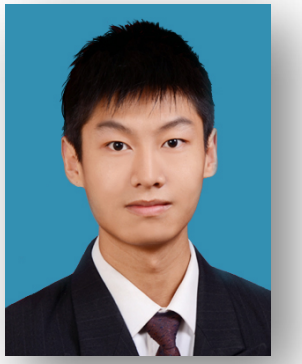
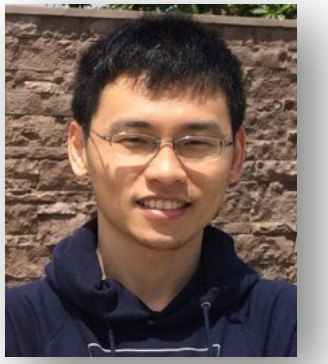
中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Baidu 百度

ICCV 2019
Seoul, Korea

Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition





Wenhao Wu *Dongliang He* *Xiao Tan* *Shifeng Chen* *Shilei Wen*

Motivation

How to Improve the Accuracy of Recognition for Untrimmed Videos?

Motivation

How to Improve the Accuracy of Recognition for Untrimmed Videos?

Improving basic recognition model & sampling method

Motivation

ICCV 2019
Seoul, Korea

How to Improve the Accuracy of Recognition for Untrimmed Videos?

Improving basic recognition model & sampling method

Direction 1: Basic recognition model

End to end:
3D CNN (I3D, P3D, S3D,…)
2D CNN (TSN, TRN, TSM, …)

Two stage (Local feature integration):
LSTM, Attention Cluster, …

Motivation

ICCV 2019
Seoul, Korea

How to Improve the Accuracy of Recognition for Untrimmed Videos?

Improving basic recognition model & sampling method

Direction 1: Basic recognition model

End to end:
3D CNN (I3D, P3D, S3D,…)
2D CNN (TSN, TRN, TSM, …)

Two stage (Local feature integration):
LSTM, Attention Cluster, …

Direction 2: Sampling method

Hand-crafted sampler:
Uniformly sampling,
Dense sampling

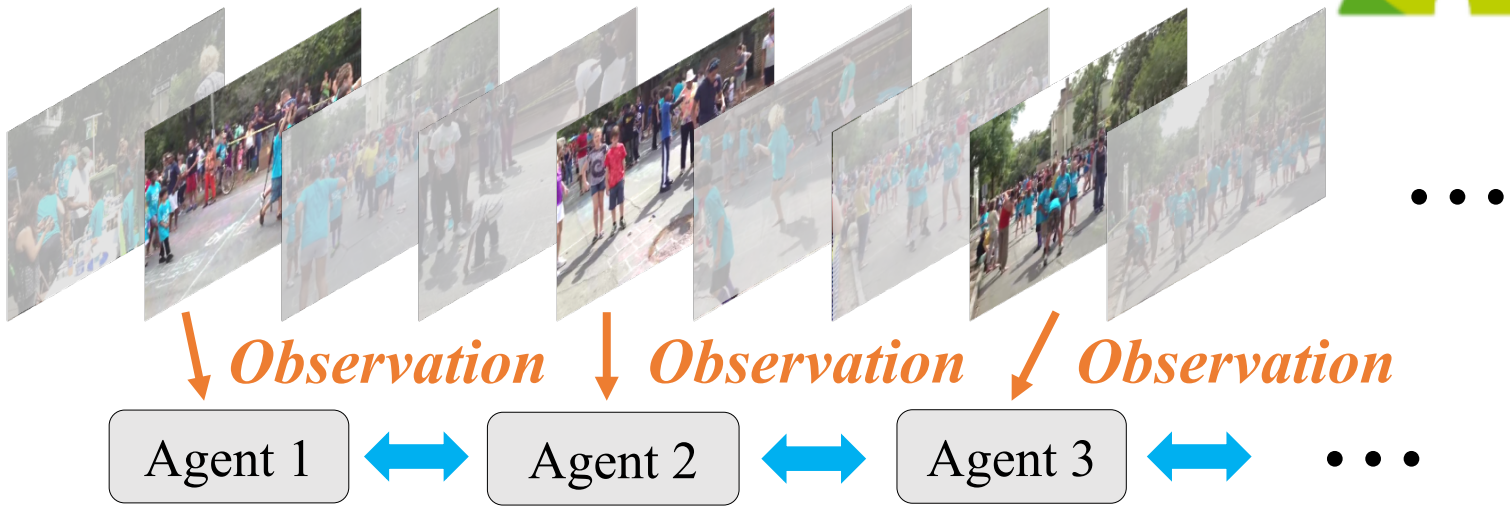
In this paper, we focus on learning-based sampler.

