2 shows only papers that focused on fewer than four types of sports; other papers are included in a general category. Typically, golf is considered both a field sport and a posture sport. Some track-and-field sports such as the long jump and pole vaulting are posture sports, whereas other sports such as the hammer throw and javelin throw are classified as racket sports. The remainder of this section focuses on reviewing methods for sports genre classification.

Brezeale and Cook [23] reviewed studies on video classification and identified the use of features in three modalities: 1) text; 2) audio; and 3) visual. Videos can be classified according to genres, such as movies, news, commercials, sports, and music, and subgenres. In the last decade, various sports genre classification methods based on different classifiers, such as hidden Markov models (HMMs) [24]–[26], naïve Bayesian classifiers (NBCs) [26], decision trees [27], and support vector machines (SVMs) [28], [29], have been presented. Schemes combining different classifiers have also been considered. You *et al.* [30] combined NBC and HMM algorithms to classify ball match video genres and detect events in football, basketball, and volleyball. Duan *et al.* [31] combined an NBC and SVM to categorize game videos into five typical ball sports, namely, tennis, basketball, volleyball, soccer, and table tennis. Zhang *et al.* [32] adopted the bag-of-visual-words (BoW) model with k-nearest neighbor classifiers to identify video genres. In their method, to achieve a generic framework for analyzing sports videos, an unsupervised probabilistic latent semantic analysis-based approach was used to characterize each frame of a video sequence into one of the following four groups: 1) close-up view; 2) midview; 3) long view; and 4) outer-field view.

For mobile video application, Cricri *et al.* [33] used a multimodal approach with four auxiliary data sets—sensor-embedded mobile devices, spatial and spatiotemporal visual information, and audio data, for confirming the sports type. To integrate all possible modality combinations in fusion models, the auxiliary input data can be used for video type classification. For home video application, Sugano *et al.* [34] extracted the salient low-level features from unedited video. The random forest [35] and the bagging [36] algorithms were used to classify the home videos into five genres, namely, sports, travel, family and pets, event, and entertainment.

### C. Overview of Sports Video Analysis

Sports video analysis aims to extract critical information from video content. Fig. 3 provides an overview of sports video analysis. The general procedure of a content-aware video analysis system includes feature extraction, information reasoning, and knowledge arrangement. Based on the concept of the content pyramid, content-aware video analysis can be performed using different level approaches. Information reasoning is used to understand low-level, midlevel, and high-level video data. Using model analytics, the information can be converted to knowledge. In other words, model analytics entails transforming data in the feature domain to knowledge in the application domain. The methodologies of model

	Sports													
	Field							Racket			Posture *2	Others *3		
	Baseball	Basketball	Football *1	F1 Racing	Golf	Hockey	Soccer	Track & field	Volleyball	Badminton	Table Tennis	Tennis		
2002-	[147][152][159][198][211][226]	[157][186][198]	[81][152][193]	[164]								[102]	[152]	
2003	[160]		[145][170]		[160]		[26][67][160][169][191]							
2004	[161][166]	[183]	[168]			[27]	[153]					[65]		[27]
2005	[146][204]	[95]					[165]							
2006	[135][195]	[135]					[45][74]					[59]	[135]	
2007	[206]	[150][156][206][225]					[179][187][150]		[173]	[178]		[150][178]		
2008	[158][205]	[39][205][224][228]		[68]			[39][177][224][228]		[20]	[20]		[20][39][60][205]		[68]
2009	[48][194][213][229]	[53]			[207]	[175]	[51][52][57][75][188][189]				[155]	[73][158]		[56]
2010		[30][154]	[154]	[30]			[17][44][148][154]		[86]	[30]		[148]		
2011	[42][100]	[79][196]					[18][72][162][196]		[87][122]			[131]	[122]	
2012	[185][214]	[65][102]					[64][167][192][212]							
2013	[210]	[55][58][184][210]	[104]	[85]	[58]		[40][170][190]	[88]				[125][184]		[54]
2014		[66][174][227][211]	[211]				[19][69][80][180]			[132]		[132]		
2015	[43]	[247]	[171]				[89][163]					[60]		
2016+		[99][246]					[172][223][244]							

\*1: American football: [81], [152], [168], [170], [193]; Gaelic football: [145]; Australian football: [154]; No specified: [104], [171], [211]  
 \*2: Sumo Wrestling: [152]; Figure Skating: [135]; Diving: [122]  
 \*3: Swimming: [27], [54]; Skiing: [68]; Speed Skating: [56]

Fig. 2. Types of sports classification.

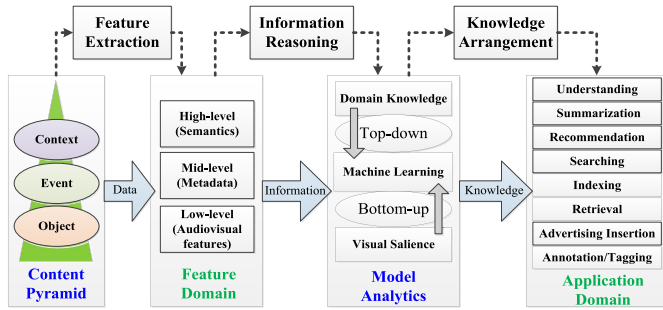


Fig. 3. Paradigm of content-aware video analysis.

analytics can be categorized into two classes: top-down and bottom-up. The top-down methodology assumes that the structure of the model is given, whereas the bottom-up methodology starts with visual saliency information treated as source features through machine learning to obtain high-level knowledge.

### III. SURVEYS OF CONTENT-AWARE VIDEO ANALYSIS WITH SEMANTIC LEVELS

According to the structure of the content pyramid, we surveyed state-of-the-art techniques from the aspect of semantic level, namely, highlight detection and event recognition, object detection and action recognition, and contextual inference and semantic analysis. The surveyed techniques are categorized in Fig. 4.

#### A. Object Detection and Action Recognition

Object-oriented sports content analysis is one of the most active research domains. This approach has been successfully applied for action recognition using intra-object posture

analysis and inter-object event determination. For single-object application, an accurate body posture and accurate movement are required to obtain the desired performance in sports such as diving, tumbling, skating, ballet, golf, and track-and-field games. For two-object application, the desired performance in sports such as boxing, sumo, wrestling, tennis, table tennis, and badminton is obtained using between-object interactions and that in sports such as soccer, rugby, hockey, basketball, baseball, and volleyball is obtained using between-group interactions. For multiple-object application, many studies focus on the group-to-group interaction for recognizing team tactics.

1) *Extracting Objects:* In sports videos, the foreground object plays a crucial role in the event scenario. Its actions and interaction with other objects can form a particular event. The objects in a video frame, such as auxiliary blobs, figures, and texts, can be either moving or stationary and either natural or superimposed. To convert low-level media features to high-level semantic labels, Naphade *et al.* [37] created a terminology named probabilistic multimedia objects (Multijects). To verify the dependence between intra-frame and inter-frame semantics, Multijects are combined with a graphical network that captures the co-occurrence probabilities of Multijects, named Multinet. A robust automatic video object (VO) extraction technique was presented in [38]. Luo *et al.* [38] reported that the dynamic Bayesian network (DBN) framework can facilitate attaining a more detailed description of VO characteristics compared with an HMM.

In a soccer game, the player and ball positions constitute the most critical information. Pallavi *et al.* [39] proposed a hybrid technique for detecting the ball position in soccer videos. This technique first classifies a shot as a medium or long shot, and motion- and trajectory-based approaches are then used for detecting medium and long shots, respectively. This method

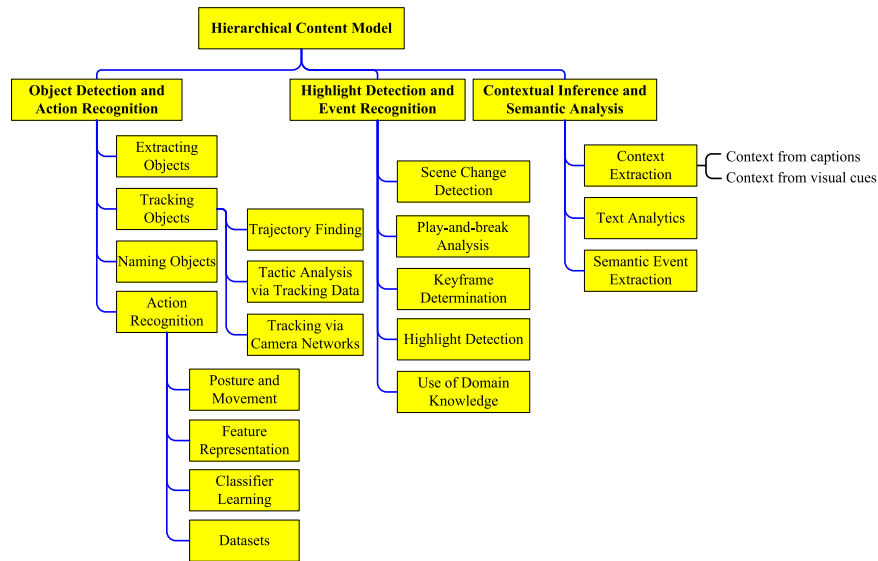


Fig. 4. Taxonomy of surveys with different semantic level applications.

does not require any preprocessing of the videos and is rather efficient. A framework based on the probabilistic analysis of salient regions was introduced in [40]. Player detection and tracking in broadcast tennis videos were presented in [41]. In [42], for a baseball game, an object segmentation algorithm was first used to retrieve the pitcher's body, and a star skeleton of the pitcher was then used to recognize the pitching style. The authors presented automatic generation of game information and video annotation through a pitching style recognition approach. Similarly, an algorithm for pitch-by-pitch video extraction proposed in [43] incorporated a pitcher localization technique, a pitch speed recognition scheme, and a motion degree measurement scheme for accurately identifying pitches.

However, there is a difficulty in extracting objects perfectly. A scene normally includes multiple objects in the playfield. The objects can partially or completely occlude each other or self-occlude because of the capturing angle. Hamid *et al.* [44] used multiple cameras to detect objects. In addition, 3D information has been used to estimate player and ball positions in broadcast soccer videos [45], [46]. This enabled the authors to measure the aerial positions without referencing the heights of other objects. Moreover, using the Viterbi decoding algorithm, they could determine the most likely ball path by considering consecutive frames. Recently, Li *et al.* [47] introduced the object bank for representing an image using a high-level image representation to encode object appearance and spatial location information.

Furthermore, studies have developed processes for quantifying the importance of VOs using the object attention model [48] and visual attention model [49], [50]. On the basis of such processes, the importance score of an object can be used to infer the importance of a frame and even a video clip. If a frame contains more high-attention objects with a high-interest contextual description, it is highly probable that the frame has higher excitement score. Regarding an event-based analysis scenario, the score of each frame is derived on the

basis of the relevant key frames that are located in an identical event.

2) *Tracking Objects*: Object tracking is a crucial technique in sports content analysis. It aims to localize the objects in a video sequence and provides diverse applications in video surveillance, HCI, and visual indexing. In sports, the object tracking technique is used to observe the continuing activities of objects (e.g., players and ball) in the playfield. However, the camera motion is a major limitation in detecting and tracking objects. Khatoonabadi and Rahmati [51] applied a region-based detection algorithm for eliminating fast camera motion effects in goal scenes and tracking soccer players. When the region-based algorithm is used to track players, either the template matching method or split-and-merge approach is applied for occlusion reasoning between players. However, players' positions can be misdetected in the split-and-merge approach. Liu *et al.* [52] presented a method for detecting multiple players, unsupervised labeling, and efficient tracking in broadcast soccer videos. The framework involves a two-pass video scan, which first learns the color of the video frame and then tests the appearance of the players. The method is generally effective, except for cases involving video blur and sudden motion. Kristan *et al.* [53] presented an algorithm for tracking multiple players in an indoor sporting environment. Considering a semicontrolled environment with certain closed-world assumptions, the algorithm relies on a particle filter. The multiplayer tracking process involves background elimination, local smoothing, and management of multiple targets by jointly inferring the players' closed worlds. In addition, Sha *et al.* [54] presented an automatic approach for swimmer localization and tracking. The approach enables large-scale analysis of a swimmer across many videos. A swimming race is divided into six states: 1) start; 2) dive; 3) underwater; 4) swim; 5) turn; and 6) end. A multimodal approach is employed to modify the individual detectors and tune to each race state. However, the location information cannot present more detailed context of the game. To solve this problem,

long-term observation is required for tracking object movement.

