COMPUTER ANIMATION AND VIRTUAL WORLDS
Comp. Anim. Virtual Worlds 2012; **23**:145–154
Published online 17 May 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cav.1452

SPECIAL ISSUE PAPER

Video event representation and inference on And-Or graph

Kai Jiang, Xiaowu Chen*, Yu Zhang and Qinping Zhao*

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China

ABSTRACT

This paper presents an approach for video event inference from dozens of actions performed by multiple players. First, we constructed an And-Or graph to describe the different configurations of the event category such as shooting in soccer matches. We considered both temporal relations and role relations for the graph and encode them as vector parameters for each pair of graph nodes. Then, we developed an inference algorithm by using bottom-up and top-down processes. We found the proposals for each node during the bottom-up step by considering three terms of energies and refined the proposals during the top-down step by measuring the action-labeling similarity and the temporal misplacement penalty. The optimal proposal of the inferring event and its score are obtained as the result. In the experiments, we tested the inference performance of the approach for the shooting events on real soccer match videos. By our approach, we can infer different kinds of shooting events in one scenario and interpret them play-by-play in a flexible way. Copyright © 2012 John Wiley & Sons, Ltd.

KEYWORDS

video event representation; video event inference; And-Or graph

***Correspondence**

Xiaowu Chen and Qinping Zhao, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China.
E-mail: {chen, zhaoqp}@vrlab.buaa.edu.cn

1. INTRODUCTION

Analyzing and understanding videos of human activities is a challenge and a hot research topic in recent years. Many applications, such as video surveillance, video broadcasting, and content-based video indexing rely much on the tasks of recognizing and parsing events from these videos. To achieve these tasks, a number of key issues, including object tracking, human action recognition, and human activity classification, are well studied in computer vision research. As humans have a strong inclination to realize what the humans are doing in the video, to tell the storyline of the video is more attractive than several words. Especially in the area of video broadcasting application, for example, the FIFA World Cup broadcasting, beyond enjoying a spectacular bicycle kick, what human want to know from the soccer match videos are the processes of meaningful events, one category of which is the *shooting* events.

We present in this paper an approach for the shooting event inference from dozens of actions performed by multiple players. First, we constructed an And-Or graph to describe the different configurations of the event category. The Leaf-nodes of the And-Or graph represent the actions performed by different roles of players. The And-nodes and Or-nodes are successive on the levels of the graph, representing the hierarchical compositions of the event. The relations between the children of And-nodes only exist at the same level. We considered both temporal relations and role relations for the graph in this paper.

Second, we developed an inference algorithm that combined a bottom-up process for proposing configurations for the shooting events together with a top-down process for refining these proposals. We found the proposals for each node by considering three terms of energies: the data term, the structure term, and the relation term. We refined the proposals by measuring the action-labeling similarity and the temporal misplacement penalty. The algorithm returns the optimal proposal for the inferring event and obtains a score as well.

In the experiments, we tested the shooting events inference performance of the approach on real soccer match videos. By our approach, we can infer different kinds of shooting events (e.g., *head shooting*, *penalty kick*)

in one scenario. The play-by-play commentaries of the events were also obtained to denote what happens at the right moment.

There are two main contributions in this paper. (i) We represented the complex video events composed of multi-agent actions and test the inference algorithm on real soccer match videos. The various configurations of the events are represented by an And-Or graph. The bottom-up and top-down processes are used for inferring the events from the actions performed by multiple players. (ii) We considered both temporal relations and role relations in the And-Or graph. As dozens of temporal segments of the same action label appear in the videos, one kind of action label can barely identify the happening of the inferring event. We encoded all the six conditions of temporal relations and four of role relations for each pair of nodes, which makes it possible to represent the relationship between the actions with the same semantic label.

This paper is organized as follows. We first review related work. Then, we construct an And-Or graph to represent the shooting event category in soccer matches and explain the components of the graph. After that, we formulate the probability distribution of the parse graph for inference problem and develop bottom-up and top-down processes for event inference. We show the experimental results of action labeling and event inference and conclude the paper in the end.

2. RELATED WORK

There have been a significant amount of video event understanding researches in various application domains. We reviewed related work on group activity understanding and sports event analysis at first and then reference the And-Or graph for analyzing and representing the video events in soccer matches.

Group activity understanding. Reference [1] presents a stochastic methodology for the recognition of various types of high-level group activities. The method maintains a probabilistic representation of a group activity, describing how individual activities of its group members must be organized. It can recognize activities such as *a group of thieves stealing an object from another group* and *a group assaulting a person*. Reference [2] presents an approach to representing and recognizing composite video events, which are specified by a scenario, on the basis of primitive events and their temporal–logical relations, to constrain the arrangements of the primitive events in the composite event. However, the soccer events we discussed in this paper are composed of human actions at a distance, which means that the events can hardly be detected by some significant actions or objects.