Overview

Untrimmed video

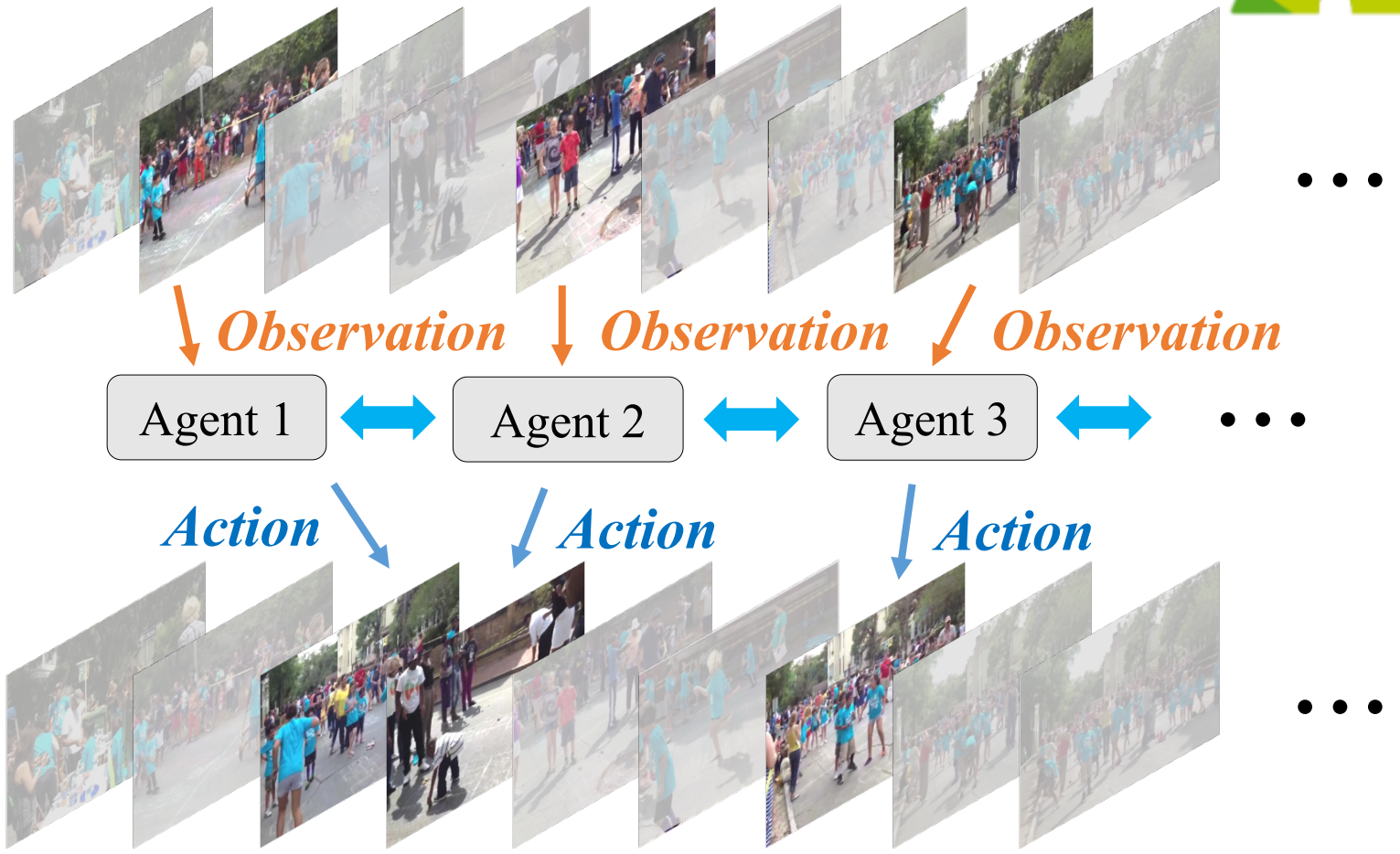
. . .

