Cost-Sensitive Top-down/Bottom-up Multi-scale Activity Recognition

Mohamed R. Amer[1], Dan Xie[2], Mingtian Zhao[2],
Sinisa Todorovic[1], and Song-Chun Zhu[2]

[1]Oregon State University, Corvallis OR
[2]University of California, Los Angeles CA

ECCV 2012

Problem – Given

- High-resolution, long video of a large scene
- People engaged in individual actions and group activities

Problem – Goal

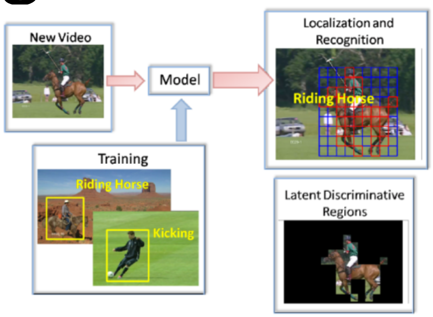
Answer WHAT, WHERE, and WHEN queries about individual actions and group activities

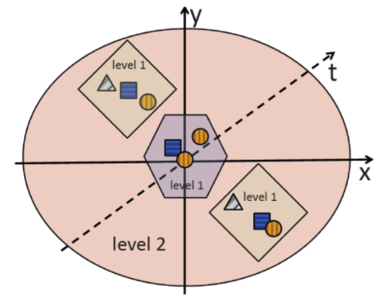
Contributions

- Multi-scale activity recognition
  - Jointly addressing activities at different scales
- Cost-Sensitive Inference
- New Dataset
  - High resolution video
  - Allows for digital zoom-in and zoom-out
  - Many co-occurring individual and group activities

Prior Work – Punctual/Repetitive Activities

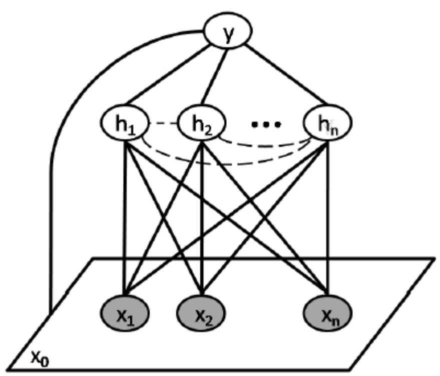
- Single Actor

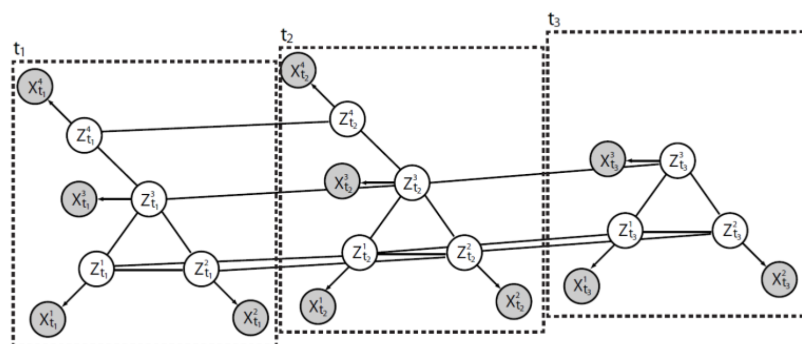




Lan et al ICCV11, Rodriguez et al. CVPR08, Kovashka & Grauman CVPR10
Laptev et al. ICCV03,ICCV07, Dollar et al. VS-PETS05 , Blank et al. ICCV05 …

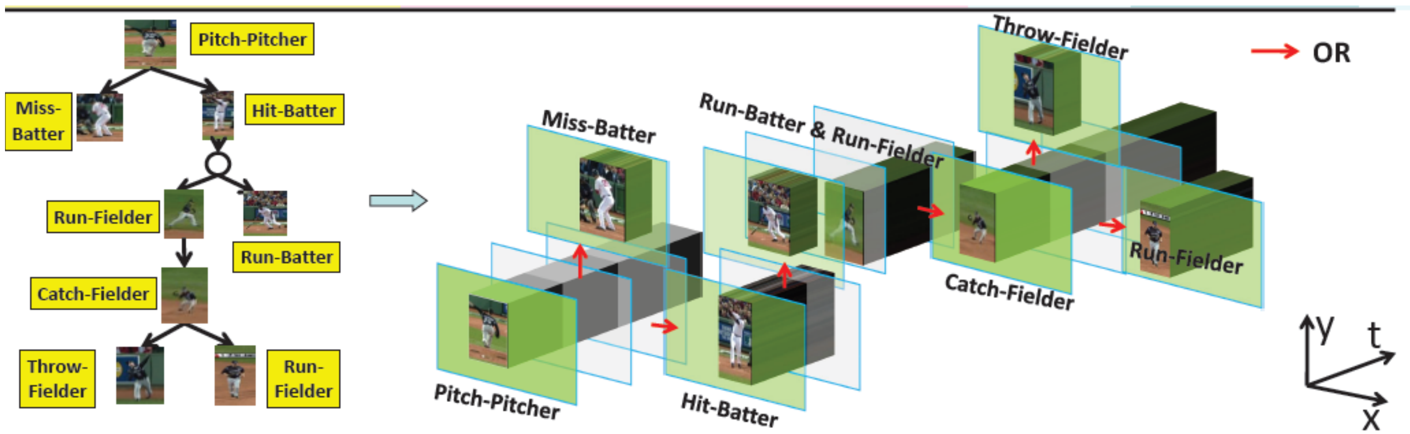
- Single Group



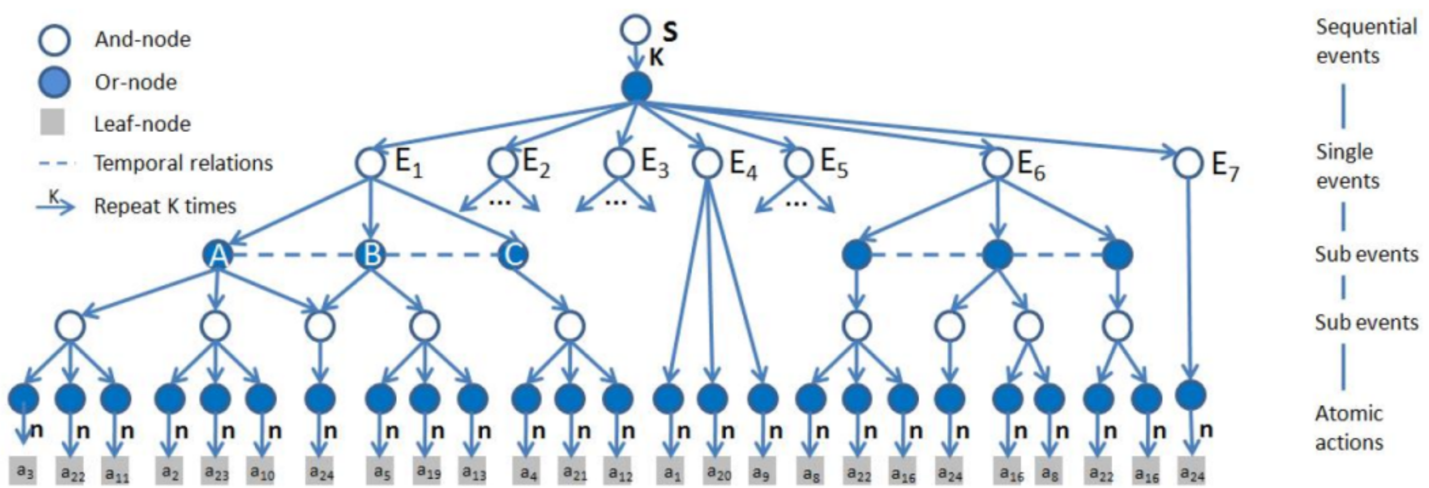


Lan et al PAMI11, Ryoo & Aggarwal ICCV09, Ryoo ICCV11, Choi et al CVPR11
Amer & Todorovic ICCV11, CVPR12 …

Prior Work – Structured Activities



Gupta et al CVPR09

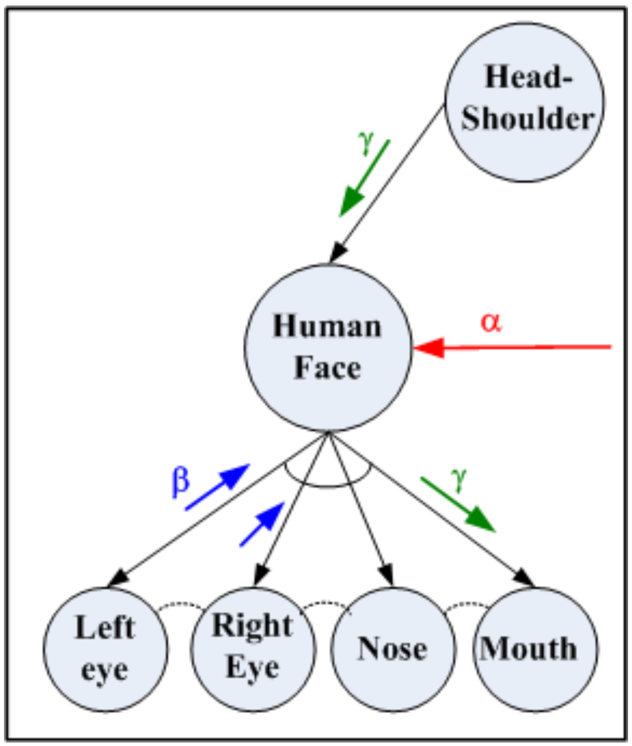


Ryoo et al ICCV09 ,Pei et al ICCV11, Brendel et al CVPR11….

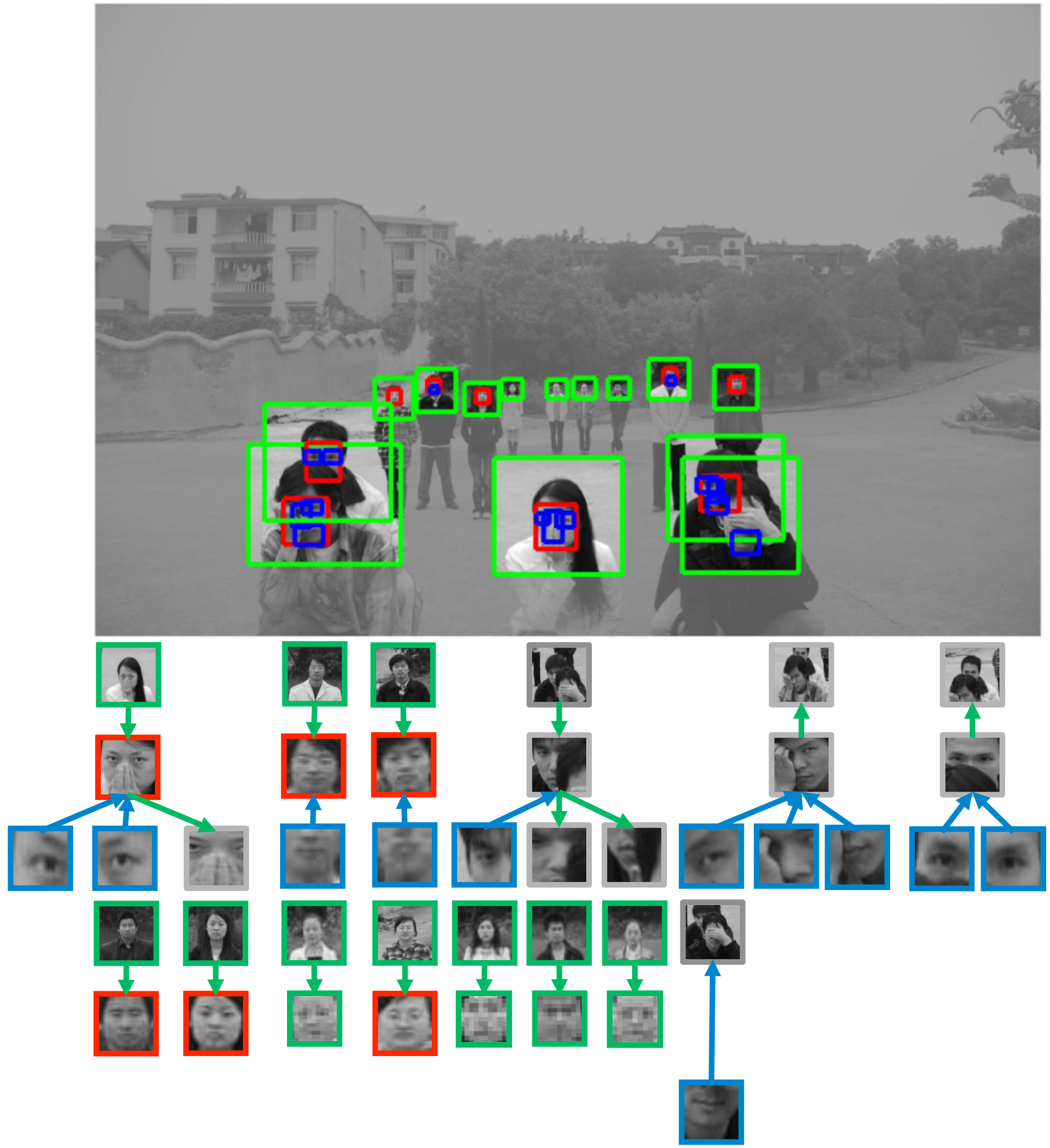
Our Approach

- **Unified hierarchical model of:**
  - People and the objects they interact with
  - Individual actions
  - Group activities
- **Cost-sensitive zooming-in/-out for:**
  - Fusing visual cues at different scales
  - Answering: What, where, when

