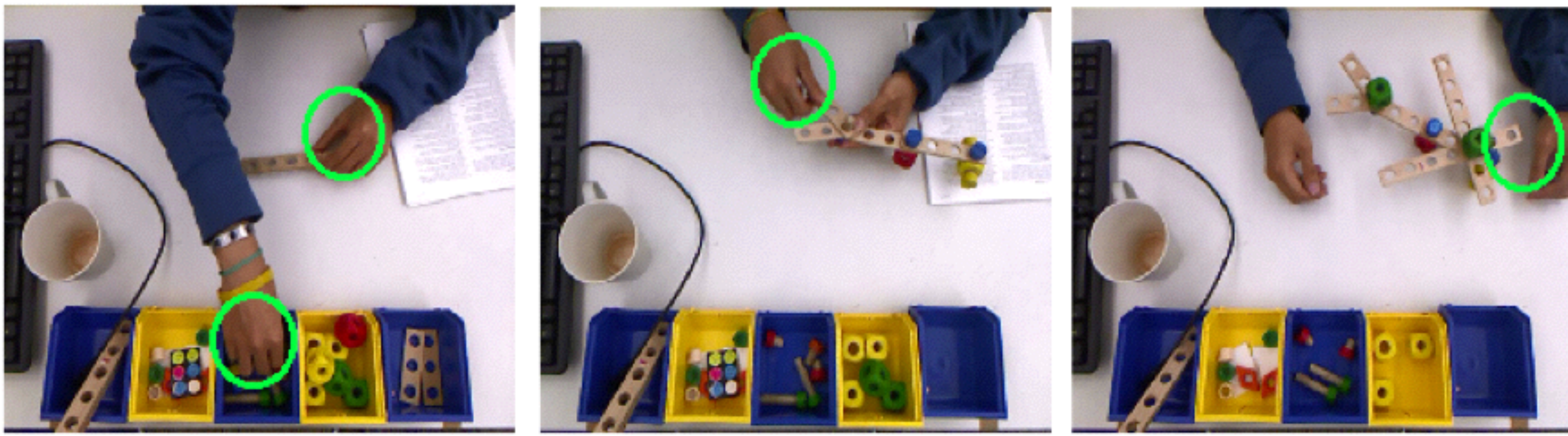
From Stochastic Grammar to Bayes Network: Probabilistic Parsing of Complex Activity



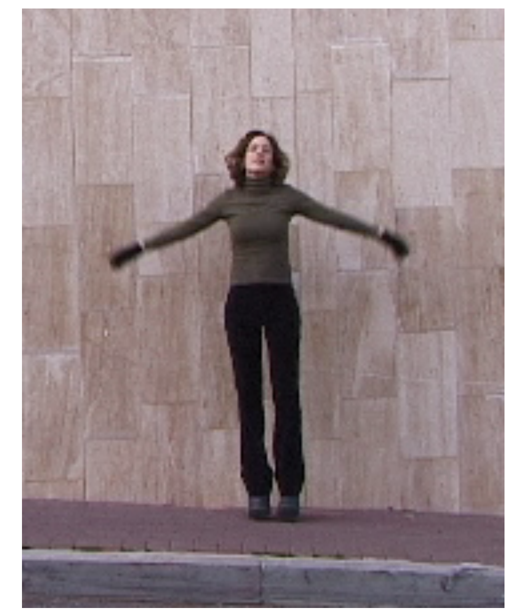
Nam Vo & Aaron Bobick (CVPR2014)

Colin Lea - Summer 2014

Data

Proposed dataset:
Toy dataset [+Human+Robot video]

Weizmann dataset
10 simple actions (concatenated)
Run, walk, skip, jump, …



GTech Egocentric Activities dataset
7 activities, 4 humans
Making sandwiches, coffee, …
All kitchen activities