Overview

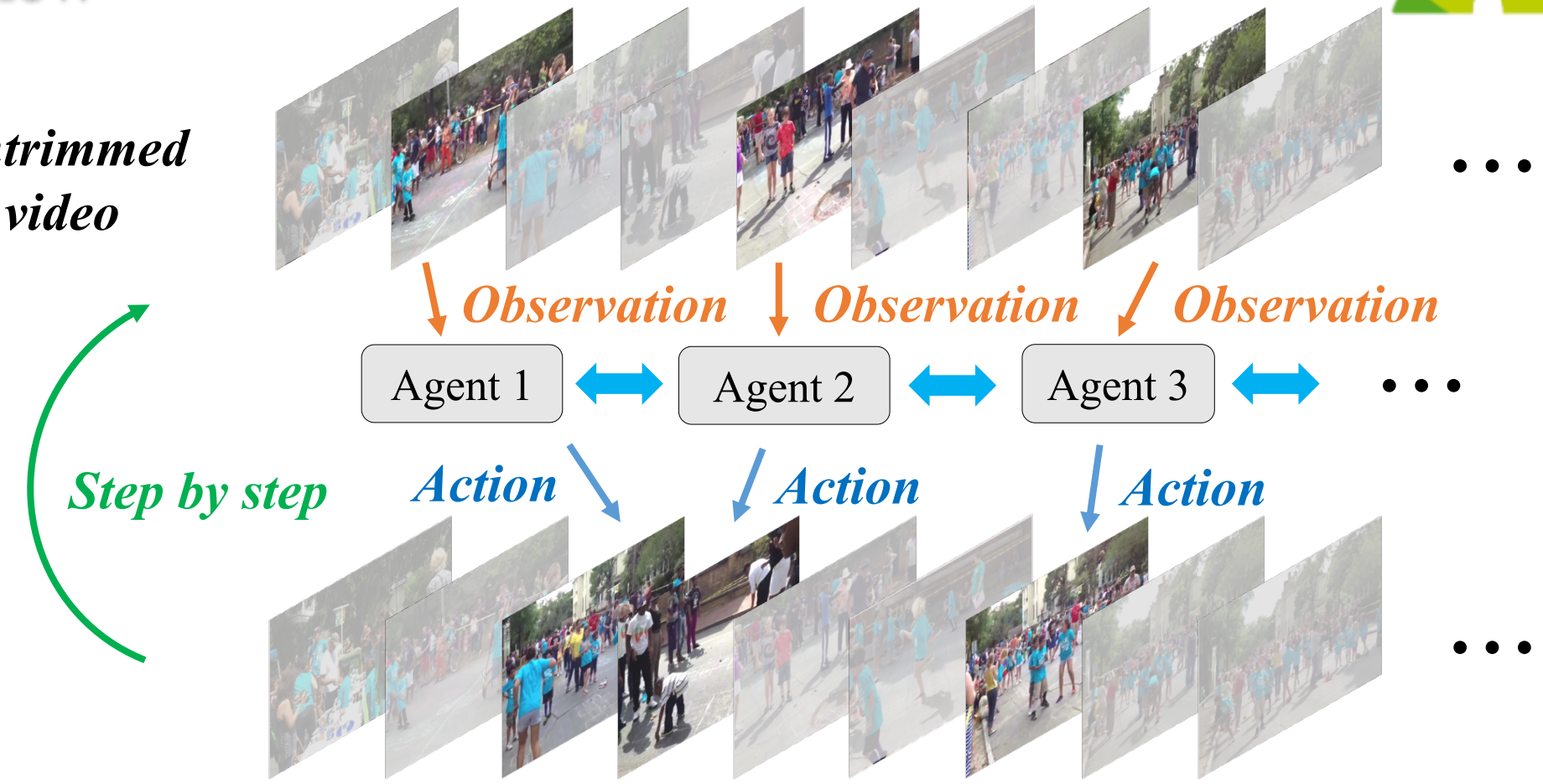

Untrimmed video
Observation
Observation
Observation
Agent 1
Agent 2