*a) Trajectory finding:* A trajectory is the path along which a moving object passes through space as a function of time. Robust solutions to trajectory-based techniques have applications in domains such as event reasoning and tactic analysis. Chakraborty and Meher [55] presented a real-time trajectory-based ball detection and tracking framework for basketball videos. Liu *et al.* [56] proposed a computer vision system that tracks high-speed nonrigid skaters over a large area. The input video sequence is fed into a registration subsystem and a tracking subsystem, and the results from both subsystems are used to determine the spatiotemporal trajectory, which can be used for event detection and sports expert analysis. However, tracking performance is lacking when skaters move in groups during long and continual complete occlusions. A possible solution was presented by Miura *et al.* [57]. They attempted to estimate ball routes and movements by tracking overlaps with players and lines in order to recognize certain scenes. When an overlap occurs, the spatiotemporal relationship between the ball and the object is examined. Ball existence near each route candidate is then examined to eliminate potential routes and determine the most likely route. Similarly, Liu *et al.* [58] presented context-conditioned motion models incorporating complex inter-object dependencies to track multiple players in team sports over long periods. Yan *et al.* [59], [60] have presented a robust tennis ball tracking method based on a multilayered data association with graph-theoretic formulation. More recently, Zhou *et al.* [61] proposed a two-layered data association approach for tennis ball tracking that was identical to Yan *et al.*'s method, but did not consider Yan's tracklet linkage, which may cause tracking failure. For addressing short-term misdetection, Wang *et al.* [62], [63] introduced a more efficient dense trajectory. They combined trajectory shape, appearance, and motion information to compute motion boundary descriptors along dense trajectories.

*b) Tactic analysis via tracking data:* For tactic analysis, Niu *et al.* [64] proposed a framework for systematically analyzing soccer tactics through the detection and tracking of field lines. The identified trajectory ultimately enables analyzing and improving soccer tactics. Comparably, the player trajectory was used for recognizing tactic patterns in basketball videos [65]. This type of framework was developed for directing focus to an open three-point attempt in a basketball game [66]. The trajectories of objects have frequently been used in event recognition [67], [68]. By employing a motion model for home and away team behaviors in soccer, Bialkowski *et al.* [69] visually summarized a soccer competition and provided indications of dominance and tactics. Recently, Zheng [70] conducted a survey on trajectory data mining, offering a thorough understanding of the field. More recently, a smart coaching assistant (SAETA) designed for professional volleyball training was introduced in [71]. SAETA relies on a sensing infrastructure that monitors both players and their environment and provides real-time automatic feedback for aerobic and technical-tactical training in a team-sports environment.

*c) Tracking via camera networks:* Another critical objective is determining how to combine multiple cameras to obtain a more accurate tracking result. Choi and Seo [72] reviewed problems regarding automatic initialization of player positions. Using SVMs, prior knowledge on the features of players can be retrieved, and depending on soccer match conditions, automatic initialization is often successful. However, reinitialization on tracking failure and guaranteeing a minimum time for initialization remain challenging. Yu *et al.* [73] presented an automatic camera calibration algorithm for broadcast tennis videos, evaluating problems regarding distortion, errors, and fluctuating camera parameters. The algorithm generates accurate camera matrices for each frame by processing the clips through the following four steps: 1) ground-feature extraction; 2) camera matrix computation; 3) camera matrix refinement; and 4) clip-wise refinement. Figueroa *et al.* [74] used four cameras to observe the positions of all players on the soccer field at all times over the entire game. Their algorithm uses paths in a graph filled with blobs, representing segmented players, and analyzes different components of the blobs, such as the area and trajectory. Ren *et al.* [75] presented a technique for estimating the trajectory of a soccer ball using multiple fixed cameras, despite constant movement and occlusion leading to size and shape discrepancies among the cameras over time. With a combination of motion information, expected appearance modeling, occlusion reasoning, and backtracking, the ball trajectory can be accurately tracked without velocity information.

*3) Naming Objects:* By taking advantage of state-of-the-art methods, detecting, tracking, and recognizing objects are feasible at present. However, the most difficult problem is identifying the detected object and the position of the athlete mentioned in subtitles. Most previous studies have focused on news videos because they are straightforward. For example, Satoh *et al.* [76] presented a system that identifies faces by associating the faces of the image with names in the captions. Everingham *et al.* [77] proposed a system that enables labeling the names of actors in TV episodes and movies. The framework integrates two procedures: 1) time-stamped character annotation by aligning subtitles and transcripts and 2) face tracking and speaker detection. An enhanced method was presented by Ramanathan *et al.* [78], who applied a bidirectional model to simultaneously assign names to the trajectories in the video and mentioned in the text. Normally, the naming process for news and movies videos is easier than that for sports videos, because stable face images and subtitles can be obtained easily. Lu *et al.* [79] used a conditional random field model to generate joint probability inferences regarding basketball player identities. A deformable part model detector was used to locate basketball players, and a Kalman-filter-based tracking-by-detection approach was employed to extract tracklets of players. Similarly, in [80], a large-scale spatiotemporal data analysis on role discovery and overall team formation for an entire season of soccer player tracking data is presented; the entropy of a set of player role distributions was minimized, and an EM approach was used to assign players to roles throughout a match. Their work can be applied future for identifying strategic patterns that team's exhibit.

Nitta *et al.* [81] focused on identifying the structure of a sports game using linguistic cues and domain knowledge to extract actors, actions, and events from a sports video and integrating the caption text and image stream. The system identifies actors who are performing an action and the action that they are performing.

However, face information is not observable at all times because the athletes usually run along an unpredictable path in 3D space. An alternative approach for name assignment to objects is based on the text or numbers printed on athletes' clothes using the video OCR [82] and scene text recognition techniques [83], [84]. To identify and track multiple players, Yamamoto *et al.* [85] proposed a tracking method that associates tracklets of identical players according to the results of player number recognition. Using a multicamera system, Pnevmatikakis *et al.* [86] and Patrikakis *et al.* [87] introduced a system that enables athlete tracking in different camera views and generates metadata such as locations and identities. Combining scene location, motion outliers, faces, and clothes enables tracking body hypotheses over time. For personalizing sports video broadcasts, Messelodi and Modena [88] presented an athlete identification module that applies scene text recognition to the embedded text from images. By applying their module to undirected and directed videos, precision rates of 98.9% and 88.7%, respectively, can be achieved. A method for soccer jersey number recognition using a deep convolution neural network (CNN) was recently presented in [89]. Two feature vector construction models, with the first treating every number as an individual class and the second treating each digit as a class, were employed and compared. With data augmentation and without dropout, the highest recognition rate achieved was 83%.

4) *Action Recognition:* Action recognition has been extensively studied in recent years. Various approaches have been proposed for obtaining an accurate human body posture and movement estimate. Comparative explorations of recent developments in action recognition have described video-based [90]–[92] and image-based [93] approaches. However, only a few descriptions focus on the sports genre. Action recognition in sports can typically be classified into three categories: 1) individuals; 2) between objects; and c) between groups. The individual class requires an accurate human body posture with articulated parts and motion parameters for evaluating the global and local gesture performance. In the between-objects class, a player interacts with the competitor in winning the match or game. Here, the players hold a racket; thus, sports in this class are called racket sports. Sports in the between-groups class are called team sports. Almost all such sports are field sports. The members of a team work together to compete with the other team. Moreover, by combining the action class and field information, the event class can be determined.

Large-scale sports video analytics with deep learning is an emerging research area in many application domains. One of the most prevalent trends is to learn features using a deep discriminatively trained neural network for action recognition. For example, Wang *et al.* [94] aggregated a hand-crafted feature [95] and feature determined through deep learning [96], using deep architectures to develop more effective descriptors.

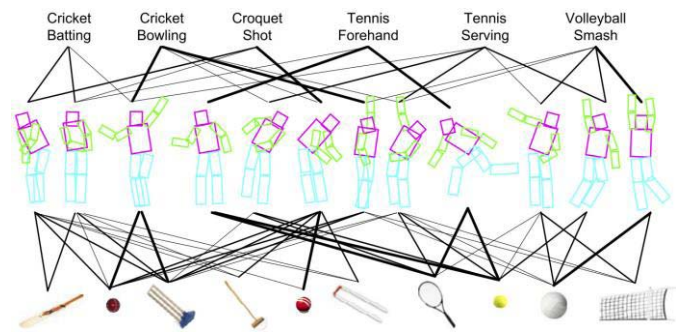


Fig. 5. Learned strength of connectivity between activities, human poses, and objects. Thicker lines indicate stronger connections, originally shown in [98].

a) *Posture and movement:* Regardless of the type of sport, the human body posture and movement of players have attracted much research attention. For example, Chen *et al.* [42] presented automatic generation of game information and video annotation through a pitching style recognition approach. In this approach, an object segmentation algorithm is first used to detect the pitcher's body. A star skeleton is then used to model the pitcher and recognize the pitching style. Yao and Fei-Fei [97] applied a conditional random field model for action analysis. They modeled the mutual context of objects and human poses in human–object interaction activities. In their extended study [98], instead of modeling the human–object interactions for each activity individually, they ascertained overall relationships among different activities, objects, and human poses. As illustrated in Fig. 5, they modeled the *mutual context* between objects and human poses in human–object interaction activities so that one can facilitate the recognition of the other. Specifically, two pieces of contextual information are considered in the mutual context model. The *co-occurrence context* models the co-occurrence statistics between objects and specific types of human poses within each activity. The types of human poses, termed “atomic poses” (shown in the center of Fig. 5), can be considered a dictionary of human poses where the human poses represented by the same atomic pose correspond to similar body part configurations. In addition, they considered the *spatial context*, which models the spatial relationship between objects and different human body parts. Pecev *et al.* [99] shifted the focus from players to referees. Educational software that enables predicting the movement of basketball referees using a multilayered perceptron neural network is beneficial in training young basketball referees.

Rather than depending on the vision-based method, Ghasemzadeh and Jafari [100] introduced a sensor-based method for tracking a baseball player's swing movements using a wearable platform that yields multidimensional physiological data collected from a body sensor network. A semi-supervised clustering technique is implemented to construct basic movement patterns, which can be used for detecting common mistakes and producing feedback. Schmidt [101] used detailed features such as angle displacements and velocities of kinematic chain articulations to analyze the movement patterns of free-throw shooters at different skill levels in a basketball video. In the experiments, several highly different

movement organizations and their associated particular functionality were observed. The individual imprint on the movement pattern was shown to be related to its own, evidently appropriate functionality, which fit well with the personal constraints of the athlete in question.

Miyamori and Iisaku [102] focused on a problem in which recognition easily becomes unstable because of partial tracking errors or lack of necessary information. They incorporated human behavior analysis and specific domain knowledge with conventional methods, developing an integrated reasoning module for robust recognition. The system was designed for tennis videos, and four typical video segments were used for reasoning object action: player behavior, ball position, player position, and curt-net lines. Kazemi *et al.* [103] addressed the problem of human pose estimation for football players, given images obtained from multiple calibrated cameras. With a bodypart-based appearance model, they used a random forest classifier to capture the variation in appearance of body parts in 2D images. The results of these 2D part detectors were then aggregated among views to produce consistent 3D hypotheses for the parts. Burenius *et al.* [104] also presented a framework for 3D pictorial structures that can be used for multiple-view articulated pose estimation. Swears *et al.* [105] proposed a modeling approach, called Granger constraints DBN, which enables modeling interactions between multiple co-occurring objects with complex activity. To simplify the measurement of temporal dependence, they used the Adaboost feature selection scheme to discriminatively constrain the temporal links of a DBN.

*b) Feature representation:* Regarding action recognition, two critical learning phases are highly required: feature representation and classifier learning. Feature representation extracts and ranks useful 2D or 3D features such as Gabor features [106], histograms of oriented gradients (HOGs) [107], gradient location-orientation histograms [108], motion history images [109], motion history volumes [110], 3D gradients [111], and other volumetric (voxel) data [112]–[115]. The common strategy is based on a predefined articulate model for matching human body parts [116]–[118].

*c) Classifier learning:* When feature vectors are extracted, a classification framework is used to recognize the types of actions. For instance, Zhang *et al.* [119] proposed an effective action recognition method based on the recently proposed overcomplete independent components analysis (ICA) model. Instead of using a pretrained classifier, they adopted the response properties of overcomplete ICA to perform classification. In this approach, a set of overcomplete ICA basis functions are learned from 3D patches from training videos for each action. The test video can be labeled as the action class whose basis functions can reconstruct the video with the smallest error. Similar to Zhang *et al.*'s method, Le *et al.* [120] presented a hierarchical invariant spatiotemporal feature learning framework based on independent subspace analysis. For modeling movement patterns, a structured codebook-based method with an SVM classifier was proposed in [121]. For modeling the temporal transition of postures, Li *et al.* [122] presented a continuous HMM with a left–right topology. In addition,

TABLE II

BENCHMARK DATA SETS AVAILABLE FOR SPORTS VIDEO ANALYSIS; IN THE “TASKS” COLUMN, “ $A_r$ ,” “ $A_l$ ,” “ $P$ ,” AND “ $C$ ,” RESPECTIVELY, REPRESENT “ACTION RECOGNITION,” “ACTION LOCALIZATION,” “POSE ESTIMATION,” AND “SPORTS TYPE CLASSIFICATION”

System	Dataset(year)	#sports/actions categories	Source	#instances	Tasks	Locations
[135]	Wang <i>et al.</i> (2006)	3 <sup>a</sup> /25	Internet	1400, 4500, 8500 images	$A_r$	[136]
[126]	UCF Sports (2008)	10 <sup>*</sup>	Broadcast TV (e.g. ESPN and BBC)	15 sequences.	$A_r, A_l$	[127]
[140]	Olympic Sports (2010)	16 <sup>*</sup>	YouTube	50 videos	$A_r$	[141]
[131]	ACASVA (2011)	2 <sup>b</sup> /3	Broadcast TV	6 sequences	$A_r, A_l$	[133]
[103]	KTH (2013)	1 <sup>c</sup> *	Staged	3D: 2400 images 2D: 5907 images	$P$	[134]
[137]	Sport-1M (2014)	487 <sup>*</sup>	YouTube	~1.1 million videos	$C$	[138]
[142]	SVW (2015)	44/30	Smartphone & Tablet	4200 videos	$A_r, A_l$	[143]

<sup>a</sup>Figure skating, baseball, and basketball

<sup>b</sup>Tennis and badminton

<sup>c</sup>Football

\*No significant distinctions between sports types and actions categories

Taylor *et al.* [123] extended the gated restricted Boltzmann machine (GRBM) [124] from 2D images to 3D videos for learning spatiotemporal features and named the extension the convolutional GRBM method.