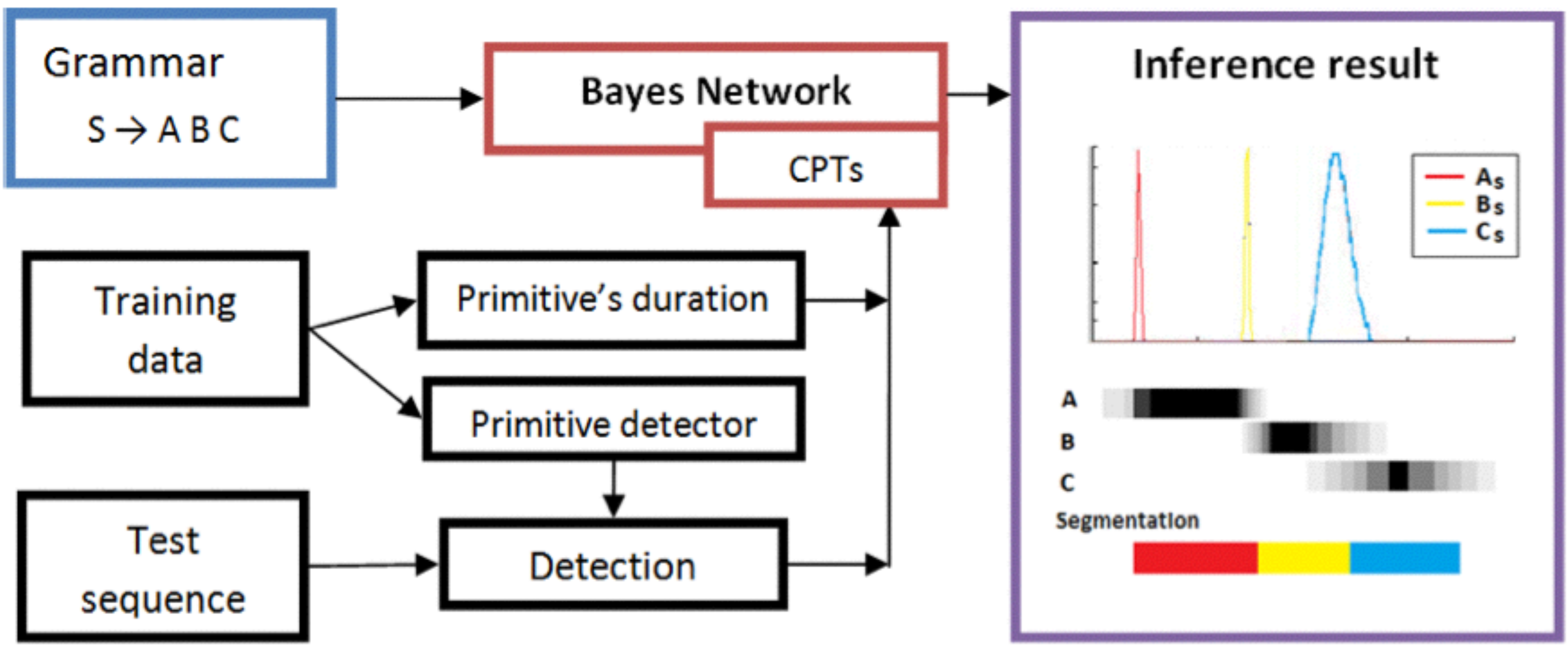
Overview

Goals:
Recognize and *predict* actions *(start, stop, type)*
in a complex activity.

Approach:
- Temporal Hierarchy w/ (Context Free Grammar)
- Generate Bayes Net
- Black box features

Contributions:
- Inference on And/Or Graph
- Start/End times are random variables
- Code+data: http://www.cc.gatech.edu/~nvo9/sin/

Overview



Grammar
S → A B C
Bayes Network
CPTs
Inference result
As
Bs
Cs
A
B
C
Segmentation
Training data
Primitive's duration
Primitive detector
Test sequence
Detection

Aside: Context Free Grammar

G = (N, T, R, S)

N is a non-terminal

T is a terminal

R is a Rule

S is the starting symbol

Example:

V=A, B, C

T =x, y, z

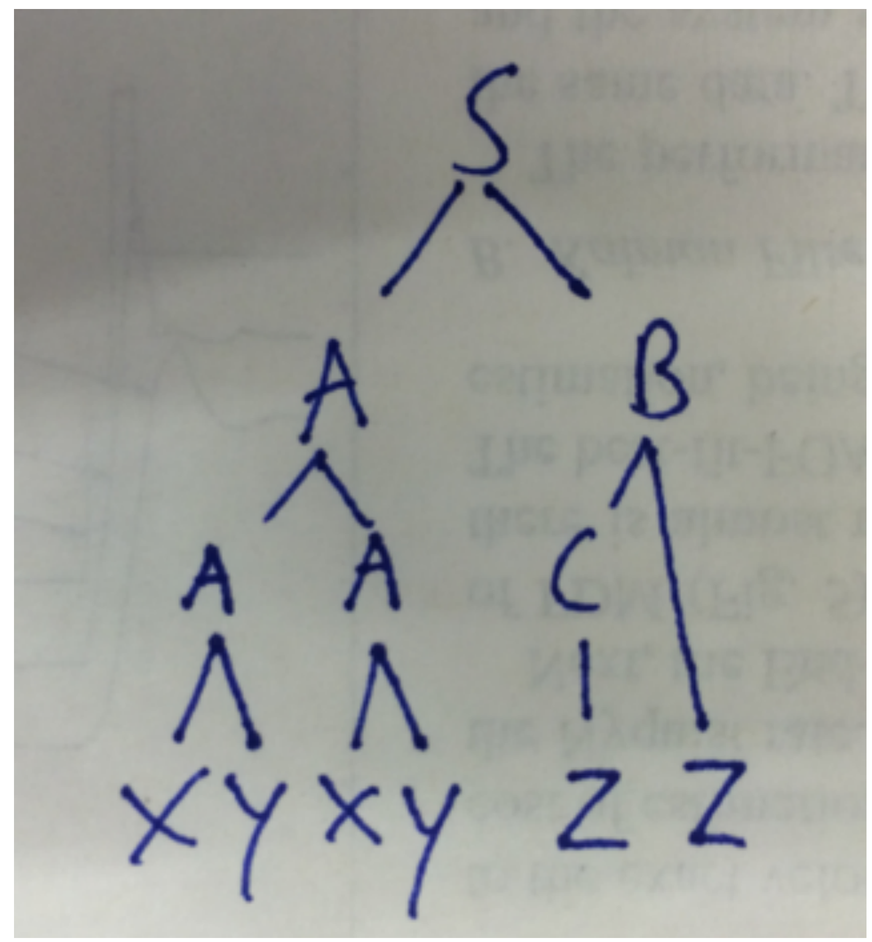
<u>Rules</u>:

S -> AB | C

A -> AA | xy

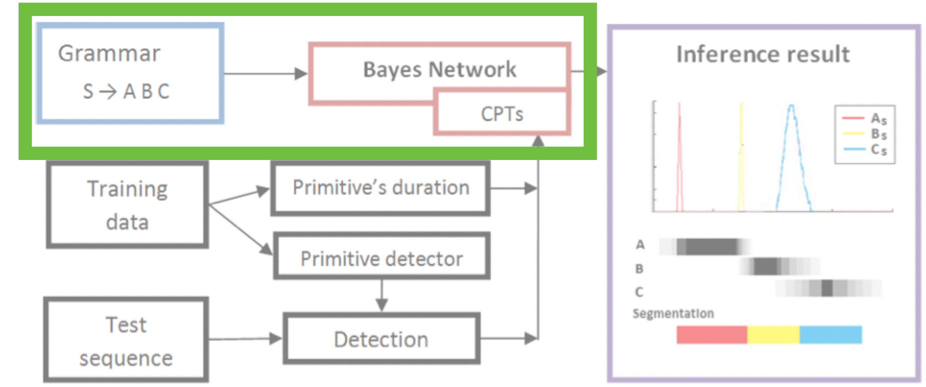
B -> yz | Cz

C -> z



For stochastic/prob/weighted grammar add weights to rules

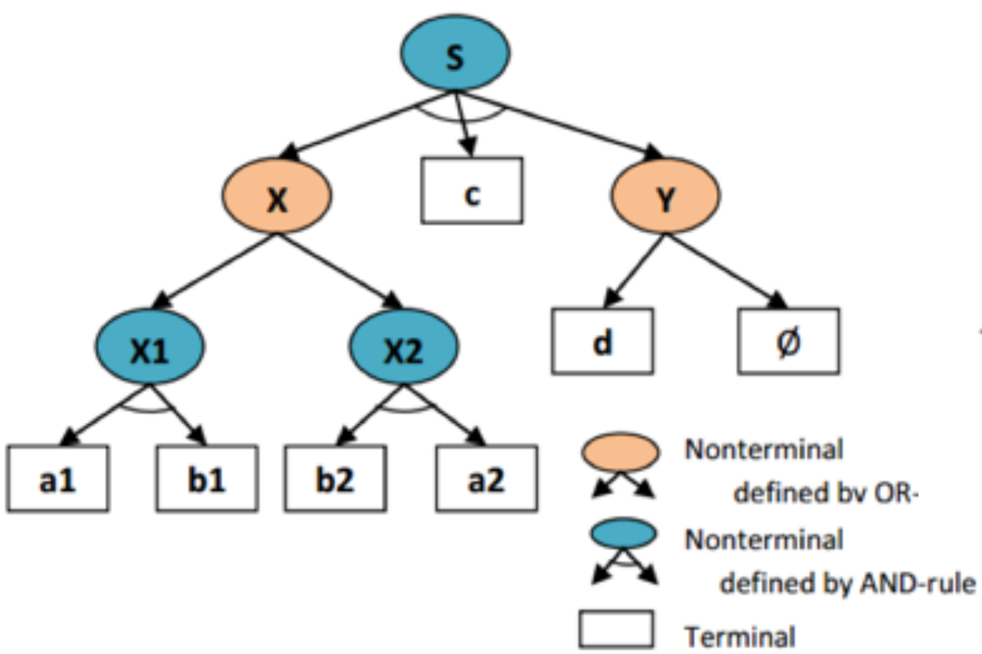
CFG -> Bayes Net

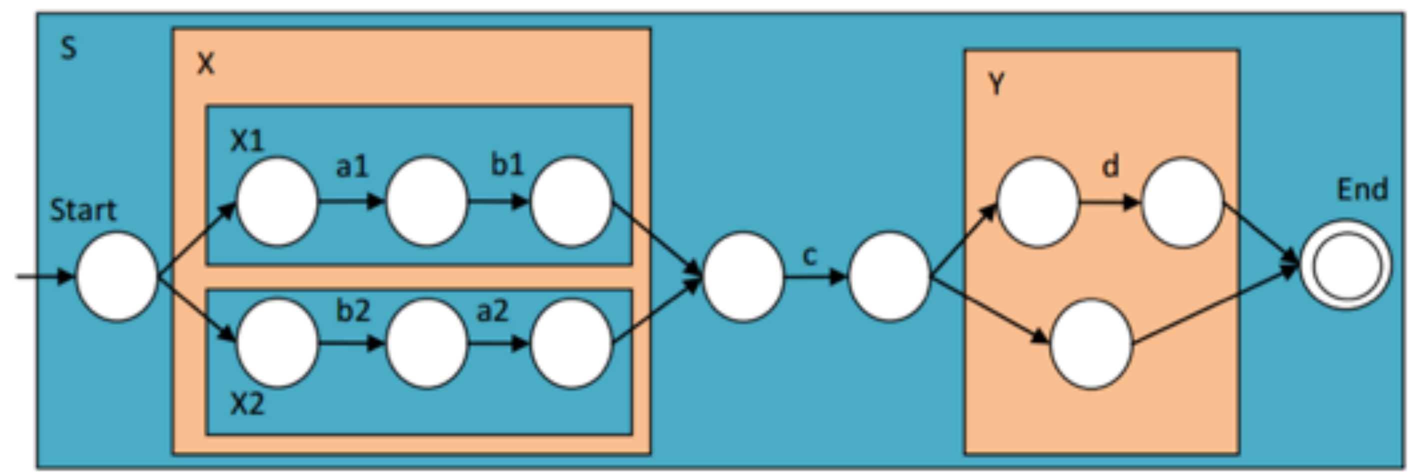


Three rules:

1. Rules must be "And" or "Or" and probabilistic
2. Symbols can only appear once on RHS of rule
3. Symbols cannot be of arbitrary length
   (e.g. S -> SA | A not allowed)

S → X c Y
X → X1 | X2
X1 → a1 b1
X2 → b2 a2
Y → d | Ø





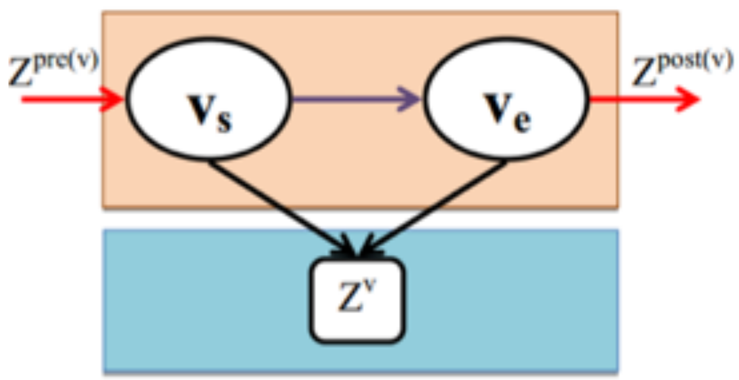
CFG And-Or Graph Bayes Net

Bayes Net

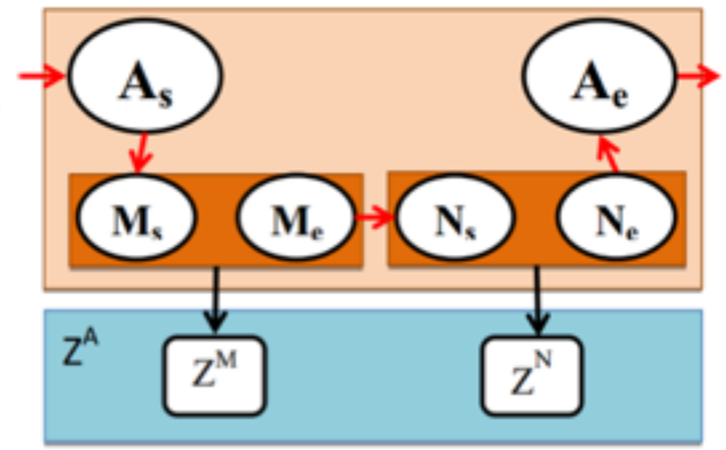
Node types: (**Primitive**: v, **And**: A->MN, **Or**: A->M | N)

