Frontiers of
Human Activity Analysis

J. K. Aggarwal
Michael S. Ryoo
Kris M. Kitani

THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

ETRI

Carnegie Mellon
THE ROBOTICS INSTITUTE

Overview

Human Activity Recognition

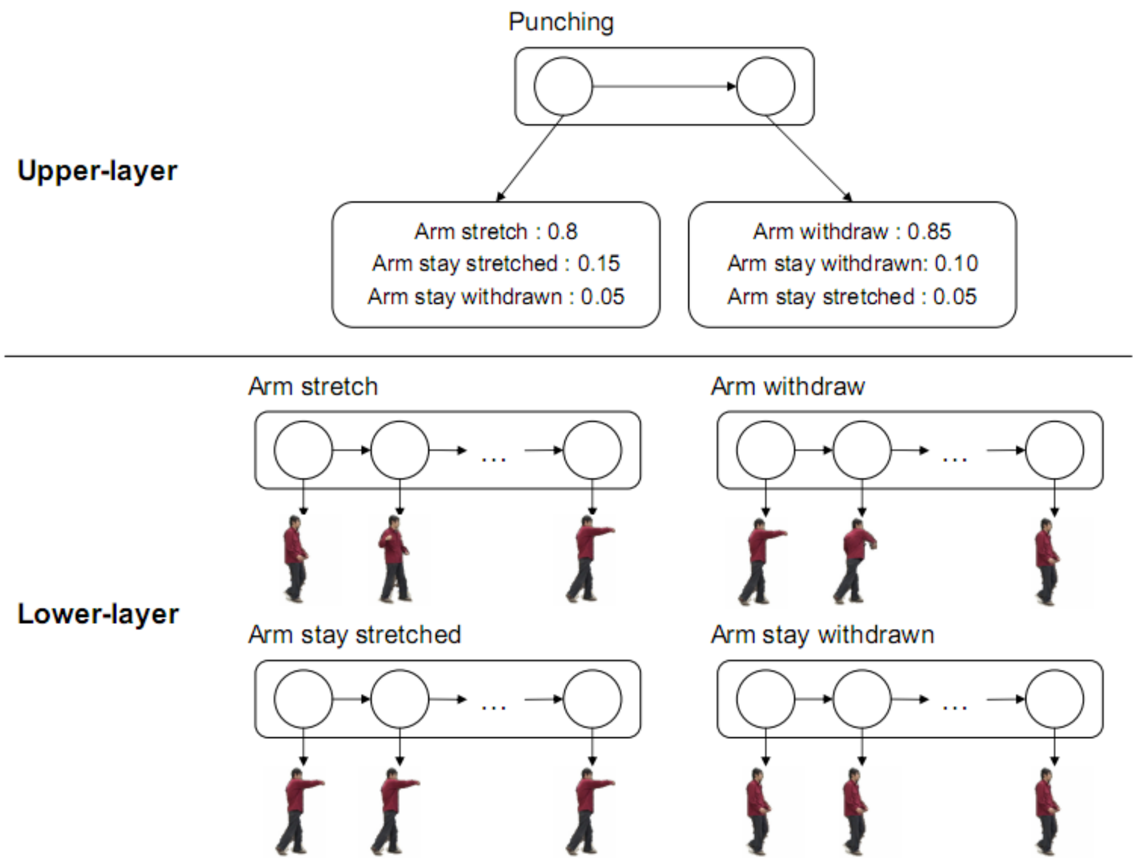
- **Single-layer**
  - Space-time
  - Sequential
- **Hierarchical**
  - Statistical
  - Syntactic
  - Descriptive

Motivation

How do we interpret a sequence of actions?

Hierarchy

Hierarchy implies decomposition into sub-parts



Now we'll cover…

Human Activity Recognition

- **Single-layer**
  - Space-time
  - Sequential
- **Hierarchical**
  - **Statistical**
  - **Syntactic**
  - Descriptive

Syntactic Approaches

Syntactic Models

Activities as strings of symbols.

s t r i n g s o f s y m b o l s

What is the underlying structure?

Early applications to Vision

Tsai and Fu 1980.
Attributed Grammar-A Tool for Combining Syntactic and Statistical Approaches to Pattern Recognition.

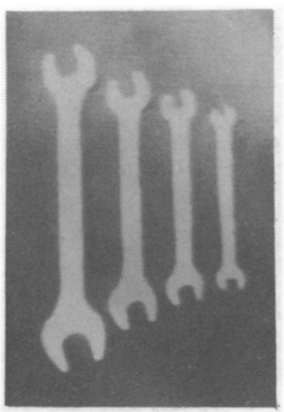


Fig. 6. Set of I wrenches.

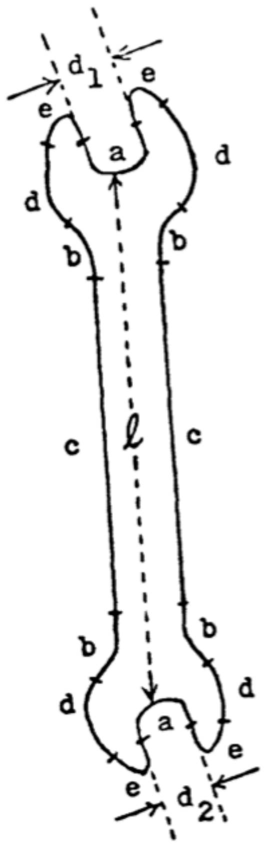


Fig. 8. I wrench and its boundary primitives.

$$z_I = aedbcbdeaedbcbde$$

Hierarchical syntactic approach

- Useful for activities with:
  - Deep hierarchical structure
  - Repetitive (cyclic) structure
- Not for
  - Systems with a lot of errors and uncertainty
  - Activities with shallow structure

Basics

Context-Free Grammar

$$\mathbf{G} = \langle S, \mathbf{T}, \mathbf{N}, \mathbf{P} \rangle$$

Generic Language	Natural Languages
Start Symbol (S)	Sentences
Set of Terminal Symbols (**T**)	Words
Set of Non-Terminal Symbols (**N**)	Parts of Speech
Set of Production Rules (**P**)	Syntax Rules

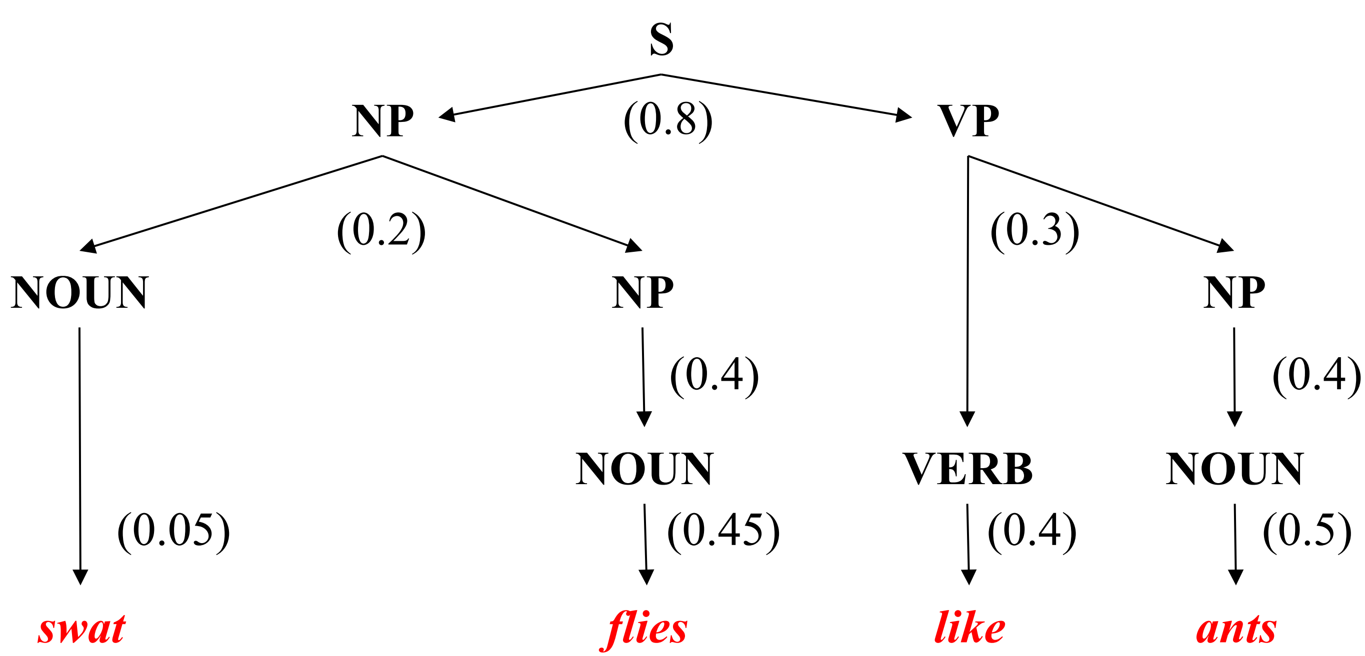
Parsing with a grammar

S → NP VP	(0.8)	PP → PREP NP	(1.0)
S → VP	(0.2)	PREP → *like*	(1.0)
NP → NOUN	(0.4)	VERB → *swat*	(0.2)
NP → NOUN PP	(0.4)	VERB → *flies*	(0.4)
NP → NOUN NP	(0.2)	VERB → *like*	(0.4)
VP → VERB	(0.3)	NOUN → *swat*	(0.05)
VP → VERB NP	(0.3)	NOUN → *flies*	(0.45)
VP → VERB PP	(0.2)	NOUN → *ants*	(0.5)
VP → VERB NP PP	(0.2)		

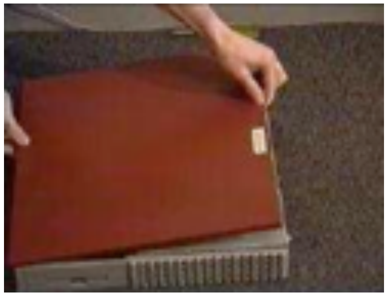
swat *flies* *like* *ants*

Parsing with a grammar

S → NP VP	**(0.8)**	**PP → PREP NP**	**(1.0)**
S → VP	**(0.2)**	**PREP →** ***like***	**(1.0)**
NP → NOUN	**(0.4)**	**VERB →** ***swat***	**(0.2)**
NP → NOUN PP	**(0.4)**	**VERB →** ***flies***	**(0.4)**
NP → NOUN NP	**(0.2)**	**VERB →** ***like***	**(0.4)**
VP → VERB	**(0.3)**	**NOUN →** ***swat***	**(0.05)**
VP → VERB NP	**(0.3)**	**NOUN →** ***flies***	**(0.45)**
VP → VERB PP	**(0.2)**	**NOUN →** ***ants***	**(0.5)**
VP → VERB NP PP	**(0.2)**		

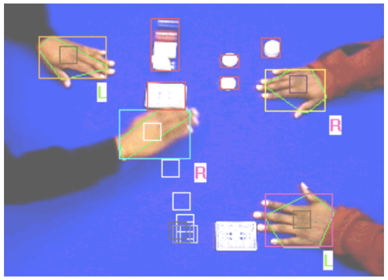


Video analysis with CFGs



The “Inverse Hollywood problem”:
From video to scripts and storyboards via causal analysis.
Brand 1997

Action Recognition using Probabilistic Parsing.
Bobick and Ivanov 1998



Recognizing Multitasked Activities from Video using Stochastic Context-Free Grammar.
Moore and Essa 2001