Our Approach – Related Prior Work

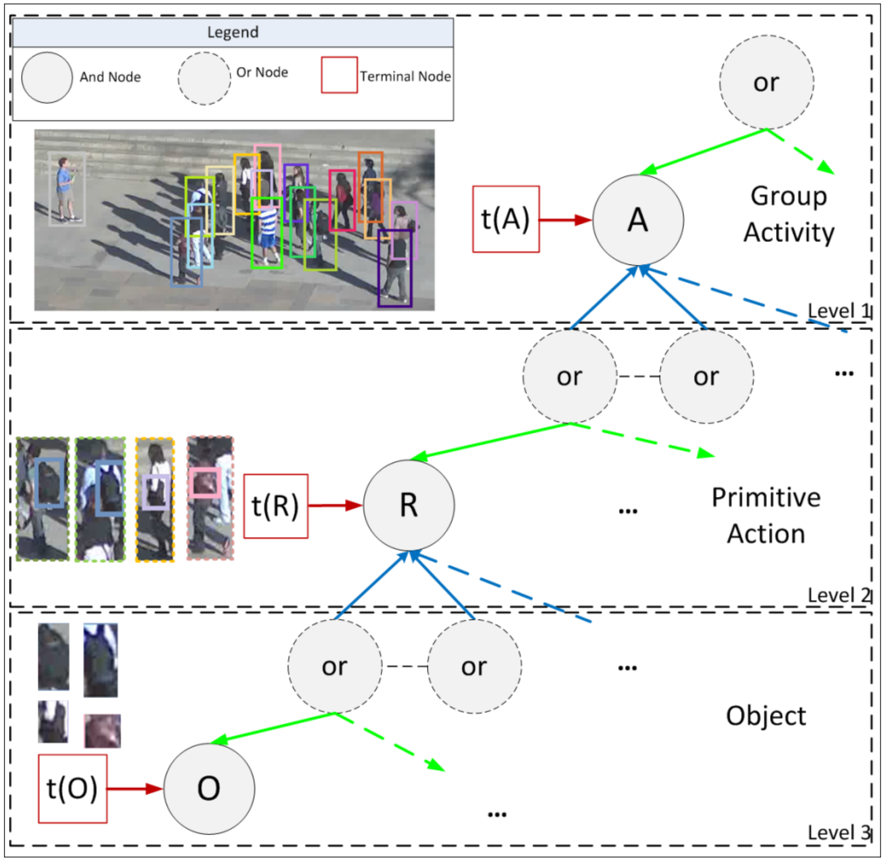


Wu & Zhu IJCV11



Model: And-Or Graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$$



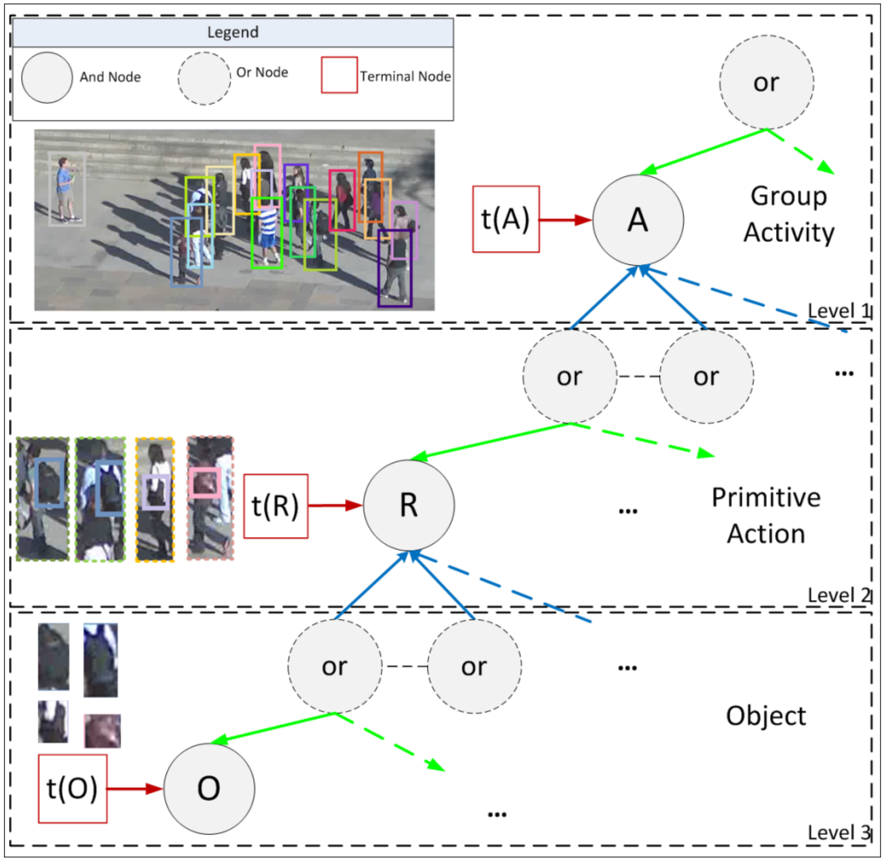
Model: And-Or Graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$$

$\mathcal{V}$: Graph nodes $(\mathcal{V}_{NT}, \mathcal{V}_T)$

$\mathcal{V}_{NT}$: Non-terminal nodes such as A, R, O

$\mathcal{V}_T$: Terminal nodes such as t(A), t(R), t(O)



Model: And-Or Graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$$

$\mathcal{V}$: Graph nodes $(\mathcal{V}_{NT}, \mathcal{V}_T)$

$\mathcal{V}_{NT}$: Non-terminal nodes such as A, R, O

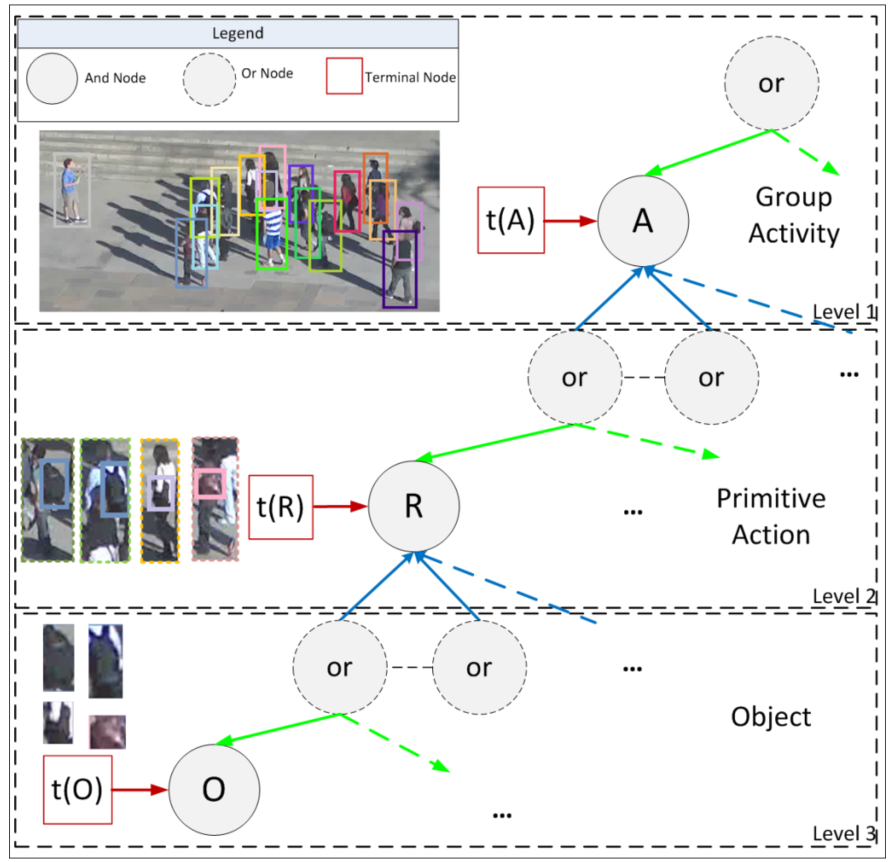
$\mathcal{V}_T$: Terminal nodes such as t(A), t(R), t(O)

$\mathcal{E}$: Graph edges $(\mathcal{E}_{\text{rel}}, \mathcal{E}_{\text{dec}}, \mathcal{E}_{\text{switch}})$

$\mathcal{E}_{\text{rel}}$: Relation edges

$\mathcal{E}_{\text{dec}}$: Decomposition edges

$\mathcal{E}_{\text{switch}}$: Switching edges



Model: And-Or Graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$$

$\mathcal{V}$: Graph nodes $(\mathcal{V}_{NT}, \mathcal{V}_T)$

$\mathcal{V}_{NT}$: Non-terminal nodes such as A, R, O

$\mathcal{V}_T$: Terminal nodes such as t(A), t(R), t(O)

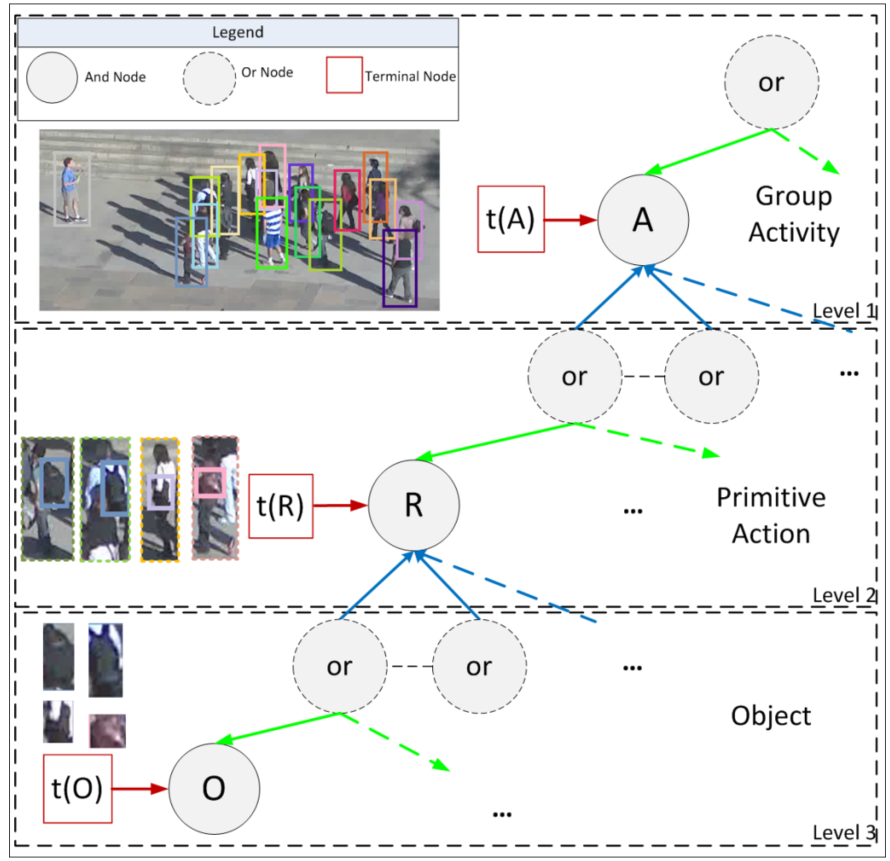
$\mathcal{E}$: Graph edges $(\mathcal{E}_{\text{rel}}, \mathcal{E}_{\text{dec}}, \mathcal{E}_{\text{switch}})$

$\mathcal{E}_{\text{rel}}$: Relation edges

$\mathcal{E}_{\text{dec}}$: Decomposition edges

$\mathcal{E}_{\text{switch}}$: Switching edges

$\mathcal{P}$: Probability over all parse graphs



Model: And-Or Graph

$$W = (K, \{\mathrm{pg}_k : k = 1, 2, \ldots, K\})$$

$$p(W) = p(K) \prod_{k=1}^{K} p(\mathrm{pg}_k)$$

$$p(\mathrm{pg}) = \frac{1}{Z} \exp(-E(\mathrm{pg}))$$

$$\begin{aligned} E(\mathrm{pg}) = -\sum_l \Big[& \sum_{(\vee^l, \wedge^l) \in \mathcal{E}_{\mathrm{switch}}(\mathrm{pg})} \log p(\wedge^l | \vee^l) \\ & + \sum_{(\wedge^l, \wedge^{l-}) \in \mathcal{E}_{\mathrm{dec}}(\mathrm{pg})} \log p(X_{\wedge^l} | X_{\wedge^{l-}}) \\ & + \sum_{(\wedge_i^{l+}, \wedge_j^{l+}) \in \mathcal{E}_{\mathrm{rel}}(\mathrm{pg})} \log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) \Big] \end{aligned}$$

Inference

$$W^* = \arg\max_{W \in \Omega} p(W) p(I_\Lambda | W)$$

$$p(W) = p(K) \prod_{k=1}^{K} p(\mathrm{pg}_k)$$

$$p(I_\Lambda | W) = q(I_\Lambda) \prod_{k=1}^{K} \frac{p(I_{\Lambda_{\mathrm{pg}_k}} | \mathrm{pg}_k)}{q(I_{\Lambda_{\mathrm{pg}_k}})}$$

Inference

$$\mathrm{pg}^* = \arg \max_{\mathrm{pg} \in \Omega(\mathrm{pg})} \left[\log p(\mathrm{pg}) + \log \frac{p(I_{\Lambda_{\mathrm{pg}}} | \mathrm{pg})}{q(I_{\Lambda_{\mathrm{pg}}})} \right]$$

$$p(\mathrm{pg}) = \frac{1}{Z} \exp(-E(\mathrm{pg})), \quad Z = \sum_{\mathrm{pg}} \exp(-E(\mathrm{pg}))$$

$$\begin{aligned} E(\mathrm{pg}) = -\sum_l \Big[& \sum_{(\vee^l, \wedge^l) \in \mathcal{E}_{\mathrm{switch}}(\mathrm{pg})} \log p(\wedge^l | \vee^l) \\ & + \sum_{(\wedge^l, \wedge^{l-}) \in \mathcal{E}_{\mathrm{dec}}(\mathrm{pg})} \log p(X_{\wedge^l} | X_{\wedge^{l-}}) \\ & + \sum_{(\wedge_i^{l+}, \wedge_j^{l+}) \in \mathcal{E}_{\mathrm{rel}}(\mathrm{pg})} \log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) \Big] \end{aligned}$$

$$\frac{p(I_{\Lambda_{\mathrm{pg}}} | \mathrm{pg})}{q(I_{\Lambda_{\mathrm{pg}}})} = \sum_{t \in \mathcal{V}_T(\mathrm{pg})} \log \frac{p(I_{\Lambda_t} | t)}{q(I_{\Lambda_t})}$$

