Understanding Sports Tactics Based on Grammatical Video Representation

Anonymous CVPR submission

Paper ID 1542

Abstract

In this paper, a probabilistic framework is proposed to represent high level semantic concepts of temporal visual events. It is applied to sports videos to address the inference of common sports tactics through multilevel probabilistic parsing trees. In sports videos, a higher level concept typically consists of multiple lower level concepts. Such hierarchical relationships are usually apparent to human cognition, but their ambiguity and uncertainty are serious challenges to existing cognitive computing techniques. The proposed framework is based on latent probabilistic context-free grammar (LCFG), which assumes a generative relationship between the athletes' behaviors and the underlying tactics within sports videos. Furthermore, by jointly designing the tactic grammars with a set of discriminative intermediate features, the optimal interpretation is obtained through a modified CYK parsing algorithm. The classical pick-and-roll attacking tactic in basketball game is studied as the target tactic in our experimental work. The hierarchical parsing results demonstrate the unique capability of the proposed method in analyzing the complex compositions of such visual events.

1. Introduction

Understanding real-world temporal visual events in a video sequence is a nontrivial technical challenge. Many previous efforts have been focused on classification of segmented video clips while ignoring long-term temporal and logical relationships between consecutive concepts and sub-concepts. Such a approach fails to consider some fundamentally important aspects in human cognitive process. Firstly, semantic concepts are derived from both observations and domain knowledge. Any computational model without encapsulating such knowledge is unlikely to perform robustly in real applications. Secondly, beyond certain variations in observations, visual events are frequently constructed from key concepts and sub-concepts according to certain underlying temporal and logical structures. Simply modeling these conceptual primitives as independent random occurrences or some first-order Markov process is inadequate to explore their inherent hierarchy.

In order to address these challenges in a unified framework, in this paper, we propose a latent probabilistic context-free grammar (LCFG) to better analyze temporal visual events in real-world video sequences. The set of domain knowledge is described in grammatical rules. The context-free property simplifies the model construction and parsing. The derived parsing tree is able to reveal the logical hierarchy of the concepts in a probabilistic manner.

In this paper, the proposed method is specifically applied to basketball videos to demonstrate its ability of describing real-world sports tactics. In achieve this, we first introduce a set of discriminative tactics features that can bridge higher level concepts with low level athlete appearance and behavior. Then we use LCFG and its derived inference framework to hierarchically construct a single tactic model.

2. Related Works

High level visual semantic analysis is frequently application specific. However, such task generally involves multiple stages and components that may have been well studied within different context. This section summarizes some of the previous works that are related to the various aspects of tactics modeling in sports videos.

2.1. Sports Video Analysis

In recent years, there are significant amount of studies on sports video analysis from both academia and industry. The advances in low level image processing and computer vision have brought us many essential tools, which include video segmentation [19], single or multiple athletes tracking [18, 17, 10, 23], player identifications [14], and court rectification [16, 10] etc. Some works directly use low level athlete features to extract high level sports semantics [28, 29, 7] and collect statistics for individual games [22].

Most existing sports analysis systems so far are concentrated on collecting athletes' statistics and performances. For example, there are commercial applications collecting basketball stats for professional games.

In contrast to most conventional computer vision applications, sports videos generally have more dynamics in motion, objects, and scene. Most of sports videos are captured in motion, with crowded background, and heavy occlusions among athletes. Therefore detecting and tracking athletes in sports game is very challenge, which requires significant amount of specialized design efforts[24].

On the other hand, sports videos also have certain inherent structures. The games are played in pre-defined standard courts. The behaviors of athletes are not only driven by physical properties and sports rules, but also guided by the formations and tactics according to their personal skills and team intentions. It is therefore desirable, especially for team sports, to have more sophisticated tools that can potentially explore the collaboration, formation and tactic among athletes.

2.2. Space Time Model

In most video analysis works, space time models are essential in exploring the underlying spatial and temporal relationships. Hidden Markov Models (HMMs) and its variants are popular choices in such applications[2, 21, 12, 9]. The spatial appearance features may includes color or gradient histogram, optical flow, and bag-of-feature etc. These features can be simply aggregated to form higher dimensional feature vectors, which frequently produce enhanced performance. However, high dimensionality typically requires more modeling and computation complexity. In this paper, only color histogram is used to build the feature vector for athletes in games.

In common HMMs, the latent state variables are mostly restricted to a first-order Markov relationship. When applying HMMs to higher level concept modeling, there are two major limitations. Firstly, the expressiveness of an HMM is too shallow to characterize any complex domain knowledge. All latent states are equal entities without any hierarchy, and any prior knowledge on this process can only be encapsulated in the transition and prior probabilities. Secondly, the latent states are essentially clusters in feature spaces. There is no guarantee of true semantic meaning of these states. Hence, this generated state sequence is frequently hard to interpret.