Agent 3

Overview

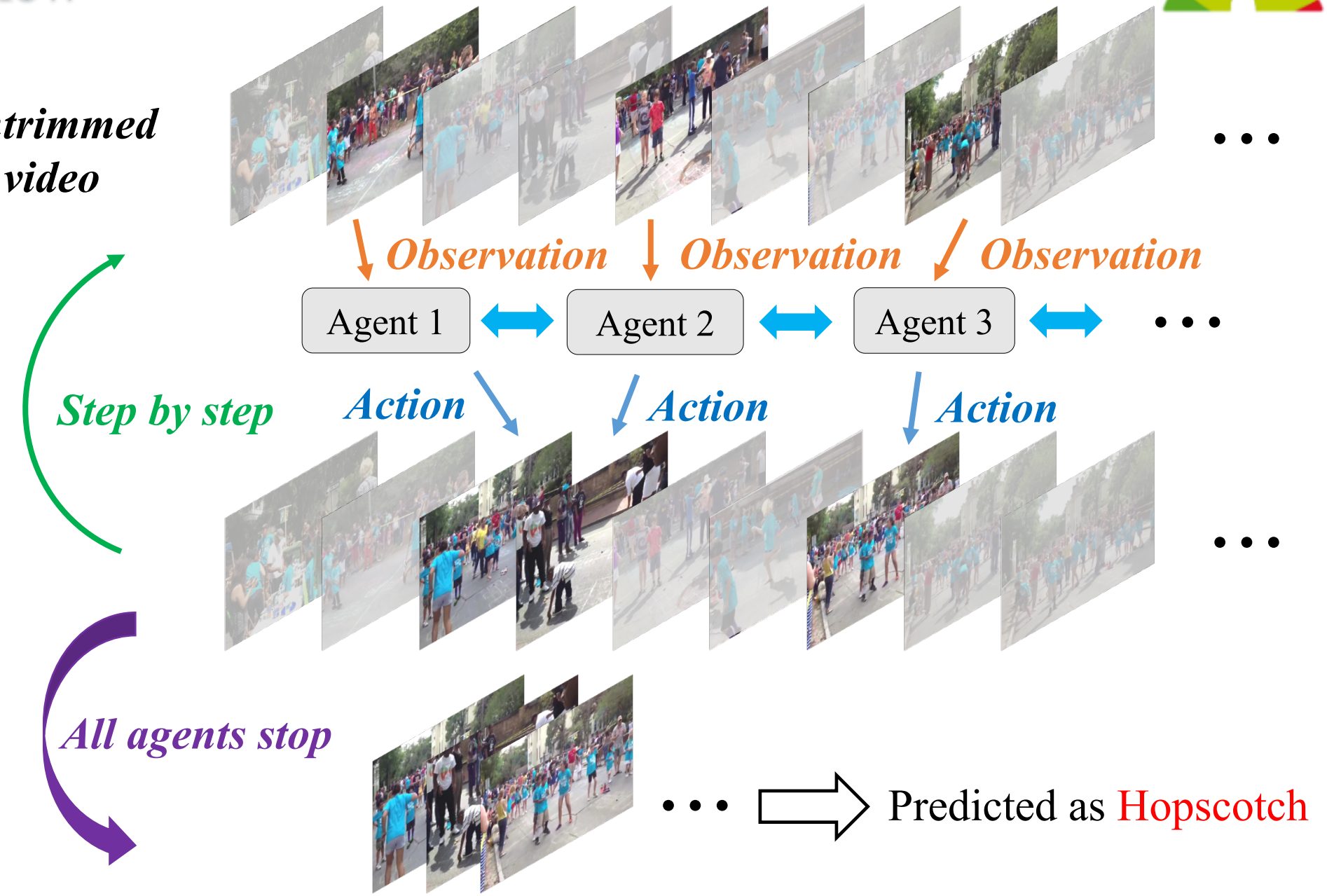
Untrimmed video



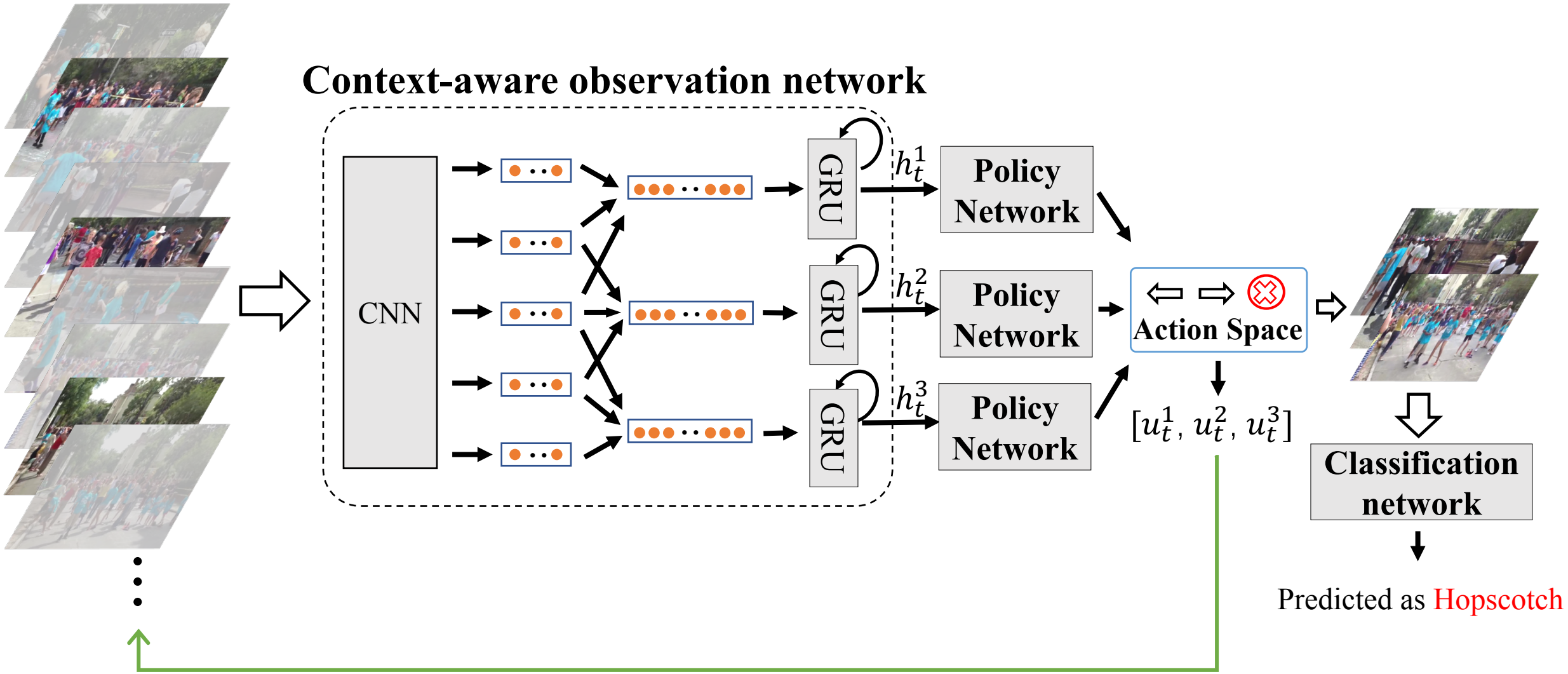
Overview


Untrimmed video
Observation
Observation
Observation
Agent 1
Agent 2
Agent 3
Step by step
Action
Action
Action