Sports event analysis. Reference [3] presents an approach to learn a visually grounded storyline model of videos directly from weakly labeled data. The authors consider that the storyline of a video describes the causal relationships between actions. Reference [4] presents a soccer video highlighting approach to find the important events during the soccer match. They use a statistical event model and linear temporal relations between the component actions. Reference [5] presents a multiple objects detection and tracking approach to track players in broadcast sports video. Other works such as Reference [6] can hardly tell how the soccer events are going. In spite of watching at a distance, the human vision system have the capability to understand with no difficulty the ongoing events. This relies on both visual tracking and human experience on these events. For example, one player moving quickly and others gathering toward him most likely indicate that a shooting event is going to happen. As human visual system seems to rely on decomposition for understanding, the whole story of an event can be naturally divided into three periods: the starting, ongoing, and ending periods, which denote the temporal interval before, during, and after the event, respectively. Each period is composed of several human actions performed by various players, which can be represented by an And-Or graph.

And-Or graph representation. As a form of tree-like graph used in problem solving and problem decomposition, And-Or graph is used widely for image and video understanding. For image understanding, Reference [7] formulates an And-Or graph representation capable of describing the different configurations of deformable articulated objects such as horses. The methodology of the bottom-up and top-down processes is useful for us. For video understanding, Reference [8] presents an event parsing algorithm based on stochastic context sensitive grammar (SCSG) for understanding events, inferring the goal of agents, and predicting their plausible intended actions. The SCSG represents the hierarchical compositions of events and the temporal relations between the sub-events. We referenced the structure of the And-Or graph used in this paper, while the semantics, the attributes, and the parameters of the nodes are different.

3. VIDEO EVENT REPRESENTATION

The structure of an And-Or graph is represented by a five-tuple [7,8]

$$G = \left\{S, V^{\mathrm{N}}, V^{\mathrm{T}}, R, P\right\} \tag{1}$$

S is the root node for a meaningful event category. $V^{\mathrm{N}} = V^{\mathrm{and}} \cup V^{\mathrm{or}}$ is the set of non-terminal nodes, which contains And-nodes and Or-nodes. Each And-node, representing an event or sub-event, is decomposed into sub-events or atomic actions as its children nodes. All the children nodes of the event or sub-event must occur when it

DOI: 10.1002/cav

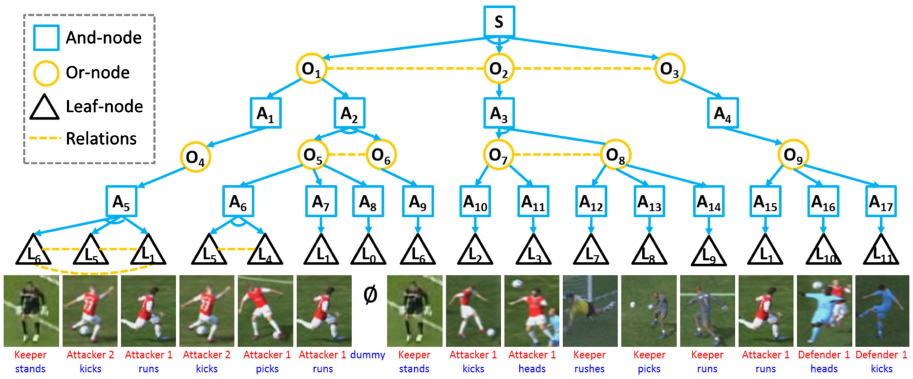


Figure 1. One of the And-Or graph representation for the *shooting* events. S is the root node, representing all kinds of shooting happening during the soccer match. $\{A_1, A_2, ...\}$ and $\{O_1, O_2, ...\}$ denote nodes of V^{and} and V^{or}. $\{L_1, L_2, ...\}$ denote terminal nodes of V^{T}. The And-nodes and the Or-nodes are alternate in rows. The Leaf-nodes represent the atomic actions performed by different players.

happens. Each Or-node is a switch to all the conditions of an event or sub-event, with a distribution to indicate the frequency of each condition. V^{T} is the set of terminal nodes, representing atomic human actions. R is the set of relations between the children of And-nodes. We considered both temporal relations and role relations in this paper. P is the probability model defined on the graph.

One of the And-Or graph representation for the shooting event is illustrated in Figure 1. The And-nodes, Or-nodes, and Terminal-nodes are depicted by rectangles, circles, and triangles, respectively. The structure of the graph is defined manually on the basis of the case study. The parameters and relations for all nodes are learned from the training data.

3.1. Non-Terminal Nodes

The non-terminal nodes V^{N} are composed of a set of And-nodes and a set of Or-nodes. Each node $v^{\text{a}} \in V^{\text{and}}$ represents a certain type of sub-event of shooting. Each $v^{\text{o}} \in V^{\text{or}}$ has a prior distribution $z_{v^{\text{o}}}$ to indicate the frequency of choosing one of its children. The root node $S \in V^{\text{and}}$ represents all types of the shooting events during the soccer match. It has three children, illustrated as $\{O_1, O_2, O_3\}$, representing the *starting*, *ongoing*, and *ending* of the shooting events, respectively.