*d) Data sets:* Unlike [125], the current survey focused on reporting sports action data sets that do not explicitly consider comprehensive action data sets. Here, we reported seven sports action data sets. Two of these data sets presented the image source with action markers, and the remaining five comprised a collection of video clips. A comprehensive list of action data sets for sports with corresponding details is provided in Table II.

The UCF sports data set [126] contains 150 video sequences with a resolution of  $720 \times 480$  concerning ten disciplines, namely, diving, golf swinging, kicking, lifting, horseback riding, running, skating, baseball swinging (split into around high bars and on the pommel or floor), and walking. All video clips were collected from sports broadcasts by networks such as ESPN and the BBC. The data set exhibits high intra-class similarity between videos and large variation in inter-class viewpoint changes and noise. It was presented by Rodriguez *et al.* [126] and is available on the Internet [127]. Since the release of this data set in 2008, the average recognition rate has increased, being 85.6% in 2009 [128], 87.3% in 2010 [129], 86.8% in 2011 [120], 95% in 2012 (used extra training data) [130], and 89.7% in 2014 [121]. This data set is the most widely used among the six sports action data sets.

A tennis data set called ACASVA, published by de Campos *et al.* [131], was used to compare the BoW and space–time–shape approaches for action recognition in videos. The tennis game videos were reproduced from TV broadcasts in standard resolution, and Yan *et al.*'s ball tracker [60] was used to detect relevant instances. The data set and MATLAB software package [132] are available in [133].

The KTH Multiview Football data set [103] is a 3D data set containing 2400 images of football players obtained from three views at 800 time frames with 14 annotated body joints



and 5907 images with the annotated 2D pose of the player. The data set is available in [134].

Wang *et al.* [135] presented a data set containing ten labels in figure skating clusters, seven labels in baseball clusters, and eight labels in basketball clusters. The data set can be acquired from [136].

Karpathy *et al.* [137] constructed an immense data set named Sports-1M, which consists of greater than 1.1 million annotated YouTube videos for 487 sports. The data set is available in the GitHub repository [138]. Compared with the UCF sports data set, Sports-1M entails greater difficulty in feature recognition. The highest recognition rate is approximately 74.4% with a two-stream CNN and temporal feature pooling [139].

The fifth set is an Olympic sports data set [140] containing 50 videos from each of the following 16 classes: high jump, long jump, triple jump, pole vault, discus throw, hammer throw, javelin throw, shot put, basketball layup, bowling, tennis serve, diving (split into platform and springboard), weightlifting (split into snatch and clean and jerk), and vault (gymnastics). This data set is available in [141].

A more recently introduced data set, sports videos in the wild (SVW) [142], contains 4200 videos covering 30 sports categories and 44 actions. The videos in this data set were captured by users of the leading sports training mobile app Coach's Eye while practicing a sport or watching a game. This data set is available in [143].

### B. Highlight Detection and Event Recognition

Because an increasing amount of research is focused on large-scale content-based multimedia mining, there is a lack of systematic surveys specifically designed for sports videos. This section mainly reviews the event-oriented state-of-the-art studies according to the type of data source, methodology, sports genre, desired event, and application field. The reviewed articles are listed in Table III.

1) *Scene Change Detection*: An event is composed of a group of video scenes or shots. The shot boundary detection technique aims to divide a long video sequence into video segments. Gong *et al.* [144] proposed temporal segmentation for partitioning a video sequence into camera shots. Each shot contains identical semantic concepts. Several studies have used scene change detection to perform event classification. The simplest approach to detecting transitions of shots is seeking discontinuities in the visual content of a video sequence. For instance, Sadlier *et al.* [145], Liang *et al.* [146], Chang *et al.* [147], and Poppe *et al.* [148] have proposed color-based shot boundary detection schemes for segmenting a video. Poppe *et al.* [148] used color histogram differences for detecting scene changes. During the broadcast of baseball games, multiple broadcast cameras were mounted at fixed locations in the stadium. Chang *et al.* [147] first built statistical models for each type of scene shot with histogram products. In addition, Sadlier and O'Connor [149] employed their own algorithm to solve the problem caused by the high-tempo nature of field sports; during live action segments, the broadcast director has little opportunity to use shot transition types other than hard shot cuts. In [150], an event-based

segmentation method was presented and a motion entropy criterion was employed to characterize the level of intensity of relevant object motion in individual frames. Because global motion usually changes largely in shot boundaries, Wu *et al.* [151] detected abrupt augmentations of the magnitude of accelerations in the frame sequence and marked them as semantic clip boundaries.

2) *Play-and-Break Analysis*: Designing a generic model for analyzing all types of sports videos is difficult, because the event is a domain-dependent concept. Nevertheless, a compromised approach, called play-and-break analysis, has been presented. This analysis roughly models sports games. Li and Sezan [152] used both deterministic and probabilistic approaches to detect critical events, called plays. An iterative algorithm was used to detect plays until a nonplay shot was identified. On the basis of this play information, games could be summarized. Xie *et al.* [153] applied statistical techniques for analyzing the structure of soccer videos. Plays and breaks (i.e., nonplay) are two mutually exclusive states of soccer videos. First, experimental observations were used to create a salient feature set. HMMs and dynamic programming algorithms were then applied for classification and segmentation. Tjondronegoro and Chen [154] pioneered the use of a play-break segment as a universal scope of detection and a standard set of features that can be applied to different sports. Specially, a racket sports game consists of many play and break events. In [155], the play event in the racket sports, which has a unique structure characteristic, was referred to as "rally." That study segmented a racket sports video into rally and break events and ranked the rally events according to their degrees of excitement.

3) *Keyframe Determination*: Keyframe detection is a crucial operation performed after the retrieval of shot boundaries. A keyframe is used to represent the status of each segment [144], [146]. Liang *et al.* [146] proposed a framework for detecting data changes in superimposed captions and employing rule-based decisions to detect meaningful events in a baseball video. They considered only the number of outs, number of scores, and base-occupation situation, and thus, they could simply select any frame as the representative shot (segment) if the caption information remained unchanged. In addition, numerous studies have focused on scoreboard detection [156]–[158]. First, a scoreboard subimage was located using the difference in consecutive frames and the gradient information of each frame. Subsequently, the changed pixels in the scoreboard subimage were evaluated [156]. Nepal *et al.* [157] used the displayed scoreboard as the second key event, and developed temporal models based on the pattern of occurrence of the key events observed during the course of the game. Regarding the structure of a baseball video, the critical events conventionally occur between pitch scene blocks [146], [152], [158].

4) *Highlight Detection*: Highlights are critical events in a sports game that must be extracted. Highlight extraction requires specific features and analyzers. A video consists of audio and visual features. Several previously proposed frameworks have applied audio features instead of classifiers, such as SVMs [148], [155], [159], HMMs [160], Bayesian

TABLE III  
COMPARISONS OF THE WORK BEING DONE IN THE DOMAIN OF SPORTS EVENT ANALYSIS

Refs.	Data sources <sup>a</sup>	Methodologies	Sport genres	Event examples	Application views
[69]	Mid-v	Expectation-Maximization (EM) method, Hungarian algorithm	Soccer	Possession condition, attacking and defending deep	Possession analysis, game formation analysis, tactic analysis
[144]	Visual	A priori model comprises four major components: a soccer court, a ball, players, and motion vectors.	Soccer	Kick and goal at different place	Content-based video retrieval
[145]	Audio, Visual	Grass Coloured Pixel Ratio Characteristic (GCPRC)	Gaelic Football	Goals-points	Event detection
[146]	Semantic	Color-based shot boundary algorithm	baseball	Outs, scores, base-occupation	Summary, indexing for sports video
[147]	Visual	Hidden Markov model (HMM), probabilistic model	Baseball	Home run, catch, hit, infield play	Highlight detection and classification
[148]	Visual, Audio	Music Analysis Retrieval and Synthesis for Audio Signals (MARSYAS), Support Vector Machine (SVM)	Tennis, soccer	Goals	Summarization, automatic annotation
[149]	Audio, Visual	SVM	Soccer, rugby, hockey, Gaelic football	Goals, play, break	Event detection, shot boundary detection
[150]	Visual	Entropy-based Motion Analysis, homoscedastic error model	Basketball, soccer, tennis	Basketball : scoring, miss, free throw Soccer : goal, free kick, corner Tennis : fault, ace, volley	Video segmentation, event detection
[151]	Semantic, Low-v	Global Motion Estimation (GME), RBF neural networks and decision-tree classifier, predefined finite-state machine model	Non-specified	High jump, long jump, javelin, weight, throwing, and dash	Events recognition in sports video
[152]	Low-v	HMM	American football, baseball, sumo wrestling	Play-break	Event detection, video summary
[153]	Visual	HMM	Soccer	Goal, corner kick	Video browsing, events detection
[154]	Audio, Visual	Binary logo model	Soccer, Australian Football League(AFL), basketball	Soccer: goal, shoot, foul AFL: goal, behind, mark, tackle Basketball: goal, free throw, foul	Content-based video retrieval, event detection
[155]	Audio, Mid-v	SVM, Radial Basis Function (RBF),	Tennis, table tennis	Cheer, silence, excited speech	Summarization, event detection on racket sports video
[156]	Text, Mid-v	Connected Component Analysis (CCA)	Basketball	Score change	Real-time score detection
[157]	Audio, Visual, Text	Temporal model	Basketball	Goals	Content-based video retrieval, sports video analysis
[158]	Visual	Bayesian Belief Network (BBN), SVM	Baseball	Strike out, homerun, sacrifice fly, hit, hit and run, fly out or ground out, double play, walk, steal base, wild pitch, passed by or balk, others	Scoreboard recognition, shot transition classifier, scoreboard identification
[159]	Audio	SVM	baseball	Hit the ball, exciting speech	Highlight extraction, summarization
[160]	Audio	Entropic Prior Hidden Markov Models (EP-HMM)	Baseball, golf, soccer	Applause, cheering, music, speech	Highlight extraction
[161]	Visual	Multi-level Semantic Network (MSN), BBN	Baseball	Highlights	Semantic feature extraction, highlight detection
[164]	Audio, Visual, Text	Dynamic Bayesian Networks (DBNs)	Formula 1 races	start, fly out, passing	Highlight detection
[165]	Semantic	DBN, temporal intervening network	Soccer	Goal, corner kick, penalty kick, card	Highlight extraction
[166]	Audio, Visual	Maximum entropy model (MEM)	Baseball	Home run, outfield hit, outfield out, infield hit, infield out, strike out, walk	Highlight detection, machine learning, video content analysis
[167]	Visual, Semantic	Hidden Markov Model under normal (NHMM) and enhanced (EHMM) observations	Soccer	Placed, foul, shoot, goal, highlight	Video retrieval and personalized sports video browsing, shot classification
[168]	Audio, Text	Recognition of the textual overlays	American football	Touchdown, field goal	Personalized video abstraction, highlights detection
[169]	Visual	Model highlights using finite state machines	Soccer	Forward pass, shot on goal, turnover, placed kick, corner, free kick, penalty, kick off, counterattack	Semantic annotation, highlights detection
[170]	Visual	Camera motion parameters	American football	Short and long pass, short and long running, quarterback sacks, punt, kick off	Event detection
[171]	Low-v	SVM, Camera label estimation	American football	Goal, foul, corner, substitution	Event tagging, spatial segmentation
[173]	Visual	Physics-based method of ball trajectory extraction	Volleyball	Serve, Receive, Spike	Ball tracking, action detection
[175]	Visual	Switching Probabilistic Principal Component Analysis (SPPCA) , boosted particle filter (BPF), Histograms of Oriented Gradients (HOG), Sparse Multinomial Logistic Regression (SMLR)	hockey	N/A	Objects tracking, action recognition
[176]	Low-v	HMM	Basketball, football, racing, boxing, soccer	Commercial, slow motions	Slow motion replay segments detection, Video summarization highlights generation
[177]	Audio, Visual, Temporal	Subspace-based data mining	Soccer	Corner, goal	Semantic event detection, concept detection
[178]	Audio, Visual, Semantic, Attention	Support Vector Regression (SVR)	Tennis, badminton	Overhead-swing, left-swing, right-swing	Affective analysis, highlight ranking, video browsing, action recognition
[183]	Visual, Text, Semantic	Combination of domain-independent global model-based filtering methods with domain-specific object-level spatiotemporal constraints	Tennis, basketball	N/A	Summarization and browsing
[186]	Low-v	Decision-tree learning method, Entropy-based inductive tree-learning algorithm	Basketball	Left court, right court, middle court, other	Video indexing, classification, summarization
[187]	Low-v	Advanced temporal pattern analysis and multimodal data mining method, decision tree learning algorithm	Soccer	Goal	Event detection, summarization, browsing and retrieval
[188]	Visual	Probabilistic Support Vector Classification (PSVC)	Soccer	Goal, shot, corner, free kick	Object tracking, tactic analysis
[189]	Visual	Circle Hough Transform (CHT)	Soccer	Goal	Real-time goal detection
[190]	Visual, Text, Semantic	Temporal Neighboring Pattern Similarity (TNPS) measure, Probabilistic Latent Semantic Analysis (PLSA), Conditional Random Field Model (CRFM).	Soccer	Goal	Summarization, event detection, provision of augmented information
[191]	Visual, Text, Audio, Semantic	Time Interval Maximum Entropy (TIME)	Soccer	Goal, yellow card, substitution	Event detection in video documents, indexing
[192]	Visual, Text	HMM	Soccer	Goal, red card, yellow card, penalty, game start, game end, half time, disallowed goals	Summarization, event detection
[193]	Visual, Semantic	Closed caption (CC) text	American football	Touchdown, field goal	Event-based video indexing

<sup>a</sup>Low-v: low-level visual features; Mid-v: mid-level visual features

belief networks (BBNs) [161]–[163], DBNs [164], [165], maximum entropy models [166], and multilevel semantic networks, which are an extension of the BBN. For instance, Petkovic *et al.* [164] proposed a robust audiovisual feature extraction scheme and a text detection method, and used classifiers for integrating multimodal evidence from different media sources to determine the highlight events of a Formula 1 race.