2.3. Grammatical Model

Context-free grammar (CFG) models have been used in gesture[5], action[20, 19] and event recognitions[11, 15, 27]. As the extension of the Markov process, CFG can explicit express context-free logic in grammar rules. For high level concepts modeling, especially in team sports, the objects of interests (i.e. athletes) are collaborating based on a variety of well understood sports tactics. In this work, we will encode such tactics into a set of CFG grammar rules. By combining the traditional space time modeling and the syntactic event construction, we propose a unified probabilistic framework for describing and extracting high level concepts from both observations and domain knowledge.

3. Sports Video Preparation

Our primary concern is syntactic modeling of high level sports concepts. But as in typical computer vision applications, significant amount of video pre-processing is required to support such high level analysis. Hence, in this section, we summarize the preparation works and methods that we specifically designed for broadcast sports videos.

3.1. Court Rectification

Commercial broadcast sports videos are typically presented through several cameras in fixed positions. Those cameras are pre-positioned as close-up, court-views, and commercial-views etc. Among all those positions, court-views are most frequent and informative ones in the video. However, unlike in surveillance scenario, the background of the court-views video are usually in motion. Thus in order to obtain the absolute position and trajectory of athletes in the game, it requires a prior court rectification process. Fortunately, the courts appear in the video are well defined. Therefore, the court can be rectified by estimating the homograph perspective transformation between the court template and its appearances.

The perspective transformation aims to project a list of 2-D coordinates to a 3-D perspective space. Suppose the point $\mathbf{p} = [x, y]^T$ on the court template will be projected to its corresponding point $\mathbf{p}' = [x', y']^T$ in the perspective space. Such relationship can be determined by transformation matrix $\mathbf{H}$ through the perspective equation Eq.1,

$$\begin{bmatrix} x'w \\ y'w \\ w \end{bmatrix} = \mathbf{H} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

where w, the scale factor, is set to be 1.

By collecting adequate pairs of $(\mathbf{p}, \mathbf{p}')$, $\mathbf{H}$ can be estimated though least square method [16]. In our application, the $\mathbf{p}'$ are pre-defined on the court template. On the other hand, the $\mathbf{p}$ are hidden and corrupted in the video frame. Thus it requires a set of cleaning operations to eliminate unwanted objects and color channels. In Fig. 1, we demonstrate the methodology on a single frame. At the beginning, the non-court floor colors have been eliminated to emphasize the court structure. Then we filter out the small size contours to eliminate the objects that might be blocking the court. The estimation of homograph transformation is then performed several iterations until it converges. The series of $\mathbf{H}$ will also be used in the following steps to determine the absolute positions of the athletes from their relative positions.

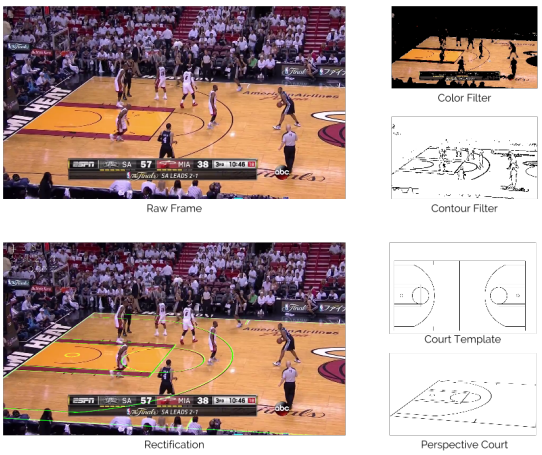

Figure 1. Court rectification through iteratively matching the frame contour and transformed court template.

3.2. Player Tracking

Detecting human objects in images and videos is a classical computer vision topic. Although athletes tend to have more variations in appearance and motion compared to normal pedestrians, the fundamental techniques are quite similar [1, 4].

Most object tracking methods are based on certain time-space models, which contain at least two individual steps. Firstly, there is a probabilistic detector to estimate the likelihood that an object of interest appears at a given position in a frame. Typically, the athelete of interest is characterized in the feature spaces which commonly includes color histograms[17], histogram of gradient[13], and even more advanced features from deformable part model[14, 26]. Secondly, the spatially detected objects are correlated through a temporal model across consecutive video frames. Such temporal correlation can be explore through mean-shift, Kalman filter, particle filter or even more advanced temporal state space solutions.

Since the athlete tracking is relatively independent to the high level syntactic modeling. We adopted color histograms with mean-shift algorithm to reduce the computational cost while processing high definition sports videos. Fig.2 shows the mean-shift athletes tracking steps and results in our application. The rectified frames which came from the prior step are used here to back-project the pre-defined two jersey color histograms.

Notice that the trajectories we obtained in this section are relative positions in the frames. In order to obtain the actual positions and trajectories on court, we have to combine the relative positions along with the previous rectification results $\mathbf{H}$ together through Eq.1.