Overview


Untrimmed video
Observation
Observation
Observation
Agent 1
Agent 2
Agent 3
Step by step
Action
Action
Action
All agents stop
Predicted as Hopscotch

Architecture



Objectives

Hybrid Loss: $$Loss = \mathcal{L}_{Cls}(\theta_p) + \lambda_2 \mathcal{L}_{MARL}(\theta_\pi)$$

Objectives

Hybrid Loss: $$Loss = \mathcal{L}_{Cls}(\theta_p) + \lambda_2 \mathcal{L}_{MARL}(\theta_\pi)$$

MARL Objective: $$\mathcal{L}_{MARL}(\theta_\pi) = \mathcal{L}_J(\theta_\pi) + \lambda_1 \, \mathcal{L}_H(\theta_\pi)$$

Reward function: $$r_t^a = p_{t,gt}^a - p_{t-1,gt}^a$$

Policy gradient: $$\mathcal{L}_J(\theta_\pi) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{a=1}^{N} \sum_{t=0}^{T_{stop}} \log \pi(u_{t,k}^a | s_{t,k}^a; \theta_\pi) \, R_t^a$$

Maximum entropy: $$\mathcal{L}_H(\theta_\pi) = -\sum_{a=1}^{N} \sum_{t=0}^{T_{stop}} \sum_{u \in \mathcal{A}} \pi(u_t^a | s_t^a; \theta_\pi) \log \pi(u_t^a | s_t^a; \theta_\pi)$$

Objectives

Hybrid Loss: $$Loss = \mathcal{L}_{Cls}(\theta_p) + \lambda_2 \mathcal{L}_{MARL}(\theta_\pi)$$

MARL Objective: $$\mathcal{L}_{MARL}(\theta_\pi) = \mathcal{L}_J(\theta_\pi) + \lambda_1 \, \mathcal{L}_H(\theta_\pi)$$

Reward function: $$r_t^a = p_{t,gt}^a - p_{t-1,gt}^a$$

Policy gradient: $$\mathcal{L}_J(\theta_\pi) = -\frac{1}{K} \sum_{k=1}^{K} \sum_{a=1}^{N} \sum_{t=0}^{T_{stop}} \log \pi(u_{t,k}^a | s_{t,k}^a; \theta_\pi) \, R_t^a$$

Maximum entropy: $$\mathcal{L}_H(\theta_\pi) = -\sum_{a=1}^{N} \sum_{t=0}^{T_{stop}} \sum_{u \in \mathcal{A}} \pi(u_t^a | s_t^a; \theta_\pi) \log \pi(u_t^a | s_t^a; \theta_\pi)$$

Classification Objective: $$\mathcal{L}_{Cls}(\theta_p) = -\sum_{c=1}^{C} y_c \log p_c$$