3.2. Terminal Nodes

The terminal nodes V^{T}, also known as Leaf-nodes, represent the atomic human actions performed by players during the soccer match. Each action is performed by a certain player in a continuous time interval. For a certain time interval, several actions happen concurrently, while a list of actions happen sequentially for a certain player. As shown in Figure 1, considering the role of the player and the semantic label of the action, we defined 12 categories of actions for the shooting events, listed in Table I. For each $v^{\text{t}} \in V^{\text{T}}$, we learned its semantic label from training data set.

The human actions $\{a_i\}$ in videos can be represented as a state attribute $\delta(a_i) = (label_{a_i}, \lambda_{a_i}, c_{a_i})$. $label_{a_i}$ and λ_{a_i} denote the set of semantic category labels of the action and the corresponding similarity with the action category for classification. There are six categories of important human actions according to our graph definition, labeled as *run*, *kick*, *head*, *pick*, *stand*, and *rush*. c_{a_i} represents the roles of the action players, marked as *team A*, *team B*, and *the keeper*.

We defined a temporal duration $\tau_v = (t_v, I_v)$ for all nodes in the And-Or graph, where t_v denotes the start anchor point of v, and I_v denotes the length of the temporal interval. The prior temporal duration $h(\tau_v) = (h(t_v), h(I_v))$ for each node v is learned as well.

3.3. Relations

The relations $R = (\xi, \chi)$ only exist between a pair of nodes of the same level, where ξ and χ represent the temporal relations and the role relations, respectively.

Table I. The atomic actions with attributes.

Node	Semantic label	Node	Semantic label
L_0	ϕ	L_6	K stands
L_1	A1 runs	L_7	K rushes
L_2	A1 kicks	L_8	K picks
L_3	A1 heads	L_9	K runs
L_4	A1 picks	L_{10}	D1 heads
L_5	A2 kicks	L_{11}	D1 kicks

A1, A2, K, and *D1* denote different roles of players. For example, *A1* and *A2* denote the *teammates*, which mean different members of the same team. *K* is the *keeper*, and *D1* is an *opponent*.

3.3.1. Temporal Relation.

There are six binary temporal relations, known as *before*, *overlaps*, *meets*, *during*, *starts*, and *finishes*, according to Allen's temporal logic theory. In this paper, we defined temporal relation $\xi(v_i, v_j)$ of two nodes v_i and v_j as a vector, representing the energies of $\xi(v_i, v_j)$ being the six conditions, respectively. The prior distribution $z_{\xi(v_i,v_j)}$ is also learned from the training data.

For temporal relation type *tr*, we computed ξ_{tr} by measuring how well the duration of v_i and v_j fit the pattern of *tr*. For example, if $t_{v_i} + I_{v_i} - t_{v_j} = 0$, it means that duration τ_{v_i} *meets* duration τ_{v_j}, and the probability ξ_{meets} is enormous. The temporal relations between nodes is composed of six conditions, each one of which has an energy for fitness measuring. That is because that for multiple agents event, the temporal relation of two actions performed by different agents is not stabilized as that by the same agent. We formulated the temporal relation ξ_{tr} as follows, where C_i is a large constant, th_i is the threshold, and $g^i(\cdot,\cdot)$ is an i-D Gaussian distribution.

Before represents that v_i finishes before v_j happens, which means $n_1\left(\tau_{v_i}, \tau_{v_j}\right) = t_{v_j} - I_{v_i} - t_{v_i} > 0$.

$$\xi_{\text{before}}(v_i, v_j) = g^1(dif_1, t_{v_j}) \tag{2}$$

where

$$dif_1 = \begin{cases} n_1\left(\tau_{v_i}, \tau_{v_j}\right), \text{if } n_1 > 0 \\ |n_1\left(\tau_{v_i}, \tau_{v_j}\right)| + C_1, \text{otherwise} \end{cases} \tag{3}$$

and g^1 is defined on t_v.

Overlaps represents that v_j happens before v_i finishes, which means $n_2\left(\tau_{v_i}, \tau_{v_j}\right) = t_{v_j} - t_{v_i} > 0$.

$$\xi_{\text{overlaps}}(v_i, v_j) = g^2\left((dif_2, dif_3), \tau_{v_j}\right) \tag{4}$$

where

$$dif_2 = \begin{cases} n_2\left(\tau_{v_i}, \tau_{v_j}\right), \text{if } n_2 > 0 \\ |n_2\left(\tau_{v_i}, \tau_{v_j}\right)| + C_2, \text{otherwise} \end{cases} \tag{5}$$

$$dif_3 = \begin{cases} -n_1\left(\tau_{v_i}, \tau_{v_j}\right), \text{if } n_1 < 0 \\ |n_1\left(\tau_{v_i}, \tau_{v_j}\right)| + C_3, \text{otherwise} \end{cases} \tag{6}$$

and g^2 is defined on τ_v.