Inference

$$pg^* = \arg \max_{pg \in \Omega(pg)} \sum_l \Big\{ \log p(\wedge^l | \vee^l)$$

$$+ \log \frac{p(t_{\wedge^l} | t)}{q(t_{\wedge^l})}$$

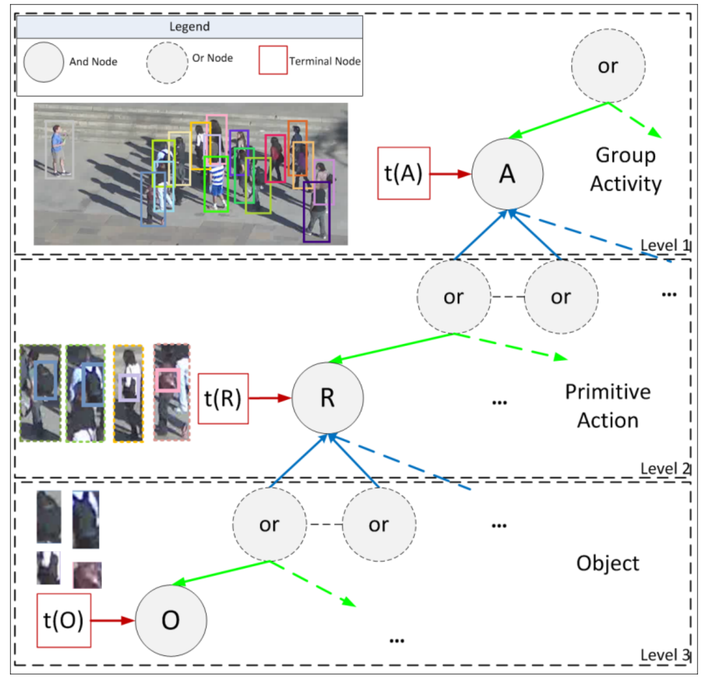
No zoom

$$+ \log \frac{p(t_{\wedge^{l-}} | t)}{q(t_{\wedge^{l-}})} + \log p(X_{\wedge^l} | X_{\wedge^{l-}})$$

zoom-out

$$+ p(N^l) \sum_{i=1}^{N^l} \Big[\log p(X_{\wedge_i^{l+}} | X_{\wedge^l}) + \log \frac{p(t_{\wedge+i^{l+}} | t)}{q(t_{\wedge_i^{l+}})} + \sum_{i \neq j} \log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) \Big] \Big\}$$

zoom-in



Inference: Structure

$$\text{pg}^* = \arg\max_{\text{pg}\in\Omega(\text{pg})} \sum_l \Big\{ \log p(\wedge^l|\vee^l)$$

$$+\log \frac{p(t_{\wedge^l}|t)}{q(t_{\wedge^l})}$$

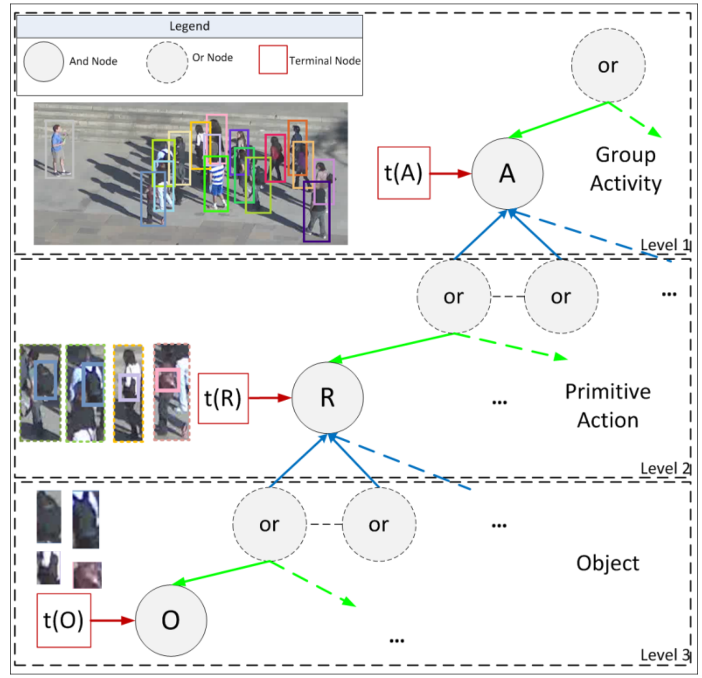
No zoom

$$+\log \frac{p(t_{\wedge^{l-}}|t)}{q(t_{\wedge^{l-}})} + \log p(X_{\wedge^l}|X_{\wedge^{l-}})$$

zoom-out

$$+p(N^l)\sum_{i=1}^{N^l}\Big[\log p(X_{\wedge_i^{l+}}|X_{\wedge^l}) + \log \frac{p(t_{\wedge+i^{l+}}|t)}{q(t_{\wedge_i^{l+}})} + \sum_{i\neq j}\log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}})\Big]\Big\}$$

zoom-in



$p(\wedge^l|\vee^l)$: is the probability of an And node given a parent Or node

Inference: α – Process

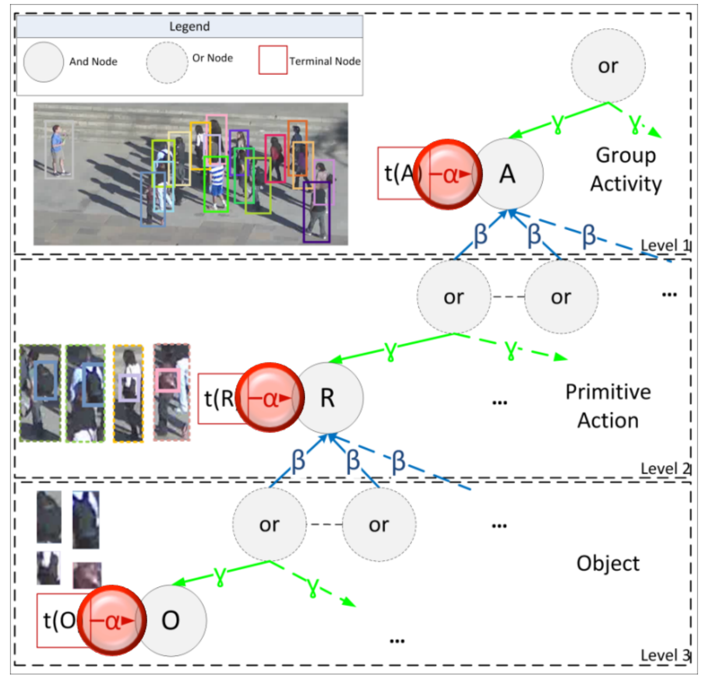
$$\text{pg}^* = \arg \max_{\text{pg}\in\Omega(\text{pg})} \sum_l \Big\{ \log p(\wedge^l | \vee^l)$$

$$+ \log \frac{p(t_{\wedge^l}|t)}{q(t_{\wedge^l})}$$

No zoom

$$+ \log \frac{p(t_{\wedge^{l-}}|t)}{q(t_{\wedge^{l-}})} + \log p(X_{\wedge^l}|X_{\wedge^{l-}})$$

zoom-out



$$+ p(N^l) \sum_{i=1}^{N^l} \Big[\log p(X_{\wedge_i^{l+}}|X_{\wedge^l}) + \log \frac{p(t_{\wedge_i^{l+}}|t)}{q(t_{\wedge_i^{l+}})} + \sum_{i\neq j} \log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) \Big] \Big\}$$

zoom-in

$p(N^l)$: is an exponential prior over the number of children

Inference: β – Process

$$pg^* = \arg \max_{pg \in \Omega(pg)} \sum_l \Big\{ \log p(\wedge^l | \vee^l)$$

$$+ \log \frac{p(t_{\wedge^l} | t)}{q(t_{\wedge^l})}$$

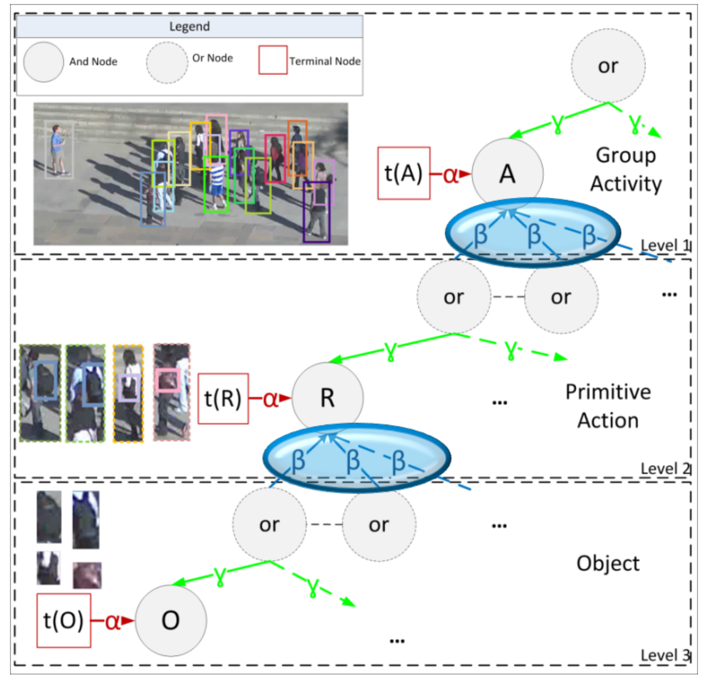
No zoom

$$+ \log \frac{p(t_{\wedge^{l-}} | t)}{q(t_{\wedge^{l-}})} + \log p(X_{\wedge^l} | X_{\wedge^{l-}})$$

zoom-out

$$+ p(N^l) \sum_{i=1}^{N^l} \Big[\log p(X_{\wedge_i^{l+}} | X_{\wedge^l}) + \log \frac{p(t_{\wedge+i^{l+}} | t)}{q(t_{\wedge_i^{l+}})} + \sum_{i \neq j} \log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) \Big] \Big\}$$

zoom-in



$p(N^l)$: is an exponential prior over the number of children

$p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}})$: is the β-process, the probability of binding two children

Inference: γ – Process

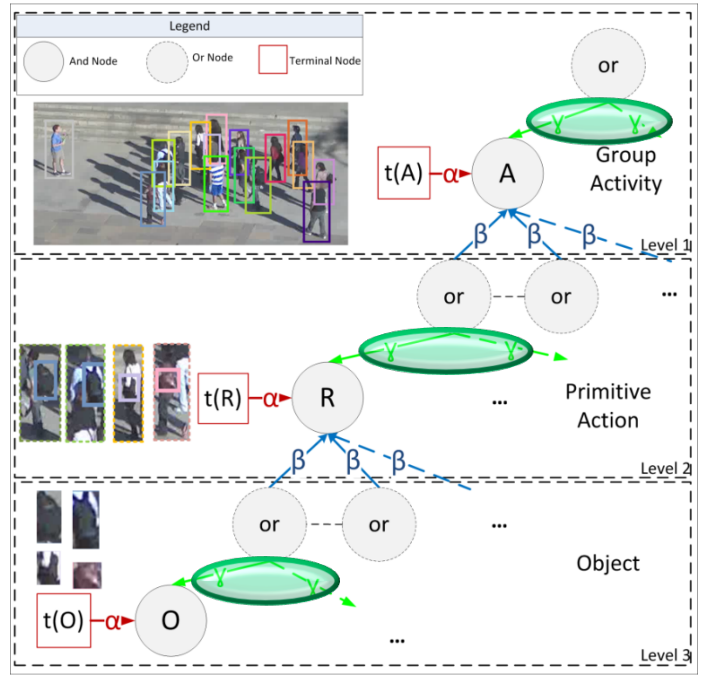
$$pg^* = \arg \max_{pg \in \Omega(pg)} \sum_l \left\{ \log p(\wedge^l | \vee^l) \right.$$

$$+ \log \frac{p(t_{\wedge^l} | t)}{q(t_{\wedge^l})}$$

No zoom

$$+ \log \frac{p(t_{\wedge^{l-}} | t)}{q(t_{\wedge^{l-}})} + \log p(X_{\wedge^l} | X_{\wedge^{l-}})$$

zoom-out



$$\left. + p(N^l) \sum_{i=1}^{N^l} \left[\log p(X_{\wedge_i^{l+}} | X_{\wedge^l}) + \log \frac{p(t_{\wedge + i^{l+}} | t)}{q(t_{\wedge_i^{l+}})} + \sum_{i \neq j} \log p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) \right] \right\}$$

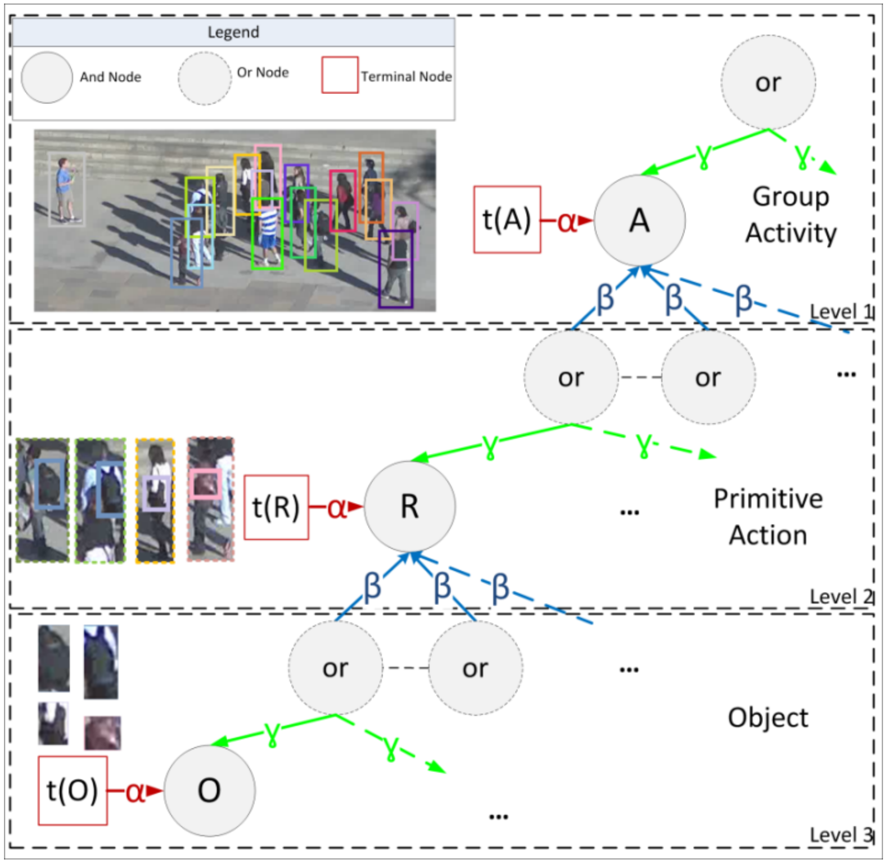
zoom-in

$p(N^l)$: is an exponential prior over the number of children

$p(X_{\wedge^l} | X_{\wedge^{l-}})$, $p(X_{\wedge_i^{l+}} | X_{\wedge^l})$: are the γ-processes, a child's likelihood given its parent

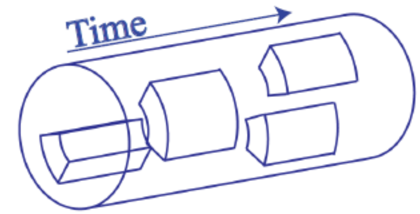
Inference – α, β, γ Processes

- **α**: running a detector of the activity
- **β**: bottom-up binding of parts of the activity
- **γ**: top-down prediction of parts from the activity



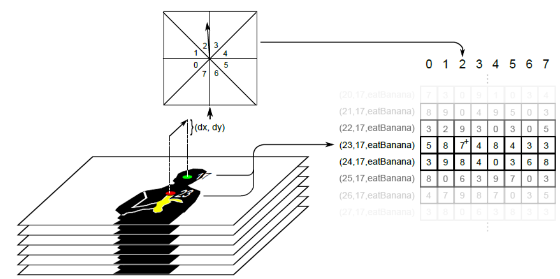
α – Process

- Group Activities:
  - – Space-Time Volume (STV)



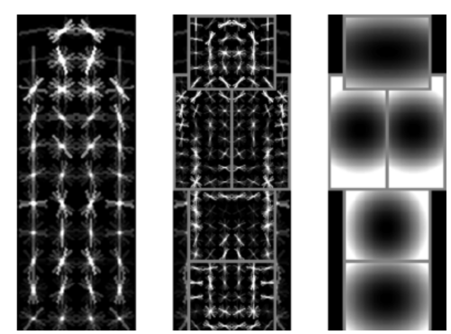
(Choi et al. CVPR2011)