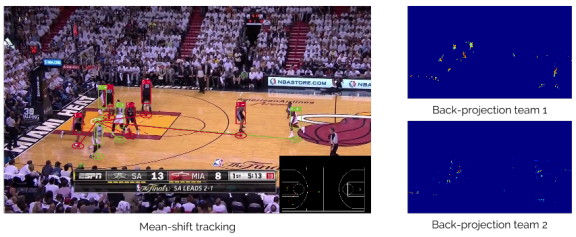

Figure 2. Mean-shift athletes tracking

4. Discriminative Tactic Features

In this section, we propose several tactic features based on certain domain knowledge. Similar to the hidden states in HMMs, the true underlying states can not be observed directly from the video sequence. Hence, it requires a probabilistic inference from the feature vectors to corresponding states. From the video pre-processing steps introduced in the previous section, we have already collected the trajectory of each athletes in a game. We will then use these low level inputs and our domain knowledge to construct some intermediate concepts.

The following features which are specifically designed for basketball sports will be used to discriminate the latent states in the grammatical model.

4.1. Defence Scores

A majority of basketball tactics aim to create miss defence (open shot) for specific attacker or mismatch the original assigned defenders during a period of time. By modeling the current defence quality, we will not only be able to measure the quality of the tactic being deployed in the game, but also construct a high level feature to distinguish the tactic. A good defence for an attacker means the defender stands in a good *distance* and *direction* towards to the attacker. The *distance* can be measured simply in Euclidean distance, and the good *direction* should be pointing from the attacker to the hoop.

We initially transform the athlete trajectories into the hoop centered polar space as $(x, y) \rightarrow (r, \theta)$. Then the *distance* as well as the *direction* can be easily measured through δr and $\delta\theta$ in polar space. In Eq.2, a Gaussian kernel is used to measure the defence quality between the preferred model parameterized by $(\hat{r}, 0)$ and the current athlete pair in the frame.

$$\mathbf{D}(\delta r, \delta d) = \exp\left(-\left(\frac{(\delta r - \hat{r})^2}{2\sigma_r^2} + \frac{\delta\theta^2}{2\sigma_\theta^2}\right)\right) \quad (2)$$

For a given attacker, the quality of the defence for each different defender positions can be visualized in Fig.3. Good defence positions are always laid between the attacker and the hoop. The more defender misses the good position,

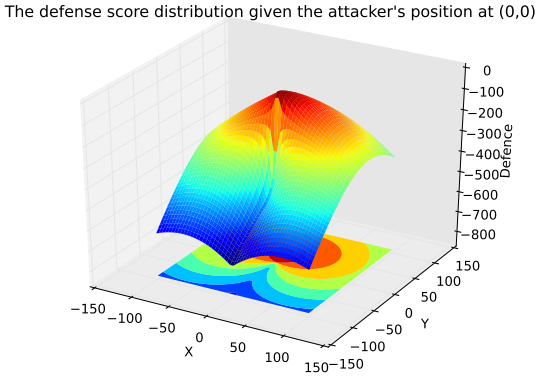

Figure 3. The heat map of defence scores for different defender positions. Attacker is assumed at position $(0, 0)$. Hoop direction is along the positive Y axis.

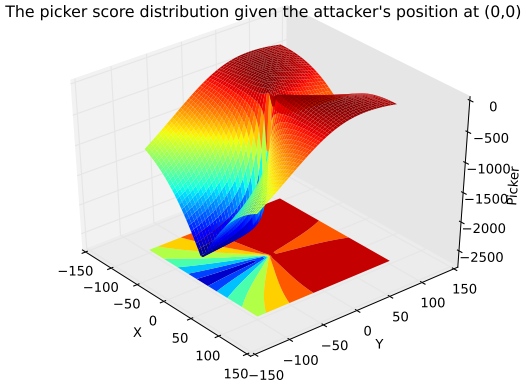

Figure 4. The heat map of picker scores for different teammate positions. Ball-handler is assumed at position $(0, 0)$. Hoop direction is along the positive Y axis.

the less the defense quality becomes, especially when the attacker has already surpassed the defender.

4.2. Picker Selection

Besides modeling the defence, the cooperation among teammates is also a common and important feature in team sports. We consider the cooperation in a pick-and-roll basketball game. The picker in the tactic is an athlete whom is picked by the ball handler to block his defender while dribbling the ball.