In sports, visual features for highlight detection can be classified into six groups: 1) captions and text; 2) playfields; 3) camera motion; 4) ball positions; 5) player positions; and 6) replays. First, an algorithm addressing text detection [167] and scoreboard recognition was presented in [168], which employed an overlay model to identify the event frame. Second, given a “playfield” frame, the region of the image that represents the playfield was roughly identified from grass color information using color histogram analysis [169]. Unlike grass field regions, audience regions have different textures. Therefore, Chao *et al.* [165] used edge density information to detect audience regions. Third, Lazarescu and Venkatesh [170] used camera motion parameters, including the number of stages in camera motion, tilt motion, and camera angle, in a method for detecting seven types of American football plays. A more recently presented system [171] identifies the type of camera view for each frame in a video. The system was designed for football games, and salient events such as goals, fouls, corners, substitutions can be distinctly tagged on the timeline.

Generally speaking, balls and players play the main roles in a sports competition. Thus, ball and player tracking techniques have typical features in sports video analysis. Kobayashi *et al.* [172] input vector time-series data consisting of player, referee, and ball positions into a data-driven stick breaking hidden Markov model for predicting key activities and dominance conditions in a soccer game. In addition, Chen *et al.* [173] addressed the obstacles and complexities involved in volleyball video analysis by developing a physics-based scheme that uses motion characteristics to extract the ball trajectory, despite the presence of numerous moving objects. In [174], temporal and spatiotemporal regions of interest were used to detect the scoring event and scoring attempts in basketball mobile videos. Furthermore, Lu *et al.* [175] used a panning, titling, and zooming camera to automatically track multiple players while detecting their action. The system involves using HOGs, a boosted particle filter, and the combination of HOG descriptions with pure multiclass sparse classifiers. Finally, slow motion replays were modeled using an HMM, which emphasizes highlights and critical events [176]. These visual cues are summarized in Table IV. In addition, Shyu *et al.* [177] proposed a subspace-based multimedia data mining scheme for event detection and used a decision tree for training event detectors. Zhu *et al.* [178] presented a multimodal approach for organizing racket sports video highlights through human behavior analysis. Support vector regression constructs a nonlinear highlight ranking model. Furthermore, some previous studies have focused on not only highlight extraction but also highlight ranking [155], [168].

5) *Use of Domain Knowledge:* Sports events are typically composed of several video shots that appear in a certain

TABLE IV  
VISUAL FEATURES FOR CONTENT ANALYSIS

Visual features	References
Captions and text	[157], [195], [200], [207], [211], [219], [220], [221]
Playfields	[144], [165], [198], [202], [203], [207], [210], [214]
Camera motion	[154], [161], [165], [170], [171]
Ball positions	[169], [172], [173], [189]
Player positions	[147], [165], [169], [172], [173], [175], [178], [186]
Replays	[148], [154], [165], [168], [176]

temporal order. Using shot boundary detection techniques, the video can be segmented into scenes. Domain knowledge (i.e., paradigm) can be used to determine the type of video shot. Xie *et al.* [153] employed domain knowledge to perform event recognition. Huang *et al.* [179] proposed a temporal intervening network for modeling the temporal actions of particular events in a soccer game, such as goal, card, and penalty events. In [180], a Bayesian network-based model was used for event detection and summarization applications in a soccer video, whose structure was estimated using the Chow–Liu tree [181], and the joint distributions of random variables were modeled using the Clayton copula [182].

Temporal video patterns typically vary depending on the type of sport. In other words, sports video analysis is a domain-specific problem. Several extant systems analyze only sports that involve an identical game structure [15], [20], such as court-net sports including tennis, badminton, and volleyball, to obtain a comprised result. Many state-of-the-art frameworks have added domain knowledge to increase the accuracy. For example, Zhong and Chang [183] presented a generic domain-independent model with spatiotemporal properties of segmented regions for identifying high-level events, such as strokes, net plays, and baseline plays. Similarly, analyzing the temporal relationships among video shots enables automatic headline detection in TV sports news [184]. Chen *et al.* [185] introduced an HMM-based ball hitting event exploration system for baseball videos. A rule-based model might require domain knowledge and employ the decision-tree learning method to calculate rules and apply them to low-level image features [186]. The investigation of decision tree algorithms and multimodal data mining schemes was the main contribution of [187].

Zhu *et al.* [188] focused on tactical analysis, extracting tactical information from attack events in broadcast soccer videos to assist coaches and professionals. In soccer, a goal event must be preceded by the offense events that were considered in [189] and [190]. D’Orazio *et al.* [189] used four high frame rate cameras placed on the two sides of the goal lines. Snoek and Worring [191] presented the time interval maximum entropy (TIME) framework for combatting multimodal indexing problems regarding representation, inclusion of contextual information, and synchronization of the heterogeneous information sources involved. The TIME framework proved to be successful, evidently reducing the viewing time of soccer videos. Regarding application, Nichols *et al.* [192] attempted to summarize events using only Twitter status updates. In addition, Babaguchi *et al.* [193] proposed an event-based video indexing method that involves

using time spans where events are likely to occur and keywords from the closed caption stream. Chen *et al.* [187] emphasized the framework of mining event patterns. Using a group of audiovisual features and domain knowledge, they applied temporal pattern analysis for semantic event mining. Gupta *et al.* [194] applied an AND-OR graph as a representation mechanism for storyline models and learned the structure and parameters of the graph from roughly labeled videos using linguistic annotations and visual data. They showed a storyline through the instantiation of AND-OR graph for a baseball game broadcast from news videos.

The temporal patterns of video shots reveal critical clues for accessing video content. Event recognition can be directly achieved using a temporal pattern. For instance, the goal event in soccer, comprising the occurrences of gate appearance, a close-up view, and replays, follows certain rules of causality. After the gate appears, the mid-distance view is shown for a short duration. The scorer celebrates with the other teammates. Afterward, the scene transitions rapidly and the first replay video is shot in slow motion. Other replay shots with different views, as well as scenes of the gate and net, follow. Finally, the scoreboard appears, the goal event concludes, and the bird's eye view of the entire field signals continuation of the match. If audio information is supplied, the loud cheering sound from the audience can serve as a useful clue for detecting the occurrence of a goal event.

For a home-run event in a baseball game, the scene begins with a typical moment before the pitch. The players are positioned in the designated locations, and thus, this scene is simple to identify by matching the color layout. If an audio signal is detected, the announcer sounds surprised after the pitch. To track the ball, the scene moves very rapidly. If the remaining time is sufficient, the auditorium is presented. Half of the audience and half of the fielder who is the closest to the outfield wall are captured, and the batter running the bases and the pitcher are shown in close-ups. Subsequently, a cheering fan is inserted into the frame. Next, several replay shots in slow motion at different angles are captured. Sometimes, the hit is caught by a high-speed camera. After the replay, the camera shoots the player who returns to home plate and captures a close-up shot of teammates clapping. When the next batter walks toward home plate, the video frame returns to the initial scene.

In a basketball game, there are fewer regular temporal patterns for event classification. The scoreboard is conventionally used to determine whether the shooter scores a two- or three-point shot. When the scoreboard changes, a close-up of the scorer is shown. Offense and defense exchange and pass the ball from the baseline. Events can be identified by players' trajectories and the path of ball passing. For example, pick and roll is a frequently used tactic in basketball games. During the game, the offense is arranged in pairs with the defense. When a pick-and-roll event occurs, the ball keeper faces the defense first. Another offensive player moves close to the ball keeper, obstructing the defense's path. Players' movement paths become discontinuous, and only the ball keeper can move freely. Finally, the ball keeper suddenly moves to the free-throw lane and passes the ball to the shooter, and shooter

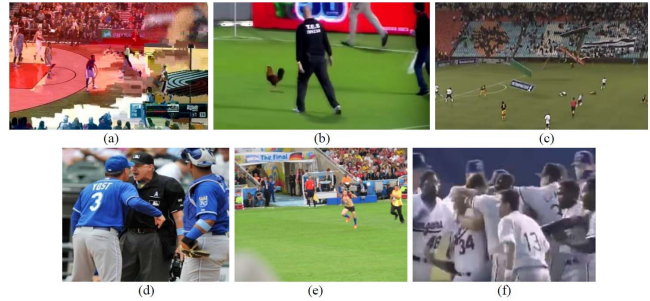


Fig. 6. Examples of unusual situation of the game. (a) Transcoding errors. (b) Animal invasion. (c) Drops result from the strong wind. (d) Arguments between the referee and coach. (e) Streaker interruption. (f) Players fighting.

shoots the ball toward the basket. Accordingly, to determine the event category in more detail, the object trajectory must be tracked precisely and motion analysis must be performed. The framework of 3D camera modeling has also been applied for event classification efficiently. Su and Hsieh [195] upgraded the feature extraction part for event classification from the image domain to the real-world domain. The real-world positions of objects (i.e., athletes) could be determined using an automatic 3D camera calibration method. At the scene level, combining a Bayesian modeling approach with player segmentation and tracking and playing-field knowledge, the system could classify events such as service, net approaches, and baseline rallies.

In summary, domain knowledge enables a reasoning system to obtain a compromised result. However, uncertainty occurring in the playfield, such as a moving bird, a strong wind, and transcoding problems, cannot be avoided, as shown in Fig. 6(a)–(c). Unpredictable interruptions caused by streakers, arguments between the referee and coach, and players fighting are shown in Fig. 6(d)–(f). These extraordinary activities are rare. Nevertheless, context-based semantic analysis can be adopted to assess unusual events.

### C. Contextual Inference and Semantic Analysis

In sports, the context is composed of event outcomes and match summaries. The semantic meanings of the context are critical for the game. The context can typically be obtained from captions or visual cues. Comparisons of recent context-oriented studies are shown in Table V.