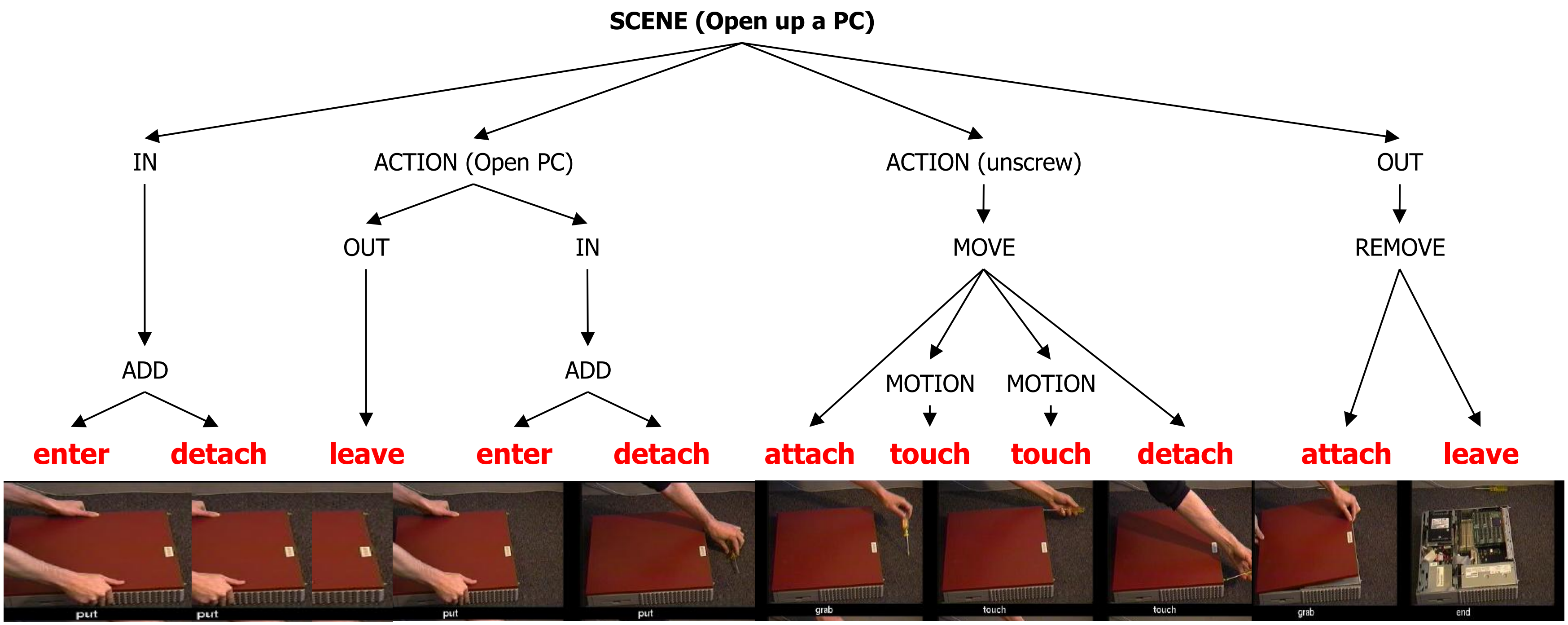
CFG for human activities

enter detach leave enter detach attach touch touch detach attach leave

scene	$\rightarrow$	in action* out
action	$\rightarrow$	motion | move | {out in}
in	$\rightarrow$	ENTER | add
out	$\rightarrow$	LEAVE | remove
add	$\rightarrow$	ENTER motion* DETACH
remove	$\rightarrow$	ATTACH motion* LEAVE
move	$\rightarrow$	ATTACH motion+ DETACH
motion	$\rightarrow$	SHIFT | TOUCH | BUMP

M. Brand. The "Inverse Hollywood Problem":
From video to scripts and storyboards
via causal analysis. AAAI 1997.

Parse tree



scene	→	in action* out
action	→	motion | move | {out in}
in	→	ENTER | add
out	→	LEAVE | remove
add	→	ENTER motion* DETACH
remove	→	ATTACH motion* LEAVE
move	→	ATTACH motion+ DETACH
motion	→	SHIFT | TOUCH | BUMP

- Deterministic low-level primitive detection
- Deterministic parsing

M. Brand. The "Inverse Hollywood Problem": From video to scripts and storyboards via causal analysis. AAAI 1997.

Stochastic CFGs

Action Recognition using Probabilistic Parsing.
Bobick and Ivanov 1998

G_{square} :

SQUARE	$\rightarrow$		RH	[0.5]
		|	LH	[0.5]
RH	$\rightarrow$		TOP UD BOT DU	[1.0]
LH	$\rightarrow$		BOT DU TOP UD	[1.0]
TOP	$\rightarrow$		LR	[0.5]
		|	RL	[0.5]
BOT	$\rightarrow$		RL	[0.5]
		|	LR	[0.5]
LR	$\rightarrow$		left-right	[1.0]
UD	$\rightarrow$		up-down	[1.0]
RL	$\rightarrow$		right-left	[1.0]
DU	$\rightarrow$		down-up	[1.0]

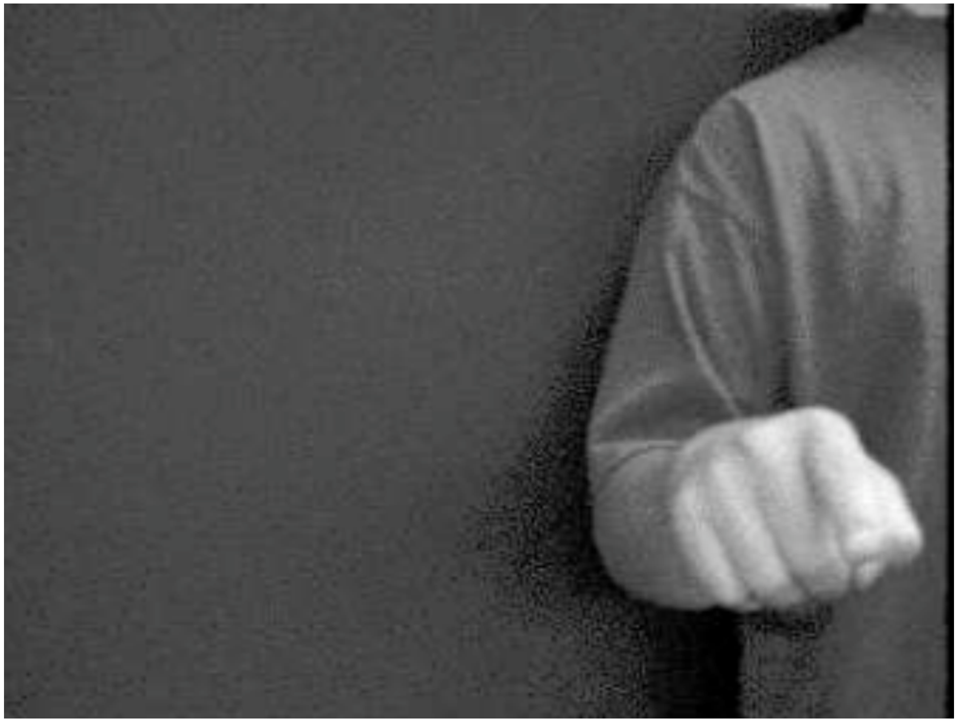
Gesture analysis with CFGs

Primitive recognition with HMMs

Action Recognition using Probabilistic Parsing. Bobick and Ivanov 1998

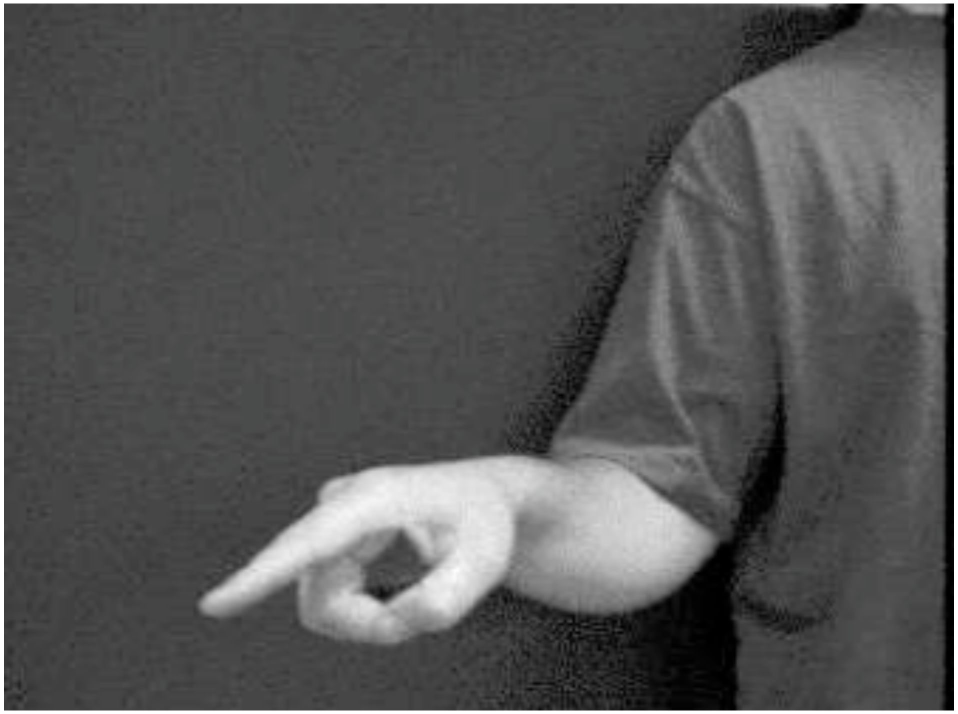
left-right

up-down



Action Recognition using Probabilistic Parsing. Bobick and Ivanov 1998

right-left



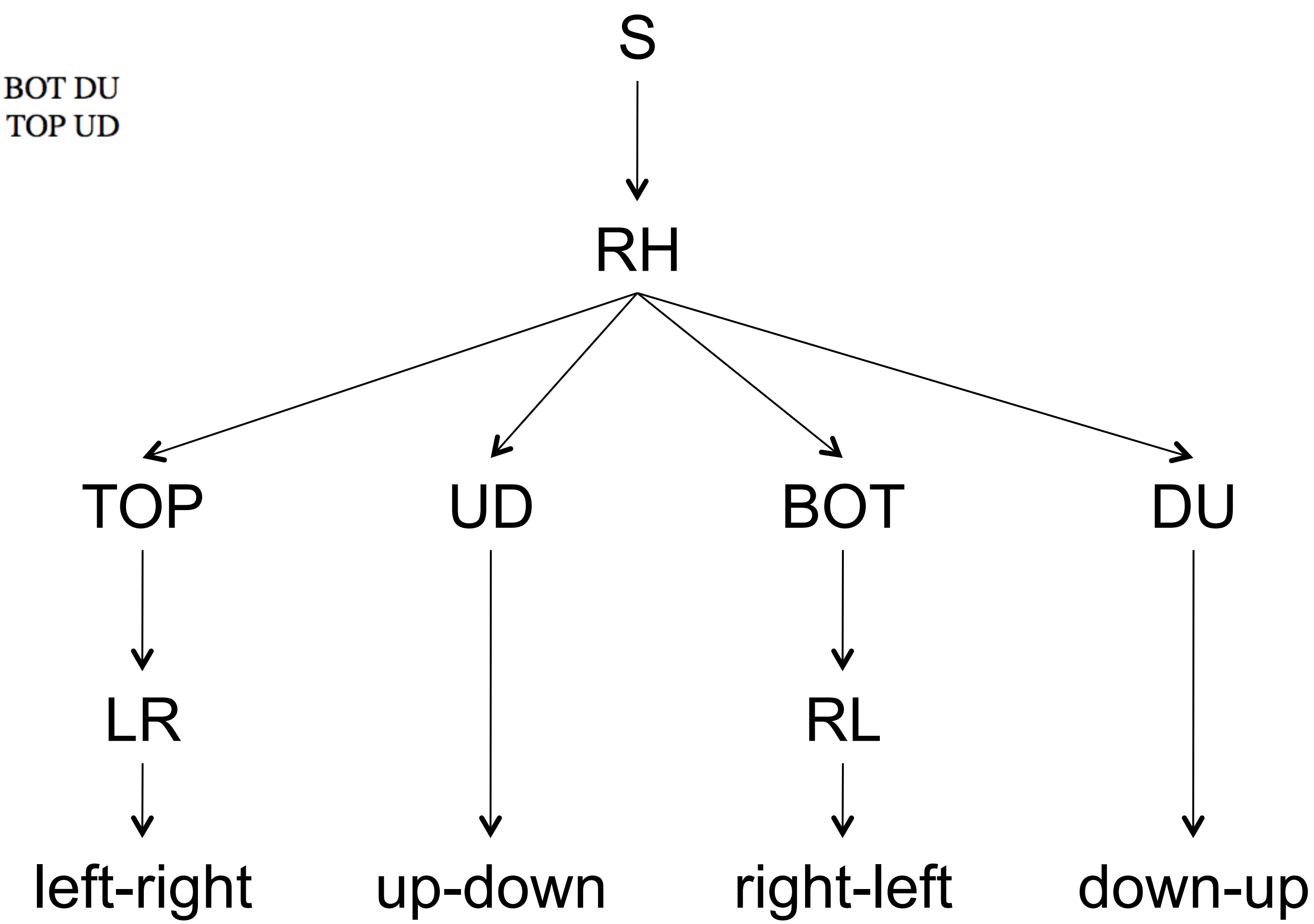
Action Recognition using Probabilistic Parsing. Bobick and Ivanov 1998

down-up

Parse Tree

G_{square} :