- Primitive Actions:
  - – Motion (STIP-HOG)/Appearance (KLT)



(Matikainen et al. ECCV2010)

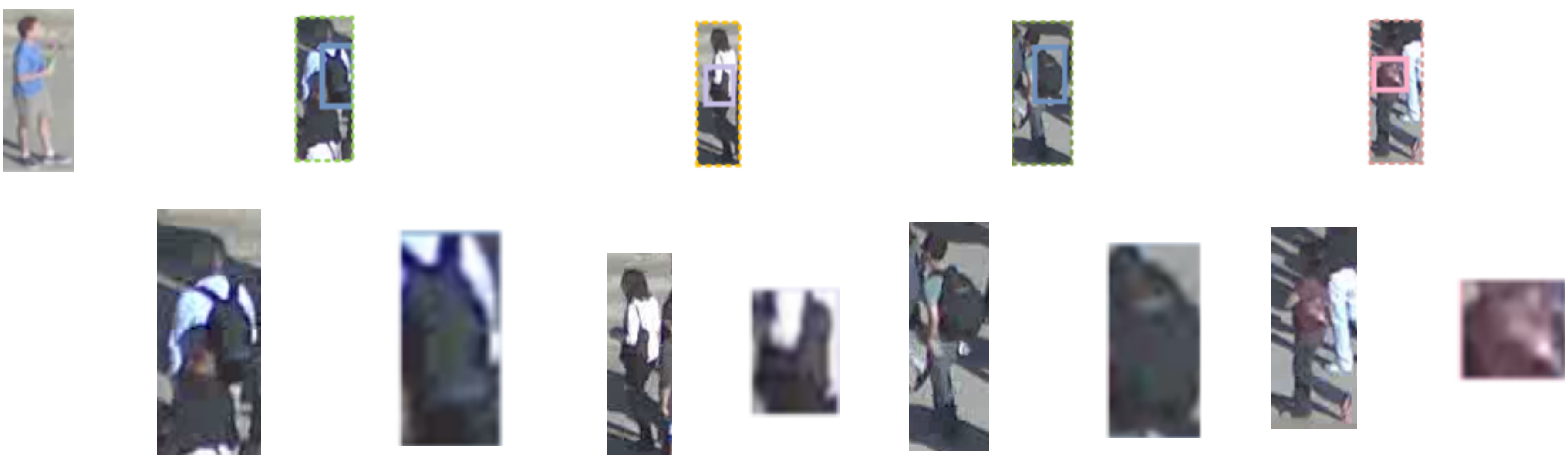
- Objects:
  - – Deformable Part-based Model (DPM)



(Felzenszwalb et al. PAMI10)

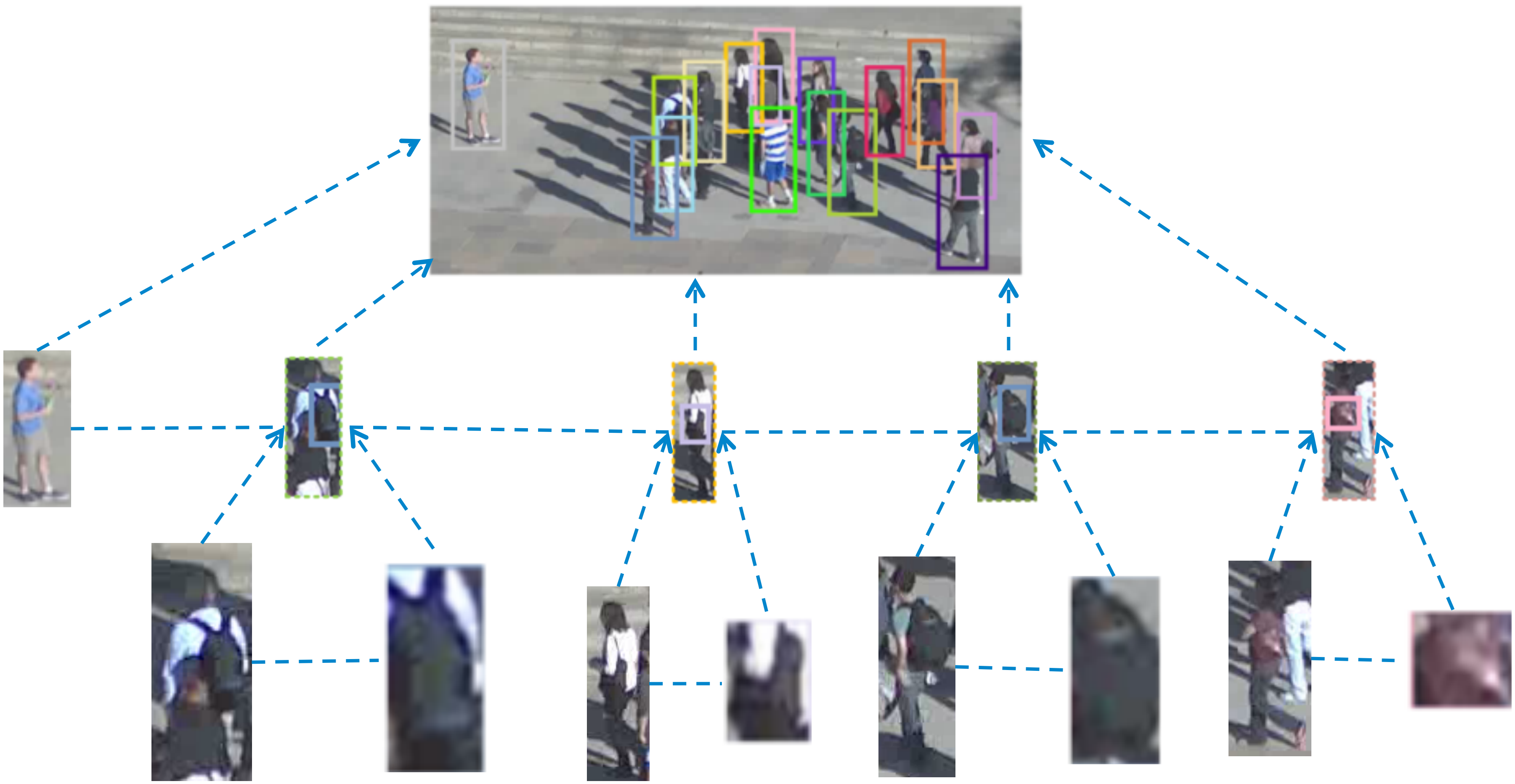
β, γ – Process

- β and γ processes are modeled as Gaussian distributions over location, scale and orientation.



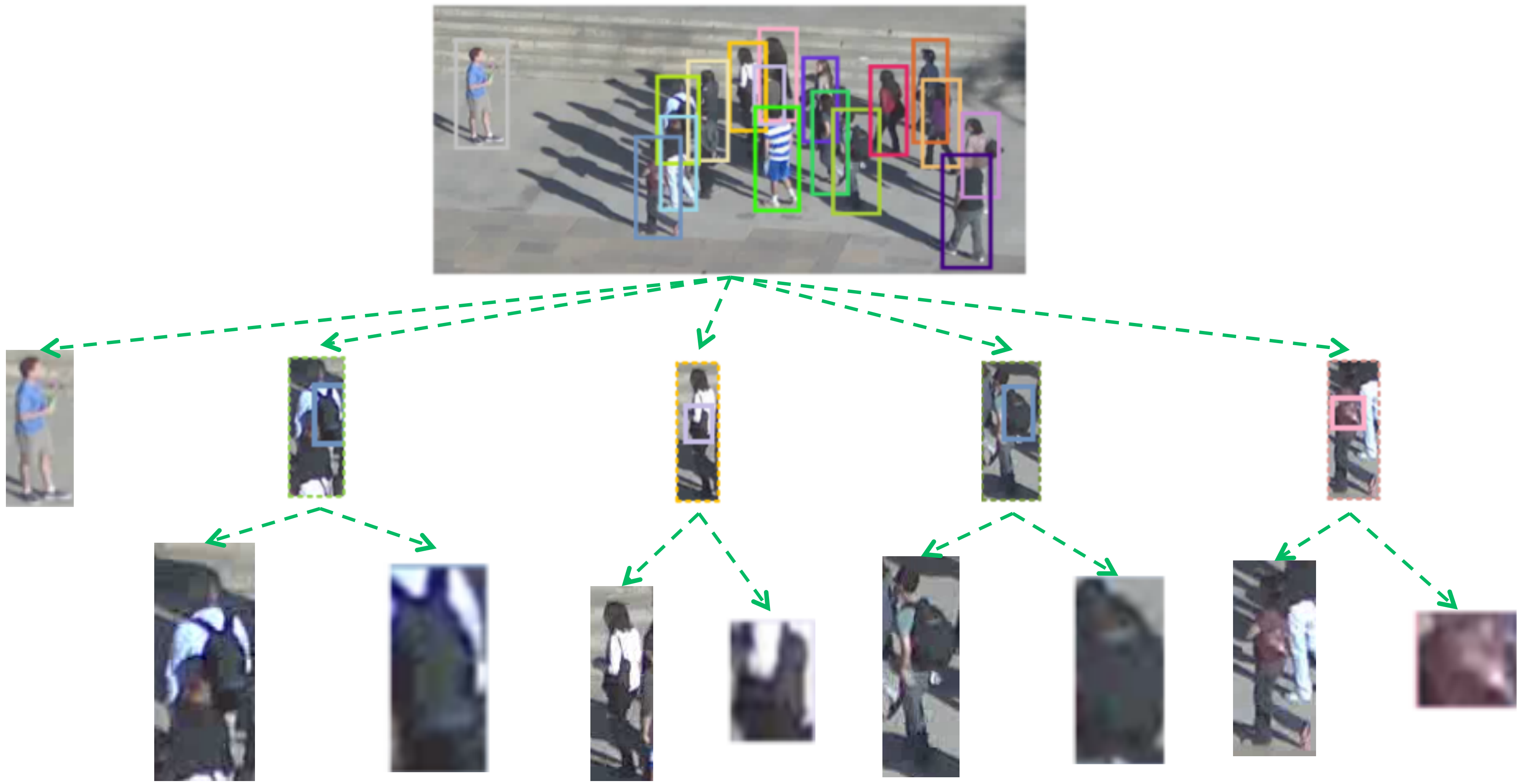
β – Process

β-Process: $p(X_{\wedge_i^{l+}}, X_{\wedge_j^{l+}}) = N(X_{\wedge_i^{l+}} - X_{\wedge_j^{l+}}; \mu_{\beta^l}, \Sigma_{\beta^l})$



γ – Process

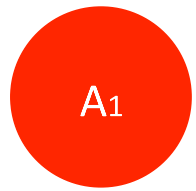
γ-Process: $p(X_{\wedge_i^{l+}} | X_{\wedge^{l}}) = N(X_{\wedge_i^{l+}} - X_{\wedge^{l}}; \mu_{\gamma^l}, \Sigma_{\gamma^l})$



Cost-Sensitive Inference

- Reinforcement Learning based Inference
  - Explore/Exploit strategy
  - Q-Learning to learn the optimal moves

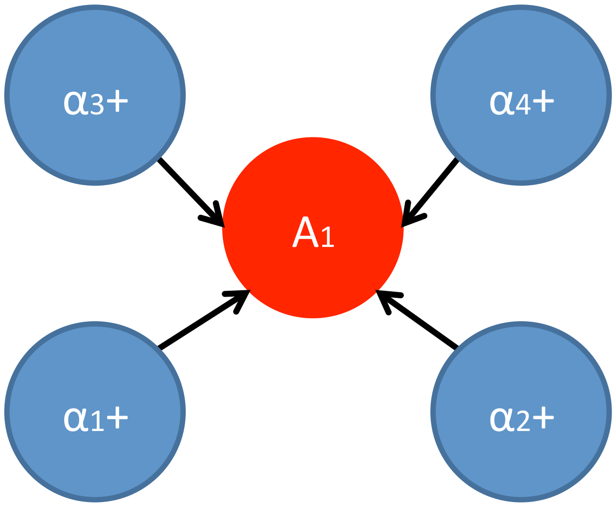
Explore/Exploit



# of detectors left= 7

$p(pg^{(t)})=0$

Explore/Exploit

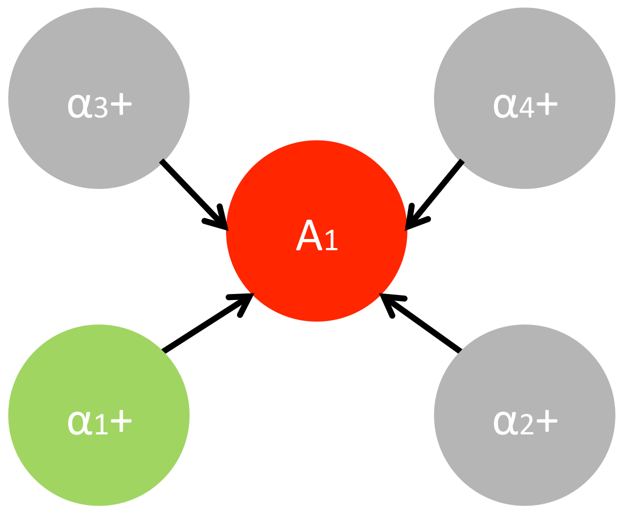


Q Table for A_1
(Exploit) α_1+
(Explore) α_2+
(Explore) α_3+
(Explore) α_4+

# of detectors left= 7

$p(pg^{(t)})=0$

Explore/Exploit



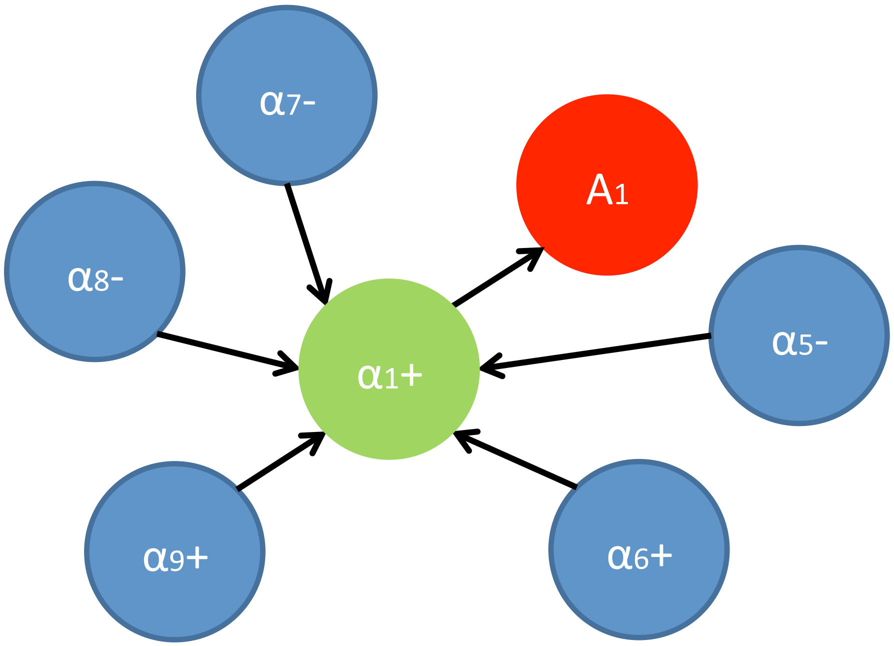
Q Table for A_1
(Exploit) α_1+
(Explore) α_2+
(Explore) α_3+
(Explore) α_4+