#### 1) Context Extraction:

a) *Context from captions:* In a broadcast sports video, a superimposed caption box (SCB) embedded in the frame is a typical means for quickly conveying the ongoing game status to viewers. To detect the SCB and read captions, Guo *et al.* [196] applied SIFT point matching and Tang *et al.* [197] used a fuzzy-clustering neural network classifier. Moreover, Su *et al.* [195] presented a caption model for identifying the meanings of the captions. Zhang *et al.* [198] used a set of modules including caption box localization, segmentation, and a Zernike feature combined with temporal transitional graph models to recognize superimposed text. Sung and Chun [199] developed a knowledge-based numeric caption recognition system to recover valuable information from an enhanced binary image using a multilayer perceptron neural network. Overall, most

TABLE V  
COMPARISONS OF CONTEXT-ORIENTED APPROACHES

Systems	Data source <sup>a</sup>	Analyzing model	Sport genres	Context examples	Application view
[195]	Temporal, Low-v	Model-based segmentation approach, OCR	Baseball, Basketball, Rugby, Soccer	N/A	Sports captions segmentation
[199]	Text	Multiplayer perceptron (MLP) network, Error back-propagation (BP) algorithm	Baseball	Numeric captions	Caption recognition
[204]	Temporal, Text, Low-v	Algorithm integrating caption rule-inference, Visual feature analysis	Baseball	Strikes, Balls, Outs, Scores, Base-occupies	Baseball video retrieval and summarization
[205]	Temporal, Text, Low-v	Superimposed caption box(SCB) color model, Probabilistic labeling algorithm	Basketball, Baseball	Inning, Scoring, Ball, Strike Out, The name of the team, Quarters	Caption template extraction and identification
[206]	Temporal, Text, Low-v	SCB color model, Connected Component Analysis (CCA), SCB interpretation	Basketball, Baseball	Quarters, The score of the home team, The score of the visiting team	Caption template extraction and identification
[223]	Temporal, audio Semantic, Text	Latent semantic index based text event detection, naïve Bayesian network	Soccer	Goal, Shot, Corner, Free Kick, Card, Foul, Offside, Substitute	Event Annotation
[224]	Audio, Text, Low-v	Gray level co-occurrence Matrix (GLCM)	Basketball, Soccer	Goal, Penalty, Red card, Shot on Goal, Attempt, Free Kick, Yellow card, Offside, Foul, Corner, Substitution, 3-point shot, 2-point shot, Free Throw, Rebound, Turnover	Video summarization, Event annotation
[225]	Temporal, Text	Hidden Markov Model(HMM)	Basketball	Shot, Jumper, Lay-up, Dunk, Block, Rebound, Foul, Free throw, Substitution	Events annotation and indexing in the video
[227]	Text	Hierarchical keywords search	Basketball, Soccer	Corner, Shot, Foul, Card, Free kick, Offside, Substitution, Goal, Jumper, Layup, Dunk, Block, Rebound, Free throw	Video annotation, retrieval, Indexing and summarization
[228]	Temporal, Text	Conditional random field model (CRFM), Expectation Maximization (EM)	Basketball, Soccer	Shot, Jumper, Layup, Dunk, Block, Rebound, Foul, Free throw, Substitution, Corner, Card, Free kick, Offside, Goal	Sports video summarization and retrieval
[229]	Audio, Temporal, Text	Basic Criterion, Greedy Criterion, Play-Cut Criterion	Baseball	Homerun, Sacrifice-hit, Two-base-hit, Single-hit, Strike-out...	Video abstraction

<sup>a</sup>Low-v: low-level visual features

caption recognition systems focus on recognizing the text through optical character recognition [200], [201] and employing a template matching algorithm [202], [203] to confirm the caption according to either numbers or text. However, recognizing symbol captions remains challenging, and thus, caption interpretation has emerged as a critical research topic. Lie *et al.* [204] applied visual feature analysis to develop a system that can classify baseball events into 11 semantic categories, which include hitting and nonhitting as well as infield and outfield events. Shih and Huang [205] performed caption interpretation to analyze the SCB in sports videos. They divided the captions into three categories: 1) proportion; 2) inverse; and 3) specific types. Similarly, Shih and Huang [206] presented a method that interprets the SCB where the SCB template cannot be determined *a priori*. A framework designed for golf videos was proposed in [207].

*b) Context from visual cues:* Contextual annotation validates essential insight into the video content while providing access to its semantic meanings. For sports video analysis, the extraction of external metadata from closed captions has been employed in many applications such as video indexing, retrieval, and summarization. In comparison with visual features, contextual information presents a semantic inference regarding the video content. Steinmetz [208] conducted comprehensive research using a context model to perform semantic analysis of video metadata. The research involved multiple knowledge bases. In the semantic inference process, the context is produced dynamically, considering the characteristics of the video metadata and the confidence values. The research findings can be applied in analyzing textual information originating from different sources and verifying different characteristics. Growing evidence indicates that, by examining the semantic context, video analysis can

be effectively used to improve content-based video processing procedures such as keyframe extraction [209], [210], sports video resizing [211], concept-based video indexing [212], and video search [48], [213]. In addition, a mapping mechanism between the context and content designed for baseball videos was presented by Chiu *et al.* [214]. The mechanism focuses on aligning the webcast text and video content. They proposed an unsupervised clustering method called hierarchical agglomerative clustering for detecting the pitch segment in baseball videos. In addition, they applied a modified genetic algorithm to align the context of webcast text and video content.

Most notably, the trend for reading text in images and videos has shifted to the use of advanced learning techniques. For example, Jaderberg *et al.* [215] presented a word detection and recognition system and used a multiway classification model with a deep CNN. Li *et al.* [216] employed a fine-tuned deep CNN to extract visual features, and fed these features into recurrent neural networks (RNNs) to generate representative descriptions for video clips. Similarly, Venugopalan *et al.* [217] proposed a video-to-text generator that uses deep RNNs for the entire pipeline from pixels to sentences.

*2) Text Analytics:* For text analysis, Chen and Yuille [218] used the AdaBoost algorithm to generate a strong classifier and detect texts. Lyu *et al.* [219] presented a multilingual video text detection, localization, and extraction method, specifically for English and Chinese. Xi *et al.* [220] proposed a text information extraction system for news videos. For detecting and tracking the text, multiframe averaging and binarization were applied for recognition. Lienhart and Wernicke [221] presented a method for text localization and segmentation for images and videos, and for extracting information used for semantic indexing. Noll *et al.* [222] revealed the effective

combination of local (the measures are derived from the local attributes of the features, e.g., the angles of a corner) and context similarities for object recognition. A Hough-based matching algorithm was introduced for analyzing corner feature similarities. More recently, Wang *et al.* [223] transformed an event detection problem into a synchronization problem. They annotated soccer video events using match reports with inexact timestamps. The game start time was first detected for a crawled match report, and the NBC was applied to identify the optimal match event.

3) *Semantic Event Extraction*: For semantic event extraction, conventionally, parse trees are first used to extract semantic events and then hierarchically arranged to provide input for a logic processing engine to generate a summary [224]. Zhang *et al.* [225] employed a multimodal framework to extract semantic events in basketball videos. They combined text analysis, video analysis, and text or video alignment for semantics extraction, event moment detection, and event boundary detection. Zhang and Chang [226] developed a system for baseball video event detection and summarization using superimposed caption text detection and recognition. Chen and Chen [227] extracted semantic events from sports webcast text using an unsupervised scheme. A filtering technique is implemented to remove unrelated words, and the remaining words are sorted into categories where keyword extraction is executed to recognize crucial text events. Similarly, a system capable of analyzing and aligning webcast text and broadcast video for semantic event detection was introduced for the automatic clustering of text events and extraction of keywords from webcast text [228]. Nitta *et al.* [229] proposed a method for automatically generating both dynamic and static video abstracts from broadcast sports videos. The method relies on metadata for semantic content and player influence and user preferences for personalization.

#### IV. CHALLENGES AND FUTURE DIRECTIONS

In the past decade, many research achievements in sports video content analysis have been witnessed and have highly influenced the direction of prospective research. In this section, we highlight the prominent challenges identified in the literature and categorize the potential directions in sports video content analysis for future research into three classes: 1) sports video content analysis in the media cloud; 2) large-scale learning techniques for deeper context discovery; and 3) robust visual feature extraction for precise human action recognition.

##### A. Sport Video Content Analysis in Media Cloud

Recently, many portable devices for video recording, such as the camera handset, personal video recorder, and tablet, have been developed. Consequently, a user device can be a content producer. Users are interested in sharing their videos through social networks, and this has become a cultural phenomenon. Currently, the focus of major research efforts in sports video content analysis such as sports type classification [33] and salient event detection [174] has shifted to mobile videos. According to [230], content delivery networks will carry 62%

of Internet traffic worldwide by 2019. The number of devices connected to Internet protocol networks will be three times as high as the global population in 2019. Therefore, research on video caching, transcoding, and adaptive delivery in the media cloud is urgently required [231]. The major challenge is minimizing the overall operational cost regarding the usage of buffering, computing, and bandwidth resources. For example, a mobile video–audio mash-up system called MoVieUp [232] was presented for collating videos shared by client users. It operates in the cloud to aggregate recordings captured by multiple devices from different view angles and different time slices into a single montage. This system was designated for concert recordings. Sports videos do not have a continuous and smooth audio stream for stitching together a video–audio mash-up. In sports applications, the mobile devices normally serve as transceivers to avoid tasks with high computational cost. How to combine sports recordings captured by multiple mobile devices remains an open question. The trend of crowd sourcing raises substantial challenges to cloud service providers. In the future, a synthetic 3D montage of the current sports field could be constructed using breakthrough object and scene alignment and synthesis techniques.

*Scalability* is among the most crucial challenges for cloud mobile media networks [233]. Previous studies have generally addressed the scalability in temporal motion smoothness and spatial resolution instead of content-aware scalability. Client users demand to receive the status of a desired sports competition. Fine-grained video content transcoding is configured under different channel conditions associated with the bandwidth resources and end devices. The client users receive various details on the sports content at varying degrees of precision such as score statistics, audio broadcasts, player portraits, highlight images, an edited video segment, and an entire video broadcast. Another future objective is to combine semantic Web tools and technologies [234] to overcome the aforementioned challenges [235].

##### B. Large-Scale Machine Learning Techniques for Deeper Information Discovery

With the proliferation of media devices, the media content that we access conceals extremely valuable information and knowledge. The era of big data is considerably changing what we know and how we know it in the information world. Big data analytics enables discovering the latent semantic concepts embedded in video content. Sports data analysis is becoming large scale and diversified. Advanced developments in machine learning lead to solutions to real-world large-scale problems.

Since 2006, with the surge of deep learning, research on visual understanding under the paradigm of data-driven learning reached a new height [236]. Deep learning methods such as CNNs, restricted Boltzmann machines, autoencoder, and sparse coding have been adopted and demonstrated to be successful for various computer vision applications. One of the most salient achievements is action and activity recognition through deep learning, which has been performed widely in recent years [94], [96]. Several studies focused on sports video analysis have employed deep learning. For

example, CNNs have demonstrated superiority in modeling high-level visual semantics [89], [215], and RNNs have exhibited promise in modeling temporal dynamics in videos [216], [217].

Nevertheless, there is substantial room for development in sports video analysis with deep learning. Different sports have domain-specific semantic concepts, structures, and features. The challenge of autoencoding and transforming the domain-specific features and models to bridge the semantic gap for sports videos has yet to be overcome. The problem of knowledge interpolation between domains, called domain adaptation (DA), is an emerging research area in the field of machine learning. In brief, training and testing data may come from different domains. Determining how to solve the target-transfer learning problem remains challenging. Numerous approaches to DA have been presented [237], [238], [239]. The research paradigm in sports video analysis will face the same problem in the future. Learning with exchangeable information from different types of sports videos is very challenging and represents a promising research direction. For example, does a model trained using the previous year's data remain valid for adapting to a new season? Furthermore, despite the sports belonging to different categories, do the tactics in a soccer game remain valid for a baseball game? In addition, given the similarity in game rule, can the offense or defense in a table tennis game be applied to a tennis game?

### C. Robust Visual Feature Extracting for Precise Human Action Recognition

With the invention of high-precision sensors and vision-based action acquisition techniques, precise body features can be obtained for modeling key human objects using an effective and high-accuracy scheme. Currently, action recognition has been a mainstream research topic not only in sports [94], [137], [240], but also in movies [241], healthcare [92], video surveillance [242], and building monitoring [243]. A direction for research extension is to develop a robust visual feature extraction scheme for understanding human actions in detail. A sports action video typically depicts multiple human actions simultaneously. To interpret a complex video, an appropriate approach is to adopt multiple dense detailed labels, instead of a single description [240]. Using robust feature extraction approaches, researchers can accurately analyze the context of a sport. Regarding video dynamics, Li *et al.* [248] encoded dynamics of deep features for action recognition. They found that explicit modeling of long-range dynamics is more important than short- and medium-range ones for action recognition.

Recently, several contributions have focused on extracting precise trajectories of objects, such as dense trajectories [63], improved trajectories [95], and multiobject trajectory using spatiotemporal convolution kernels [244] to determine human actions in detail. Ma *et al.* [249] used rich feature hierarchies of CNNs as target object representations and learned a linear correlation filter on each CNN layer to improve tracking accuracy and robustness. When these methods are used in tracking VOs, an intelligent system is required to perform deeper information discovery and to determine the relationships between trajectory data and actions. A recent review

article [245] reported the impact of revamped deep neural networks on action recognition. For sports applications, deep learning was used to analyze basketball trajectories and predict whether a three-point shot is successful [246]. In addition, deep neural networks were used to construct classifiers that can recognize NBA offensive plays [247].

Although we have already seen numerous examples of successful applications of feature extraction and action recognition methods, many open problems remain because of the diversity of game structures among sports domains. Developing a unified framework that enables processing data from diverse sports is still challenging. The tradeoff between commonality and robustness must be overcome. Ultimately, the prospective goal of action recognition in sports is to develop a machine that can read, write, listen to, and speak a voice-over to broadcast sports videos directly.

## V. CONCLUSION

We conducted a comprehensive survey of existing approaches on sports content analysis. Several approaches were reviewed from the viewpoint of content-aware scalability. First, we introduced the fundamentals of content analysis, such as the concept of the content pyramid, the categorization of the sports genre, and an overview of the sports video analytics. Second, we reviewed the state-of-the-art studies conducted in this decade according to the content hierarchical model. The methods of content-aware analysis were discussed with respect to object-, event-, and context-oriented groups. Finally, we reported the prominent challenges identified in the literature and categorized the potential future directions in sports video content analysis. We believe that our survey can advance the field of research on content-aware video analysis for sports.

## ACKNOWLEDGMENT

The author would like to thank the Associate Editor and reviewers for their valuable comments, and also the authors who contributed figures to this survey. Thanks also go to Jeremy Liu (UCSD), Jung-Yu Chen (YZU), Wan-Ting Chen (YZU), and Shu-Ting Huang (YZU) for helping me to improve upon an earlier version of this paper.