Z = label s = start e = end

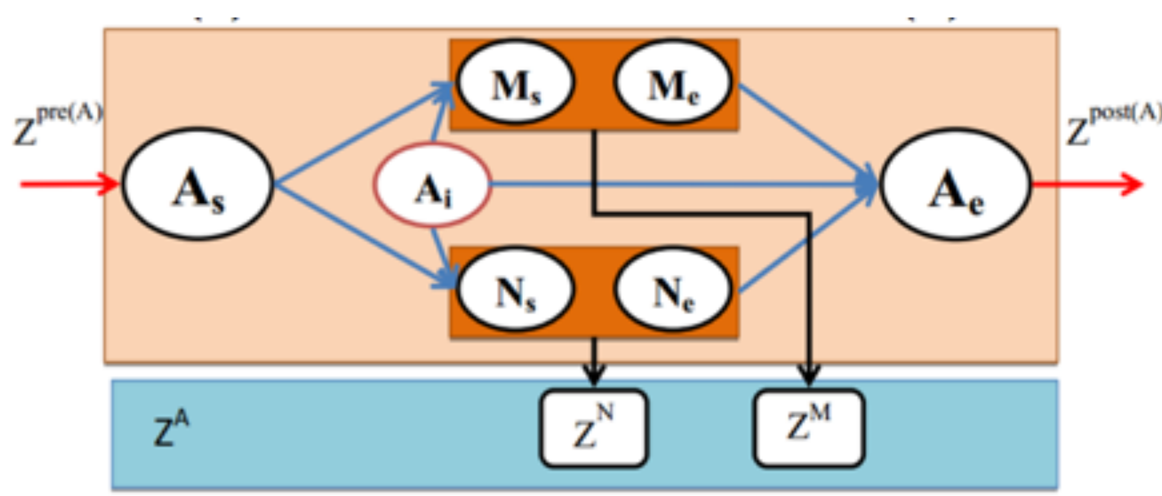
Primitive



And



Or



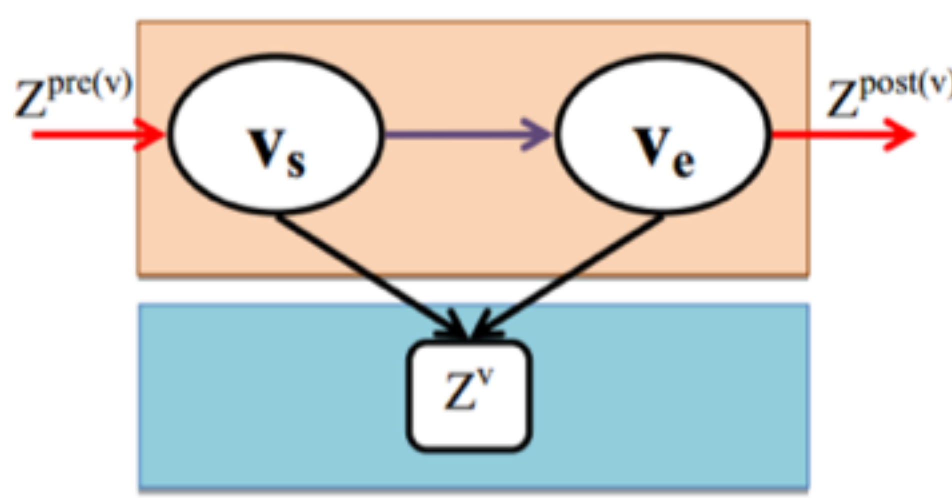
And: Ns = Me

Bayes Net

Node types: (**Primitive**: v, **And**: A->MN, **Or**: A->M | N)

Conditional probability: P(ve | vs) = N(ve-vs; mu, sigma)

Likelihood: P(Zv | vs, ve) = hv * Fv[vs, ve]
    Blackbox (e.g. probabilistic output of SVM)
    They use hand position, indicators from environment



Inference

Inputs:

$P(v_e \mid v_s)\ \forall\ v$

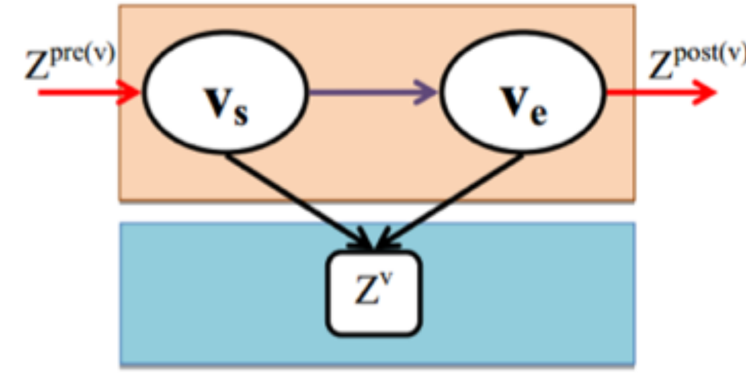
$P(Z^v \mid v_s, v_e)\ \forall\ v$

Priors: $P(\exists M|\exists A)$; $P(\exists S)=1$; $P(S_s|\exists S)$; $P(Z^e|S_e, \exists S)$

1. Forward Step
2. Backward Step
3. Compute posteriors
4. Compute happening probability

Complexity: $O(KT^2)$

Inference (forward)



$P(A_s, Z^{pre(A)} \mid \exists A)$ and $P(A_e, Z^{pre(A),A} \mid \exists A)$ for every A.

1. For **primitives** v. Given $P(v_s, Z^{pre(v)} \mid \exists v)$:

P(start,end,actions_{prev} | v) = P(start, actions_{prev} | v) P(duration) P(Likelihood)

$$P(v_s, v_e, Z^{pre(v),v} \mid \exists v) = P(v_s, Z^{pre(v)} \mid \exists v) P(v_e \mid v_s) P(Z^v \mid v_s, v_e)$$

Marginalize over starting positions

$$P(v_e, Z^{pre(v),v} \mid \exists v) = \sum_{t=1}^{T} P(v_s = t, v_e, Z^{pre(v),v} \mid \exists v)$$

2. For **Ands** A->MN

$$P(M_s = t, Z^{pre(M)} \mid \exists M) = P(A_s = t, Z^{pre(A)} \mid \exists A)$$
$$P(N_s = t, Z^{pre(N)} \mid \exists N) = P(M_e = t, Z^{pre(M)}, M \mid \exists M)$$
$$P(A_e = t, Z^{pre(A),A} \mid \exists A) = P(N_e = t, Z^{pre(N),N} \mid \exists N)$$