Meets represents that v_i finishes the same time as v_j happens. **During** represents that v_j happens before v_i and finishes after it, which means $n_2\left(\tau_{v_i}, \tau_{v_j}\right) < 0$ and $n_3\left(\tau_{v_i}, \tau_{v_j}\right) = t_{v_j} + I_{v_j} - t_{v_i} - I_{v_i} > 0$. **Starts** represents that v_i and v_j happens at the same time, which means $|n_2\left(\tau_{v_i}, \tau_{v_j}\right)| < th_1$. **Finishes** represents that v_i and v_j finishes at the same time, which means $|n_3\left(\tau_{v_i}, \tau_{v_j}\right)| < th_2$. The definitions of the last four temporal relations are similar to those of the first two relations.

3.3.2. Role Relation.

The role relation represents the restrains of the players of a pair of actions. For a pair of non-terminal nodes v_i and v_j, the role relation $\chi(v_i, v_j)$ is calculated by multiplying the role relations between all the Leaf-nodes in the topology of the sub-graph rooted by v_i and v_j.

The role relations between Leaf-nodes include both unary and binary relations. The unary relations $\text{ur}(A)$ represent that player A is a *ball-player* or *keeper*, whereas the binary relations $\text{br}(A, B)$ explain that players A and B are *teammates* or *opponents*. We enumerated the role relations between Leaf-nodes as *the-same-person*, *teammates*, *opponents*, and *player-and-keeper*. The role relation $\chi\left(v_i^{\text{t}}, v_j^{\text{t}}\right)$ of two nodes v_i^{t} and v_j^{t} is a four-element vector, representing the energies of $\chi\left(v_i^{\text{t}}, v_j^{\text{t}}\right)$ being the four conditions, respectively. The prior distribution $z_{\chi\left(v_i^{\text{t}}, v_j^{\text{t}}\right)}$ is learned from the training data. For role relation type *rr*, we calculated χ_{rr} as a binarization by the attribute c_{a_i} in δ_{a_i} and c_{a_j} in δ_{a_j} for a_i and a_j.

We have constructed the And-Or graph, and the components of the graph are described earlier. Then, we introduce the algorithm of video event inference on the graph.

4. VIDEO EVENT INFERENCE

We implemented the shooting event inference by finding a parse graph pg^S, which is an instance of the And-Or graph, obtained by switching each Or-node to one of its conditions and specifying the attributes of all the And-nodes and Leaf-nodes.

4.1. Probability Distribution of Parse Graph

We considered pg^S as a graph rooted by node S. For adopting the bottom-up and top-down algorithms for event inference, we defined the energy of sub-graph pg rooted by a certain node as follows.

$$\begin{aligned} \varepsilon(pg, d) = & \sum_{v^{\text{t}} \in V^{\text{T}}(pg)} E(v^{\text{t}}, d) + \sum_{v^{\text{n}} \in V^{\text{N}}(pg)} E(v^{\text{n}}) \\ & + \sum_{v^{\text{a}} \in V^{\text{and}}(pg)} \sum_{x,y \in V_{v^{\text{a}}}^{\text{child}}} E(x, y) \end{aligned} \tag{7}$$

where d is the input data, and $V_{v^{\text{a}}}^{\text{child}}$ is the set of children of v^{a}.

So, the probability for pg^S under the input data is of the following form

$$p(pg^S, d) = \frac{1}{Z} \exp\left\{-\varepsilon(pg^S, d)\right\} \tag{8}$$

where Z is the partition function. The formulations of $\varepsilon(pg)$ are listed in the succeeding texts.

DOI: 10.1002/cav

The first term of $\varepsilon(pg)$ is the data term that measures the accordance of a certain action a_i fitting the terminal node v^{t} in $V^{\mathrm{T}}(pg)$. It is calculated by

$$E(v^{\mathrm{t}}, a_i) = \max_{t_{a_i}} \xi_{\mathrm{same}}(\tau_{v^{\mathrm{t}}}, \tau_{a_i}) \tag{9}$$

where $\tau_{a_i} = \left(t_{a_i}, h(I_{v^{\mathrm{t}}})\right)$ is a segment of a_i with the same length as the prior temporal duration of v^{t}. xi_{same} means that the two temporal intervals begin at the same time and end at the same time, too. The calculation of ξ_{same} is similar to the ones formulated in the temporal relation.

The second term is the structure term for holding the structure of the graph with a link from a parent node to a child node. For an And-node, it means the similarity between the temporal duration of the And-node and the total duration of its children. For an Or-node, it means the similarity between the temporal duration of the Or-node and its child in this parse graph with the probability of switching to this condition. The second term is calculated by

$$E(v^{\mathrm{n}}) = \begin{cases} \xi_{\mathrm{same}}\left(\tau_{v^{\mathrm{a}}}, \tau_{V_{v^{\mathrm{a}}}^{\mathrm{child}}}\right) & \text{if } v^{\mathrm{a}} \\ z_{v^{\mathrm{o}}} \xi_{\mathrm{same}}\left(\tau_{v^{\mathrm{o}}}, \tau_{c_{v^{\mathrm{o}}}(pg)}\right) & \text{if } v^{\mathrm{o}} \end{cases} \tag{10}$$

where $\tau_{V_{v^{\mathrm{a}}}^{\mathrm{child}}}$ is the temporal interval of the children set of v^{a}, which is calculated as the first start point t and the longest duration I (interval between the first start point and the last end point in $V_{v^{\mathrm{a}}}^{\mathrm{child}}$). $c_{v^{\mathrm{o}}}(pg)$ is the child of v^{o} in pg.