## REFERENCES

- [1] M. Marchi and J. Albert, *Analyzing Baseball Data With R*. (Chapman & Hall/CRC The R Series). New York, NY, USA: Taylor & Francis, Oct. 2013.
- [2] J. Albert and J. Bennett, *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*, 1st ed. New York, NY, USA: Copernicus Books, Apr. 2003.
- [3] T. Tango, M. Lichtman, and A. Dolphin, "The book: playing the percentages in baseball," in *CreateSpace Independent Publishing Platform*. New York, NY, USA: Copernicus Books, Apr. 2014.
- [4] S. M. Shea and C. E. Baker, "Basketball analytics: Objective and efficient strategies for understanding how teams win," in *CreateSpace Independent Publishing Platform*. Washington, DC, USA: Potomac Books, Nov. 2013.
- [5] S. M. Shea, *Basketball Analytics: Spatial Tracking*. CreateSpace Independent Publishing Platform, Dec. 2014.
- [6] D. Oliver, *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington, DC, USA: Potomac Books, Nov. 2004.
- [7] *SAP Web Site*. [Online]. Available: <http://go.sap.com/index.html>
- [8] *Vizrt Web Site*. [Online]. Available: <http://www.vizrt.com>
- [9] *Vizrt in Wikipedia Web Site*. [Online]. Available: <http://en.wikipedia.org/wiki/Vizrt>

- [10] *MLBAM Web Site*. [Online]. Available: <http://www.mlbam.com>
- [11] *Sportvision Web Site*. [Online]. Available: <http://www.sportvision.com>
- [12] *Brooks Baseball Web Site*. [Online]. Available: <http://www.brooksbaseball.net>
- [13] *Sportradar Web Site*. [Online]. Available: <http://sportradar.com>
- [14] J. R. Wang and N. Parameswaran, "Survey of sports video analysis: Research issues and applications," in *Proc. Pan-Sydney Area Workshop Vis. Inform. Process.*, 2003, pp. 87–90.
- [15] A. Kokaram *et al.*, "Browsing sports video: Trends in sports-related indexing and retrieval work," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 47–58, Mar. 2006.
- [16] C. B. Santiago, A. Sousa, M. L. Estriga, L. P. Reis, and M. Lames, "Survey on team tracking techniques applied to sports," in *Proc. IEEE Int. Conf. Auto. Intell. Syst.*, Jun. 2010, pp. 1–6.
- [17] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognit.*, vol. 43, no. 8, pp. 2911–2926, Aug. 2010.
- [18] S. F. de Sousa Júnior, A. de A. Araújo, and D. Menotti, "An overview of automatic event detection in soccer matches," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2011, pp. 31–38.
- [19] A. Rehman and T. Saba, "Features extraction for soccer video semantic analysis: Current achievements and remaining issues," *Artif. Intell. Rev.*, vol. 41, no. 3, pp. 451–461, 2014.
- [20] J. Han, D. Farin, and P. H. N. D. With, "Broadcast court-net sports video analysis using fast 3-D camera modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1628–1638, Nov. 2008.
- [21] H.-C. Shih and C.-L. Huang, "Content-based scalable sports video retrieval system," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2005, pp. 1553–1556.
- [22] H.-C. Shih and C.-L. Huang, "Content-based multi-functional video retrieval system," in *Proc. IEEE Int. Conf. Consum. Electron.*, Jan. 2005, pp. 383–384.
- [23] D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 3, pp. 416–430, May 2008.
- [24] J. Wang, C. Xu, and E. Ching, "Automatic sports video genre classification using pseudo-2D-HMM," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2006, pp. 778–781.
- [25] X. Gibert, H. Li, and D. Doermann, "Sports video classification using HMMS," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2003, pp. 345–348.
- [26] A. Ekin and A. M. Tekalp, "Shot type classification by dominant color for sports video segmentation and summarization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2003, pp. 173–176.
- [27] E. Jaser, J. Kittler, and W. Christmas, "Hierarchical decision making scheme for sports video categorisation with temporal post-processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2004, pp. 908–913.
- [28] X. Yuan, W. Lai, T. Mei, X. S. Hua, X. Q. Wu, and S. Li, "Automatic video genre categorization using hierarchical SVM," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 2905–2908.
- [29] L. Li, N. Zhang, L. Y. Duan, Q. Huang, J. Du, and L. Guan, "Automatic sports genre categorization and view-type classification over large-scale dataset," in *Proc. ACM Multimedia*, 2009, pp. 653–656.
- [30] J. You, G. Liu, and A. Perkis, "A semantic framework for video genre classification and event analysis," *Signal Process., Image Commun.*, vol. 25, no. 4, pp. 287–302, 2010.
- [31] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1066–1083, Dec. 2005.
- [32] N. Zhang *et al.*, "A generic approach for systematic analysis of sports videos," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, May 2012, Art. no. 46.
- [33] F. Cricri *et al.*, "Sport type classification of mobile videos," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 917–932, Jun. 2014.
- [34] M. Sugano, T. Yamada, S. Sakazawa, and S. Hangai, "Genre classification method for home videos," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2009, pp. 1–5.
- [35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [36] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
- [37] M. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijets): A novel approach to video indexing and retrieval in multimedia systems," in *Proc. 5th IEEE Int. Conf. Image Process.*, Oct. 1998, pp. 536–540.
- [38] Y. Luo, T. D. Wu, and J. N. Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks," *Comput. Vis. Image Understand.*, vol. 92, nos. 2–3, pp. 196–216, 2003.
- [39] V. Pallavi, J. Mukherjee, A. K. Majumdar, and S. Sural, "Ball detection from broadcast soccer videos using static and dynamic features," *J. Vis. Commun. Image Represent.*, vol. 19, no. 7, pp. 426–436, Oct. 2008.
- [40] M. Leo, P. L. Mazzeo, M. Nitti, and P. Spagnolo, "Accurate ball detection in soccer images using probabilistic analysis of salient regions," *Mach. Vis. Appl.*, vol. 24, no. 8, pp. 1561–1574, 2013.
- [41] Y. C. Jiang, K. T. Lai, C. H. Hsieh, and M. F. Lai, "Player detection and tracking in broadcast tennis video," in *Proc. 3rd Pacific Rim Symp. (PSIVT)*, vol. 5414, 2009, pp. 759–770.
- [42] H. T. Chen, C. L. Chou, W. J. Tsai, S. Y. Lee, and J. Y. Yu, "Extraction and representation of human body for pitching style recognition in broadcast baseball video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–4.
- [43] K. L. Hua, C. T. Lai, C. W. You, and W. H. Cheng, "An efficient pitch-by-pitch extraction algorithm through multimodal information," *Inform. Sci.*, vol. 294, pp. 64–77, Feb. 2015.
- [44] R. Hamid, R. K. Kumar, M. Grundmann, K. Kim, I. Essa, and J. Hodgins, "Player localization using multiple static cameras for sports visualization," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 731–738.
- [45] Y. Liu, D. Liang, Q. Huang, and W. Gao, "Extracting 3D information from broadcast soccer video," *Image Vis. Comput.*, vol. 24, no. 10, pp. 1146–1162, Oct. 2006.
- [46] Y. Liu, D. Liang, Q. Huang, and W. Gao, "Self-calibration based 3D information extraction and application in broadcast soccer video," in *Proc. Asia Conf. Comput. Vis.*, vol. 3852, 2006, pp. 852–861.
- [47] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vis.*, vol. 107, no. 1, pp. 20–39, Mar. 2014.
- [48] H. C. Shih, J. N. Hwang, and C. L. Huang, "Content-based attention ranking using visual and contextual attention model for baseball videos," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 244–255, Feb. 2009.
- [49] M. Wischniewski, A. Belardinelli, W. X. Schneider, and J. J. Steil, "Where to look next? Combining static and dynamic proto-objects in a TVA-based model of visual attention," *Cognit. Comput.*, vol. 2, no. 4, pp. 326–343, 2010.
- [50] C. Bundesen, "A theory of visual attention," *Psychol. Rev.*, vol. 97, no. 4, pp. 523–547, 1990.
- [51] S. H. Khatonabadi and M. Rahmati, "Automatic soccer players tracking in goal scenes by camera motion elimination," *Image Vis. Comput.*, vol. 27, no. 4, pp. 469–479, Mar. 2009.
- [52] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang, "Automatic player detection, labeling and tracking in broadcast soccer video," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 103–113, Jan. 2009.
- [53] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "Closed-world tracking of multiple interacting targets for indoor-sports applications," *Comput. Vis. Image Understand.*, vol. 113, no. 5, pp. 598–611, May 2009.
- [54] L. Sha, P. Lucey, S. Morgan, S. Sridharan, and D. Pease, "Swimmer localization from a moving camera," in *Proc. IEEE Int. Conf. Digit. Image Comput., Technol. Appl.*, Nov. 2013, pp. 1–8.
- [55] B. Chakraborty and S. Meher, "A real-time trajectory-based ball detection-and-tracking framework for basketball video," *J. Opt.*, vol. 42, no. 2, pp. 156–170, Jun. 2013.
- [56] G. J. Liu, X. L. Tang, H. D. Cheng, J. H. Huang, and J. F. Liu, "A novel approach for tracking high speed skaters in sports using a panning camera," *Pattern Recognit.*, vol. 42, no. 11, pp. 2922–2935, Nov. 2009.
- [57] J. Miura, T. Shimawaki, T. Sakiyama, and Y. Shirai, "Ball route estimation under heavy occlusion in broadcast soccer video," *Comput. Vis. Image Understand.*, vol. 113, no. 5, pp. 652–653, May 2009.
- [58] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1830–1837.
- [59] F. Yan, A. Kostin, W. Christmas, and J. Kittler, "A novel data association algorithm for object tracking in clutter with application to tennis video analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 634–641.
- [60] F. Yan, W. Christmas, and J. Kittler, "Layered data association using graph-theoretic formulation with application to tennis ball tracking in monocular sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1814–1830, Oct. 2008.