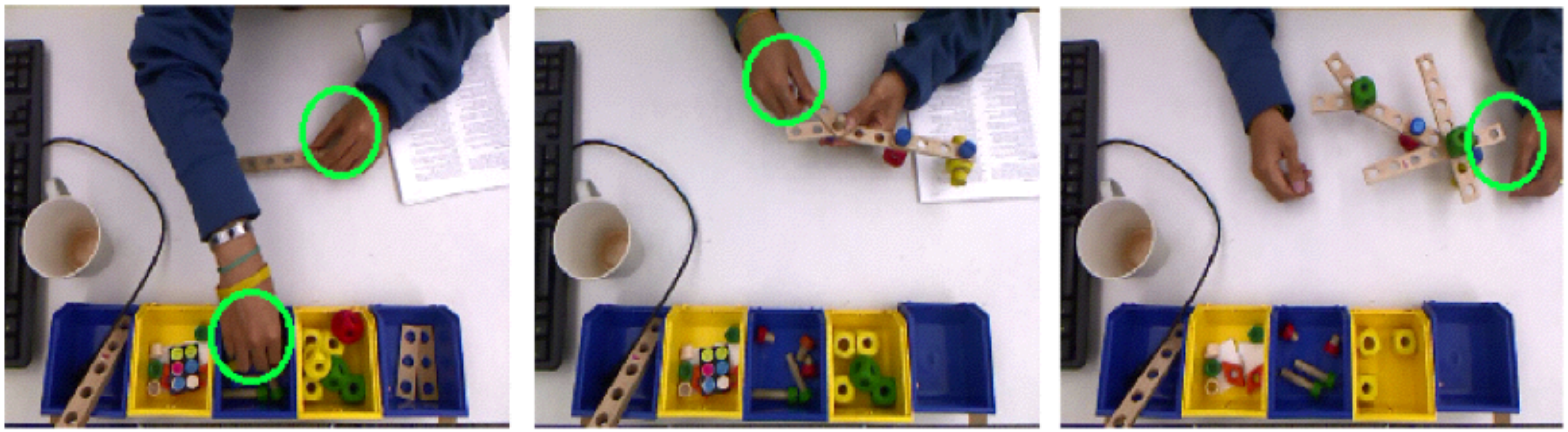
3. For **ORs** A-> M | N

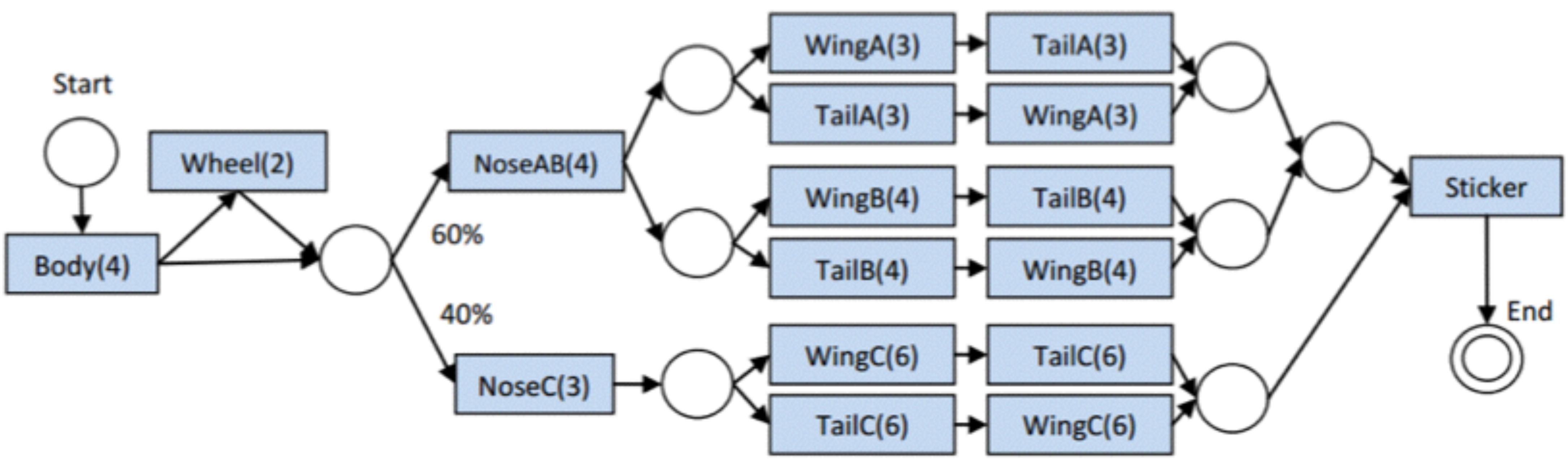
$$P(M_s = t, Z^{pre(M)} \mid \exists M) = P(A_s = t, Z^{pre(A)} \mid \exists A)$$
$$P(N_s = t, Z^{pre(N)} \mid \exists N) = P(A_s = t, Z^{pre(A)} \mid \exists A)$$

$P(N_e, Z^{pre(N),N} \mid \exists N)$, then:

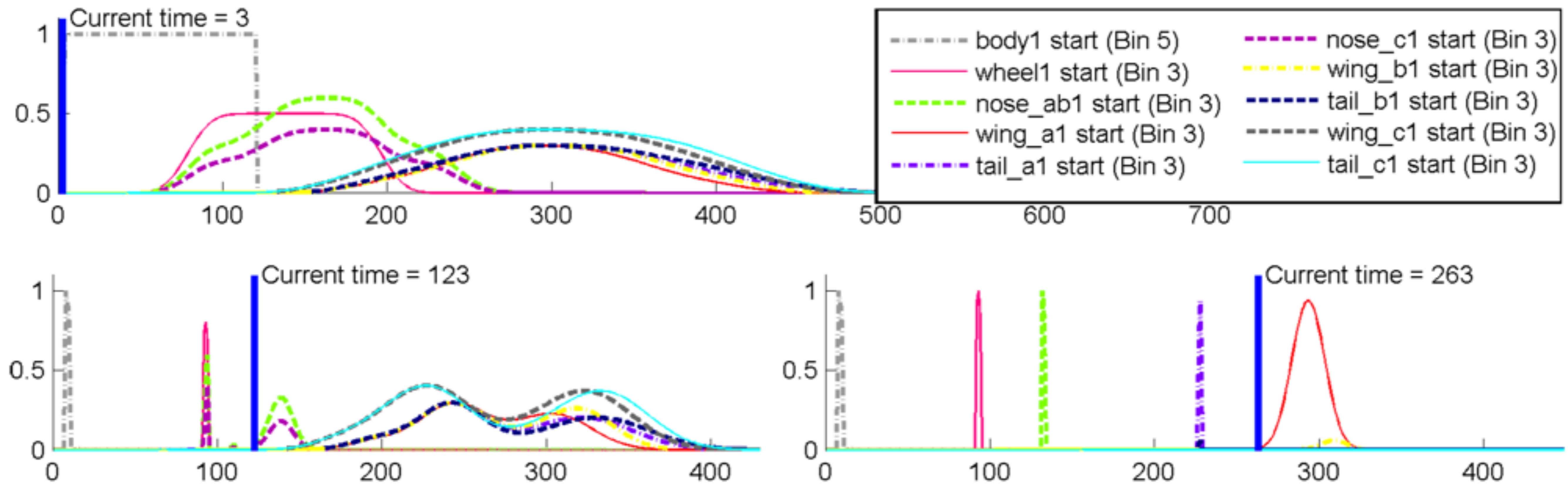
$$P(A_e = t, Z^{A,pre(A)} \mid \exists A) = P(\exists M \mid \exists A) P(Z^N \mid !N) P(M_e = t, Z^{M,pre(M)} \mid \exists M)$$
$$P(\exists N \mid \exists A) P(Z^M \mid !M) P(N_e = t, Z^{N,pre(N)} \mid \exists N)$$

