Monte Carlo Tree Search for Scheduling Activity Recognition

Mohamed R. Amer, Sinisa Todorovic, Alan Fern
Oregon State University, Corvallis, OR, USA

Song-Chun Zhu
University of California, Los Angeles, CA, USA

Motivation

Querying surveillance videos requires running many detectors of:

- group activities,
- individual actions,
- objects.

Running all the detectors is inefficient as they may provide little to no information for answering the query.

Problem Statement

Given: a large video with high resolution and multiple activities happening at the same time, and a query.
Goal: perform cost-sensitive parsing to answer the query.

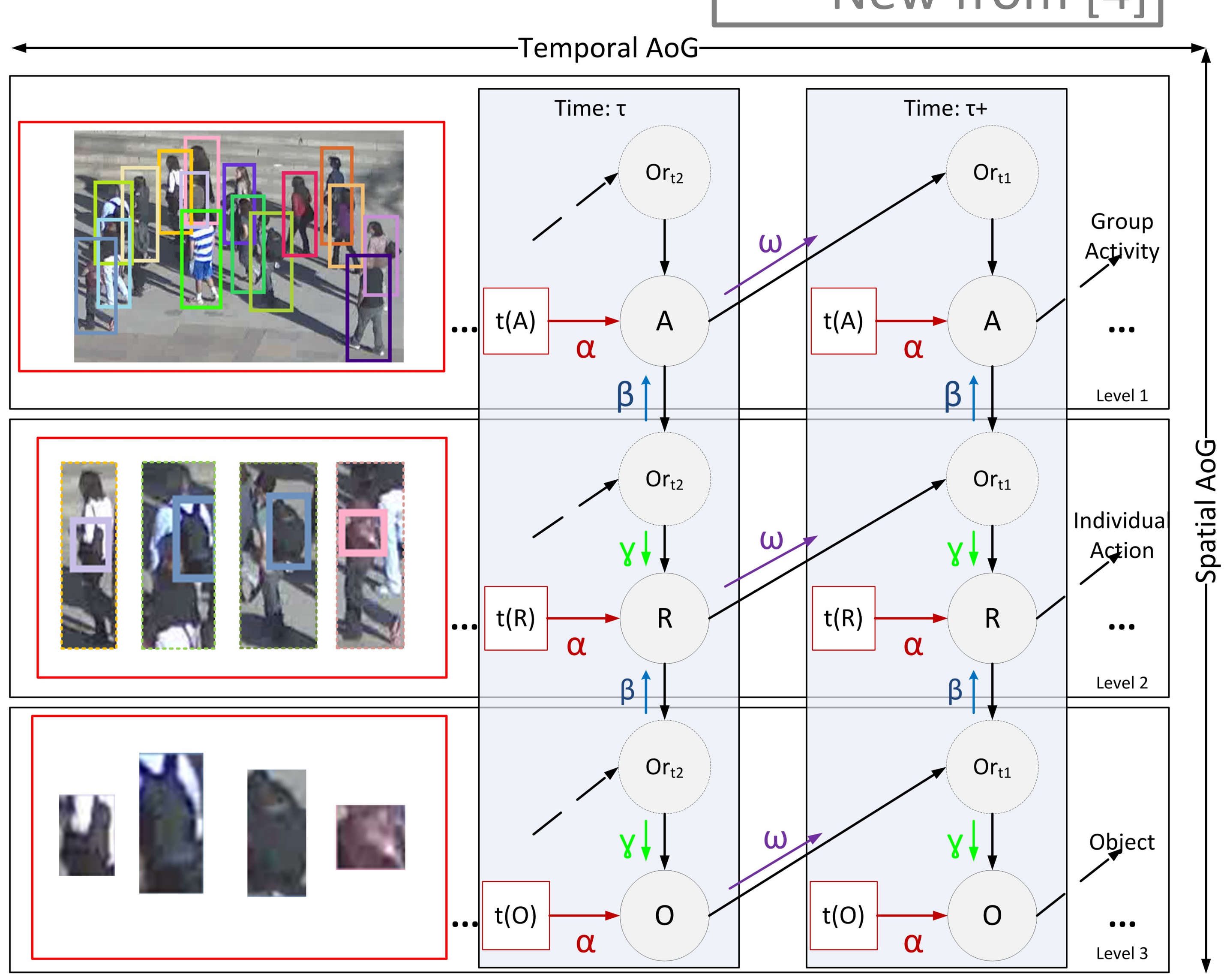
Approach

Parsing spatiotemporal And-Or Graphs can be defined in terms of ($\alpha, \beta, \gamma, \omega$):