SQUARE	$\rightarrow$	RH
	$\mid$	LH
RH	$\rightarrow$	TOP UD BOT DU
LH	$\rightarrow$	BOT DU TOP UD
TOP	$\rightarrow$	LR
	$\mid$	RL
BOT	$\rightarrow$	RL
	$\mid$	LR
LR	$\rightarrow$	left-right
UD	$\rightarrow$	up-down
RL	$\rightarrow$	right-left
DU	$\rightarrow$	down-up

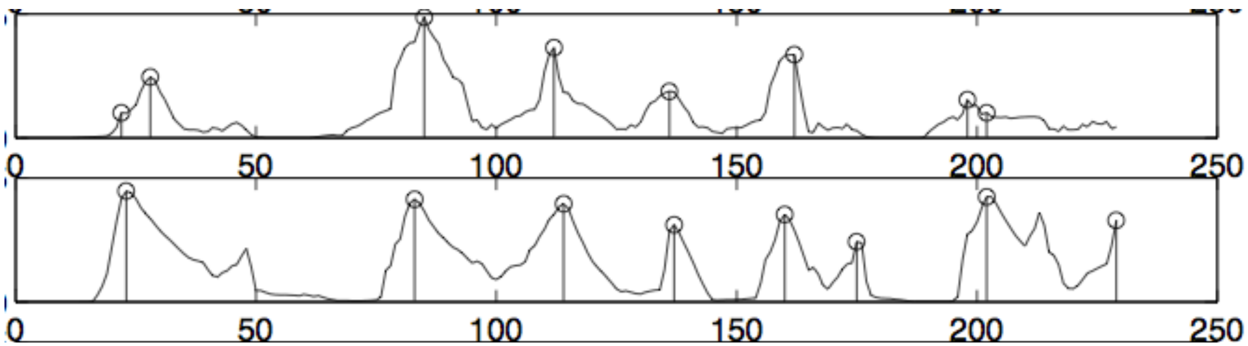


Errors

Likelihood value over time (not discrete symbols)

HMM a

HMM b



Errors are inevitable…

but the grammar acts as a top-down constraint

Action Recognition using Probabilistic Parsing. Bobick and Ivanov 1998

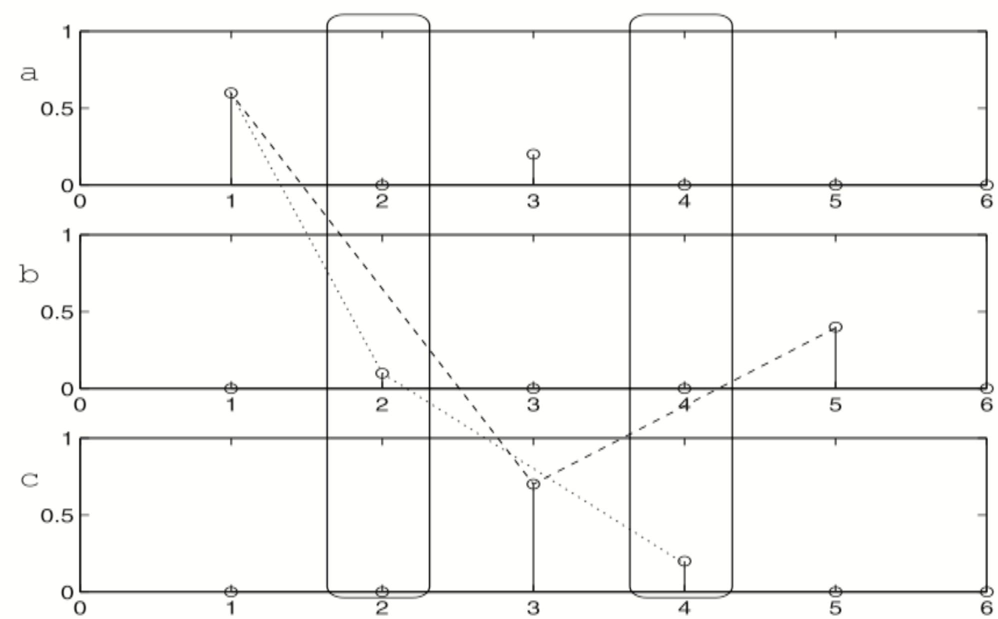
Dealing with uncertainty & errors

- Stolcke-Early (probabilistic) parser
- SKIP rules to deal with **insertion** errors

HMM a

HMM b

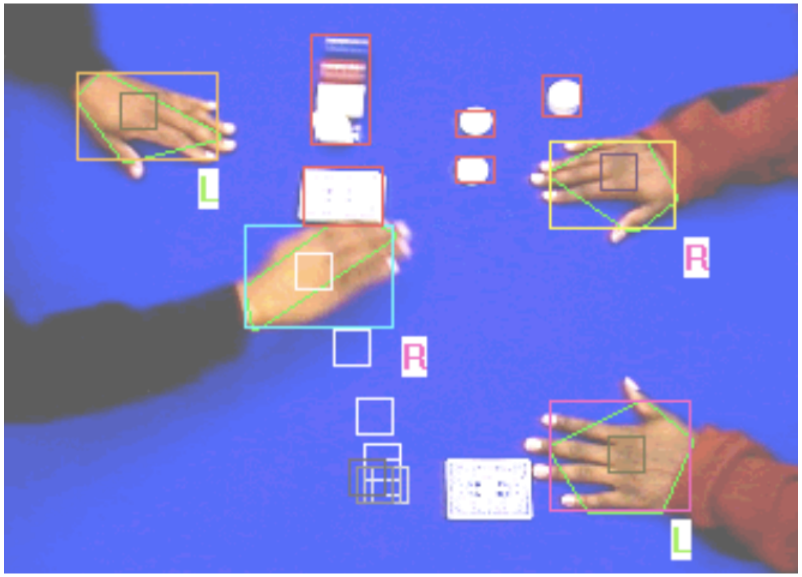
HMM c

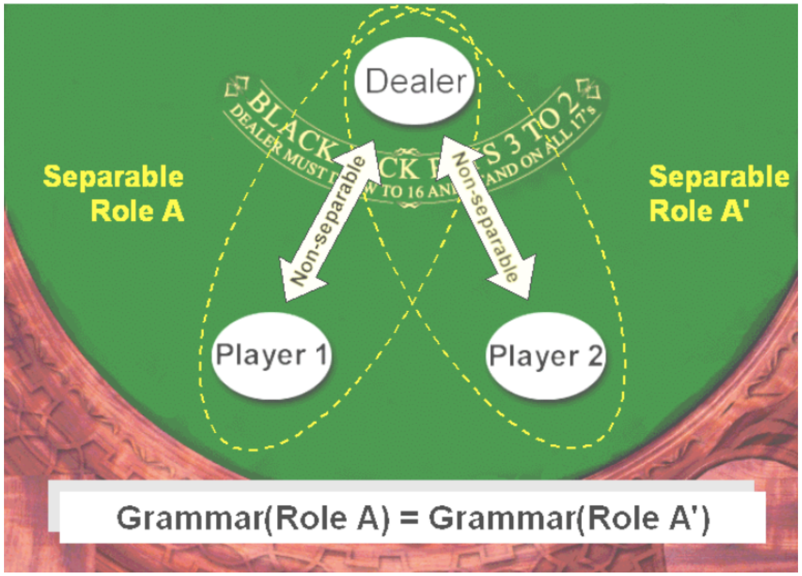


Action Recognition using Probabilistic Parsing. Bobick and Ivanov 1998

SCFG for Blackjack

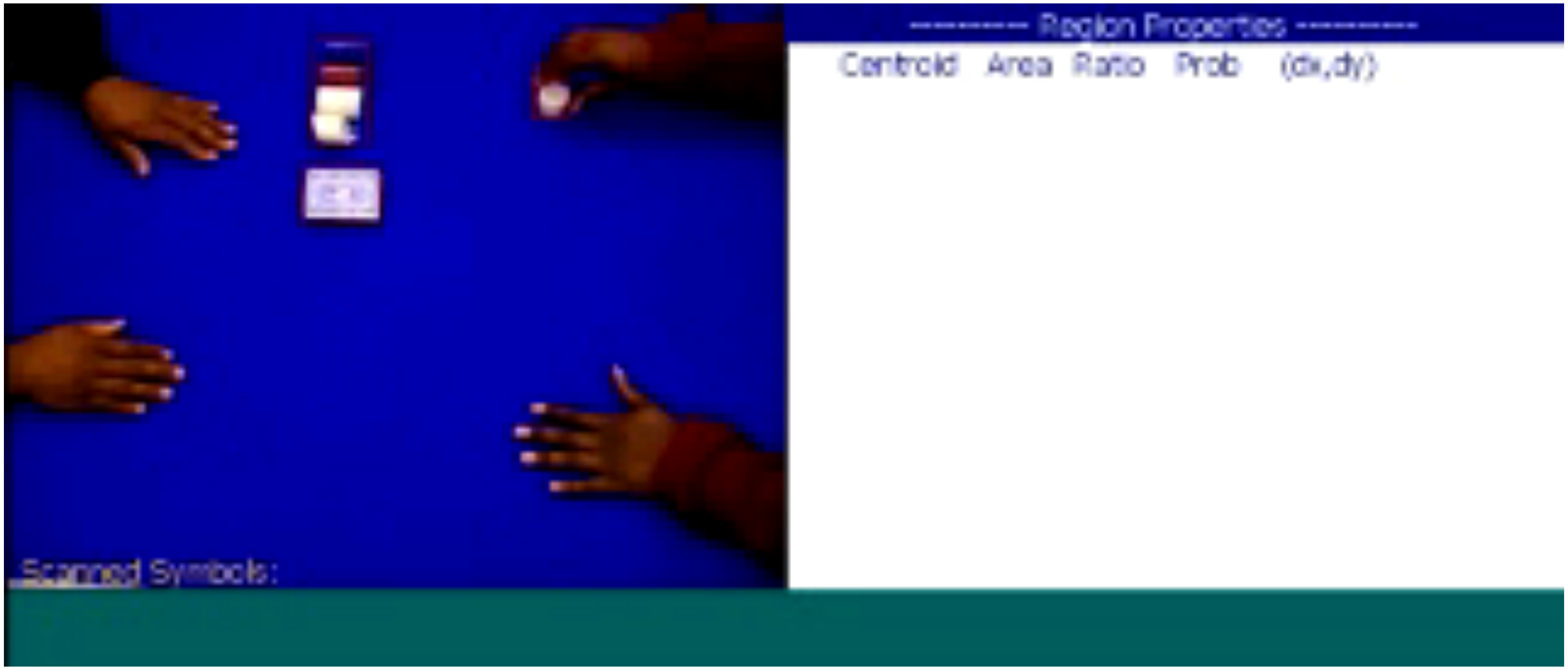
Recognizing Multitasked Activities from Video using
Stochastic Context-Free Grammar.
Moore and Essa 2001