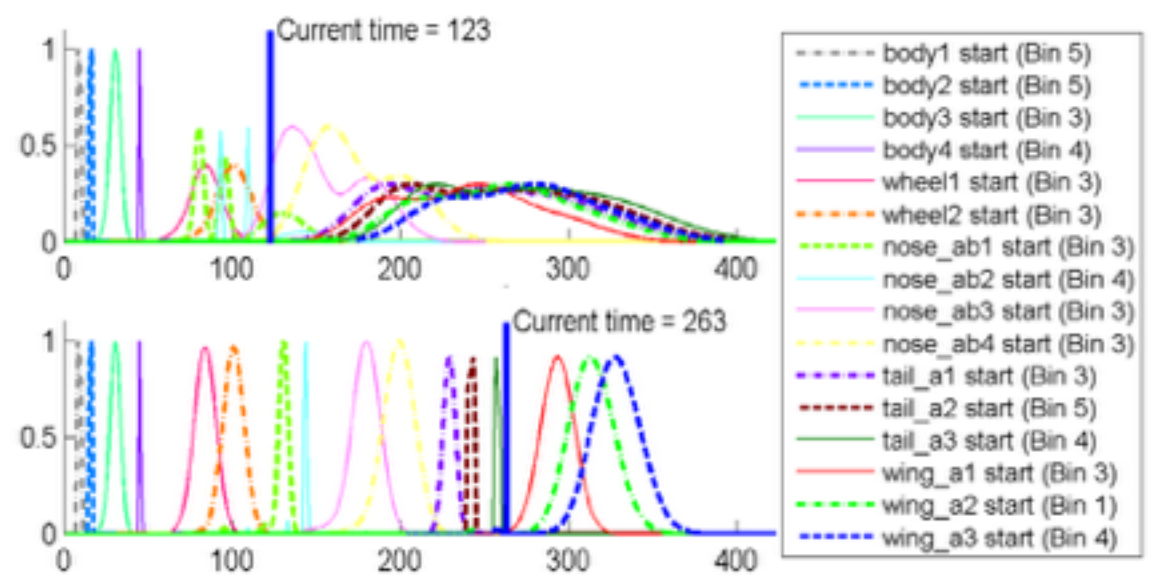
Toy Plane Dataset





Toy Plane Dataset





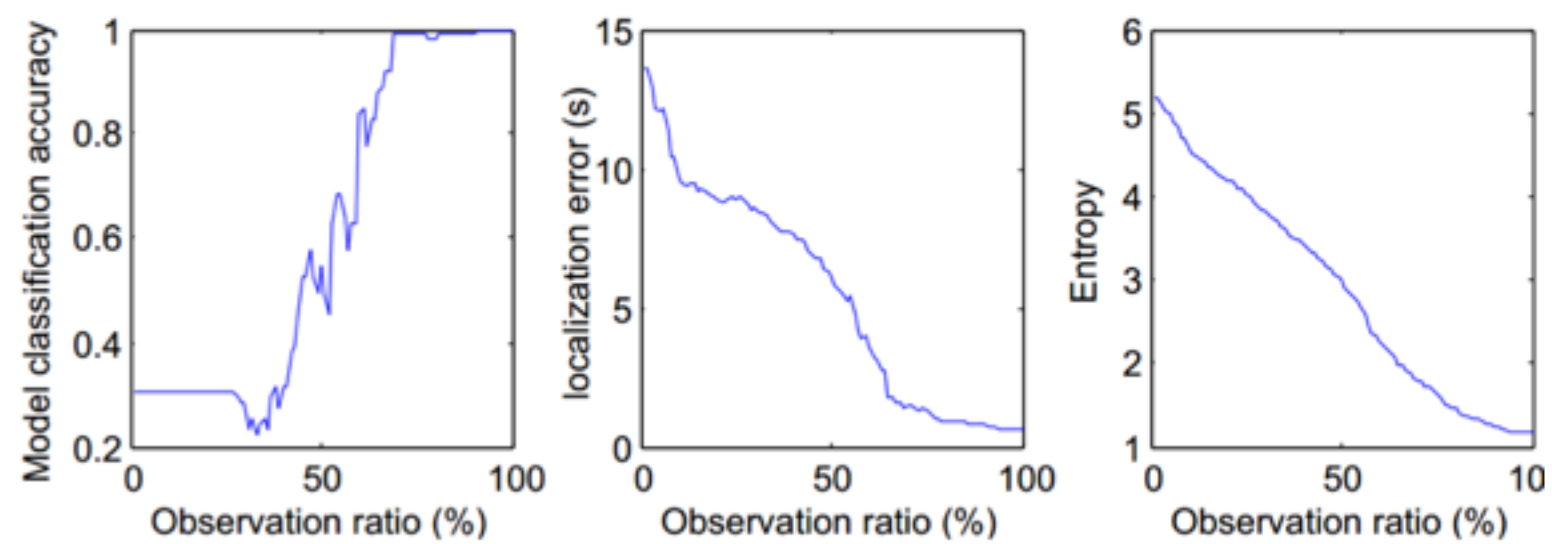
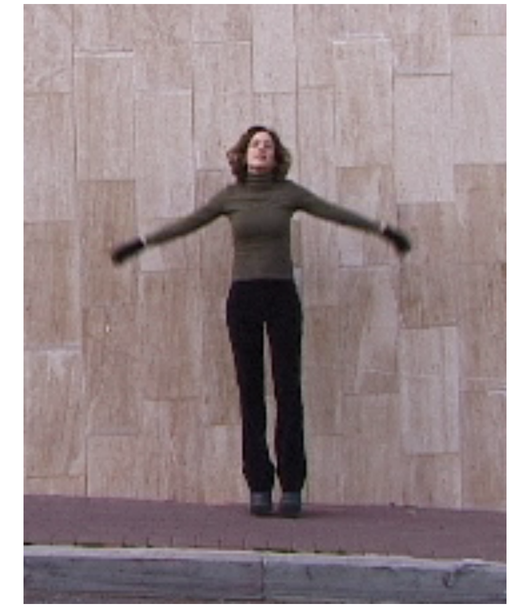