- α: detecting objects, individual actions, group activities directly from video features.
- β: predicting an activity from its detected parts by bottom-up composition.
- γ: top-down prediction of an activity using its context.
- ω: predicting an activity at a given time based on tracking across the video.

$$\log p(\boldsymbol{pg}_l^\tau) \propto \underbrace{w_\alpha^T \phi(x_l^\tau, y, h)}_{\alpha \text{ detector of activity}} + \underbrace{w_\gamma^T \phi(x_{l-}^\tau, y, h)}_{\gamma \text{ context of activity}} + \underbrace{\sum_{i,j} w_\beta^T \phi(x_{il+}^\tau, x_{jl+}^\tau, y, h)}_{\beta \text{ relations between parts of the activities}} + \boxed{\underbrace{w_\alpha^T \phi(x_l^\tau, x_l^{\tau-}, y, h)}_{\omega \text{ tracking of activity}}}$$

New from [4]



Activity Parsing as a Search Problem

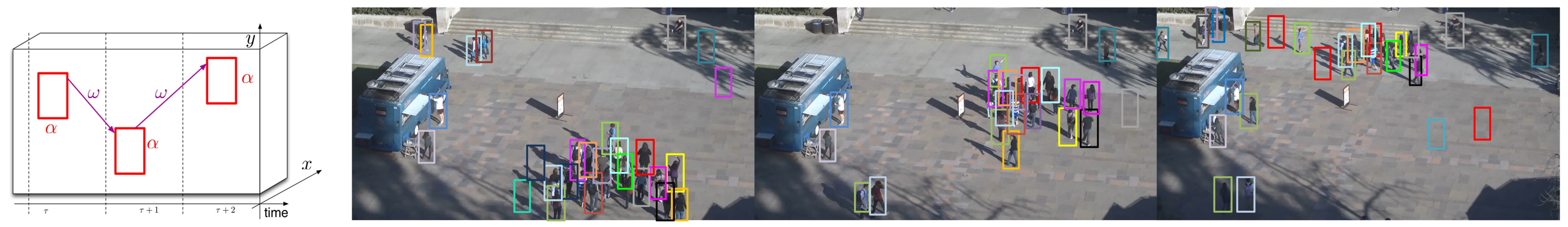
Action: $a_t \in \{process, detector, spatiotemporal\ volume\}$,
State: $S_{0:t} = (s_0 | a_0, a_1, \ldots, a_{t-1})$, New from [4]
Policy: $\Pi = \{s_0, a_{0:t} | t = 1,2,\ldots,B\}$, where B is the inference budget,
Reward: $r(\Pi) = \mathrm{I}(\log p(\boldsymbol{pg}(\Pi)) > \theta)$, where θ is an input threshold.

The cost-sensitive inference for a policy, $\hat{\Pi}$, is conducted by taking the set of actions, $a_{0:B}$, that provides the maximum utility, Q, for states, $S_{0:B}$:

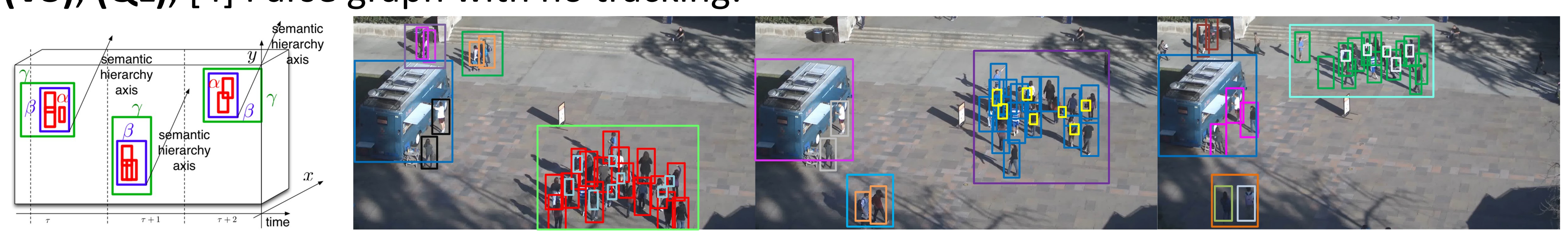
$$\hat{\Pi} = \mathrm{argmax}_a Q(S_{0:t}, a_t), \forall\, t = \{1,2,\ldots,B\},$$