Initially the picker will be standing roughly around the ball-handler's 2 and 10 o'clock positions. Similar to the way we measure the defence based on the relative positions of the athletes, we will once again use a Gaussian kernel to mimic the picker selection. Since there are generally two potential sweet points for picking the picker, we will use two component Gaussian mixture for this feature. In Eq.3, Gaussian mixture is used to here to measure the probability of a teammate positioned at $(\delta r, \delta d)$ acting as the picker given the ball-handler's initial position at $(0, 0)$. This model is parametrized by $(\mathbf{w}, \hat{\mathbf{r}}, \hat{\mathbf{d}})$, and $g(|)$ is standard Gaussian function.

$$\mathbf{Pk}(\delta r, \delta d) = \sum_{i=1}^{2} w_i g(\delta r, \delta d | \hat{r}_i, \hat{d}_i) \qquad (3)$$

For a given ball-handler, the probability of a teammate acting as the picker in this tactic is approximated and visualized in Fig.3. The closer the teammate is to those two sweet spots, the higher the probability that he will be acting as a picker will become. When the ball handler has already surpassed his teammate, this teammate will almost impossible to be selected as a picker to screen the defender.

4.3. Other Features

We have presented the methods that construct two intermediate features from the basketball domain knowledge. Although we do use other tactic features (e.g. average athletes distances) in this work, it is not necessary to enumerate all of them. The principle idea is straightforward, we are trying to build discriminative features from domain knowledge that can best distinguish the target concepts and subconcepts among each others. For example, in Fig.5, we show the features for two pairs of athletes in a real pick-and-roll scenario. Both features can well follow the behaviors and logic of the ongoing tactic. Initially, the ball-handler is looking for a good picker who has a relatively high picking score. Then, after the picker stepped in and the tactic moves on, the ball-handler attracts both defenders' attention, which gives the picker a miss defence opportunity. This opportunity is well quantified from the significant defence score drop in the tactic feature space.

5. Syntactic Method

5.1. Latent Context-free Grammar

According to formal language theory, *regular grammar* is equivalent to automata and Markov chain under certain restrictions [8]. Any rule in regular grammar essentially characterize the left to right symbol generation process. In Chomsky hierarchy, context-free grammar (CFG) is a super-set of regular grammar. It is therefore straightforward to use CFG to model any process that can be modeled through regular grammar, i.e. Markov chain. In addition, CFG as well as its probabilistic extension, probabilistic context-free grammar (PCFG), possess more representation power in data modeling. PCFG extends CFG in the

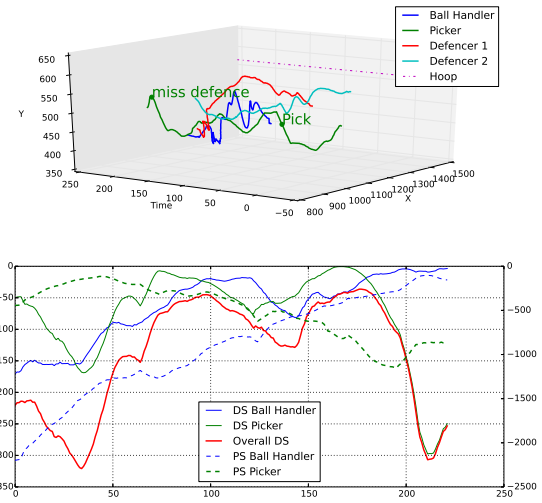

Figure 5. Likelihood evolution through time

way similar to how HMM extends regular grammar. The *context-free* property allows the grammar to create nested, long distance pairwise correlations among symbols in data strings. Traditional PCFG have been widely used in natural language processing, biomedical analysis [6, 25].

In order to apply CFG to continuous observation samples, the deterministic terminal symbols have to be modified to become latent terminal symbols. They are similar to the hidden states in HMMs. With such modification, we here introduce the latent context-free grammar (LCFG) framework. In this work, we use Gaussian terminals in LCFG. An LCFG model is defined as $\mathbf{G} = (\mathbf{N}, \mathbf{T}, \mathbf{R}, S, p)$ where,

- $\mathbf{N}$ is a finite set of non-terminal symbols.
- $\mathbf{T}$ is a finite set of parameterized Gaussian multivariate terminals.
- $\mathbf{R}$ is a finite set of the rules of the form

$$X \to Y_1\ Y_2\ \ldots\ Y_n,$$

  where $X \in \mathbf{N}$ and $Y \in (\mathbf{N} \cup \mathbf{T})$ for $i = 1, 2, \ldots, n$.
- $S \in \mathbf{N}$ is the start non-terminal of the model.
- p is the set of probabilities of every production rules $(\alpha \to \beta) \in \mathbf{R}$. For any $X \in \mathbf{N}$, there is naturally a probability constraint,

$$\sum_{\substack{(\alpha \to \beta) \in \mathbf{R} \\ \alpha = X}} p(\alpha \to \beta) = 1$$

Compared with the traditional PCFG, when the production rules have a terminal symbol on the right hand side, the Gaussian density $f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used to model the probability of the data sample $\mathbf{x}$ given terminal $T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

A fundamental assumption in language modeling is that the output symbols are generated by a series derivations starting from the initial non-terminal S. LCFG is a probabilistic generative model. Instead of using sequential states paths as seen in HMMs, CFMs use parsing trees to track the symbol derivation processes. Similar to HMM, there are three basic problems for this dynamic data model.