# of detectors left = 6

$p(pg^{(t+1)})=0.2$

$p(pg^{(t)})=0$

Explore/Exploit

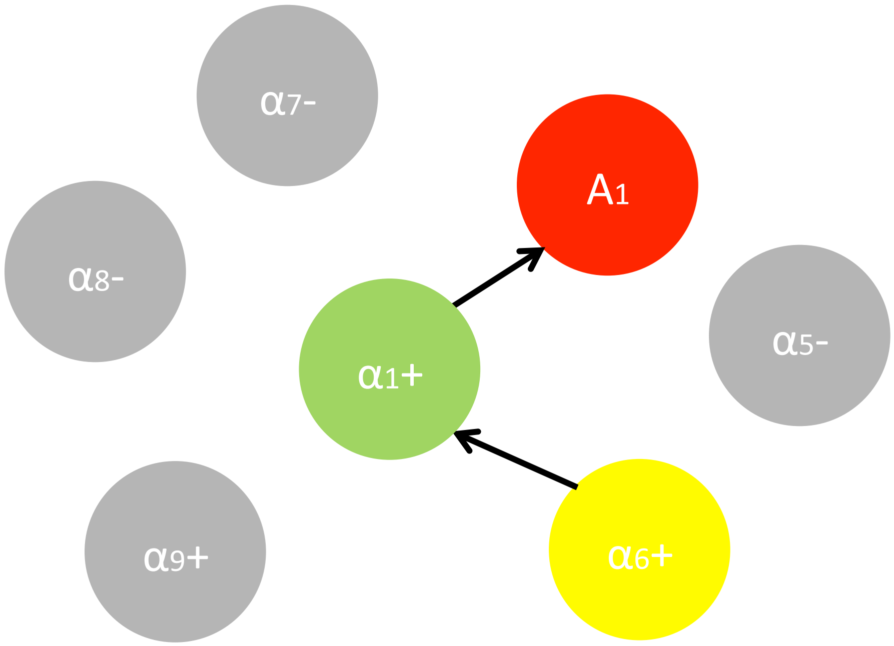


Q Table for α_1
(Exploit) α_5-
(Explore) α_6+
(Explore) α_7-
(Explore) α_8-
(Explore) α_9+

# of detectors left = 6

$p(pg^{(t)})=0.2$

Explore/Exploit



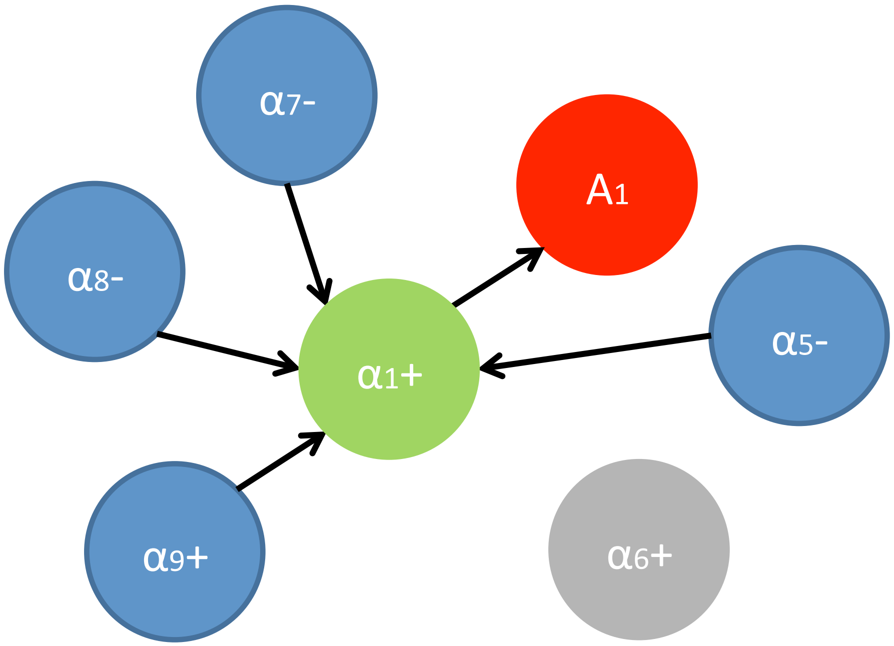
Q Table for α_1
(Exploit) α_5-
(Explore) α_6+
(Explore) α_7-
(Explore) α_8-
(Explore) α_9+

# of detectors left = 5

$p(pg^{(t+1)})=0.2$

$p(pg^{(t)})=0.2$

Explore/Exploit

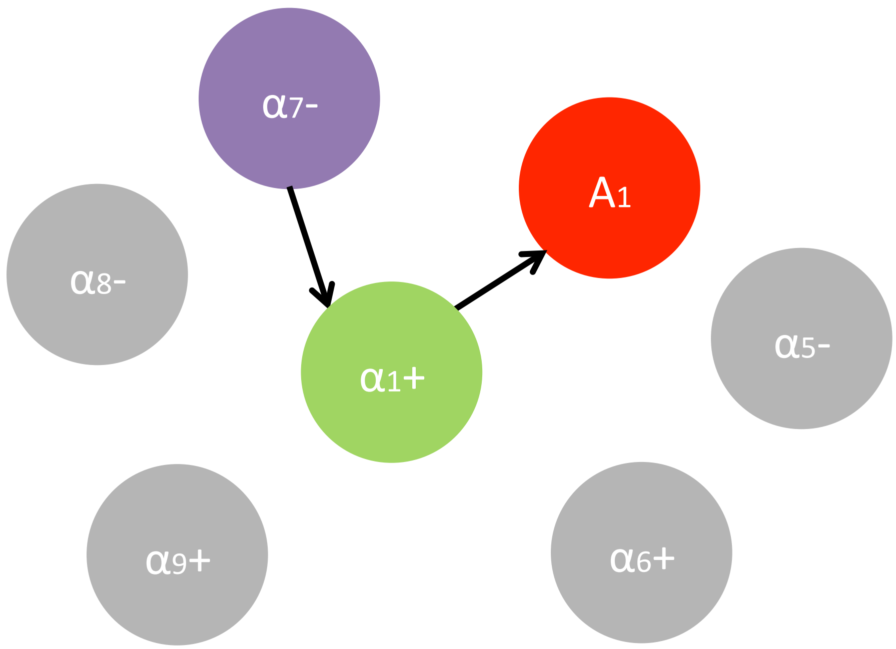


Q Table for α_1
(Exploit) α_5-
(Explore) α_6+
(Explore) α_7-
(Explore) α_8-
(Explore) α_9+

# of detectors left = 5

$p(pg^{(t)})=0.2$

Explore/Exploit



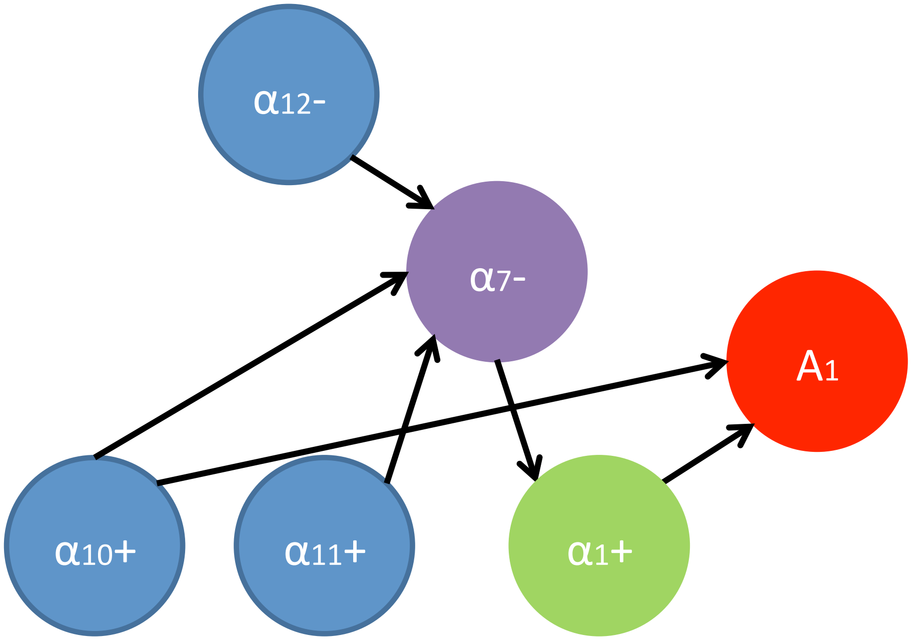
Q Table for α_1
(Exploit) α_5-
(Explore) α_6+
(Explore) α_7-
(Explore) α_8-
(Explore) α_9+

# of detectors left = 4

$p(pg^{(t+1)})=0.4$

$p(pg^{(t)})=0.2$

Explore/Exploit

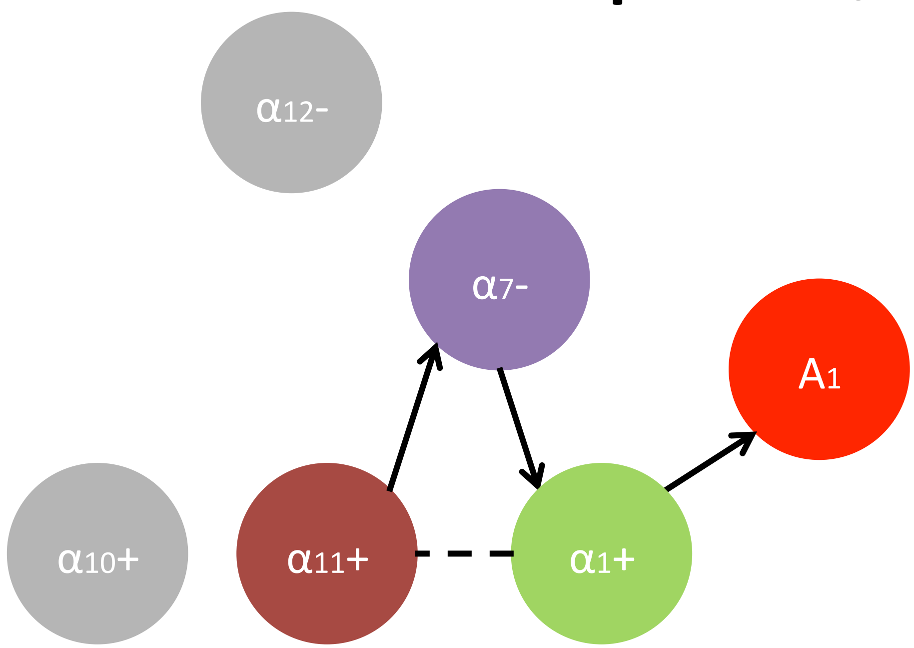


Q Table for α_7
(Exploit) α_{10}+
(Explore) α_{11}+
(Explore) α_{12}-

# of detectors left = 4

$p(pg^{(t)})=0.4$

Explore/Exploit

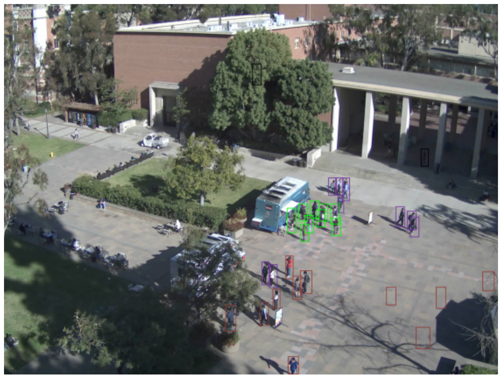


Q Table for α_7
(Exploit) α_{10}+
(Explore) α_{11}+
(Explore) α_{12}-

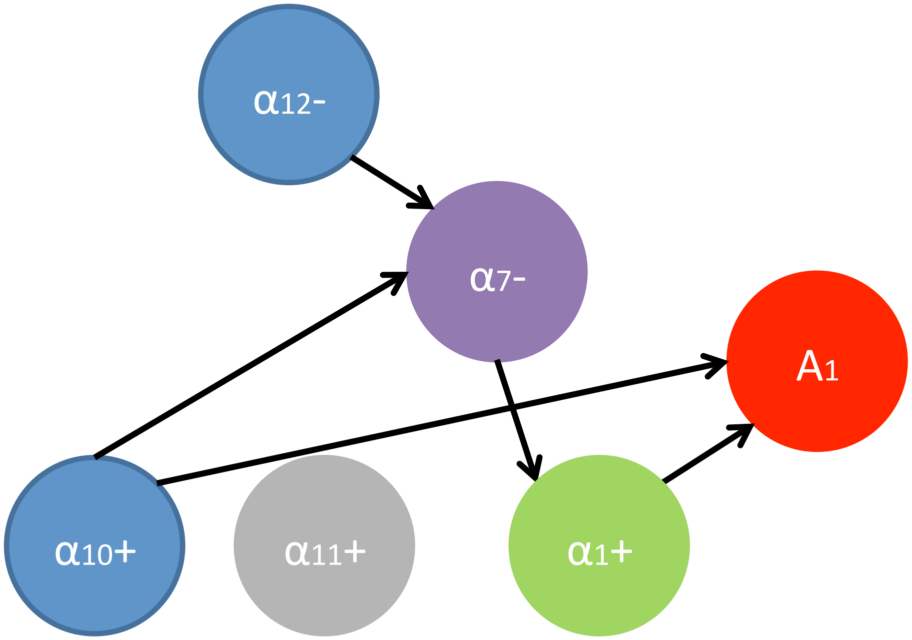
# of detectors left = 3

$p(pg^{(t+1)})=0.4$

$p(pg^{(t)})=0.4$



Explore/Exploit

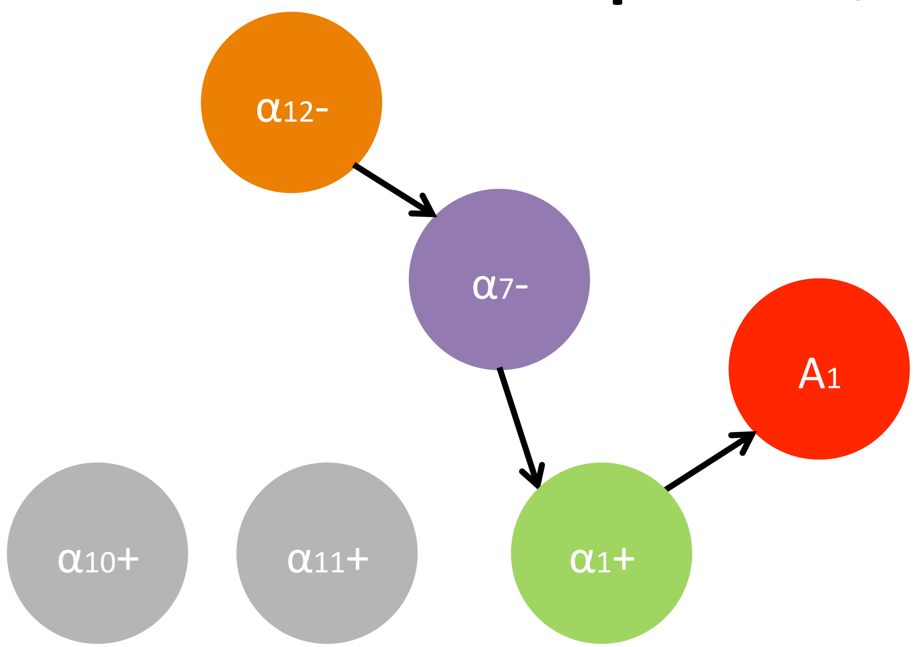


Q Table for α_7
(Exploit) $\alpha_{10}+$
(Explore) $\alpha_{11}+$
(Explore) $\alpha_{12}-$