- Deals with more complex activities
- Deals with more error types

extracting primitive actions



Game grammar

Production Rules			Description
S	$\rightarrow AB$	[1.0]	Blackjack → "play game" "determine winner"
A	$\rightarrow CD$	[1.0]	play game → "setup game" "implement strategy"
B	$\rightarrow EF$	[1.0]	determine winner → "eval. strategy" "cleanup"
C	$\rightarrow HI$	[1.0]	setup game → "place bets" "deal card pairs"
D	$\rightarrow GK$	[1.0]	implement strategy → "player strategy"
E	$\rightarrow LKM$	[0.6]	eval. strategy → "dealer down-card" "dealer hits" "player down-card"
	$\rightarrow LM$	[0.4]	eval. strategy → "dealer down-card" "player down-card"
F	$\rightarrow NO$	[0.5]	cleanup → "settle bet" "recover card"
	$\rightarrow ON$	[0.5]	→ "recover card" "settle bet"
G	$\rightarrow J$	[0.8]	player strategy → "Basic Strategy"
	$\rightarrow Hf$	[0.1]	→ "Splitting Pair"
	$\rightarrow bfffH$	[0.1]	→ "Doubling Down"
H	$\rightarrow l$	[0.5]	place bets
	$\rightarrow lH$	[0.5]	
I	$\rightarrow ffI$	[0.5]	deal card pairs
	$\rightarrow ee$	[0.5]	
J	$\rightarrow f$	[0.8]	Basic strategy
	$\rightarrow fJ$	[0.2]	
K	$\rightarrow e$	[0.6]	house hits
	$\rightarrow eK$	[0.4]	
L	$\rightarrow ae$	[1.0]	Dealer downcard
M	$\rightarrow dh$	[1.0]	Player downcard
N	$\rightarrow k$	[0.16]	settle bet
	$\rightarrow kN$	[0.16]	
	$\rightarrow j$	[0.16]	
	$\rightarrow jN$	[0.16]	
	$\rightarrow i$	[0.18]	
	$\rightarrow iN$	[0.18]	
O	$\rightarrow a$	[0.25]	recover card
	$\rightarrow aO$	[0.25]	
	$\rightarrow b$	[0.25]	
	$\rightarrow bO$	[0.25]	

Symbol	Domain-Specific Events (Terminals)
a	dealer removed card from house
b	dealer removed card from player
c	player removed card from house
d	player removed card from player
e	dealer added card to house
f	dealer dealt card to player
g	player added card to house
h	player added card to player
i	dealer removed chip
j	player removed chip
k	dealer pays player chip
l	player bets chip

Recognizing Multitasked Activities from Video using Stochastic Context-Free Grammar. Moore and Essa 2001

Dealing with errors

- Ungrammatical strings cause parser to fail
- Account for errors with multiple hypothesis
  - Insertion, deletion, substitution
- Issues
  - How many errors should we tolerate?
  - Potentially exponential hypothesis space
  - Ungrammatical strings: vision problem or illegal activity?

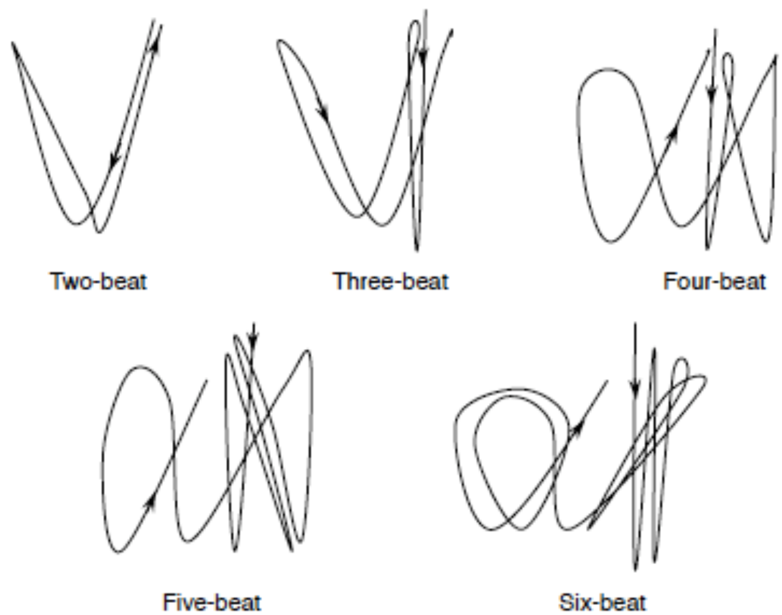
Observations

- CFGs good for structured activities
- Can incorporate uncertainty in observations
- Natural contextual prior for recognizing errors

- Not clear how to deal with errors
- Assumes ‘good’ action classifiers
- Need to define grammar manually

Can we learn the grammar from data?

Heuristic Grammatical Induction



1. Lexicon learning
   - Learn HMMs
   - Cluster HMMs
2. Convert video to string
3. Learn Grammar

Unsupervised Analysis of Human Gestures. Wang et al 2001

COMPRESSIVE

a b c d a b c d b c d a b a b

length occurrence new rule new symbol

$$\arg\max_{\lambda} \Delta DL = \arg\max_{\lambda}\{M \times N - (M+1) - N\}$$

substring

deletion of substring

insertion of new rule

substring	M	N	ΔDL
ab	2	4	1
cd	2	3	0
bcd	3	3	2
abcd	3	2	1

On-Line and Off-Line Heuristics for Inferring Hierarchies of Repetitions in Sequences. Nevill-Manning 2000.

example

$S \rightarrow a\ b\ c\ d\ a\ b\ c\ d\ b\ c\ d\ a\ b\ a\ b$

(DL=16)

$A \rightarrow b\ c\ d$

$S \rightarrow a\ A\ a\ A\ A\ a\ b\ a\ b$

(DL=14)

Repeat until compression becomes 0.

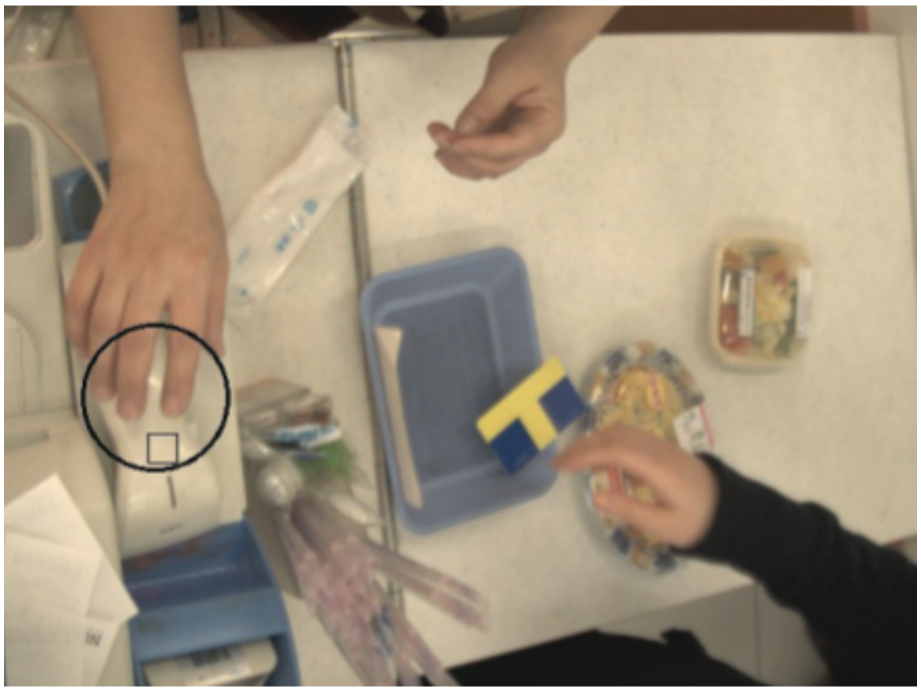
Critical assumption

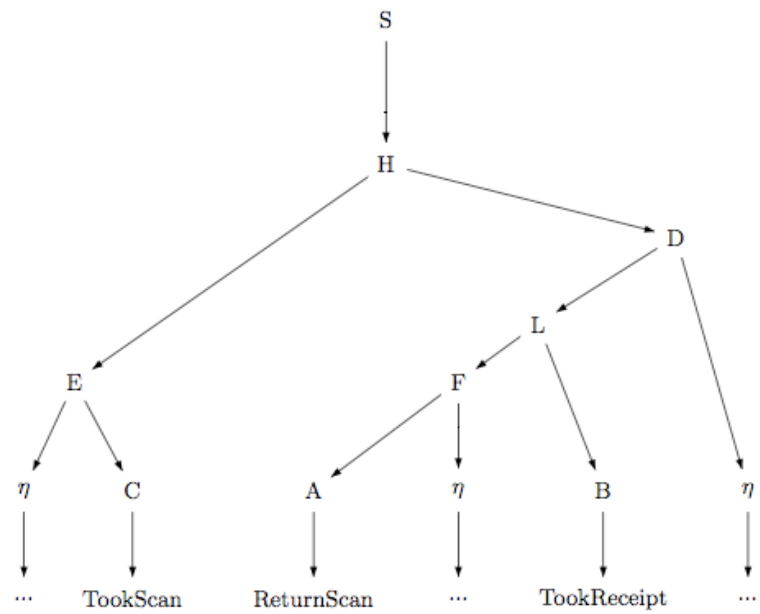
- No uncertainty
- No errors
  - insertions
  - deletions
  - substitution

Can we learn grammars despite errors?

Learning with noise

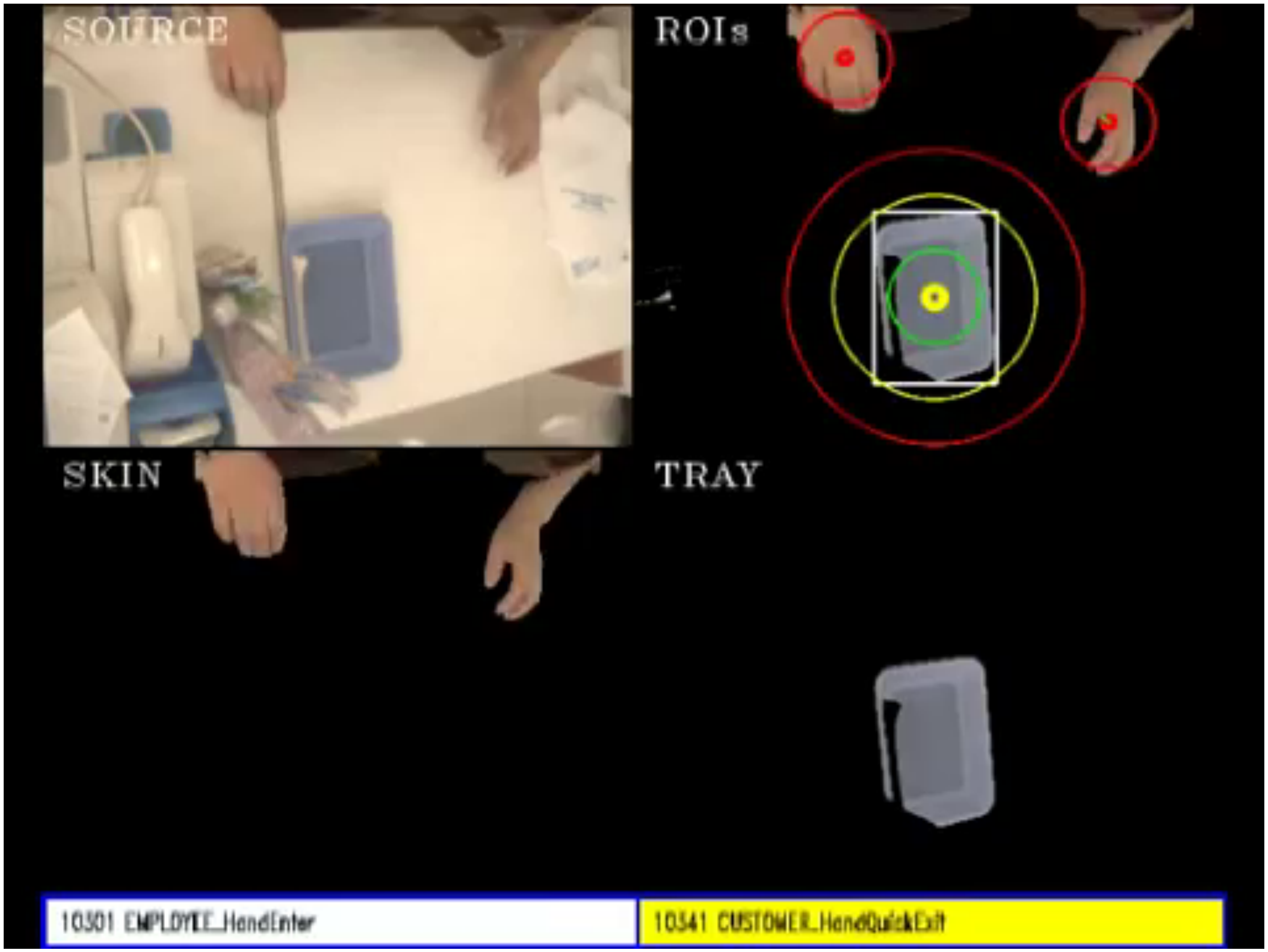
Can we learn the basic structure of a transaction?





Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

extracting primitives



Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Underlying structure?

$$D \rightarrow \mathtt{a\ x\ b\ y\ c\ a\ b\ x\ c\ y\ a\ b\ c\ x}$$

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Underlying structure?

$$D \rightarrow a\ x\ b\ y\ c\ a\ b\ x\ c\ y\ a\ b\ c\ x$$

$$D \rightarrow a\quad b\quad c\ a\ b\quad c\quad a\ b\ c$$

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Underlying structure?

$$D \rightarrow a\ x\ b\ y\ c\ a\ b\ x\ c\ y\ a\ b\ c\ x$$

$$D \rightarrow \boxed{a\quad b\quad c}\ \boxed{a\ b\quad c}\quad \boxed{a\ b\ c}$$

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Underlying structure?

$$D \rightarrow a\ x\ b\ y\ c\ a\ b\ x\ c\ y\ a\ b\ c\ x$$

$$D \rightarrow a\quad b\quad c\ a\ b\quad c\quad a\ b\ c$$

$$\boxed{A \rightarrow a\ b\ c} \qquad \boxed{D \rightarrow A\ A\ A}$$

Simple grammar

Efficient compression

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Information Theory Problem (MDL)

$$\hat{G} = \arg\min_{G} \{ \underbrace{DL(G)}_{\text{Model complexity}} + \underbrace{DL(D|G)}_{\text{Data compression}} \}$$

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Information Theory Problem (MDL)

$$\hat{G} = \arg\min_{G} \{ \underbrace{DL(G)}_{\text{Model complexity}} + \underbrace{DL(D|G)}_{\text{Data compression}} \}$$

$$\begin{aligned}
\underbrace{DL(G)}_{\text{Model complexity}} &= -\log p(G) \\
&= -\log p(\theta_S, G_S) \\
&= -\log p(\theta_S | G_S) - \log p(G_S) \\
&= \underbrace{DL(\theta_S | G_S)}_{\text{Grammar parameters}} - \underbrace{DL(G_S)}_{\text{Grammar structure}}
\end{aligned}$$

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Information Theory Problem (MDL)

$$\hat{G} = \arg\min_{G} \{ DL(G) + DL(D|G) \}$$

Model complexity — Data compression

$$\begin{aligned} DL(G) &= -\log p(G) \\ &= -\log p(\theta_S, G_S) \\ &= -\log p(\theta_S|G_S) - \log p(G_S) \\ &= DL(\theta_S|G_S) - DL(G_S) \end{aligned}$$

Model complexity

Grammar parameters — Grammar structure

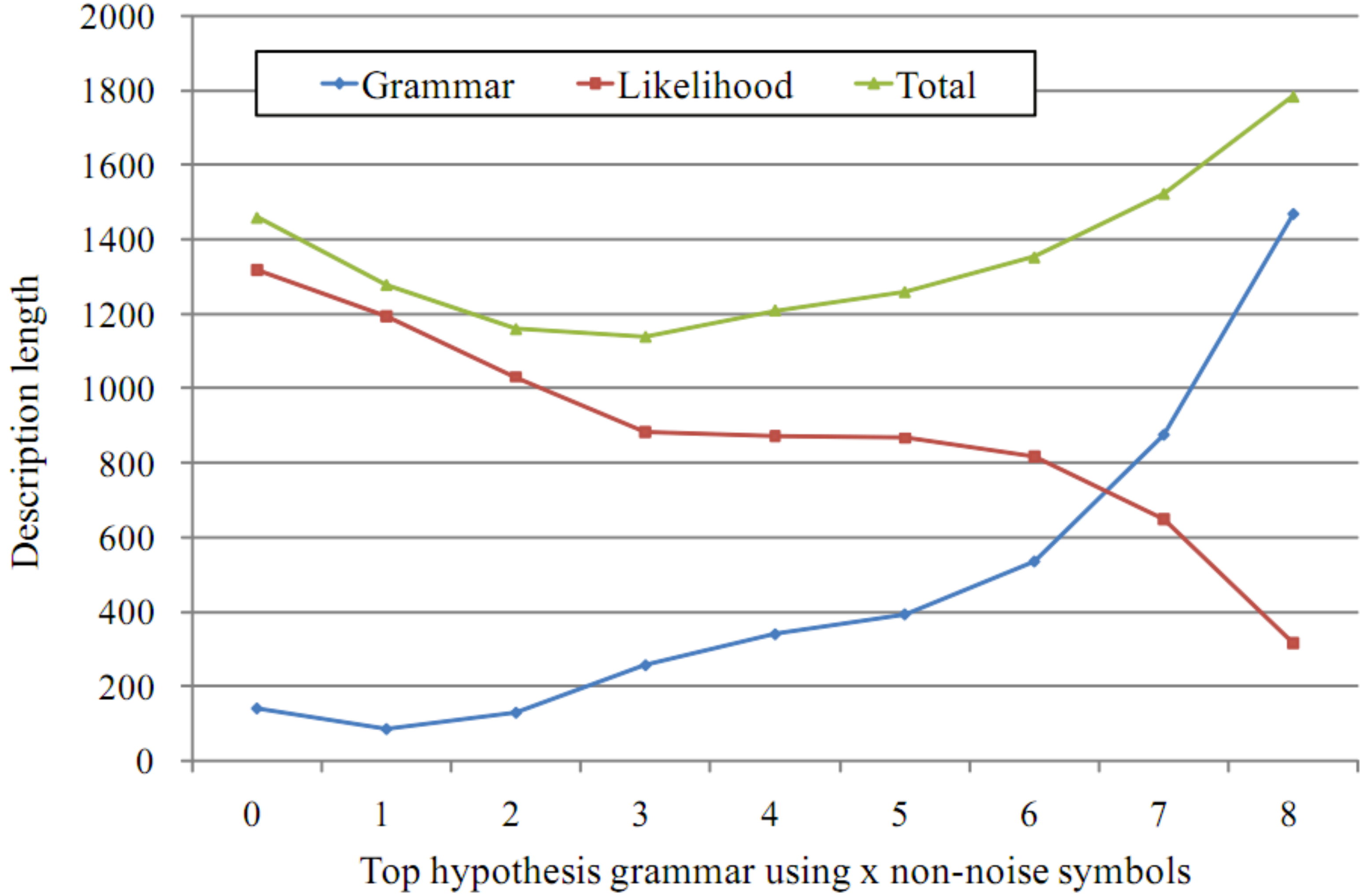
$$DL(D|G) = -\log p(D|G)$$

Data compression

Likelihood
(inside probabilities)

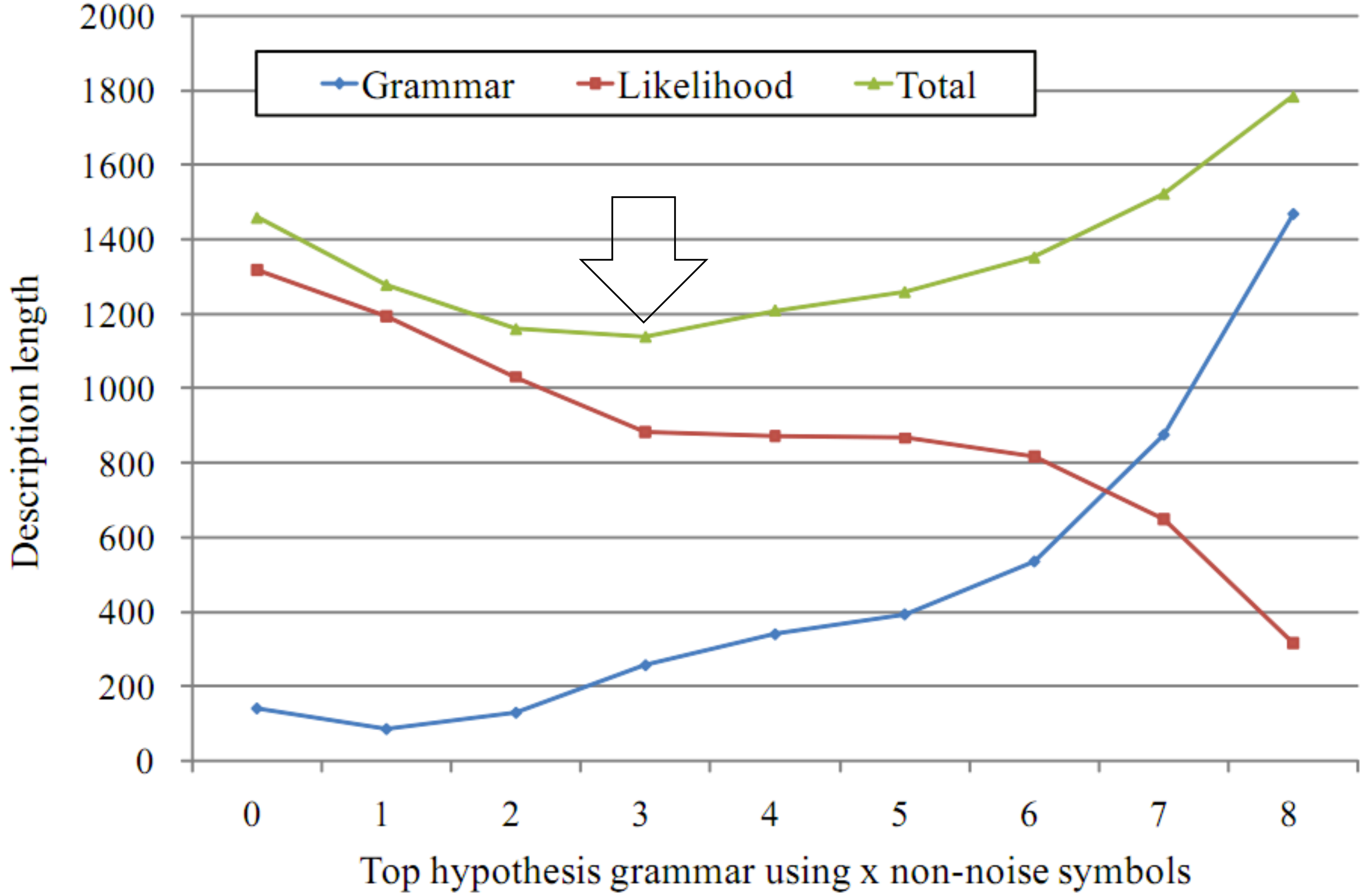
Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Minimum Description Length



Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Minimum Description Length

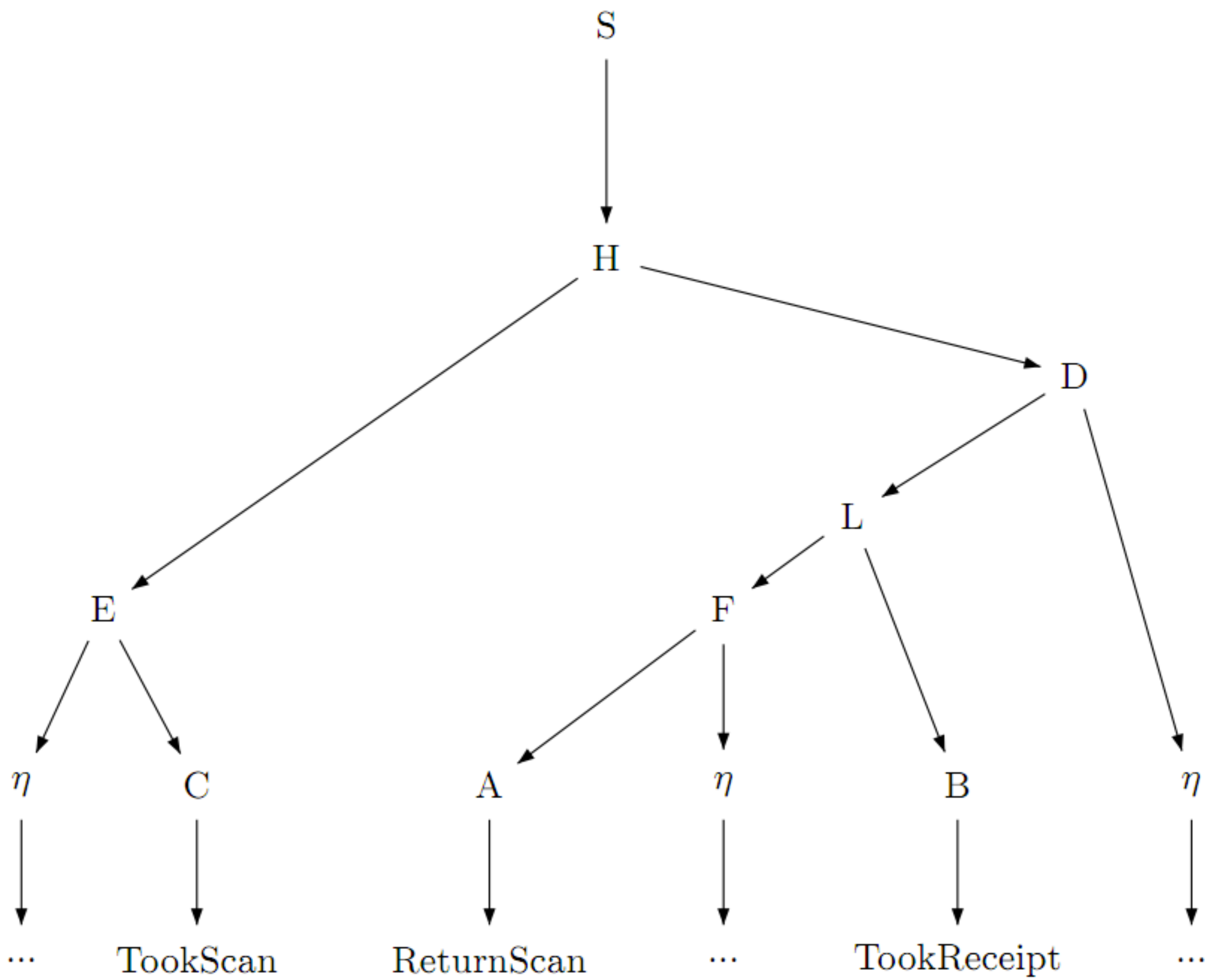


Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

$\mathbf{S} \to$ D					(0.02)
$\mathbf{S} \to$ H					(0.16)
$\mathbf{S} \to$ G					(0.18)
$\mathbf{S} \to$ N	η				(0.04)
$\mathbf{S} \to$ J					(0.13)
$\mathbf{S} \to$ Q					(0.05)
$\mathbf{S} \to \eta$					(0.02)
$\mathbf{S} \to$ N					(0.02)
$\mathbf{S} \to$ R					(0.05)
$\mathbf{S} \to$ J	B				(0.02)
$\mathbf{S} \to$ M	L				(0.04)
$\mathbf{S} \to$ M	A	H			(0.02)
$\mathbf{S} \to$ C	K				(0.04)
$\mathbf{S} \to$ C	A	M	F		(0.02)
$\mathbf{S} \to$ O	F				(0.02)
$\mathbf{S} \to$ M					(0.02)
$\mathbf{S} \to$ O	L				(0.02)
$\mathbf{S} \to$ O					(0.02)
$\mathbf{S} \to$ P					(0.05)
$\mathbf{S} \to$ I					(0.04)
$\mathbf{S} \to$ K					(0.04)
A $\to$ EMP_ReturnedScanner					(1.00)
B $\to$ EMP_TookReceipt					(1.00)
C $\to$ EMP_TookScanner					(1.00)

D $\to$ L	η		(1.000)
E $\to$ η	C		(1.000)
F $\to$ A	η		(1.000)
G $\to$ C	D		(1.000)
H $\to$ E	D		(1.000)
I $\to$ $*$	B	η	(1.000)
J $\to$ C	F		(1.000)
K $\to$ $*$	D		(1.000)
L $\to$ F	B		(1.000)
M $\to$ C	$*$		(1.000)
N $\to$ E	A	B	(1.000)
O $\to$ E	$*$		(1.000)
P $\to$ E	I		(1.000)
Q $\to$ E	K		(1.000)
R $\to$ E	L		(1.000)
$\eta \to \eta$	η		(0.309)
$\eta \to$ CUS_AddMoney			(0.153)
$\eta \to$ CUS_MovedTray			(0.006)
$\eta \to$ CUS_RemMoney			(0.003)
$\eta \to$ EMP_HandReturn			(0.080)
$\eta \to$ EMP_Interaction			(0.275)
$\eta \to$ EMP_MovedTray			(0.028)
$\eta \to$ EMP_RemMoney			(0.147)

Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.



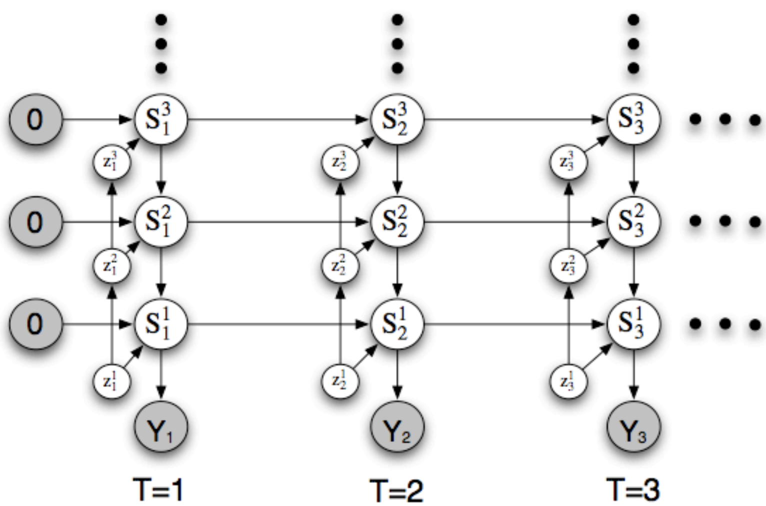
Recovering the basic structure of human activities from noisy video-based symbol strings. Kitani et al 2008.

Conclusions

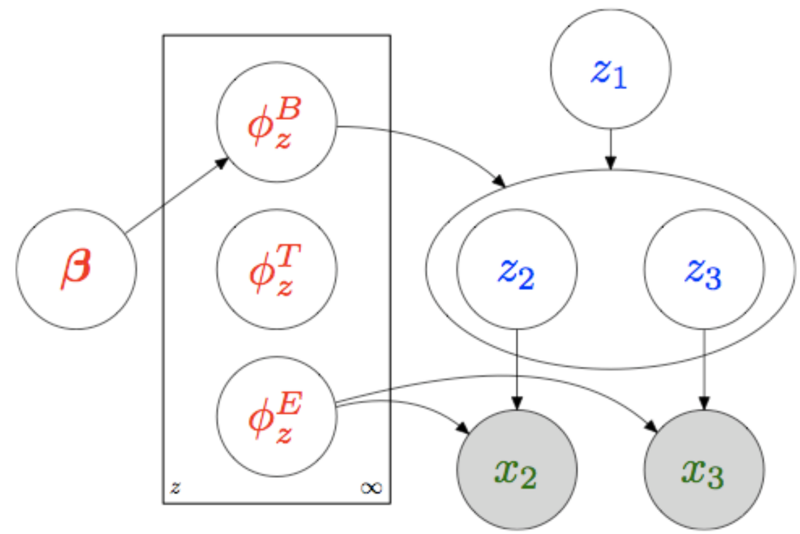
- Possible to learn basic structure
- Robust to errors
  (insertion, deletion, substitution)

- Need a lot of training data
- Computational complexity

Bayesian Approaches



Infinite Hierarchical Hidden Markov Models.
Heller et al 2009.



The Infinite PCFG using Hierarchical Dirichlet Processes.
Liang et al 2007.

Take home message

Hierarchical Syntactic Models

- Useful for activities with:
  - Deep hierarchical structure
  - Repetitive (cyclic) structure

- Not for
  - Systems with a lot of errors and uncertainty
  - Activities with weak structure

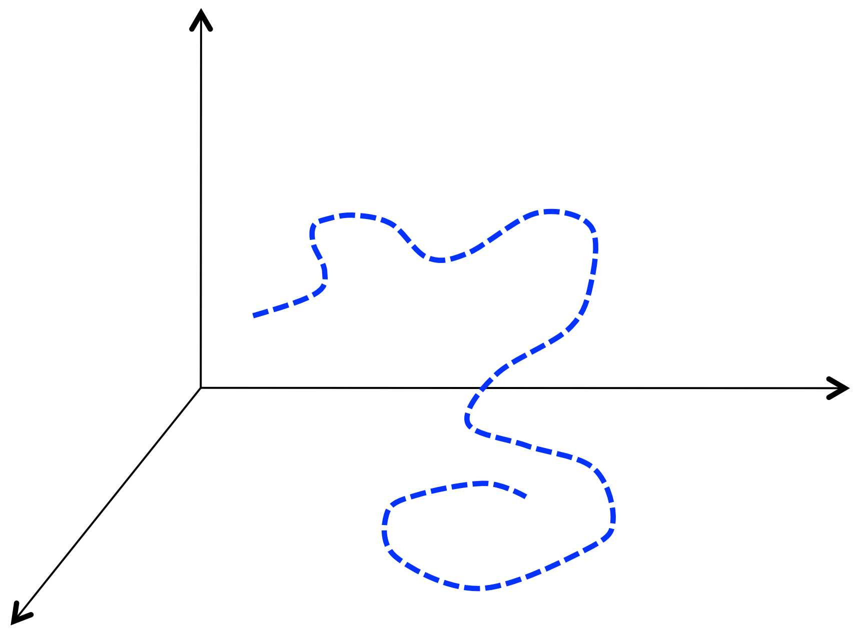
Statistical Approaches

Using a hierarchical statistical approach

- Use when
  - Low-level action detectors are noisy
  - Structure of activity is sequential
  - Integrating dynamics
- Not for
  - Activities with deep hierarchical structure
  - Activities with complex temporal structure

Statistical (State-based) Model

Activities as a stochastic path.



What are the underlying dynamics?

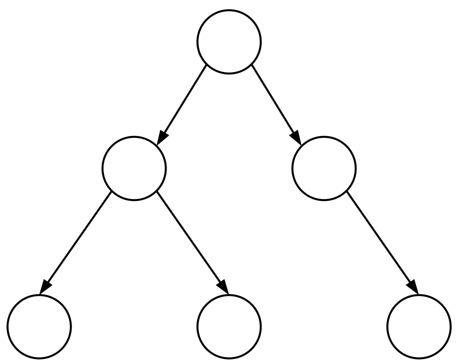
Characteristics

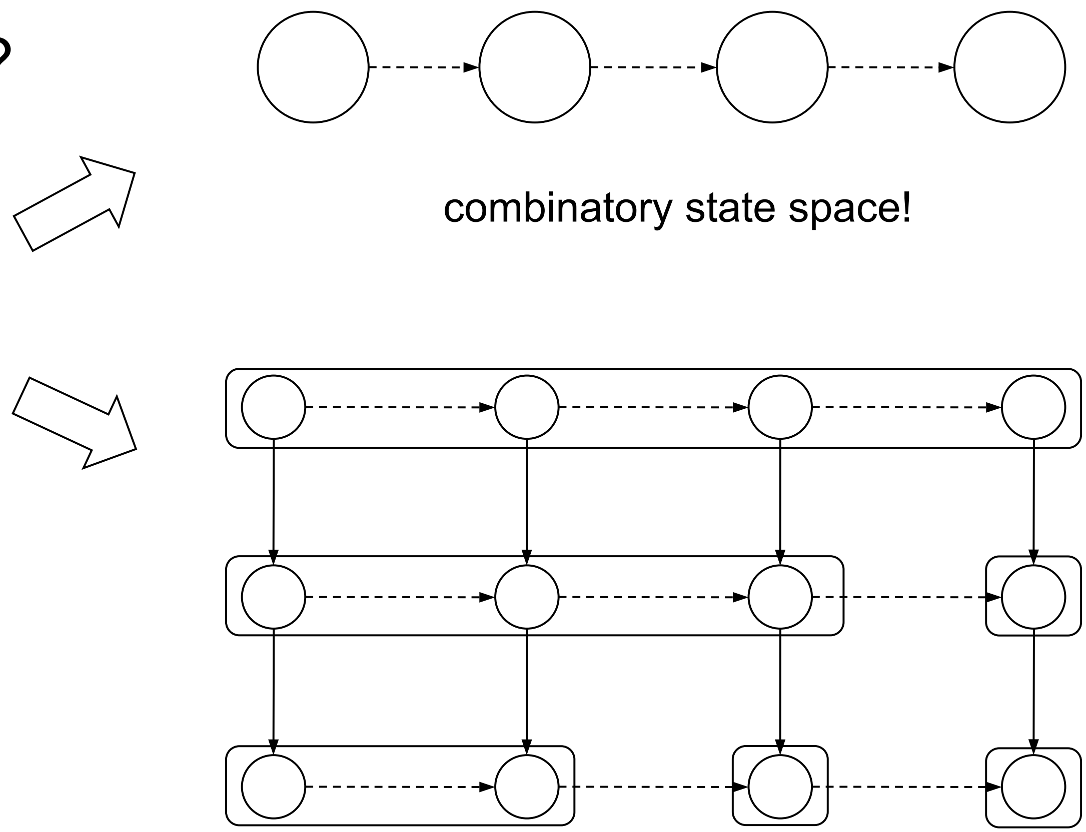
- Strong Markov assumption
- Strong dynamics prior
- Robust to uncertainty