- Alignment: what is an optimal alignment of an observation to a parameterized LCFG?
- Scoring: what is the probability of an observation given a parameterized LCFG?
- Learning: how to estimate the probability parameters for an LCFG?

In this work, we are focusing on the first problem, i.e. what and how to find the optimum alignment of an observations series $\mathbf{O}$ given a parameterized $\mathbf{G}$. This is equivalent to the decoding problem in HMM. LCFG as well as HMM are special cases of probabilistic graph models where the conditional dependence are structured either by a tree structure, or by a chain structure. In HMM, the Viterbi algorithm allows the belief recursively propagats from left to right. Similarly, in LCFG, the CYK algorithm [3] is specially designed to dynamically calculate the best alignment among all possible parsing trees for a given model to generate $\mathbf{O}$.

We modify the conventional CYK's initialization step to accommodate the emission distributions in LCFG. More specifically, we define a quantity $\alpha(i, j, v)$ for a parsing subtree rooted at non-terminal v covering partial observations $\mathbf{o}_i, \ldots, \mathbf{o}_j$. In order to recursively find the best alignment, three essential steps are involved from bottom to top.

Initialization for $i = 1$ to L, $v \in \mathbf{N}$

$$\alpha(i, i, v) = \max_t \{\log p(t|v) + \ log f(\mathbf{o}_i|t)\}$$

where L is the length of the observations $\mathbf{O}$.

Iteration for $i = 1$ to $L - 1$, $j = i + 1$ to L, and $v \in \mathbf{N}$

$$\alpha(i, j, v) = \max_{y,z} \max_{k=i}^{j-1} \{\alpha(i, k, y) + \alpha(k + 1, j, z) + \log p(v \to y\ z)\} \quad (4)$$

This is a recursive bottom-up iteration step that propagates the belief between non-terminal and slices of the observations.

Termination The iteration will eventually cover the entire observations as well as the complete grammar. The best alignment $\hat{\pi}$ will be given as,

$$\mathbf{P}(\mathbf{O}, \hat{\pi}|\mathbf{G}) = \alpha(1, L, S)$$

where S is the start non-terminal of the model.

5.2. Tactic LCFG

A sports tactic is a concept composed of a hierarchically connected sequence of athlete formations and trajectory states. The behaviors of the athletes vary dramatically, but they are guided by the logics of the intended tactics. In order to emphasize such consecutive logical process, we use LCFG as a syntactic method to model the behaviors of these athletes.

Take pick-and-roll as an example, we could use the following context-free rules to describe the tactic.

$$\begin{aligned} \text{PR} &\rightarrow \text{Pick Roll } [1.0] \\ \text{Pick} &\rightarrow \text{picker block } [1.0] \\ \text{picker} &\rightarrow \text{picker } [p] \mid \nu\ [1-p] \\ \text{block} &\rightarrow \text{block } [p] \mid \nu\ [1-p] \\ \text{Roll} &\rightarrow \text{roll } [p] \mid \nu\ [1-p] \end{aligned}$$

Fig.6 shows the hierarchy when LCFG is applied on a specific sports tactic. The lower level features are essentially from raw trajectories. The latent level composes of many discriminative features which are designed based on basketball domain knowledge. By applying this model, we can resolve the ambiguity among the latent states from trajectories while preserving the high level logic through the syntactic tree structure.

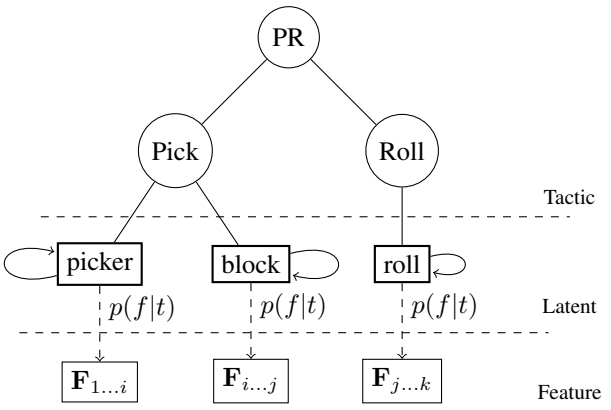

Figure 6. LCFG model for pick-and-roll tactic

Similar to Viterbi decoding in HMM, under LCFG, the most-likely interpretations will be constructed from the probabilistic parsing tree through the modified CYK algorithm. Beyond the one dimensional relationship in HMMs, the LCFG parsing tree encapsulates rich hierarchical information containing the relationships of concepts and subconcepts relationship as well as the interpretation likelihood.

Specifically in this work, the sports tactics can be effectively identified based on the pre-defined tactic features and grammar rules. The issues of ambiguity and the syntactic restrictions are simultaneously addressed, and the proposed LCFG is able to quantitatively seek the best interpretation among all possibilities.