# of detectors left = 3

$p(pg^{(t)})=0.4$

Explore/Exploit



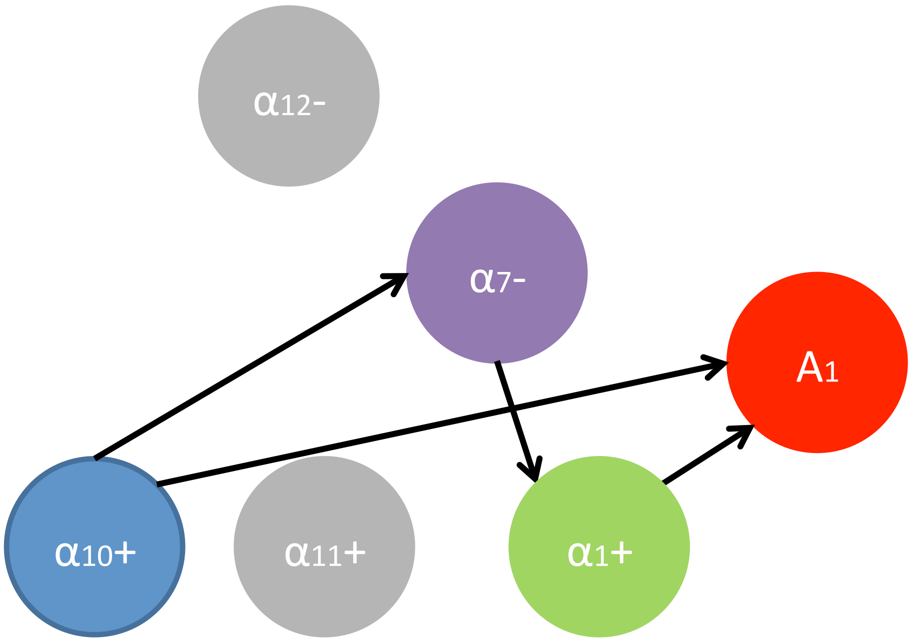
Q Table for α_7
(Exploit) $\alpha_{10}+$
(Explore) $\alpha_{11}+$
(Explore) $\alpha_{12}-$

# of detectors left = 2

$p(pg^{(t+1)})=0.4$

$p(pg^{(t)})=0.4$

Explore/Exploit

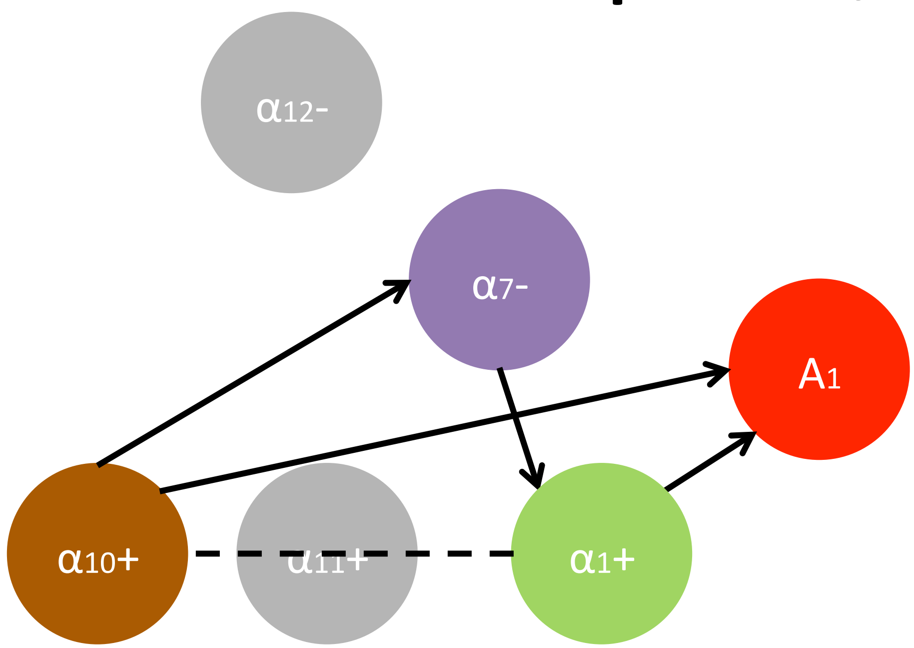


Q Table for α_7
(Exploit) α_{10}+
(Explore) α_{11}+
(Explore) α_{12}-

# of detectors left = 2

$p(pg^{(t)})=0.4$

Explore/Exploit



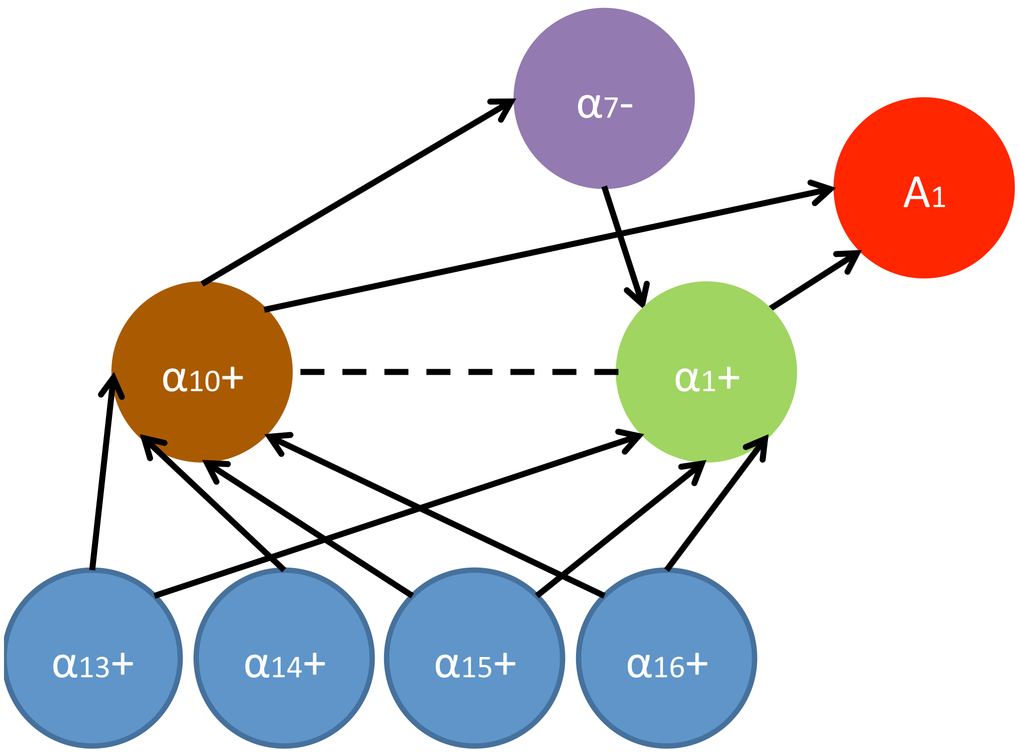
Q Table for α_7
(Exploit) α_{10}+
(Explore) α_{11}+
(Explore) α_{12}-

# of detectors left = 1

$p(pg^{(t+1)})=0.5$

$p(pg^{(t)})=0.4$

Explore/Exploit

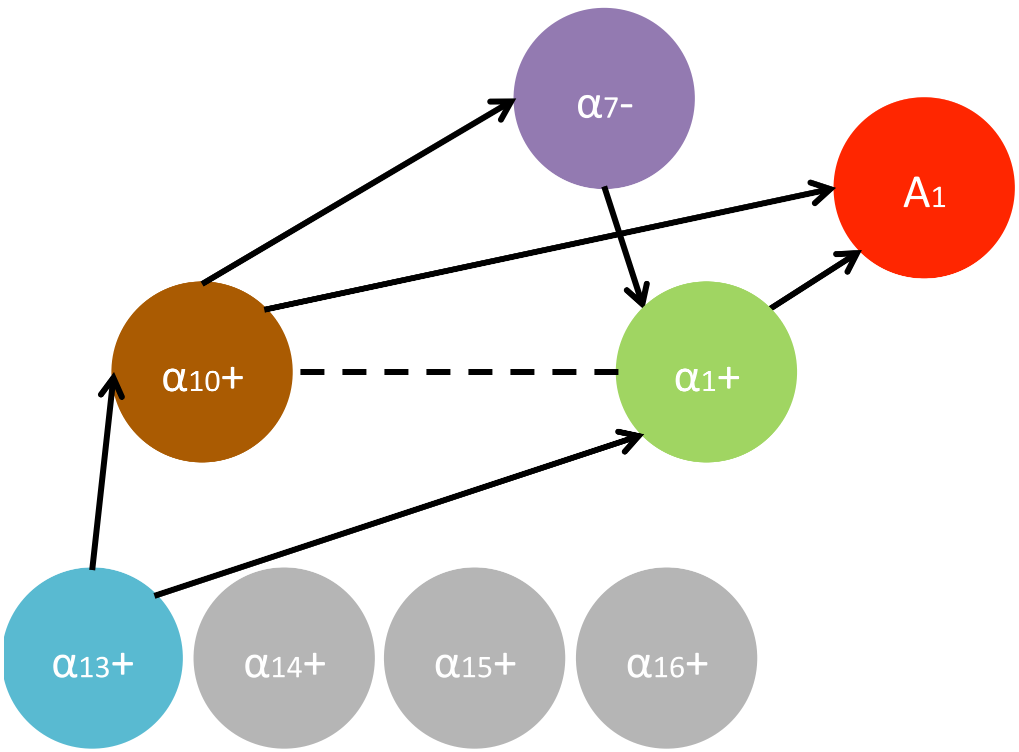


Q Table for α_{10}
(Exploit) $\alpha_{13}+$
(Explore) $\alpha_{14}+$
(Explore) $\alpha_{15}+$
(Explore) $\alpha_{16}+$

\# of detectors left = 1

$p(pg^{(t)})=0.5$

Explore/Exploit



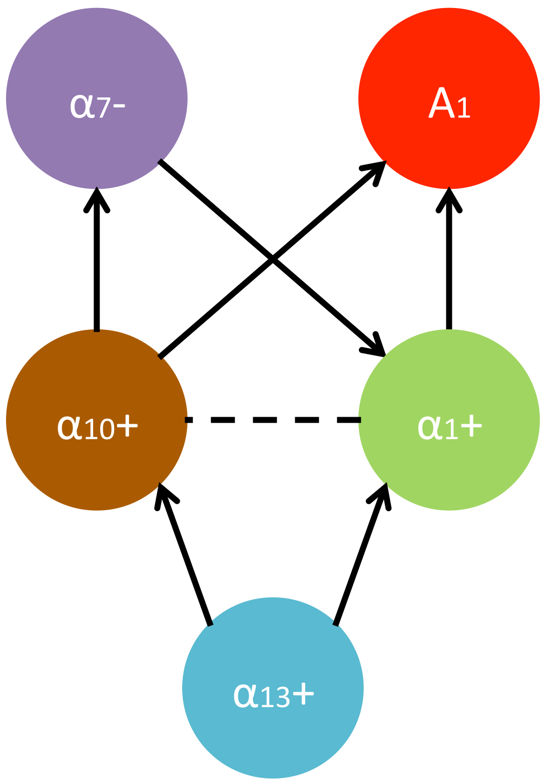
Q Table for α_{10}
(Exploit) α_{13}+
(Explore) α_{14}+
(Explore) α_{15}+
(Explore) α_{16}+