The third term of $\varepsilon(pg)$ is the relation term that measures the temporal relations and the role relations between the children of And-nodes. It models all the six conditions of temporal relations and the four conditions of role relations, calculated by

$$E(x, y) = z_{\xi} \cdot \xi(x, y) + z_{\chi} \cdot \chi(x, y) \tag{11}$$

where $\cdot$ means the dot multiplication of two vectors.

Now, we have formulated a complete probability distribution for the parse graph pg. The energy of each node v in its sub-graph pg^v can be calculated by Equation (7). This mechanism makes it possible to adopt an iterative processing algorithm for event inference from a set of human actions.

4.2. Bottom-Up and Top-Down Processing

Given a set of human actions $A^{\Lambda} = \{a_i\}$ as the input data in video sequence Λ, the task of video event inference is to find the parse graph pg^{S*} that maximizes the posterior probability defined by

$$pg^{S*} = arg\max\, p\left(pg^S, A^{\Lambda}\right) \tag{12}$$

where $p(pg^S, A^{\Lambda})$ is defined in Equation (8). A parse graph is a proposal for A^{Λ}.

The pseudo code of the algorithm of bottom-up and top-down process for video event inference is shown in Algorithm 1. In the bottom-up step, we collected the proposals for each node of each level from the bottom to the top. Because of the increase in the proposals of an And-node, we pruned the proposals by the energy defined in Equation (7). In the top-down step, for each valid proposal of the root node S, we measured the fitness of the proposal $df(pp^S)$ by the input data A^{Λ}.

$$df\left(pp^S\right) = -\log\left(\lambda_{a_v^{pp}}\right) + \psi\left(h(\tau_v), \tau_{a_v^{pp}}\right) \tag{13}$$

where ψ is the temporal misplacement penalty function for the temporal segment a_v^{pp}, defined similarly as in [9].

5. EXPERIMENTS

We implemented our algorithm and tested it on the data set chosen from HD videos of real soccer matches. Specifically, we manually chose some clips from 2010 FIFA World Cup DVDs. These clips are taken in 30 frames per second in the resolution of 1080×720, varying from short to long.

In order to obtain a better result for action labeling, we preprocessed the videos in the following steps. First, we used a modified version of particle filter tracker [10] for human tracking. Second, we classified the player-centric figures by measuring color histogram to obtain the role of the player. Last, the player-centric video clips of human actions are passed into the action-labeling step. This process took about 10 min for a 200-frame video.

5.1. Action Labeling

In the action-labeling step, we followed the experimental steps proposed in [11]. We first constructed an action data set consisting about 108 tracked sequences, approximately 9600 frames. In order to collect enough actions of *kick*, *head*, or *pick* acted in different views and styles, we added in some sequences cut from FIFA football game videos. The sequences in the data set are tracked through the aforementioned steps manually, supplemented by flipping some of the sequences.

We specified 12 action labels for classification (Figure 2) in total and eight labels in use: *run*, *walk*, *stand*, *slide*, *pick*, *kick*, *rush*, and *head*. Optic flows are half-rectified and blurred as features for frame-to-frame matching. A blurry kernel is convolved by the frame-to-frame similarity matrix to obtain final motion-to-motion similarity matrix. In order to classify instantaneous action such as *kick* and *head*, we computed the motion descriptor with seven frames of temporal extent.

Given a video clip obtained from the preprocessing step, we first calculated optic flow on each frame and labeled the frame to *stand* if the variation of the optic flow is below a threshold. Then, we followed the algorithm proposed in

Algorithm 1 Video event inference.

Require:

The And-Or Graph, G;

The set of human actions, A^{Λ};

Ensure:

1: For each node v^t at level 0 (the level of terminal nodes):

2: Collect the proposal union: $PP_{v^t} = \{a_k\}$, where $a_k \in A^{\Lambda}$, k is the number of the proposal and $label_{a_k}$ is the same as the semantic label of v^t;

3: For each level $l = 1$ to N:

4: For each node v at level l:

5: If v is an Or-node, collect the proposal union: $PP_v = \{pp_k\}$, where pp_k is the optimal proposal for each child of v;

6: If v is an And-node, do 7-9:

7: Collect the proposal union: $PP_v = \{pp_k\}$, where pp_k is the combination of proposals by choosing one from each children of v at once;

8: For each proposal pp_k, calculate $\varepsilon(pg^v, pp_k)$ as its score, and prune the proposals whose score below a threshold;

9: Make the proposal of the maximal energy as the optimal proposal for v, the rest as the candidates;

10: For each proposal pp^S of the root node S, calculate $df(pp^S)$;

11: **return** pp^S as the proposal of the maximal $df(pp^S)$ and its score;

Figure 2. The action categories classified by our algorithm.