- Modifications to account for
  - Hierarchical structure
  - Concurrent structure

Hierarchical activities

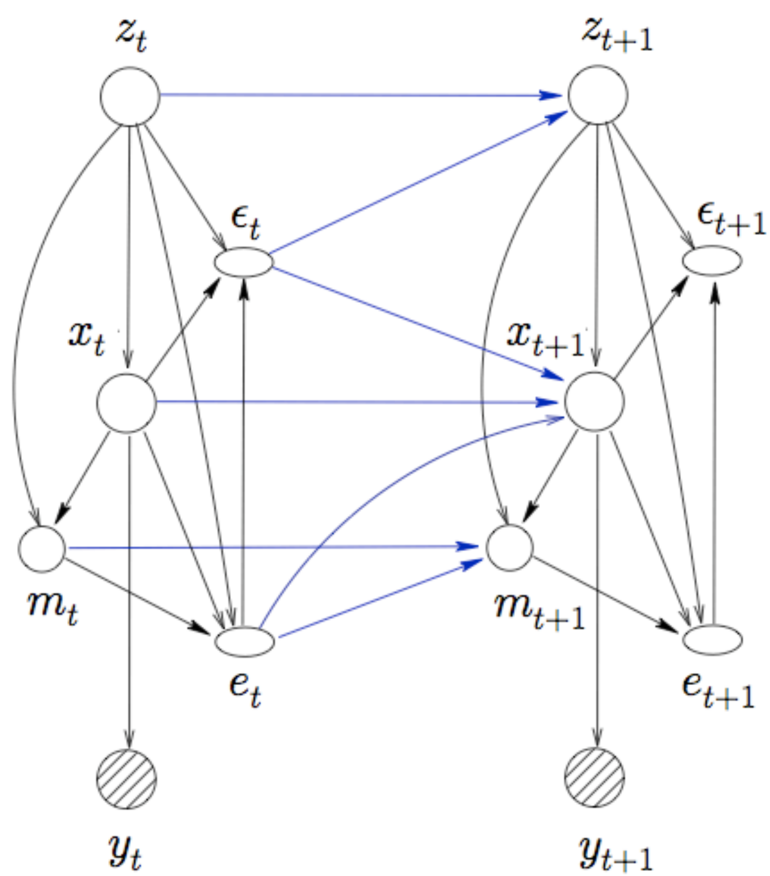
Problem:
How do we model
hierarchical activities?





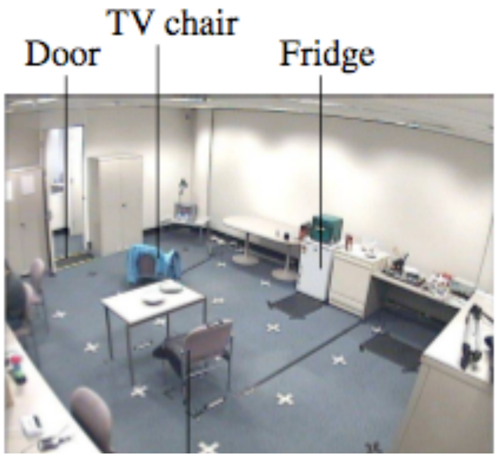
Solution:
"**stack**" actions for
hierarchical activities

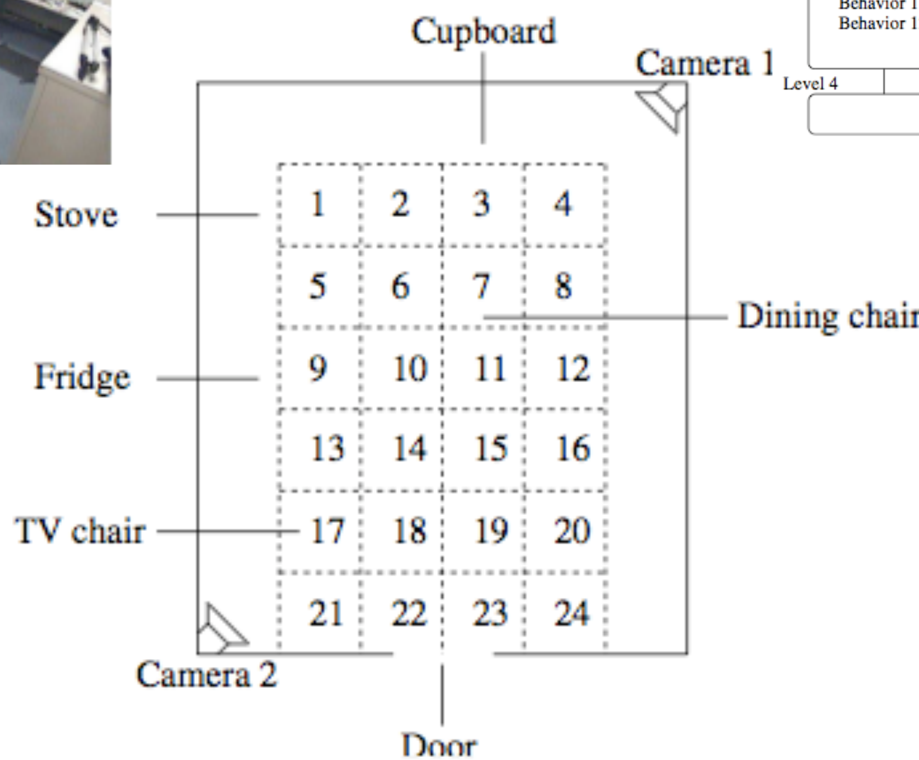
Hierarchical hidden Markov model

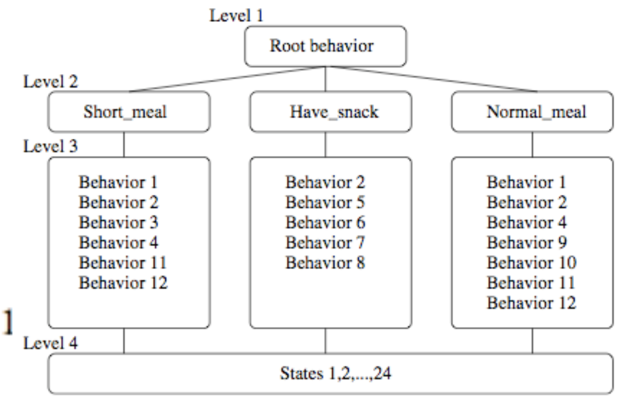


Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models. Nguyen et al 2005

Context-free activity grammar

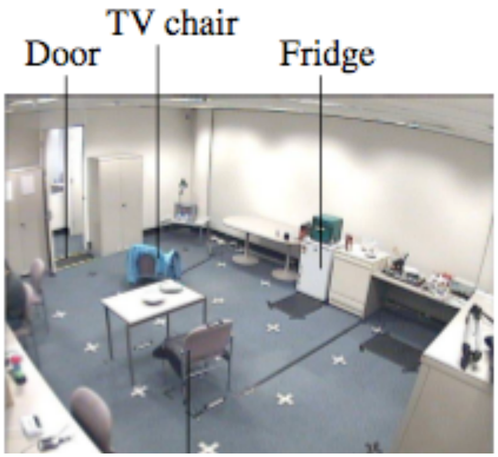


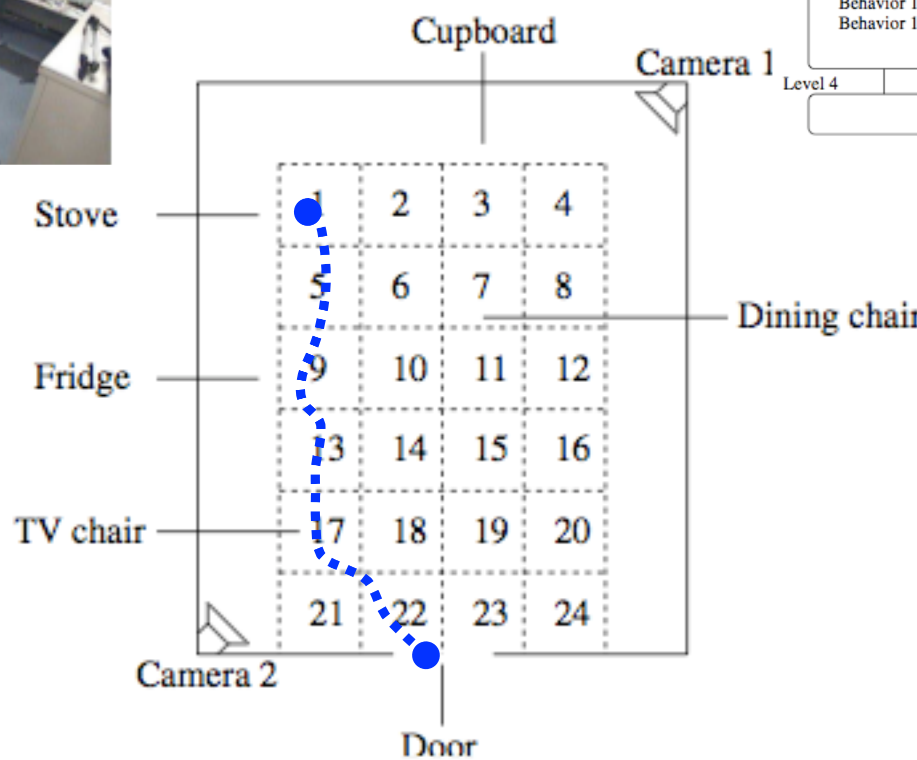


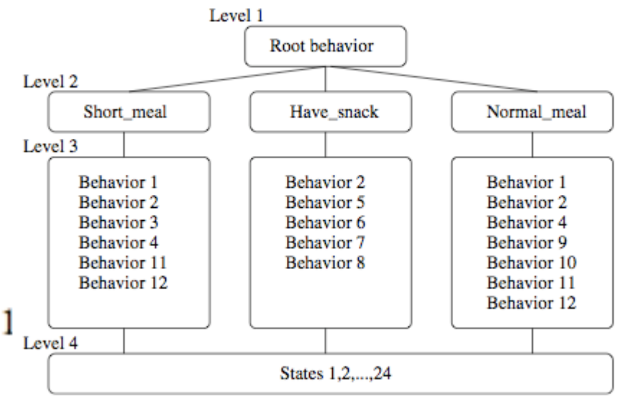


Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models. Nguyen et al 2005

Context-free activity grammar







Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models. Nguyen et al 2005

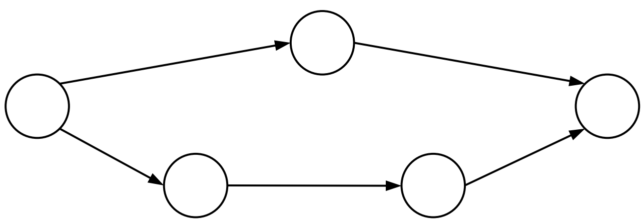
Observations

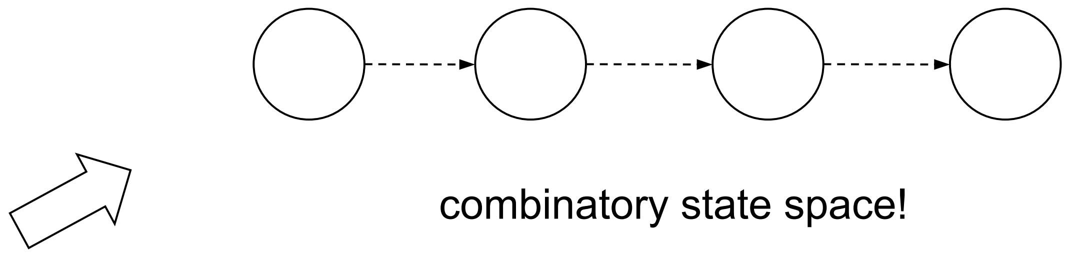
- Tree structures useful for hierarchies
- Tight integration of trajectories with abstract semantic states