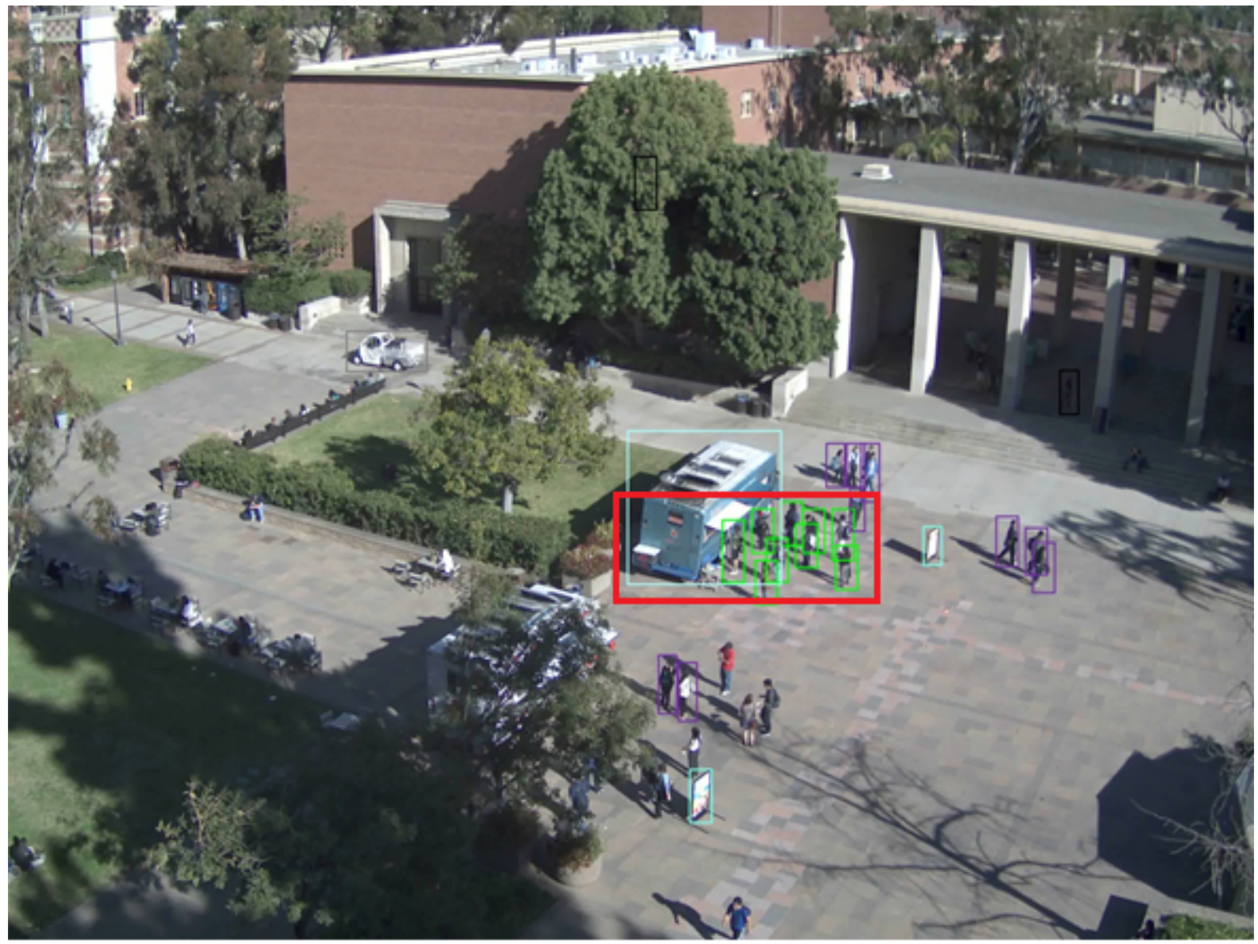
# of detectors left = 0

$p(pg^{(t+1)})=0.6$

$p(pg^{(t)})=0.5$

Explore/Exploit





$p(pg^*)=0.6$

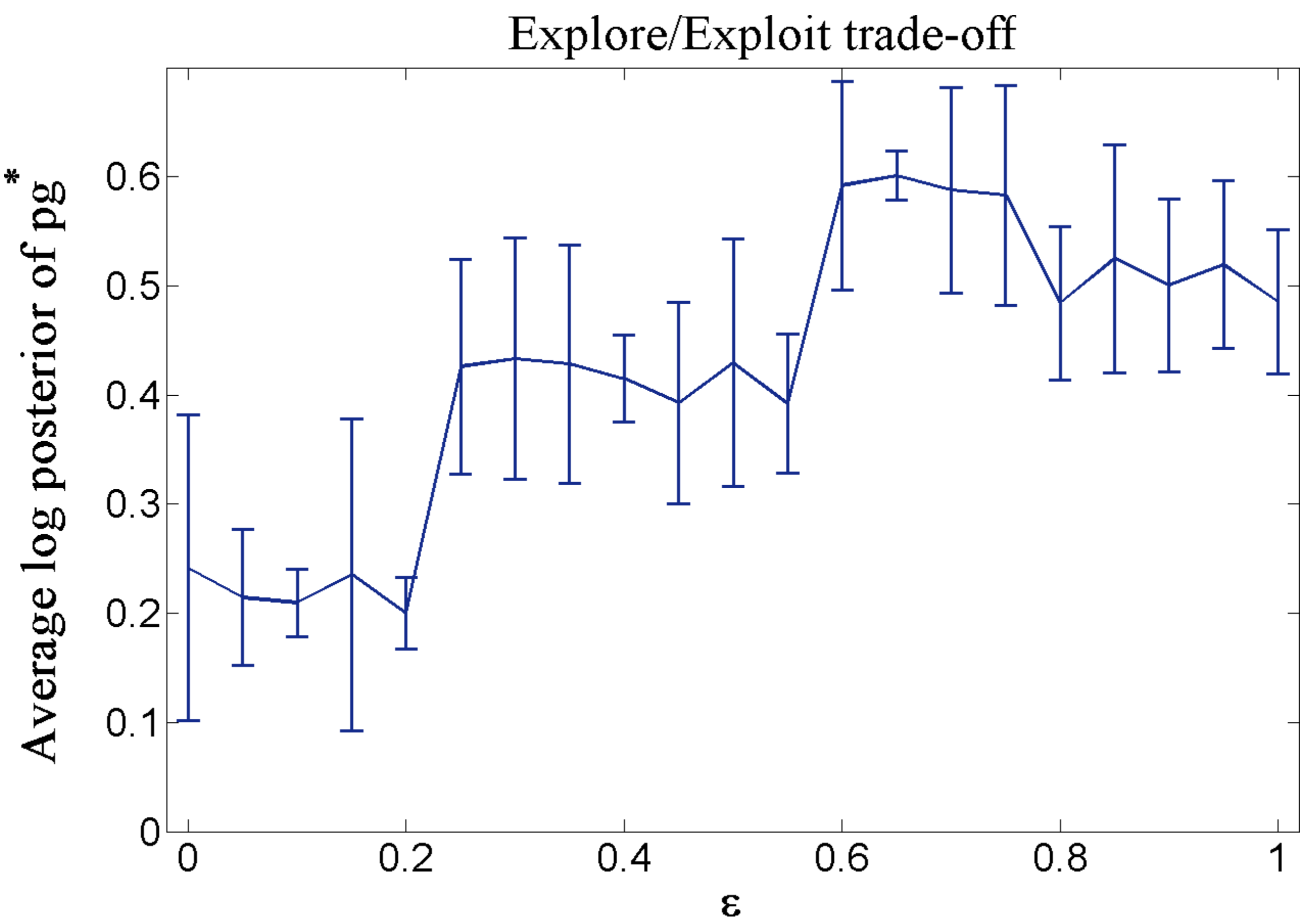
Q-Learning

- States: $\mathbb{S} = \{s\}$
  - Query
  - Current node in the And-Or graph
- Moves: $\mathbb{M} = \{m\}$
  - Run detectors applicable to the current state
- Reward: $\mathbb{R}$
  - Reward the move that increments the log posterior

$$\mathbb{R}_t(s, m; q) = \frac{1}{\left(1 + \exp^{-\left(\log p(\mathrm{pg}_t|\mathbb{M}) - \log p(\mathrm{pg}_t|\mathbb{M} \cup \{m\})\right)}\right)}$$

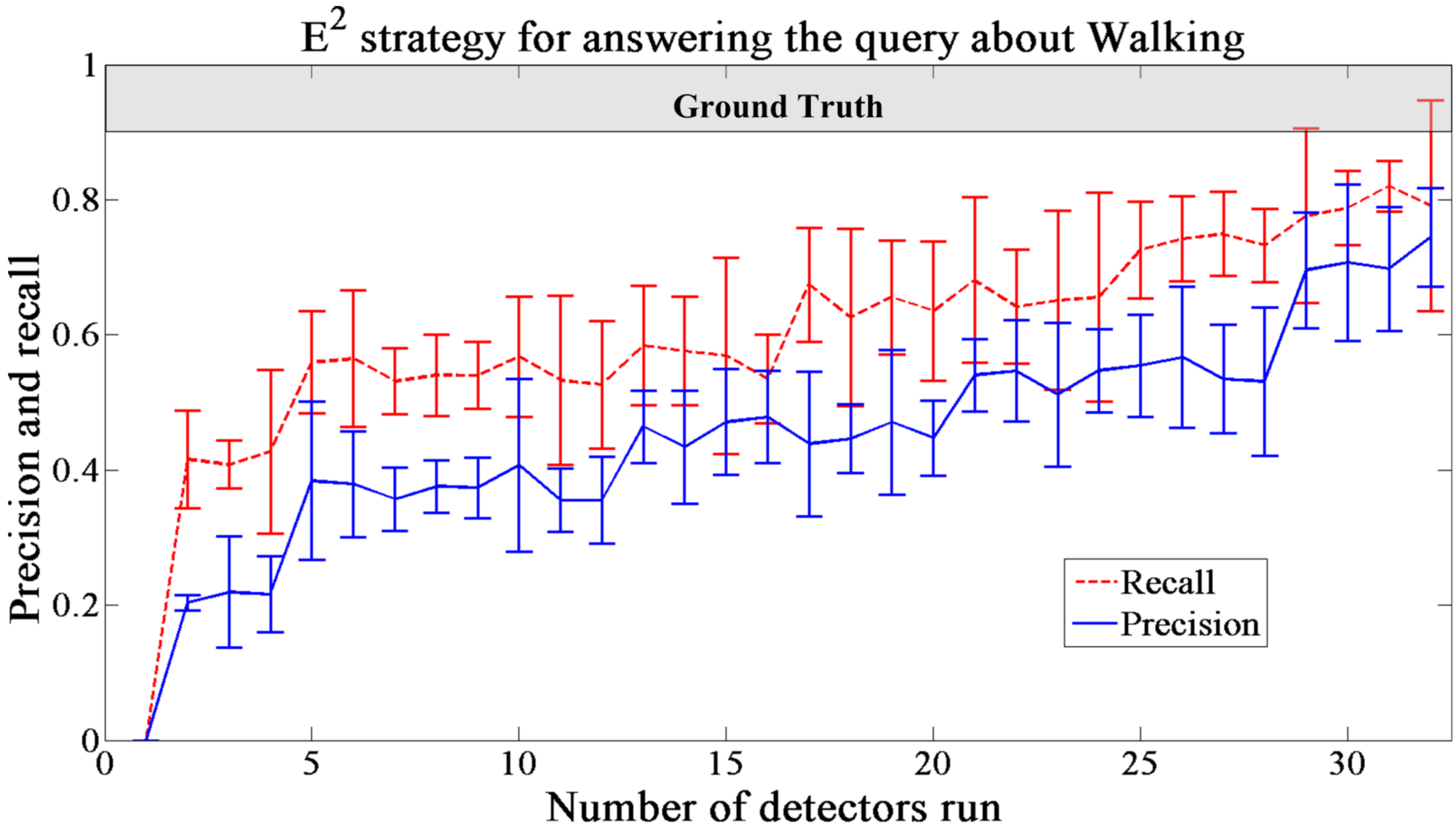
- Transitions: Deterministic simulator.

Varying the Explore/Exploit trade-off



Random Greedy

Varying the Number of Detectors Run



New Dataset

- Footage: 106min
- Frame Rate: 30 fps
- Resolution: 2560x1920 pixels
- Annotations:
  - Group (activities, formation)
  - Individual (actions, poses, facing direction)
  - Objects

Domain Knowledge

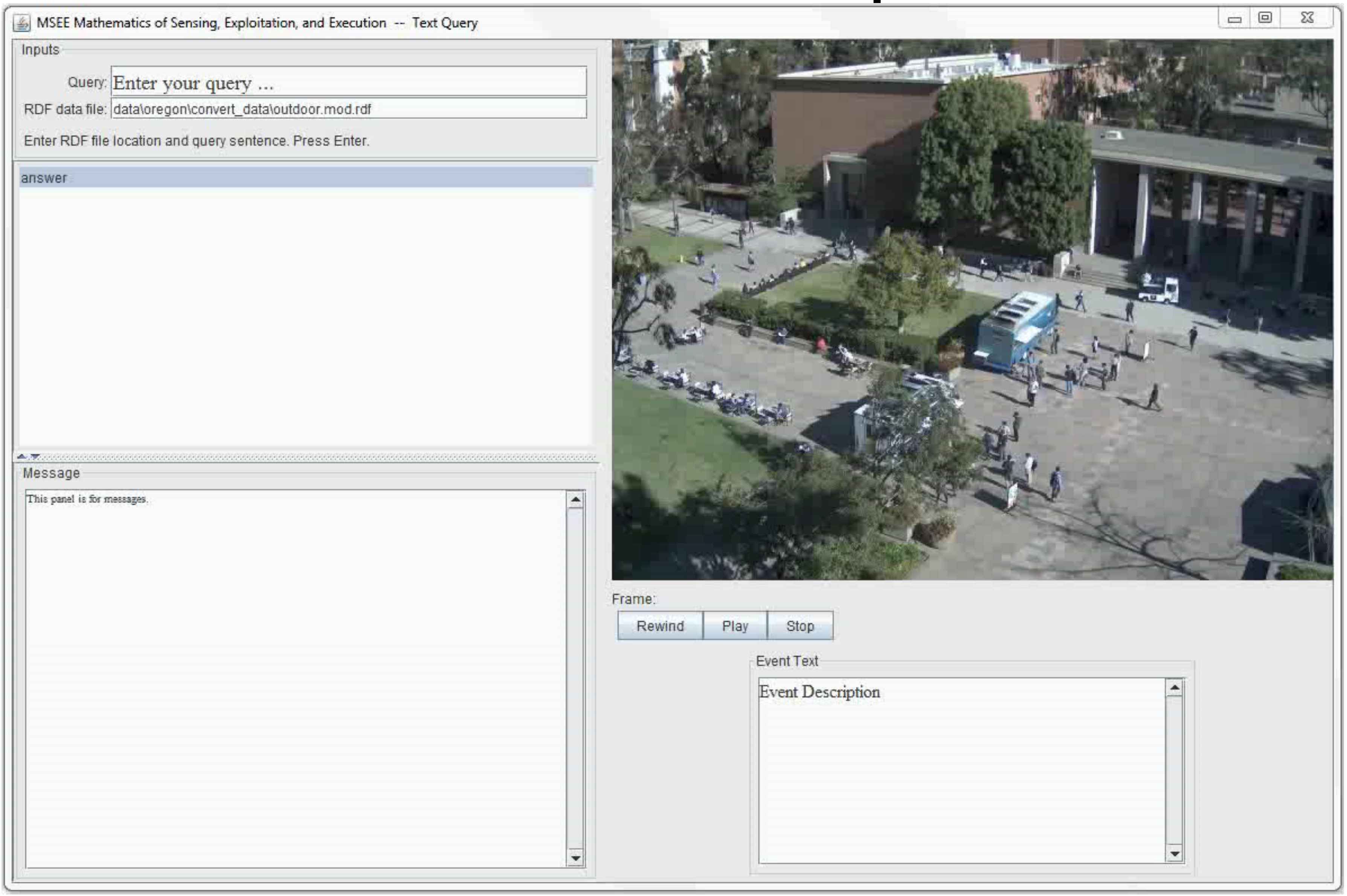
- 6 Group Activities:
  - Walking together, Queuing, Campus tour, ...
- 10 Individual Actions:
  - Walking, Sitting, Riding a bike, …
- 17 Objects:
  - Food truck, Vending machine, Bike, Backpack, ...