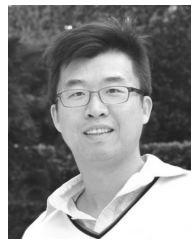


- [61] X. Zhou, L. Xie, Q. Huang, S. J. Cox, and Y. Zhang, "Tennis ball tracking using a two-layered data association approach," *IEEE Trans. Multimedia*, vol. 17, no. 2, pp. 145–156, Feb. 2015.
- [62] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3169–3176.
- [63] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.
- [64] Z. Niu, X. Gao, and Q. Tian, "Tactic analysis based on real-world ball trajectory in soccer video," *Pattern Recognit.*, vol. 45, no. 5, pp. 1937–1947, May 2012.
- [65] H. T. Chen, C.-L. Chou, T.-S. Fu, S.-Y. Lee, and B.-S. P. Lin, "Recognizing tactic patterns in broadcast basketball video using player trajectory," *J. Vis. Commun. Image Represent.*, vol. 23, no. 6, pp. 932–947, Aug. 2012.
- [66] P. Lucey, A. Bialkowski, P. Carr, Y. Yue, and I. Matthews, "How to get an open shot: Analyzing team movement in basketball using tracking data," in *Proc. MIT SLOAN Sports Anal. Conf.*, Feb. 2014, pp. 1–10.
- [67] X. Yu, C. Xu, H. W. Leong, Q. Tian, Q. Tang, and K. W. Wan, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," in *Proc. ACM Multimedia*, 2003, pp. 11–20.
- [68] A. Hervieu, P. Bouthemy, and J.-P. Le Cadre, "A statistical video content recognition method using invariant features on object trajectories," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1533–1543, Nov. 2008.
- [69] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, and I. Matthews, "Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors," in *Proc. MIT SLOAN Sports Anal. Conf.*, Feb. 2014, pp. 1–7.
- [70] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, May 2015, Art. no. 29.
- [71] J. Vales-Alonso, D. Chaves-Diéguez, P. López-Matencio, J. J. Alcaraz, F. J. Parrado-García, and F. J. Gonzalez-Castano, "SAETA: A smart coaching assistant for professional volleyball training," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 8, pp. 1138–1150, Aug. 2015.
- [72] K. Choi and Y. Seo, "Automatic initialization for 3D soccer player tracking," *Pattern Recognit. Lett.*, vol. 32, no. 9, pp. 1274–1282, Jul. 2011.
- [73] X. Yu, N. Jiang, L.-F. Cheong, H. W. Leong, and X. Yan, "Automatic camera calibration of broadcast tennis video with applications to 3D virtual content insertion and ball detection and tracking," *Comput. Vis. Image Understand.*, vol. 113, no. 9, pp. 643–652, May 2009.
- [74] P. J. Figueroa, N. J. Leite, and R. M. L. Barros, "Tracking soccer players aiming their kinematical motion analysis," *Comput. Vis. Image Understand.*, vol. 101, no. 2, pp. 122–135, Feb. 2006.
- [75] J. Ren, J. Orwell, G. A. Jones, and M. Xu, "Tracking the soccer ball using multiple fixed cameras," *Comput. Vis. Image Understand.*, vol. 113, no. 5, pp. 633–642, May 2009.
- [76] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE MultimediaMag.*, vol. 6, no. 1, pp. 22–35, Jan. 1999.
- [77] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Image Vis. Comput.*, vol. 27, no. 5, pp. 545–559, 2009.
- [78] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei, "Linking people in videos with 'their' names using coreference resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 95–110.
- [79] W.-L. Lu, J.-A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3249–3256.
- [80] A. Bialkowski, P. Lucey, P. Carr, Y. Yisong, S. Sridharan, and I. Matthews, "Large-scale analysis of Soccer matches using spatiotemporal tracking data," in *Proc. IEEE Conf. Data Mining*, Dec. 2014, pp. 725–730.
- [81] N. Nitta, N. Babaguchi, and T. Kitahashi, "Extracting actors, actions and events from sports video—A fundamental approach to story tracking," in *Proc. IEEE Int. Conf. Pattern Recognit.*, vol. 4, Sep. 2000, pp. 718–721.
- [82] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh, "Video OCR: Indexing digital news libraries by recognition of superimposed captions," *Multimedia Syst.*, vol. 7, no. 5, pp. 385–395, 1999.
- [83] S. Messelodi and C. M. Modena, "Automatic identification and skew estimation of text lines in real scene images," *Pattern Recognit.*, vol. 32, no. 5, pp. 791–810, 1999.
- [84] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, Oct. 2009.
- [85] T. Yamamoto, H. Kataoka, M. Hayashi, Y. Aoki, K. Oshima, and M. Tanabiki, "Multiple players tracking and identification using group detection and player number recognition in sports video," in *Proc. IEEE IECON*, Nov. 2013, pp. 2442–2446.
- [86] A. Pnevmatikakis *et al.*, "Tracking for context extraction in athletic events," in *Proc. ACM Multimedia Int. Workshop Social, Adapt. Pers. Multimedia Interaction Access*, 2010, pp. 67–72.
- [87] C. Patrikakis *et al.*, "Direct your personal coverage of large athletic events," *IEEE Multimedia Mag.*, vol. 18, no. 4, pp. 18–29, Apr. 2011.
- [88] S. Messelodi and C. M. Modena, "Scene text recognition and tracking to identify athletes in sport videos," *Multimedia Tools Appl.*, vol. 63, no. 2, pp. 521–545, Mar. 2013.
- [89] S. Gerke and K. Müller, "Soccer Jersey number recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 17–24.
- [90] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 538–552, Sep. 2012.
- [91] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understanding*, vol. 115, no. 2, pp. 224–241, Feb. 2011.
- [92] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.
- [93] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognit.*, vol. 47, no. 10, pp. 3343–3361, Oct. 2014.
- [94] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4305–4314.
- [95] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.
- [96] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. NIPS*, 2014, pp. 1–11.
- [97] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 17–24.
- [98] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1691–1703, Sep. 2012.
- [99] P. Pecev, M. Racković, and M. Ivković, "A system for deductive prediction and analysis of movement of basketball referees," *Multimedia Tools Appl.*, vol. 75, no. 23, pp. 16389–16416, 2016.
- [100] H. Ghasemzadeh and R. Jafari, "Coordination analysis of human movements with body sensor networks: A signal processing model to evaluate baseball swings," *IEEE Sensors J.*, vol. 11, no. 3, pp. 603–610, Mar. 2011.
- [101] A. Schmidt, "Movement pattern recognition in basketball free-throw shooting," *Human Movement Sci.*, vol. 31, no. 2, pp. 360–382, Apr. 2012.
- [102] H. Miyamori and S. I. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit.*, Jun. 2000, pp. 320–325.
- [103] V. Kazemi, M. Burenus, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forest," in *Proc. BMVC*, 2013, Sep. 2013.
- [104] M. Burenus, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3618–3625.
- [105] E. Swears, A. Hoogs, J. Qiang, and K. Boyer, "Complex activity recognition using granger constrained DBN (GCDBN) in sports and surveillance video," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 788–795.
- [106] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

- [107] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Pattern Recogn.*, Jun. 2005, pp. 886–893.
- [108] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.
- [109] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, "Motion history image: Its variants and applications," *Mach. Vis. Appl.*, vol. 23, no. 2, pp. 255–281, 2012.
- [110] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 249–257, 2006.
- [111] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D gradients," in *Proc. Brit. Mach. Vis. Assoc.*, 2008, pp. 995–1004.
- [112] I. Mikić, M. M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and tracking using voxel data," *Int. J. Comput. Vis.*, vol. 53, no. 3, pp. 199–223, 2003.
- [113] M. Hofmann and D. M. Gavrila, "Multi-view 3D human pose estimation in complex environment," *Int. J. Comput. Vis.*, vol. 96, no. 1, pp. 103–124, 2011.
- [114] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with physical forces," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 328–357, 2001.
- [115] G. Cheung and T. Kanade, "A real time system for robust 3D voxel reconstruction of human motions," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2000, pp. 714–720.
- [116] S. Y. Cheng and M. M. Trivedi, "Articulated human body pose inference from voxel data using a kinematically constrained Gaussian mixture model," in *Proc. IEEE Comput. Vis. Pattern Recognit. Workshop*, Jun. 2007, pp. 1–11.
- [117] B. C. Shen, H. C. Shih, and C. L. Huang, "Real-time human motion capturing system," in *Proc. IEEE Int. Conf. Image Process.*, Jun. 2005, pp. 1322–1325.
- [118] S. R. Ke, J. N. Hwang, K. M. Lan, and S. Z. Wang, "View-invariant 3D human body pose reconstruction using a monocular video camera," in *Proc. 5th ACM/IEEE Int. Conf. Distrib. Smart Cameras*, Aug. 2011, pp. 1–6.
- [119] S. Zhang *et al.*, "Action recognition based on overcomplete independent components analysis," *Inf. Sci.*, vol. 281, pp. 635–647, Oct. 2014.
- [120] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Jun. 2011, pp. 3361–3368.
- [121] W. Zhou, C. Wang, B. Xiao, and Z. Zhang, "Action recognition via structured codebook construction," *Signal Process. Image Commun.*, vol. 29, no. 4, pp. 546–555, Apr. 2014.
- [122] H. Li, J. Tang, S. Wu, and Y. Zhang, "Automatic detection and analysis of player action in moving background sports video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 351–364, Mar. 2010.
- [123] G. Taylor, R. Fergus, Y. Lecun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 140–153.
- [124] R. Memisevic and G. Hinton, "Unsupervised learning of image transformations," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
- [125] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Understand.*, vol. 117, no. 6, pp. 633–659, Jun. 2013.
- [126] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
- [127] *UCF Sports Action Data Set*. [Online]. Available: [http://csrcv.ucf.edu/data/UCF\\_Sports\\_Action.php](http://csrcv.ucf.edu/data/UCF_Sports_Action.php)
- [128] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 124.1–124.11.
- [129] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2046–2053.
- [130] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1234–1241.
- [131] T. de Campos *et al.*, "An evaluation of bags-of-words and spatio-temporal shapes for action recognition," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2011, pp. 344–351.
- [132] A. Khan, D. Windridge, and J. Kittler, "Multilevel Chinese takeaway process and label-based processes for rule induction in the context of automated sports video annotation," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1910–1923, Oct. 2014.
- [133] *ACASVA Dataset*. [Online]. Available: <http://www.cvssp.org/acasva/Downloads>
- [134] *KTH Computer Vision Group*. [Online]. Available: <http://www.csc.kth.se/cvapg/>
- [135] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori, "Unsupervised discovery of action classes," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1654–1661.
- [136] *SFU VM Lab*. [Online]. Available: [http://www2.cs.sfu.ca/research/groups/VML/people\\_cluster/](http://www2.cs.sfu.ca/research/groups/VML/people_cluster/)
- [137] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
- [138] *Sorts-1M-Dataset*. [Online]. Available: <https://github.com/gtoderici/sports-1m-dataset>
- [139] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4694–4702.
- [140] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 392–405.
- [141] *Stanford Olympic Sports Dataset*. [Online]. Available: <http://vision.stanford.edu/Datasets/OlympicSports/>
- [142] S. M. Safdarnejad, X. Liu, L. Udupa, B. Andrus, J. Wood, and D. Craven, "Sports videos in the wild (SVW): A video dataset for sports analysis," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, May 2015, pp. 1–7.
- [143] *MSU SVW Dataset*. [Online]. Available: <http://www.cse.msu.edu/~liuxm/sportsVideo/>
- [144] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc. IEEE Int. Conf. Multimedia Comput. Syst.*, May 1995, pp. 167–174.
- [145] D. A. Sadlier, N. O'Connor, S. Marlow, and N. Murphy, "A combined audio-visual contribution to event detection in field sports broadcast video. Case study: Gaelic football," in *Proc. 3rd IEEE Int. Symp. Signal Process. Inf. Technol.*, Dec. 2003, pp. 552–555.
- [146] C.-H. Liang, W.-T. Chu, J.-H. Kuo, J.-L. Wu, and W.-H. Cheng, "Baseball event detection using game-specific feature sets and rules," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2005, pp. 3829–3832.
- [147] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden Markov models," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2002, pp. 609–612.
- [148] C. Poppe, S. De Bruyne, and R. Van de Walle, "Generic architecture for event detection in broadcast sports video," in *Proc. 3rd Int. Workshop Autom. Inf. Extraction Media Prod.*, 2010, pp. 51–56.
- [149] D. A. Sadlier and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1225–1233, Oct. 2005.
- [150] C.-Y. Chen, J.-C. Wang, J.-F. Wang, and Y.-H. Hu, "Event-based segmentation of sports video using motion entropy," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2007, pp. 107–111.
- [151] C. Wu, Y.-F. Ma, H.-J. Zhang, and Y.-Z. Zhong, "Events recognition by semantic inference for sports video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2002, pp. 805–808.
- [152] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *Proc. IEEE Workshop CBAIVL*, Dec. 2001, pp. 132–138.
- [153] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recognit. Lett.*, vol. 25, no. 7, pp. 767–775, May 2004.
- [154] D. W. Tjondronegoro and Y.-P. P. Chen, "Knowledge-discounted event detection in sports video," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 5, pp. 1009–1024, Sep. 2010.
- [155] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 415–424, Mar. 2009.