6. Experimental Results

Our data set is collected from various sources. Most of them came from an entire commercial broadcasting basketball video (NBA - 2014 Miami Heat vs San Antonio Spurs) on YouTube. In order to evaluate our approach, we pre-separate the data into video chunks by adopt the method used in [14]. And the initial court and athlete key points are also manually provided.

6.1. Tracking Results

The actual sports videos vary significantly. Thus there is no single perfect detection and tracking method exists. The purpose of athlete tracking in this experiment is to support high level tactic recognition. So to keep the computation simple, we use color histogram and mean-shift tracking method to obtain the trajectories of athletes. In Fig.7, four identified athletes are acting pick-and-roll tactic in the game. Their trajectories are mapped into the absolute court space.

6.2. Tactics Recognition

Based on the athletes tracking results, we apply the proposed tactic LCFG method to parse the video sequence. The classical pick-and-roll tactics can be decomposed through the following list of context-free rules,

$$\begin{aligned} \text{S} &\rightarrow \text{NULL PR } [1.0] \\ \text{PR} &\rightarrow \text{NT1 Roll } [1.0] \\ \text{NT1} &\rightarrow \text{Pick Block } [1.0] \\ \text{NULL} &\rightarrow \text{Null NULL } [1.0] \mid \text{Null } [\varepsilon] \\ \text{Pick} &\rightarrow \text{pick Pick } [1.0] \mid \text{pick } [\varepsilon] \\ \text{Block} &\rightarrow \text{block Block } [1.0] \mid \text{block } [\varepsilon] \\ \text{Roll} &\rightarrow \text{defMiss Roll } [0.5] \mid \text{defMiss } [\varepsilon] \\ \text{Roll} &\rightarrow \text{defGood Roll } [0.5] \mid \text{defGood } [\varepsilon] \\ \text{pick} &\rightarrow \text{pick0 } [0.5] \mid \text{pick1 } [0.5] \end{aligned}$$

Except the non-terminal symbols on the left-hand side, each terminal symbol corespondents to a discriminative tactic feature. The feature *pick0* and *pick1* denote the actual picker selection in the sequence. We also define a *NULL* state to

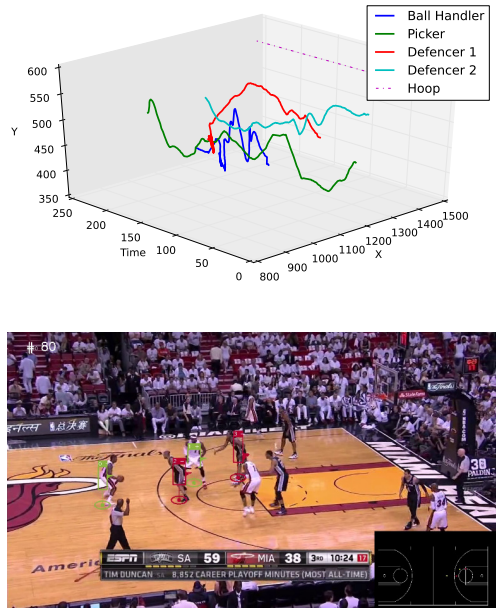

Figure 7. Player tracking.

model that both players are not acting as a picker. Practically, a cut-off parameter β is introduced in Eq.5 to discriminate *NULL* from other picker states.

$$\mathbf{P}(\text{NULL}) = \beta - \max(\mathbf{P}(\text{picker0}),\ \mathbf{P}(\text{picker1})) \quad (5)$$

Similar logic is applied on the defence score. The state *defMiss* is introduced in Eq.6 to indicate bad defence based on defence quality *defence* and parameter α.

$$\mathbf{P}(\text{defMiss}) = \alpha - \mathbf{P}(\text{defence}) \quad (6)$$

The block terminal is introduced to indicate the picker blocking the way of the defender. It is based on the average Euclidean distances between the athletes.

For selected athletes, the input of LCFG are prepared by stacking the absolute trajectories in Euclidean space, and the relative trajectories in hoop centered polar space together. The process of parsing the data in our pick-and-roll LCFG model is shown in Fig.8. Each parsing tree is a self-explaining concept and contains consecutive and hierarchical structured intermediate states. In Fig.9, the optimal separations along with key-frames, and interpretations differ with pickers, and defence variations are extracted through our modified CYK parsing methods.

In our experiments, we parsed 2324 frames among 17 video clips. Each clip have different frame length and the contents differ in the pick-and-roll tactic variations. The

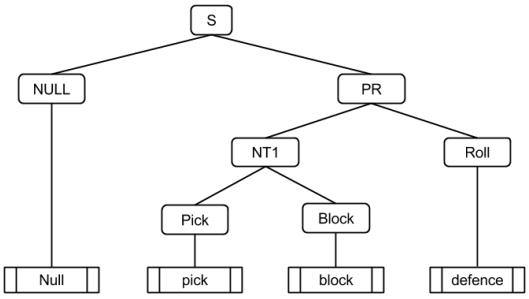

Figure 8. Pick-and-roll parsing tree concept.