[11] to give each of the rest frames a label by action classification; the average precision of classification is shown in Figure 3. So, each frame has its action label right now. The labeled frames were clustered into action sequences by Markov chains, each sequence only having one action label. As the algorithm was tested in a much better condition (high-resolution videos with little camera variation and a larger data set), we obtained a better classification result. Thus, a set of action sequences were obtained, with the annotations of the role, the action label, and corresponding temporal extent. These annotations were directly passed into the event inference step.

DOI: 10.1002/cav

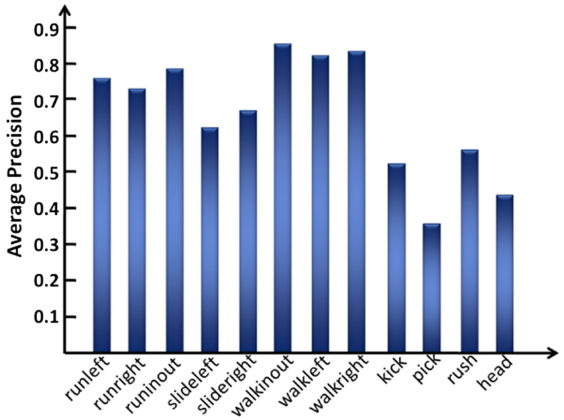


Figure 3. The precision for action categories classification, which gives each frame an action label. Walk-related actions have the highest classification precisions.

5.2. Event Inference

After the action-labeling process, our event inference system behaves automatically with the annotations of action sequences as input data. We collected annotations of about 30 videos processed using the aforementioned step, including goal or shooting events of different kinds and in various temporal alignments. We split the data into training set and testing set and constructed the And-Or graph manually using the annotations in a training set (future work will include learning graphs from the training set).

The attributes of every node were determined by statistical data and refined from the training set.

Figure 4 shows the result of an inference example of *general kick*. Different from other works related to shooting event recognition on the basis of camera shot analysis or ball detection, our system only explores inner relationships between sub-actions of humans. This example clip consists of 215 frames, involving more than 10 actors behaving simultaneously. After action labeling, the annotations of 57 actions were put into the graph. Among these actions, about 10 of them were incorrectly classified. The parsing process took less than 1 s in a computer with a 3.2-GHz CPU. At the top level of the graph, it collects about 540 valid proposals. Each proposal is within the variation allowed by the graph on all aspects and denotes one interpretation of the event. For example, the starting process contains three actions: one attacker kicks, another attacker runs, and the keeper stands. However, it is also reasonable according to the graph that only the keeper stands, no matter what the other two attackers are doing. Although the vital information of the ball is absent, we successfully parsed the video by inferring the event due to tight constraints on either temporal aspects or the relationships between sub-actions themselves. As shown in Figure 5, we can infer different kinds of shooting events (e.g., *heading*, *free kick*, *penalty kick*) by our approach. The play-by-play commentaries of the events were also obtained to denote what happens at the right moment.

Our approach encourages proposals with highly rigid as well as more detailed configuration (which means a more complex topology of the proposal). Thus, the system can

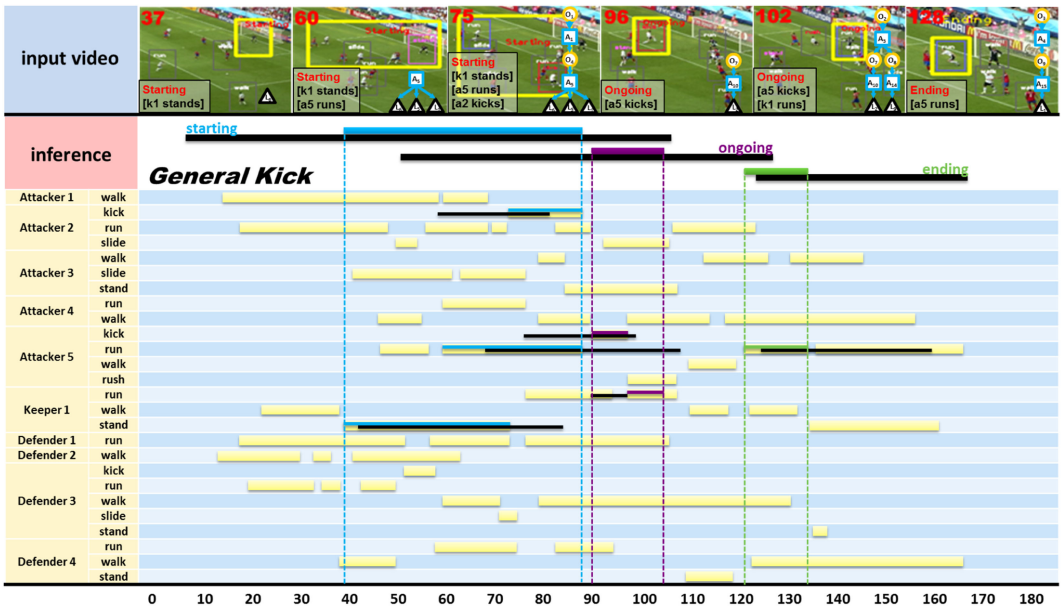


Figure 4. An inference example of *general kick* event. The blocks in the table denote temporal extents of tracks, sorted by roles and actions. Black lines are temporal extents expected by the graph. The blue, purple, and green lines are the temporal extents selected by the graph during the starting, ongoing, and ending periods, respectively. Images extracted from frames on the top show the significant change of the graph (e.g., a new branch is successfully parsed).