Result: Experiments

- *Performance comparison of different backbones on the ActivityNet dataset.*

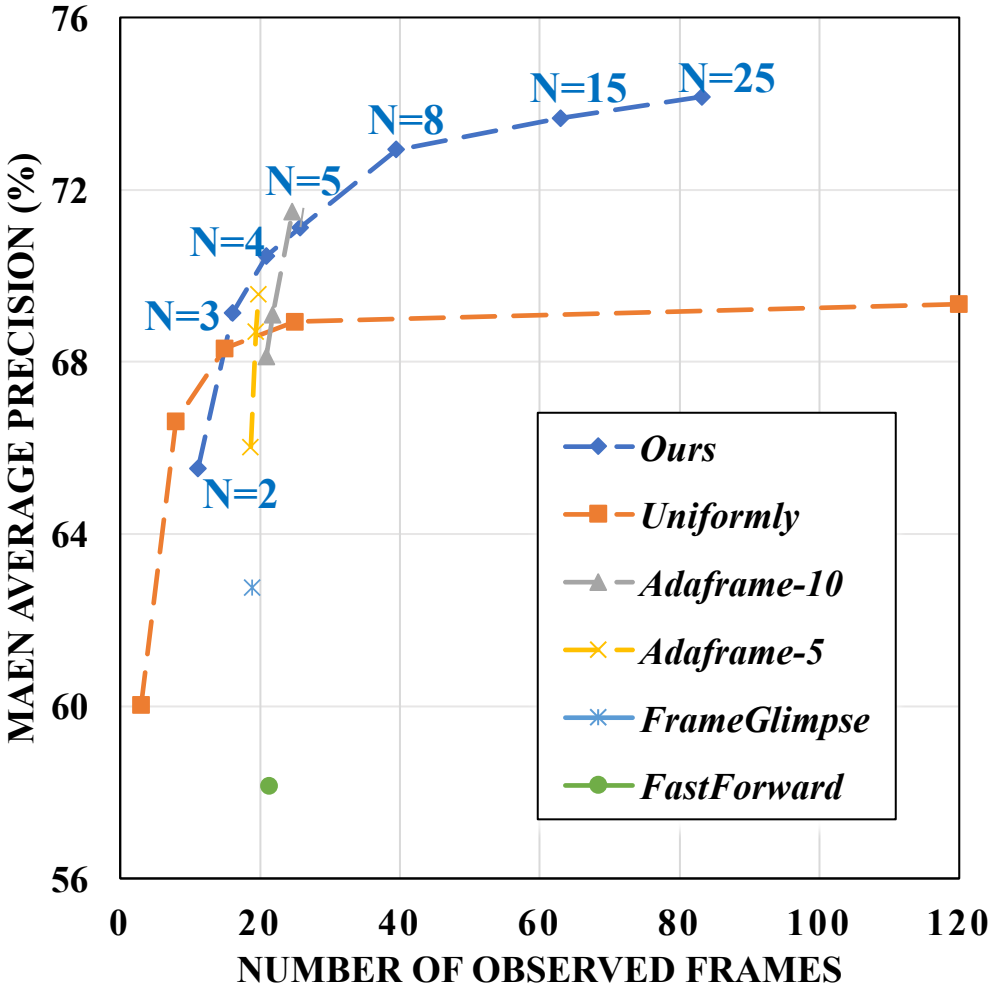
Architecture	ActivityNet v1.2				ActivityNet v1.3			
	R25	U25	All	Ours	R25	U25	All	Ours
C3D	62.06	62.89	63.00	**64.13**	59.73	60.68	60.83	**62.00**
BN-Inception	78.76	80.02	80.50	**81.99**	75.08	76.48	77.33	**78.32**
ResNet-101	80.73	81.94	82.26	**83.76**	78.69	79.96	80.64	**81.54**
Inception-V3	81.90	82.66	83.25	**85.01**	79.27	80.33	80.86	**82.34**
ResNet-152	82.72	83.71	84.07	**85.70**	80.69	82.08	82.53	**83.81**

R25: Randomly sampling 25 frames **U25**: Uniformly sampling 25 frames

All: Using all frames of video (1FPS) **Ours**: Using our method to sample 25 frames

Result: Experiments

mAP vs. number of observed frames.