Figure 7. Result on our toy assembly dataset: (a) accuracy of model classification, (b) average localization error and (c) entropy of all actions' start

Weizmann Dataset

<u>Concatenate random videos together</u>

10x: Walk, Run, Jump, Gallop sideways, Bend
Jumping Jack, Skip, One-hand wave,
Two-hands wave, Jump in place



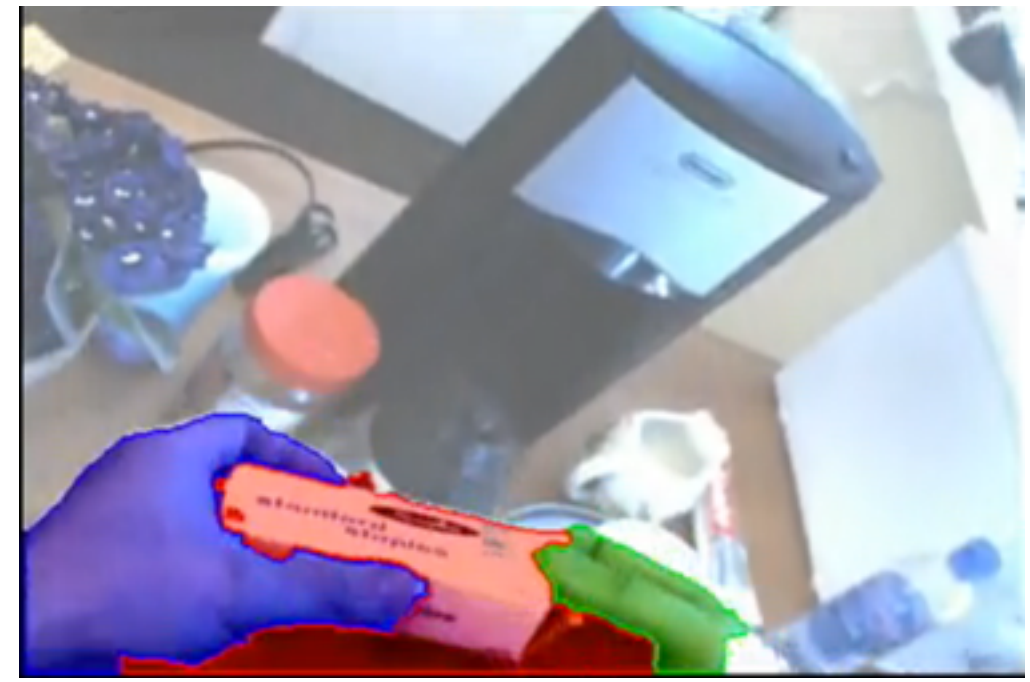
Method	Segmentation Accuracy
[14]	69.7%
[6]	87.7%
Ours	93.6%

$$S \rightarrow AAAAAAAAAAAAAAAAAAAAA...$$

$$A \rightarrow walk \mid run \mid jump \mid side \mid bend \mid$$
$$wave1 \mid wave2 \mid jump \mid jumpjack \mid skip \mid \emptyset$$

GTech Egocentric Activities

<u>Long activities</u>
Cheese sandwich, sweet tea
coffee, coffee with honey, hotdog
peanut butter sandwich
peanut butter and jelly



$$S \rightarrow Activity1 \mid Activity2 \mid \ldots$$
$$Activity1 \rightarrow Sequence1 \mid Sequence2 | \ldots$$
$$Sequence1 \rightarrow p\_action1\ p\_action2\ p\_action3 \ldots$$
$$\ldots$$

Takeaways

Good

- Removes typical Markov assumption
- Streaming method available (also doable with other PGM methods)

Bad

- Rules must be defined per task
- Restrictions on CFG's grammar