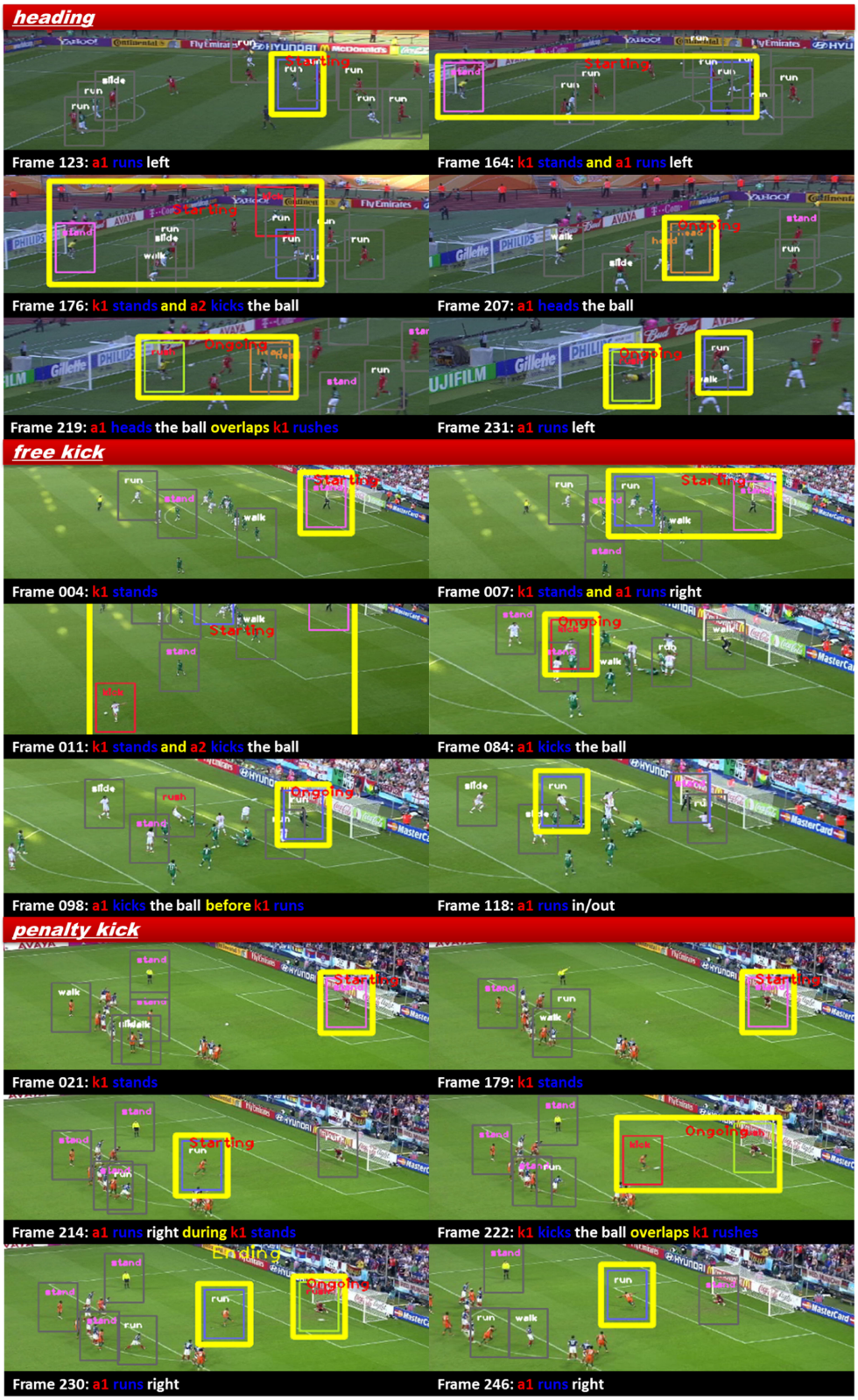


Figure 5. More inference results. We can infer different kinds of *shooting* events (e.g., *heading*, *free kick*, *penalty kick*) by our approach. The play-by-play commentaries of the events were also obtained to denote what happens at the moment.