$$if\ r(\hat{\Pi}) = \begin{cases} 1 & accept, \\ 0 & reject. \end{cases}$$

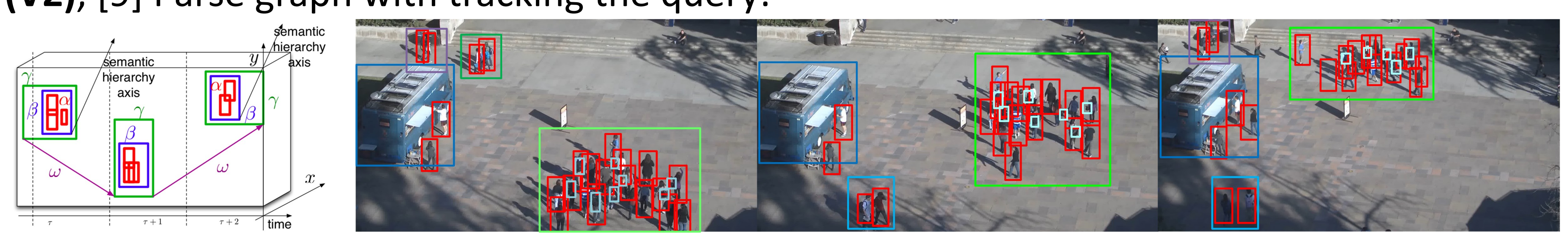
[6] Tracking detections:



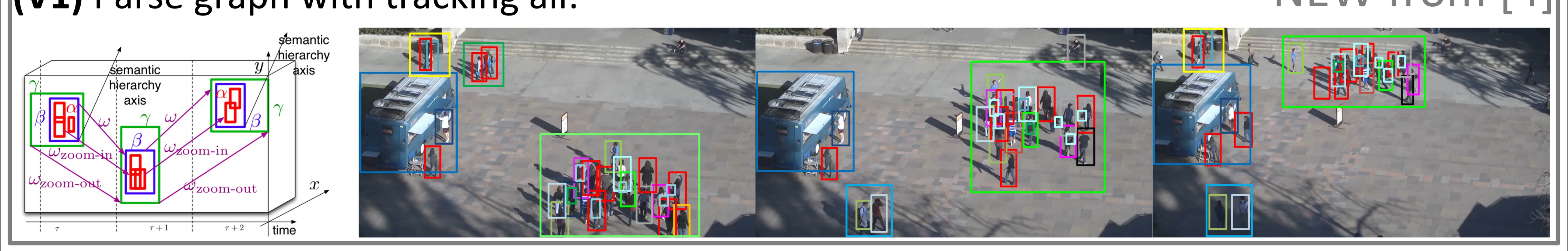
(V3), (QL), [4] Parse graph with no tracking:



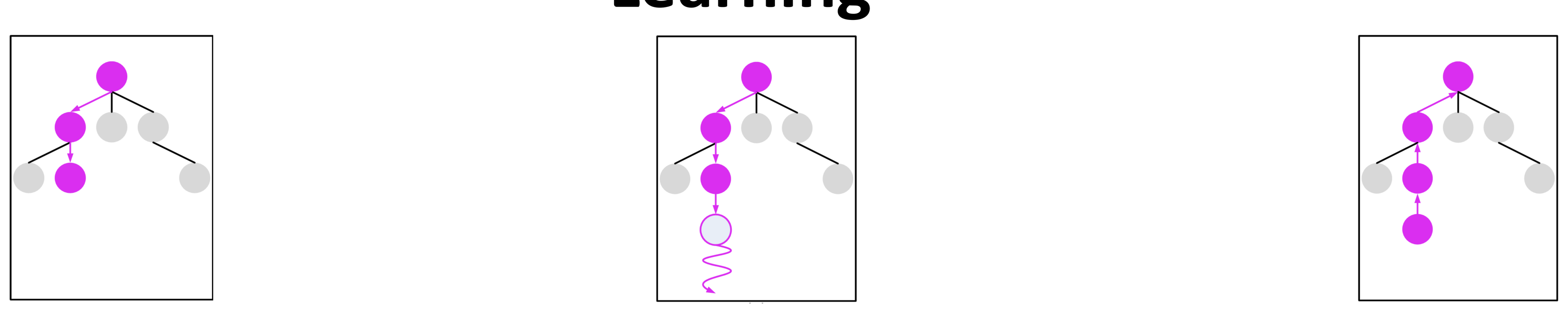
(V2), [9] Parse graph with tracking the query:



(V1) Parse graph with tracking all: NEW from [4]



Learning



Selection Simulation Backpropagation

Selection: select state $S_{0:t}$.
Simulation: update the expected utility for the simulations, given $r(\Pi_{sim})$.
Backpropagation: propagate the expected utility back to the root node by updating all the nodes on the path using:

$$Q'(S_{0:t}, a_t) \leftarrow Q'(S_{0:t}, a_t) + C(r(\Pi_{sim}))$$
$$\forall\, S_{0:t} \in \Pi_{\mathrm{sim}}, t = \{1,2,\ldots,B\}$$

Results

Classification accuracy on UCLA Courtyard dataset

Variant	Line	Tour	Disc.	Sit	Walk	Wait	Avg	Time
V1(5)	75.1	77.2	76.2	81.4	80.1	73.2	77.2	15
V2(5)	73.2	76.1	74.9	78.3	76.1	68.3	74.5	15
V3(5)	68.9	71.2	72.8	73.2	75.6	61.3	70.5	15
QL(5)	64.1	65.4	68.3	66.5	69.8	63.1	66.2	25

Limited budget

Variant	Line	Tour	Disc.	Sit	Walk	Wait	Avg	Time
V1(∞)	80.4	83.5	81.5	87.2	88.6	80.1	83.7	170
V2(∞)	77.4	82.2	77.2	84.2	79.3	72.9	78.8	170
V3(∞)	74.8	73.5	77.1	75.8	80.1	71.0	75.4	170
QL(∞)	68.0	70.2	75.1	71.4	78.6	72.6	72.7	230

Infinite budget

Classification accuracy on New Collective Activity dataset

Class	**V3**(∞)	**QL**(∞)	[7]	**V1**(∞)	**V2**(∞)	[6]
Gathering	48.1	44.2	50.0	48.9	42.8	43.5
Talking	81.3	76.9	72.2	86.5	82.4	82.2
Dismissal	55.3	50.1	49.2	84.1	81.2	77.0
Walking	89.1	84.3	83.2	92.5	89.9	87.4
Chasing	95.9	91.2	95.2	96.5	95.3	91.9
Queuing	96.7	92.2	95.9	97.2	96.1	93.4
Avg	77.7	74.8	77.4	84.2	80.1	79.2
Time	130s	150s	N/A	180s	170s	N/A

Classification accuracy on Collective Activity dataset

Class	**V3**(∞)	[4]	[2]	[12]	[7]	**V1**(∞)	**V2**(∞)	[6]	[9]
Walk	78.1	74.7	72.2	80	57.9	83.4	79.3	65.1	61.5
Cross	79.4	77.2	69.9	68	55.4	81.1	80.0	61.3	67.2
Queue	95.3	95.4	96.8	76	63.3	97.5	96.3	95.4	81.1
Wait	81.5	78.3	74.1	69	64.6	83.9	82.4	82.9	56.8
Talk	98.1	98.4	99.8	99	83.6	98.8	98.4	94.9	93.3
Avg	86.5	84.8	82.5	78.4	64.9	88.9	87.2	80	72
Time	120	165	55	N/A	N/A	180	150	N/A	N/A

References

[4] M. Amer, D. Xie, M. Zhao, S. Todorovic, S-C. Zhu. "Cost-sensitive top-down/bottom-up inference for multi-scale activity recognition" ECCV12.
[6] W. Choi, S. Savarse. "A unified frame work for multi-target tracking and collective activity recognition" ECCV12.
[7] W. Choi, K. Shahid, S. Savarse. "What are they doing?: Collective activity classification using spatio-temporal relationship among people." ICCV-W09
[9] S. Khamis, V. Morariu, L.S. Davis. "Combining per-frame and per-track cues for multi-person action recognition." ECCV12

Acknowledgment

NSF IIS 1018490, ONR MURI N00014-10-1-0933, DARPA MSEE FA 8650-11-1-7149