- [156] G. Miao, G. Zhu, S. Jiang, Q. Huang, C. Xu, and W. Gao, "The demo: A real-time score detection and recognition approach in broadcast basketball sports video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2007, pp. 1691–1694.
- [157] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of 'goal' segments in basketball videos," in *Proc. ACM Multimedia*, 2001, pp. 261–269.
- [158] M. H. Hung and C. H. Hsieh, "Event detection of broadcast baseball videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1713–1726, Dec. 2008.
- [159] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia*, 2000, pp. 105–115.
- [160] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proc. IEEE Int. Conf. Multimedia Expo*, Apr. 2003, pp. 401–404.
- [161] H.-C. Shih and C.-L. Huang, "Detection of the highlights in baseball video program," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2004, pp. 595–598.
- [162] M. H. Kolekar, "Bayesian belief network based broadcast sports video indexing," *Multimedia Tools Appl.*, vol. 54, no. 1, pp. 27–54, 2011.
- [163] M. H. Kolekar and S. Sengupta, "Bayesian network-based customized highlight generation for broadcast soccer videos," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 195–209, Jun. 2015.
- [164] M. Petkovic, V. Mihajlovic, W. Jonker, and S. Djordjevic-Kajan, "Multi-modal extraction of highlights from TV Formula 1 programs," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2002, pp. 817–820.
- [165] C.-Y. Chao, H.-C. Shih, and C.-L. Huang, "Semantics-based highlight extraction of soccer program using DBN," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 1057–1060.
- [166] Y. Gong, M. Han, W. Hua, and W. Xu, "Maximum entropy model-based baseball highlight detection and classification," *Comput. Vis. Image Understand.*, vol. 96, no. 2, pp. 181–199, Nov. 2004.
- [167] X. Qian, H. Wang, G. Liu, and X. Hou, "HMM based soccer video event detection using enhanced mid-level semantic," *Multimedia Tools Appl.*, vol. 60, no. 1, pp. 233–255, 2012.
- [168] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted American football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, Aug. 2004.
- [169] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: Automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 92, nos. 2–3, pp. 285–305, Nov./Dec. 2003.
- [170] M. Lazarescu and S. Venkatesh, "Using camera motion to identify types of American football plays," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2003, pp. 181–184.
- [171] R. A. Sharma, V. Gandhi, V. Chari, and C. V. Jawahar, "Automatic analysis of broadcast football videos using contextual priors," *Signal, Image Video Process.*, vol. 11, no. 1, pp. 171–178, 2017.
- [172] G. Kobayashi, H. Hatakeyama, K. Ota, Y. Nakada, T. Kaburagi, and T. Matsumoto, "Predicting viewer-perceived activity/dominance in soccer games with stick-breaking HMM using data from a fixed set of cameras," *Multimedia Tools Appl.*, vol. 75, no. 6, pp. 3081–3119, Mar. 2016.
- [173] H.-T. Chen, H.-S. Chen, and S.-Y. Lee, "Physics-based ball tracking in volleyball videos with its applications to set type recognition and action detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2007, pp. 1097–1100.
- [174] F. Cricri, S. Mate, I. D. D. Curcio, and M. Gabbouj, "Salient event detection in basketball mobile videos," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2014, pp. 63–70.
- [175] W.-L. Lu, K. Okuma, and J. J. Little, "Tracking and recognizing actions of multiple hockey players using the boosted particle filter," *Image Vis. Comput.*, vol. 27, nos. 1–2, pp. 189–205, Jan. 2009.
- [176] H. Pan, P. van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 1649–1652.
- [177] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 252–259, Feb. 2008.
- [178] G. Zhu, Q. Huang, C. Xu, L. Xing, W. Gao, and H. Yao, "Human behavior analysis for highlight ranking in broadcast racket sports video," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1167–1182, Oct. 2007.
- [179] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic analysis of soccer video using dynamic Bayesian network," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 749–760, Aug. 2006.
- [180] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using Bayesian network and Copula," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 291–304, Feb. 2014.
- [181] R. B. Nelsen, *An Introduction to Copulas*, 2nd ed. New York, NY, USA: Springer, 2006, pp. 7–14.
- [182] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.
- [183] D. Zhong and S.-F. Chang, "Real-time view recognition and event detection for sports video," *J. Vis. Commun. Image Represent.*, vol. 5, no. 3, pp. 330–347, Sep. 2004.
- [184] K. Choros, "Automatic detection of headlines in temporally aggregated TV sports news videos," in *Proc. IEEE Int. Symp. Image Signal Process. Anal.*, Sep. 2013, pp. 147–153.
- [185] H.-T. Chen, C.-L. Chou, W.-C. Tsai, S.-Y. Lee, and B.-S. P. Lin, "HMM-based ball hitting event exploration system for broadcast baseball video," *J. Vis. Commun. Image Represent.*, vol. 23, no. 5, pp. 767–781, Jul. 2012.
- [186] W. Zhou, A. Vellaikal, and C. C. J. Kuo, "Rule-based video classification system for basketball video indexing," in *Proc. ACM Multimedia*, 2000, pp. 213–216.
- [187] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via multimodal data mining," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 38–46, Mar. 2006.
- [188] G. Zhu *et al.*, "Event tactic analysis based on broadcast sports video," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 49–67, Jan. 2009.
- [189] T. D'Orazio, M. Leo, P. Spagnolo, M. Nitti, N. Mosca, and A. Distanto, "A visual system for real time detection of goal events during soccer matches," *Comput. Vis. Image Understand.*, vol. 113, no. 5, pp. 622–632, May 2009.
- [190] T. Saba and A. Altameem, "Analysis of vision based systems to detect real time goal events in soccer videos," *J. Appl. Artif. Intell.*, vol. 27, no. 7, pp. 656–667, Aug. 2013.
- [191] C. G. M. Snoek and M. Worring, "Time interval maximum entropy based event indexing in soccer video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2003, pp. 481–484.
- [192] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using Twitter," in *Proc. ACM Int. Conf. Intell. User Interfaces*, 2012, pp. 189–198.
- [193] 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.
- [194] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2012–2019.
- [195] Y.-M. Su and C.-H. Hsieh, "A novel model-based segmentation approach to extract caption contents on sports videos," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1829–1832.
- [196] J. Guo, C. Gurrin, S. Lao, C. Foley, and A. F. Smeaton, "Localization and recognition of the scoreboard in sports video based on SIFT point matching," in *Proc. ACM MMM*, 2011, pp. 337–347.
- [197] X. Tang, X. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 961–971, Nov. 2002.
- [198] D. Zhang, R. K. Rajendran, and F.-S. Chang, "General and domain specific techniques for detecting and recognizing superimposed text in video," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, Sep. 2002, pp. 593–596.
- [199] S.-H. Sung and W.-S. Chun, "Knowledge-based numeric open caption recognition for live sportscast," in *Proc. IEEE 16th Int. Conf. Pattern Recognit.*, Aug. 2002, pp. 822–825.
- [200] R. Lienhart, "Video OCR: A survey and practitioner's guide," in *Video Mining*. Norwell, MA, USA: Kluwer, 2003, pp. 155–184.
- [201] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR research and development," *Proc. IEEE*, vol. 80, no. 7, pp. 1029–1058, Jul. 1992.
- [202] G. J. Vanderbrug and A. Rosenfeld, "Two-stage template matching," *IEEE Trans. Comput.*, vol. 26, no. 4, pp. 384–393, Apr. 1997.
- [203] H.-C. Shih and K.-C. Yu, "SPiraL aggregation map (SPLAM): A new descriptor for robust template matching with fast algorithm," *Pattern Recognit.*, vol. 48, no. 5, pp. 1707–1723, May 2015.

- [204] W.-N. Lie and S.-H. Shia, "Combining caption and visual features for semantic event classification of baseball video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2005, pp. 1254–1257.
- [205] H.-C. Shih and C.-L. Huang, "Content extraction and interpretation of superimposed captions for broadcasted sports videos," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 333–346, Sep. 2008.
- [206] H.-C. Shih and C.-L. Huang, "Semantics interpretation of superimposed captions in sports videos," in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, Oct. 2007, pp. 235–238.
- [207] C. Jung and J. Kim, "Player information extraction for semantic annotation in golf videos," *IEEE Trans. Broadcast.*, vol. 55, no. 1, pp. 79–83, Mar. 2009.
- [208] N. Steinmetz, "Context-aware semantic analysis of video metadata," Ph.D. dissertation, Dept. Faculty Math. Natural Sci., Univ. Potsdam, Potsdam, Germany, May 2014.
- [209] S.-P. Yong, J. D. Deng, and M. K. Purvis, "Wildlife video key-frame extraction based on novelty detection in semantic context," *Multimedia Tools Appl.*, vol. 62, no. 2, pp. 359–376, 2013.
- [210] H.-C. Shih, "A novel attention-based key-frame determination method," *IEEE Trans. Broadcast.*, vol. 59, no. 3, pp. 556–562, Sep. 2013.
- [211] L. Wu, Y. Gong, X. Yuan, X. Zhang, and L. Cao, "Semantic aware sport image resizing jointly using seam carving and warping," *Multimedia Tools Appl.*, vol. 70, no. 2, pp. 721–739, 2014.
- [212] M.-F. Weng and Y.-Y. Chuang, "Cross-domain multicue fusion for concept-based video indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1927–1941, Oct. 2012.
- [213] H. C. Shih, C. L. Huang, and J. N. Hwang, "An interactive attention-ranking system for video search," *IEEE Multimedia Mag.*, vol. 16, no. 4, pp. 70–80, Apr. 2009.
- [214] C.-Y. Chiu, P.-C. Lin, S.-Y. Lin, T.-H. Tsai, and Y.-L. Tsai, "Tagging webcast text in baseball videos by video segmentation and text alignment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 999–1013, Jul. 2012.
- [215] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.
- [216] G. Li, S. Ma, and Y. Han, "Summarization-based video caption via deep neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1191–1194.
- [217] S. Venugopalan, H. Xu, J. Donahu, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," in *Proc. NAACL HLT*, 2015, pp. 1494–1504.
- [218] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2004, pp. 366–373.
- [219] M. R. Lyu, J. Song, and M. Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 243–255, Feb. 2005.
- [220] J. Xi, X.-S. Hua, X.-R. Chen, L. Wenyin, and H.-J. Zhang, "A video text detection and recognition system," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2001, pp. 873–876.
- [221] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 256–268, Aug. 2002.
- [222] D. Noll, M. Schwarzingler, and W. Seelen, "Contextual feature similarities for model-based object recognition," in *Proc. IEEE 4th Int. Conf. Comput. Vis.*, May 1993, pp. 286–290.
- [223] Z. Wang, J. Yu, and Y. He, "Soccer video event annotation by synchronization of attack–defense clips and match reports with coarse-grained time information," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
- [224] M. A. Refaey, W. Abd-Almageed, and L. S. Davis, "A logic framework for sports video summarization using text-based semantic annotation," in *Proc. 3rd Int. Workshop Semantic Media Adapt. Pers.*, Dec. 2008, pp. 69–75.
- [225] Y. Zhang, C. Xu, Y. Rui, J. Wang, and H. Lu, "Semantic event extraction from basketball games using multi-modal analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2007, pp. 2190–2193.
- [226] D. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. ACM MULTIMEDIA*, 2002, pp. 315–318.
- [227] C.-M. Chen and L.-H. Chen, "A novel approach for semantic event extraction from sports webcast text," *Multimedia Tools Appl.*, vol. 71, no. 3, pp. 1937–1952, Aug. 2014.
- [228] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1342–1355, Nov. 2008.
- [229] N. Nitta, Y. Takahashi, and N. Babaguchi, "Automatic personalized video abstraction for sports videos using metadata," *Multimedia Tools Appl.*, vol. 41, no. 1, pp. 1–25, Jan. 2009.
- [230] *The Zettabyte Era: Trends and Analysis*, CISCO, document 1465272001812119, Jun. 2014.
- [231] Y. Jin, Y. Wen, and C. Westphal, "Optimal transcoding and caching for adaptive streaming in media cloud: An analytical approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1914–1925, Dec. 2015.
- [232] Y. Wu, T. Mei, Y.-Q. Xu, N. Yu, and S. Li, "MoVieUp: Automatic mobile video mashup," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1941–1954, Dec. 2015.
- [233] Y. Wen, X. Zhu, J. J. P. C. Rodrigues, and C. W. Chen, "Cloud mobile media: Reflections and outlook," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 885–902, Jun. 2014.
- [234] P. Hitzler and K. Janowicz, "Linked data, big data, and the 4th paradigm," *Semantic Web*, vol. 4, no. 3, pp. 233–235, Jul. 2013.
- [235] K. Mahmood and H. Takahashi, "Cloud based sports analytics using semantic Web tools and technologies," in *Proc. IEEE 4th Global Conf. Consum. Electron.*, Oct. 2015, pp. 431–432.
- [236] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.
- [237] S. Chopra, S. Balakrishnan, and R. Gopalan, "DLID: Deep learning for domain adaptation by interpolating between domains," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1–8.
- [238] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2288–2302, Nov. 2014.
- [239] J. Donahue et al., "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
- [240] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. (Jul. 2015). "Every moment counts: Dense detailed labeling of actions in complex videos." [Online]. Available: <https://arxiv.org/abs/1507.05738>
- [241] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 219–238, 2016.
- [242] C.-B. Jin, S. Li, T. D. Do, and H. Kim, "Real-time human action recognition using CNN over temporal images for static video surveillance cameras," in *Proc. PCM*, vol. 9315, 2015, pp. 330–339.
- [243] H. C. Shih, "A robust occupancy detection and tracking algorithm for the automatic monitoring and commissioning of a building," *Energy Buildings*, vol. 77, pp. 270–280, Jul. 2014.
- [244] K. Knauf, D. Memmert, and U. Brefeld, "Spatio-temporal convolution kernels," *Mach. Learn.*, vol. 102, no. 2, pp. 247–273, 2016.
- [245] S. Herath, M. Harandi, and F. Porikli. (May 2016). "Going deeper into action recognition: A survey." [Online]. Available: <https://arxiv.org/abs/1605.04988>
- [246] R. C. Shah and R. Romijnders. (Aug. 2016). "Applying deep learning to basketball trajectories." [Online]. Available: <https://arxiv.org/abs/1608.03793>
- [247] K. C. Wang and R. Zemel, "Classifying NBA offensive plays using neural networks," in *Proc. MIT Sloan Sports Anal. Conf.*, Mar. 2016, pp. 1–9.
- [248] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos, "VLAD3: Encoding dynamics of deep features for action recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1951–1960.
- [249] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3074–3082.



**Huang-Chia Shih** received the B.S. degree, with the highest honors, in electronic engineering from the National Taipei University of Technology, Taipei, Taiwan in 2000, and the M.S. and Ph.D. degrees in electrical engineering (EE) from the National Tsing Hua University in 2002 and 2008, respectively.

He joined the faculty at the Yuan Ze University (YZU) in 2010, where he is currently an Associate Professor in the EE Department. He served as a Visiting Scholar and Visiting Professor in the Department of EE, University of Washington, from 2006 to 2007, and in 2013, respectively, and a Visiting Professor in the John von Neumann Faculty of Informatics, Obuda University, Hungary, in 2011. He has authored or co-authored more than 50 papers in international journals and conferences. His research interests include content-based multimedia processing, pattern recognition, and human-computer interaction (HCI).

Dr. Shih was a recipient of the Youth Medal from China Youth Corps in March 2017, an Outstanding Youth Electrical Engineer Award from the Chinese Institute of Electrical Engineering in December 2015, the YZU Young Scholar Research Award from YZU in March 2015, the Kwoh-Ting Li Young Researcher Award from the ACM Taiwan Chapter in April 2014, the Pan Wen Yuan Exploration Research Award from the Pan Wen Foundation in May 2013, and was awarded the best paper in ISCE2013 and student paper (advisor) in GCCE2015 from the IEEE Consumer Electronics Society.