Comp. Anim. Virtual Worlds 2012; **23**:145–154 © 2012 John Wiley & Sons, Ltd.
DOI: 10.1002/cav

reserve as much as possible the information contained by videos, which reflects the flexibility of our approach for interpretation. Other proposals that have poor temporal orders, such as Attacker 5 kicking and Attacker 2 running in starting period, are discarded because of a relatively low matching score. Note that although several players perform kicking or heading (Attacker 2 and 5, Defender 3), only Attacker 5's action is considered valid considering its better satisfaction of temporal and role constraints defined by the graph.

6. CONCLUSION

We present in this paper an approach for video event inference on an And-Or graph from dozens of actions. The actions occur repeatedly and simultaneously, performed by various players. The complex video events are composed of these actions, without starting or ending signals.

The And-Or graph represents the different configurations of the event category. The Leaf-nodes of the And-Or graph represent human actions performed by different roles of players. The relations between a pair of nodes include six temporal relations and four role relations. The event inference algorithm contains a bottom-up process for proposing possible configurations and a top-down process for refining the results.

We test our approach on real soccer match videos. After an initial action-labeling step, the action annotations of the test video clip are collected for event inference. The bottom-up and top-down processes pick up all possible proposals and obtain the best one by the matching score. By our approach, we can infer different kinds of the shooting events (*head shooting*, *penalty kick*, etc.) in one scenario in the absence of the ball information and interpret them play-by-play in a flexible way.

The limitation of this work is that the structure of the manually defined And-Or graph is barely suitable for other event categories. In the future work, we will focus on learning the structure of the And-Or graph from training data, which makes our approach flexible for other event categories.

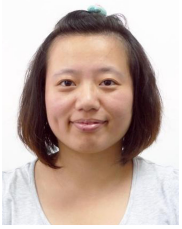
ACKNOWLEDGEMENTS

This work was partially supported by NSFC (60933006), 863 Program (2012AA011504 and 2012AA02A606), R&D Program (2012BAH07B01), Doctoral Program (20091102110019), and BUAA (YWF-12-LKGY-001).

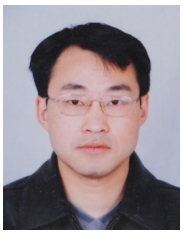
REFERENCES

1. Ryoo MS, Aggarwal JK. Stochastic representation and recognition of high-level group activities. *International Journal of Computer Vision* 2011; **93**(2): 183–200.
2. Kwak S, Han B, Han JH. Scenario-based video event recognition by constraint flow, In *CVPR*, Colorado Springs, USA, 2011; 3345–3352.
3. Gupta A, Srinivasan P, Shi J, Davis L-S. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos, In *CVPR*, Miami, USA, 2009; 2012–2019.
4. Jiang K, Chen XW, Zhao QP. Automatic compositing soccer video highlights with core-around event model, In *CAD/Graphics*, Jinan, China, 2011; 183–190.
5. Zhu G, Xu C, Huang Q, Gao W. Automatic multiplayer detection and tracking in broadcast sports video using support vector machine and particle filter, In *ICME*, Toronto, Canada, 2006; 1629–1632.
6. Qian X, Liu G, Wang H, Li Z, Wang Z. Soccer video event detection by fusing middle level visual semantics of an event clip, In *PCM (2)*, Shanghai, China, 2010; 439–451.
7. Chen Y, Zhu L, Lin C, Yuille AL, Zhang HJ. Rapid inference on a novel and/or graph for object detection, segmentation and parsing, In *NIPS*, Vancouver, B.C., Canada, 2007; 289–296.
8. Pei M, Jia Y, Zhu S-C. Parsing video events with goal inference and intent prediction, In *ICCV*, Barcelona, Spain, 2011; 487–494.
9. Niebles JC, Chen C-W, Fei-Fei L. Modeling temporal structure of decomposable motion segments for activity classification, In *ECCV*, Heraklion, Greece, 2010; 392–405.
10. Pérez P, Hue C, Vermaak J, Gangnet M. Color-based probabilistic tracking, In *ECCV*, Copenhagen, Denmark, 2002; 661–675.
11. Efros A, Berg A, Mori G, Malik J. Recognizing action at a distance, In *ICCV*, Nice, France, 2003; 726–733.

AUTHORS' BIOGRAPHIES



Kai Jiang is presently a PhD candidate in the School of Computer Science and Engineering and works in the State Key Laboratory of Virtual Reality Technology & Systems at Beihang University. Her research interests are computer vision and video event understanding.



Xiaowu Chen received his PhD degree at Beihang University in 2001. He is a professor in the School of Computer Science & Engineering and State Key Laboratory of Virtual Reality Technology & Systems, Beihang University. His research interests include computer vision, computer graphics, virtual reality, and augmented reality.

Yu Zhang is presently an undergraduate student in the School of Computer Science and Engineering at Beihang University. His research interests are computer vision and video event understanding.

Qinping Zhao is a professor in the School of Computer Science and Engineering, Beihang University. He is the founder and director of the State Key Laboratory of Virtual Reality Technology & Systems. His research interests include virtual reality and artificial intelligence.

DOI: 10.1002/cav