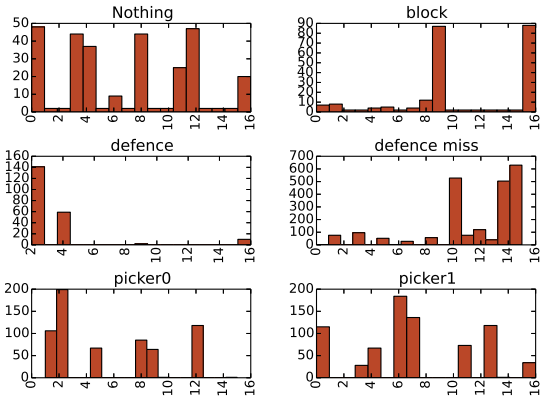

Figure 11. Parsing results across the data set [total #: 2324 frames]

high level concept distribution across our test set is shown in Fig.11.

7. Conclusion

In this paper, we propose a probabilistic framework to interpret high level concepts from temporal visual events. Specifically, we apply the proposed LCFG to commercial broadcast sports videos to represent common sports tactics. This framework is able to address two major challenges in semantic representation. Firstly, to fill the gap between lower level features and higher level concept. In contrast to conventional approaches that rely on shallow models like HMMs or deeply stacked neural networks, we introduced intermediate discriminative features that can bridge observations and domain knowledge. Secondly, to represent and extract long-term temporal relationships in a time series. In many applications, long term temporal behaviors can not be simply described in first-order statistics. They actually follow certain underlying logical order from domain knowledge. Beyond the classical Markov property in HMM and its extensions, we introduced context-free rules to represent these high level logical processes through syntactic tree structures. We evaluated our proposed frame-

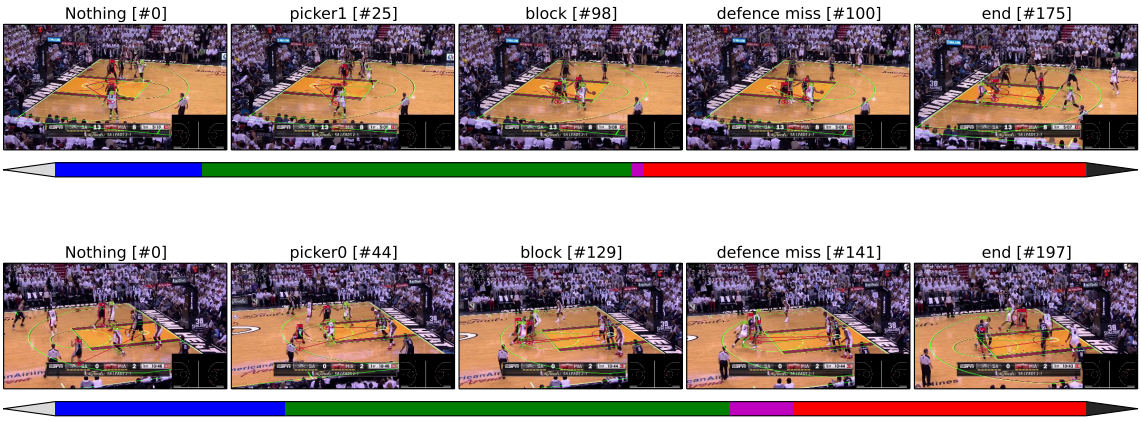

Figure 9. Parsing results with successful [miss defence] pick and roll tactic

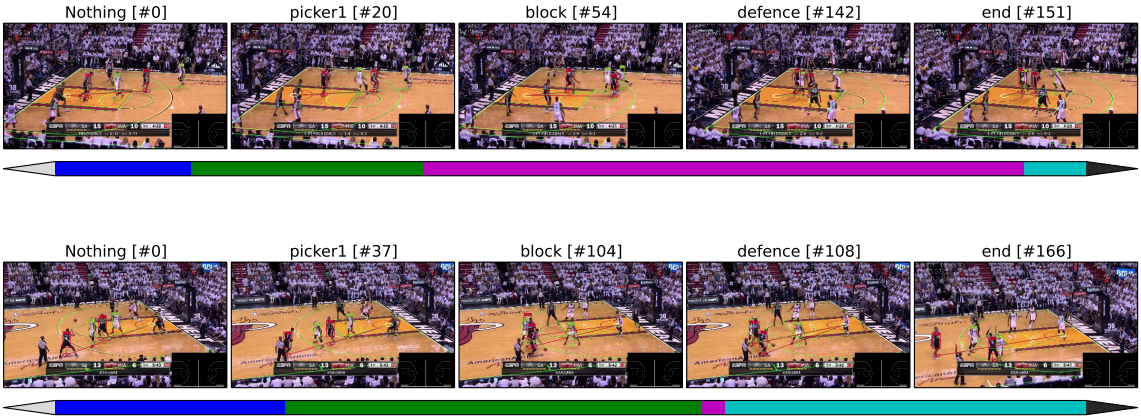

Figure 10. Parsing results with unsuccessful [defence] pick and roll tactic

work and methodology on real-world basketball video clips. The experiments results demonstrated the rich representation and interpretation power of LCFG through the probabilistic parsing trees.