New Dataset

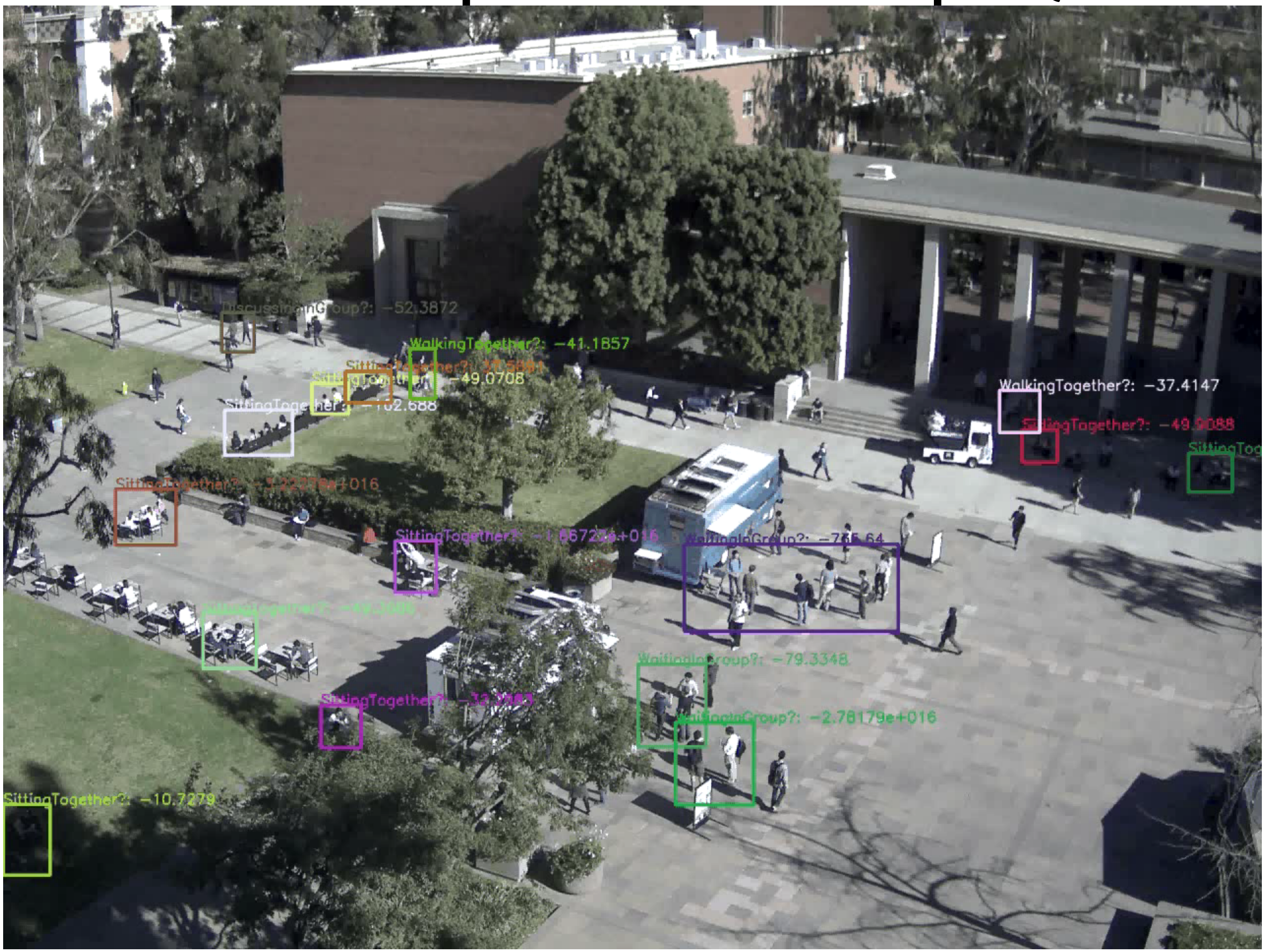
Available Datasets

Dataset	Resolution	Object	Individual	Group	Background	Instances	Poses
Our Dataset	2560x1920	Yes	Yes	Yes	Cluttered	7+	Yes
VIRAT Ground	1920x1080	Yes	Yes	No	Cluttered	4-	No
CompCollective	1440x960	No	Yes	Yes	Cluttered	4	Yes
Collective	720x480	No	Yes	Yes	Cluttered	1	Yes
UT-Interaction	720x480	No	No	Yes	Clear	2	No
KTH	160x120	No	Yes	No	Clear	1	No
Weizmann	180x144	No	Yes	No	Clear	1	No
UCF Youtube	240x500	No	Yes	No	Cluttered	1	No
UCF 50	240x500	No	Yes	No	Cluttered	1	No
Olympic Sports	360x450	No	Yes	No	Cluttered	1	No

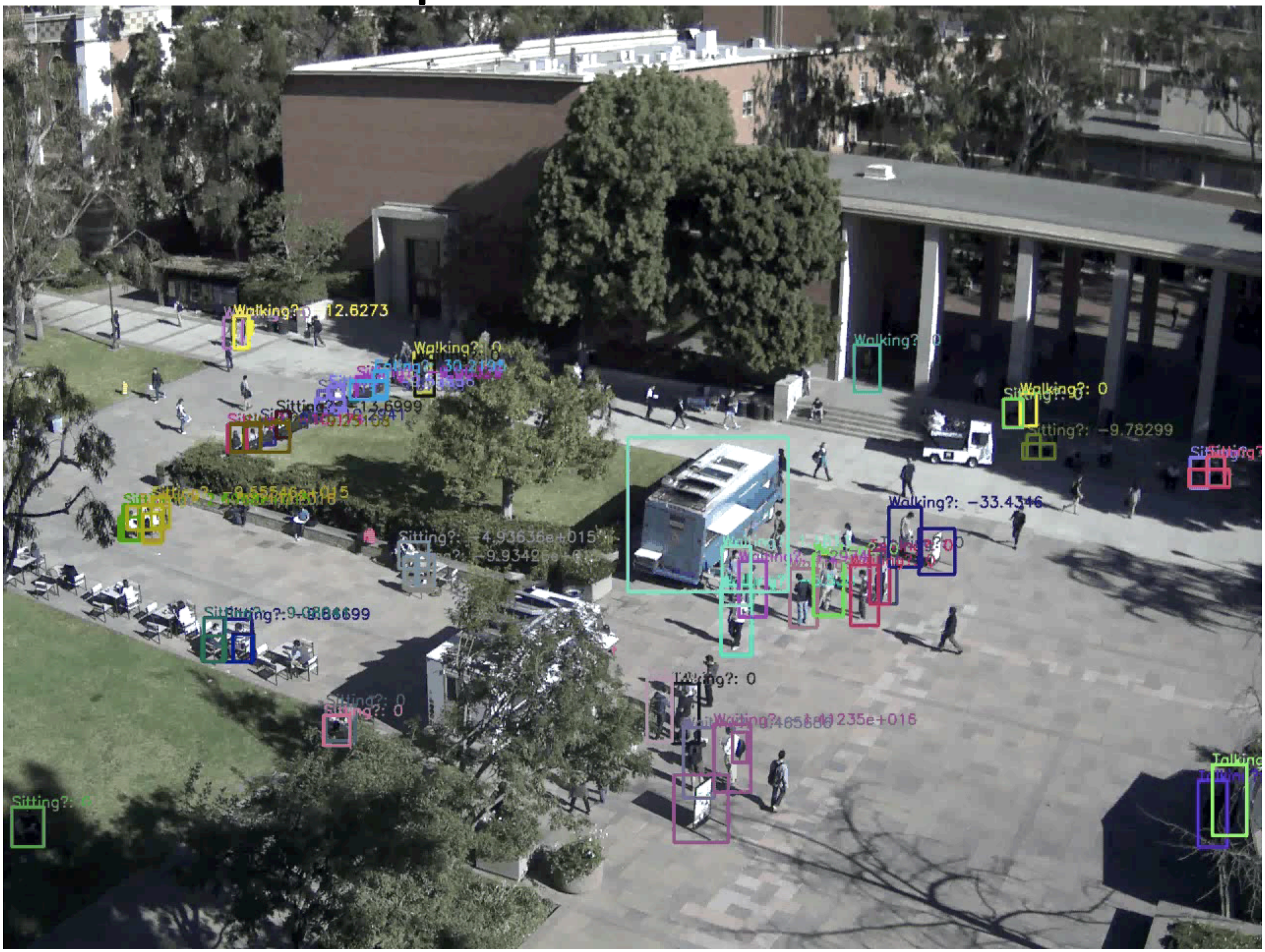
Queries Example



All Parse Graphs for Group Queries



All Parse Graphs for Individual Queries



Results – Courtyard Dataset

	Query about group activities						
E^2 **strategy**	Standing-in-line	Guided-tour	Discussing	Sitting	Walking	Waiting	**Time**
$\mathcal{B} = 1$, Precision	62.2%	63.7%	68.1%	65.3%	69.4%	61.2%	5s
$\mathcal{B} = 1$, FP	7.2%	2.3%	9.8%	12.6%	8.1%	10.4%	5s
$\mathcal{B} = 15$, Precision	65.4%	66.1%	69.0%	68.7%	70.3%	66.5%	75s
$\mathcal{B} = 15$ FP	10.1%	4.7%	11.1%	11.1%	8.7%	10.9%	75s
$\mathcal{B} =$, Precision	68.0%	70.2%	75.1%	71.4%	78.6%	72.6%	230s
$\mathcal{B} = \infty$, FP	13.6%	10.3%	17.1%	13.7%	10.1%	12.2%	230s

	Query about primitive actions										
E^2 **strategy**	Walk	Wait	Talk	Drive Car	Ride S-board	Ride Scooter	Ride Bike	Read	Eat	Sit	**Time**
$\mathcal{B} = 1$, Precision	63.3%	61.2%	58.4%	65.8%	63.5%	60.1%	56.8%	55.3%	60.9%	54.3%	10s
$\mathcal{B} = 1$, FP	12.1%	16.2%	11.4%	3.4%	10.2%	11.6%	6.2%	8.2%	2.2%	5.3%	10s
$\mathcal{B} = 15$, Precision	67.6%	63.4%	62.3%	67.2%	67.1%	65.9%	59.3%	61.2%	66.3%	59.2%	150s
$\mathcal{B} = 15$, FP	14.2%	17.1%	15.1%	7.1%	13.8%	13.2%	9.3%	10.3%	4.3%	7.1%	150s
$\mathcal{B} = \infty$, Precision	69.1%	67.7%	69.6%	70.2%	71.3%	68.4%	61.4%	67.3%	71.3%	64.2%	330s
$\mathcal{B} = \infty$, FP	18.7%	20.2%	17.9%	9.7%	17.1%	16.3%	12.3%	12.1%	7.7%	9.0%	330s

Conclusion

- New problem of Multi-scale activity recognition.

Conclusion

- New problem of Multi-scale activity recognition.
- Efficient formulation using And-Or graphs

Conclusion

- New problem of Multi-scale activity recognition.
- Efficient formulation using And-Or graphs
- Cost-sensitive inference using RL

Conclusion

- New problem of Multi-scale activity recognition.
- Efficient formulation using And-Or graphs
- Cost-sensitive inference using RL
- New dataset

ACKNOWLEDGMENTS

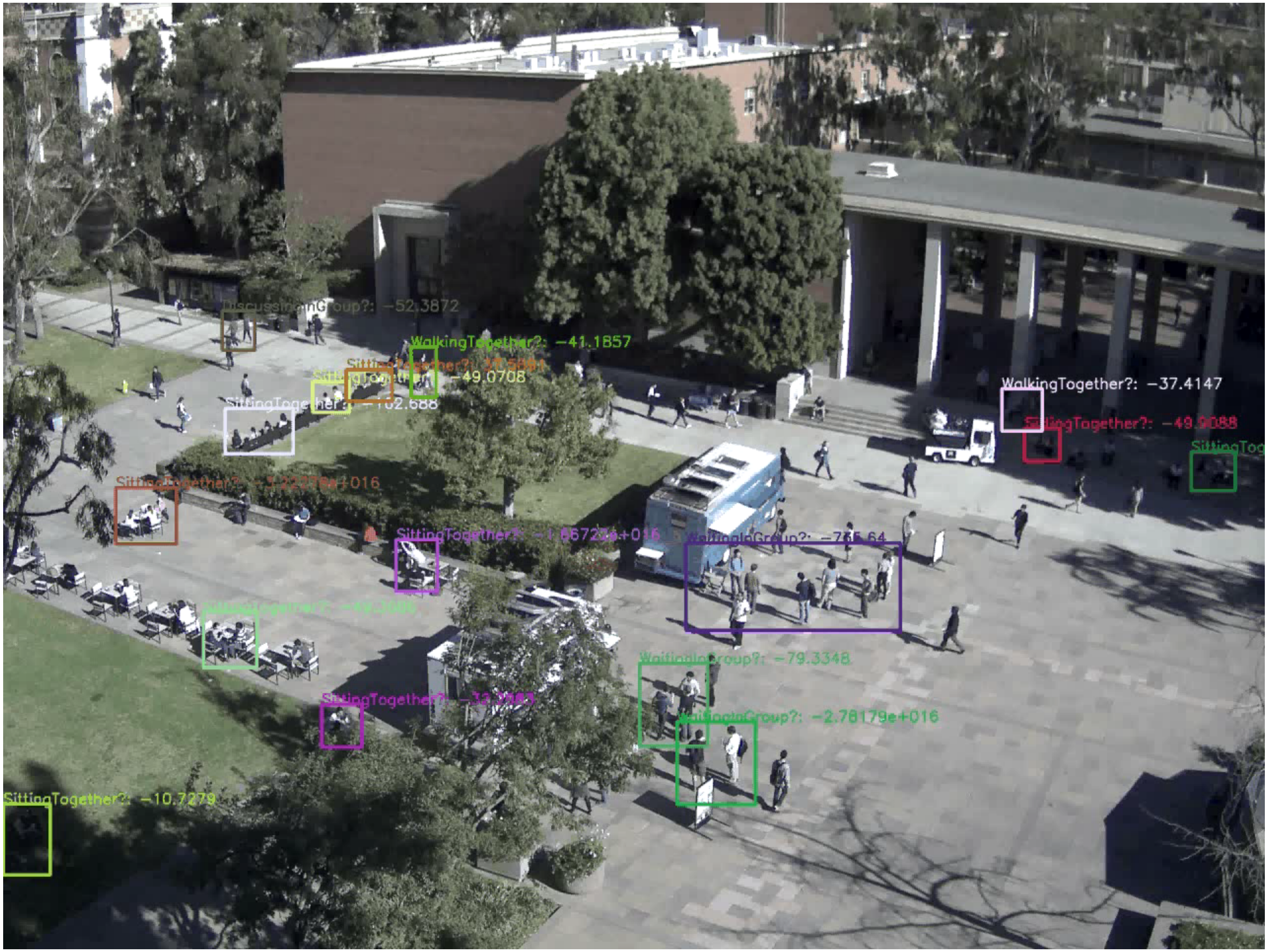
DARPA

MSEE FA 8650-11-1-7149

ONR
Office of Naval Research

MURI N00014-10-1-0933

Questions


WalkingTogether?: −41.1857
WalkingTogether?: −37.4147
SittingTogether?: −49.9088
WaitingInGroup?: −79.3348
WaitingInGroup?: −2.78179e+016
SittingTogether?: −10.7279