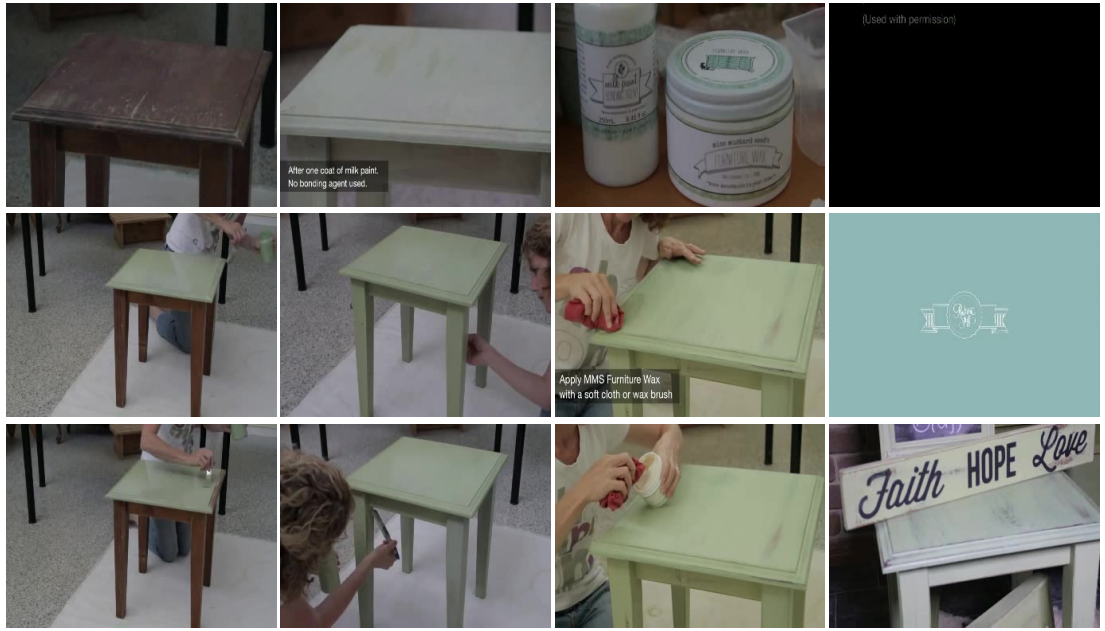
Method	Backbone	Pre-trained	top-1	mAP
IDT [43]	-	ImageNet	64.70	68.69
C3D [33]	-	Sports1M	65.80	67.68
TSN [45]*	BN-Inception	ImageNet	72.97	76.56
P3D [33]	ResNet-152	ImageNet	75.12	78.86
RRA [55]	ResNet-152	ImageNet	78.81	83.42
Ours	ResNet-152	ImageNet	**79.82**	**83.81**
TSN [45]*	BN-Inception	Kinetics	78.98	81.80
Ours	BN-Inception	Kinetics	80.22	83.52
C16	Ensemble	-	-	**90.9**
Ours	SEResNeXt152	Kinetics	**85.72**	90.05

Method	YouTube Birds	YouTube Cars
BN-Inception*	60.13	61.96
I3D [3]*	40.68	40.92
TSN [45]*	72.361	74.340
RRA [55]*	73.205	77.625
U25/All/Ours	76.56/76.77/**79.01**	76.49/76.99/**79.77**

Result: Visualization

ICCV 2019
Seoul, Korea

Painting furniture



Playing drums

The first row: *Uniformly sampling*
The second row: *Our method w/o context-aware observation*
The last row: *Our method*

Result: Ablation Study

- *Different number of agents*

	N=5	N=15	N=25	N=120	All
ResNet-152	80.19	82.99	83.81	83.72	82.53

Result: Ablation Study

- *Different number of agents*

	N=5	N=15	N=25	N=120	All
ResNet-152	80.19	82.99	83.81	83.72	82.53

- *Different context range*

	U25	All	M=0	M=1	M=2	M=4
mAP	82.08	82.53	82.99	83.81	83.80	83.72

Result: Ablation Study

- *Different number of agents*

	N=5	N=15	N=25	N=120	All
ResNet-152	80.19	82.99	83.81	83.72	82.53

- *Different context range*

	U25	All	M=0	M=1	M=2	M=4
mAP	82.08	82.53	82.99	83.81	83.80	83.72

- *Policy network transferring*

U25	All	MARL	Birds' Sampler	Cars' Sampler	ANet1.2' Sampler	R101' Sampler
82.08	82.53	83.81	82.70	82.66	83.41	83.43

中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Baidu 百度

ICCV 2019
Seoul, Korea

Thank you!

Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition

Welcome to Our Poster #26.