References

[1] H. Ben Shitrit, M. Raca, F. Fleuret, and P. Fua. Tracking multiple players using a single camera. Technical report, Springer Verlag, 2013. 3

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402 Vol. 2, Oct 2005. 2

[3] J.-C. Chappelier, M. Rajman, et al. A generalized cyk algorithm for parsing stochastic cfg. *TAPD*, 98:133–137, 1998. 5

[4] H.-T. Chen, M.-C. Tien, Y.-W. Chen, W.-J. Tsai, and S.-Y. Lee. Physics-based ball tracking and 3d trajectory reconstruction with applications to shooting location estimation in basketball video. *Journal of Visual Communication and Image Representation*, 20(3):204–216, 2009. 3

[5] Q. Chen, N. D. Georganas, and E. Petriu. Hand gesture recognition using haar-like features and a stochastic context-free grammar. *Instrumentation and Measurement, IEEE Transactions on*, 57(8):1562–1571, Aug 2008. 2

[6] R. Dowell and S. Eddy. Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC Bioinformatics*, 5(1):71, 2004. 5

[7] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin. A unified framework for semantic shot classification in sports video. *Multimedia, IEEE Transactions on*, 7(6):1066–1083, 2005. 1

[8] P. Dupont, F. Denis, and Y. Esposito. Links between probabilistic automata and hidden markov models: Probability distributions, learning models and induction algorithms. *Pattern Recogn.*, 38(9):1349–1371, Sept. 2005. 4

[9] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3265–3272, June 2011. 2

[10] M.-C. Hu, M.-H. Chang, J.-L. Wu, and L. Chi. Robust camera calibration and player tracking in broadcast basketball video. *Multimedia, IEEE Transactions on*, 13(2):266–279, 2011. 1

[11] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872, Aug 2000. 2

[12] I. Laptev and P. Perez. Retrieving actions in movies. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007. 2

[13] W.-L. Lu, K. Okuma, and J. J. Little. Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *Image and Vision Computing*, 27(1):189–205, 2009. 3

[14] W.-L. Lu, J.-A. Ting, J. Little, and K. Murphy. Learning to track and identify players from broadcast sports videos. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1704–1716, July 2013. 1, 3, 6

[15] D. Moore and I. Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *Eighteenth National Conference on Artificial Intelligence*, pages 770–776, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence. 2

[16] K. Okuma, J. J. Little, and D. G. Lowe. Automatic rectification of long image sequences. 1, 2

[17] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In T. Pajdla and J. Matas, editors, *Computer Vision - ECCV 2004*, volume 3021 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin Heidelberg, 2004. 1, 3

[18] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proceedings of the 7th European Conference on Computer Vision-Part I*, ECCV '02, pages 661–675, London, UK, UK, 2002. Springer-Verlag. 1

[19] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 612–619, June 2014. 1, 2

[20] M. Ryoo and J. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1709–1718, 2006. 2

[21] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In M. Shah and R. Jain, editors, *Motion-Based Recognition*, volume 9 of *Computational Imaging and Vision*, pages 227–243. Springer Netherlands, 1997. 2

[22] H. K. Stensland, V. R. Gaddam, M. Tennøe, E. Helgedagsrud, M. Næss, H. K. Alstad, A. Mortensen, R. Langseth, S. Ljødal, O. Landsverk, C. Griwodz, P. Halvorsen, M. Stenhaug, and D. Johansen. Bagadus: An integrated real-time system for soccer analytics. *ACM Trans. Multimedia Comput. Commun. Appl.*, 10(1s):14:1–14:21, Jan. 2014. 1

[23] M. Tamir and G. Oz. Real-time objects tracking and motion capture in sports events, Aug. 14 2008. US Patent App. 11/909,080. 1

[24] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21. 10.1007/s11263-012-0564-1. 2

[25] X. Xu and H. Man. Activation analysis on fmri time series using stochastic context-free model. In *Wireless and Optical Communication Conference (WOCC), 2014 23rd*, pages 1–6, May 2014. 5

[26] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1385–1392. IEEE, 2011. 3

[27] Z. Zhang, T. Tan, and K. Huang. An extended grammar system for learning and recognizing complex visual events. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):240–255, Feb 2011. 2

[28] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, and H. Yao. Trajectory based event tactics analysis in broadcast sports video. In *Proceedings of the 15th international conference on Multimedia*, pages 58–67. ACM, 2007. 1

[29] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 431–440. ACM, 2006. 1