- Activities are not always a single sequence
(ie. they sometimes happen in parallel)

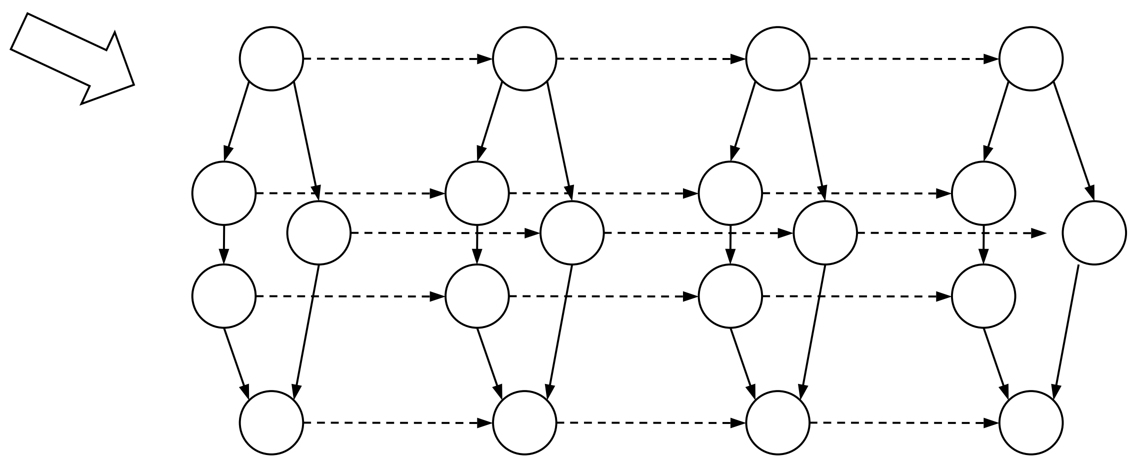
Concurrent activities

Problem:
How do we model
<u>concurrent</u> activities?



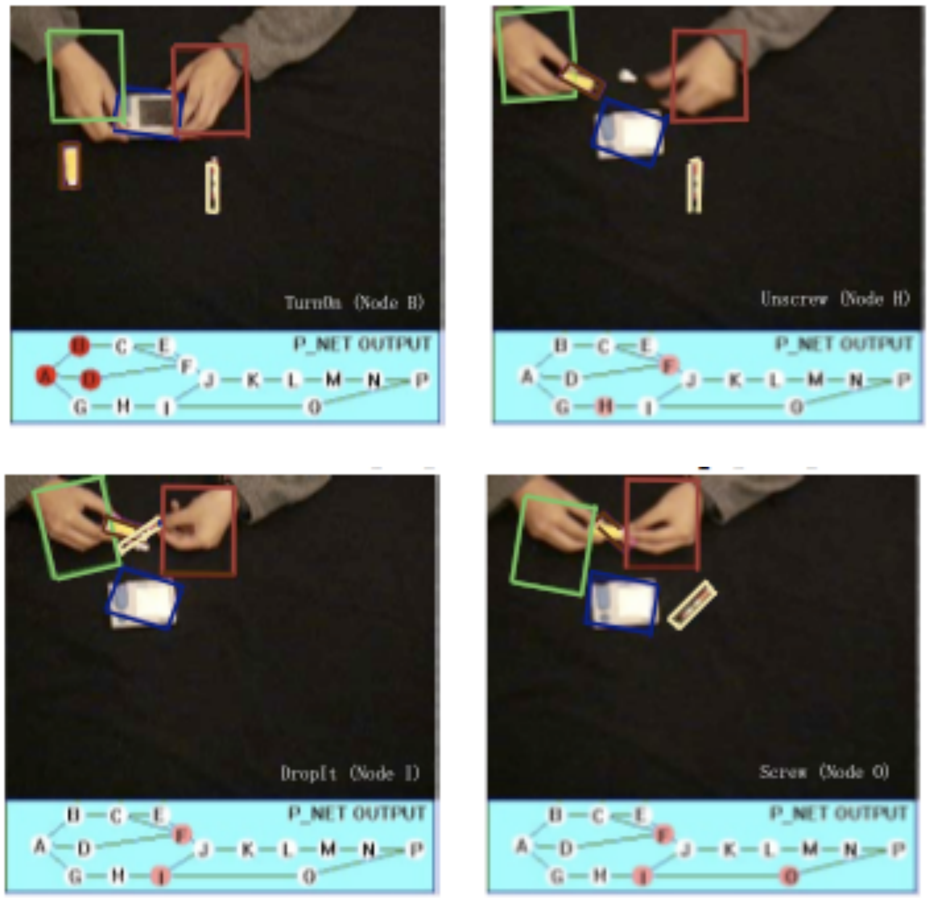


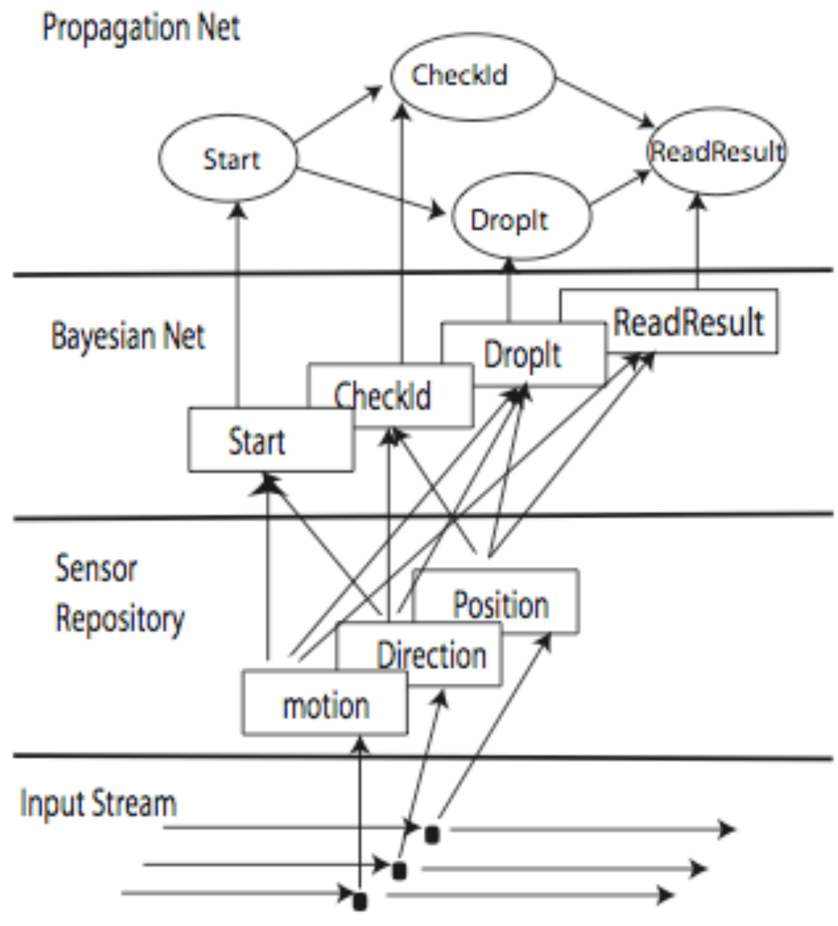
combinatory state space!



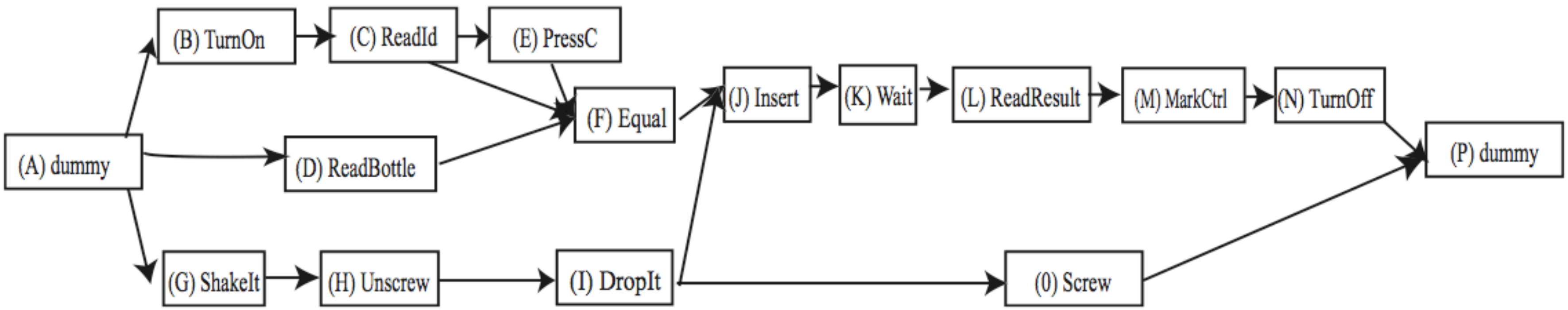
Solution:
"**stand-up**" model for
concurrent activities

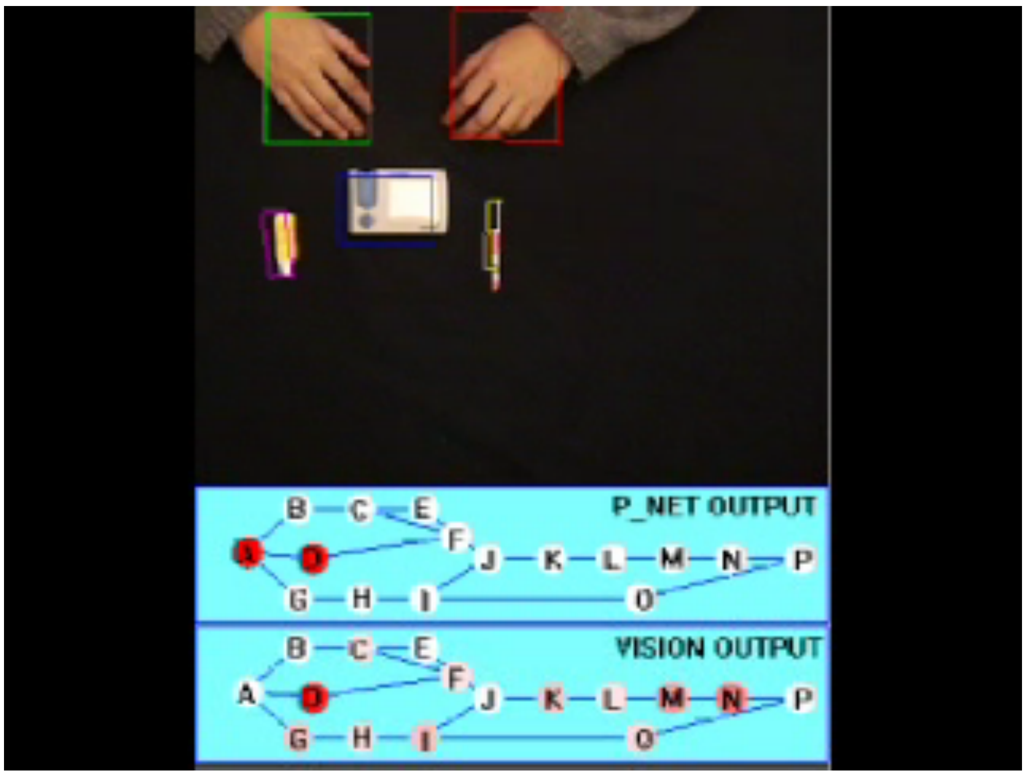
Propagation network





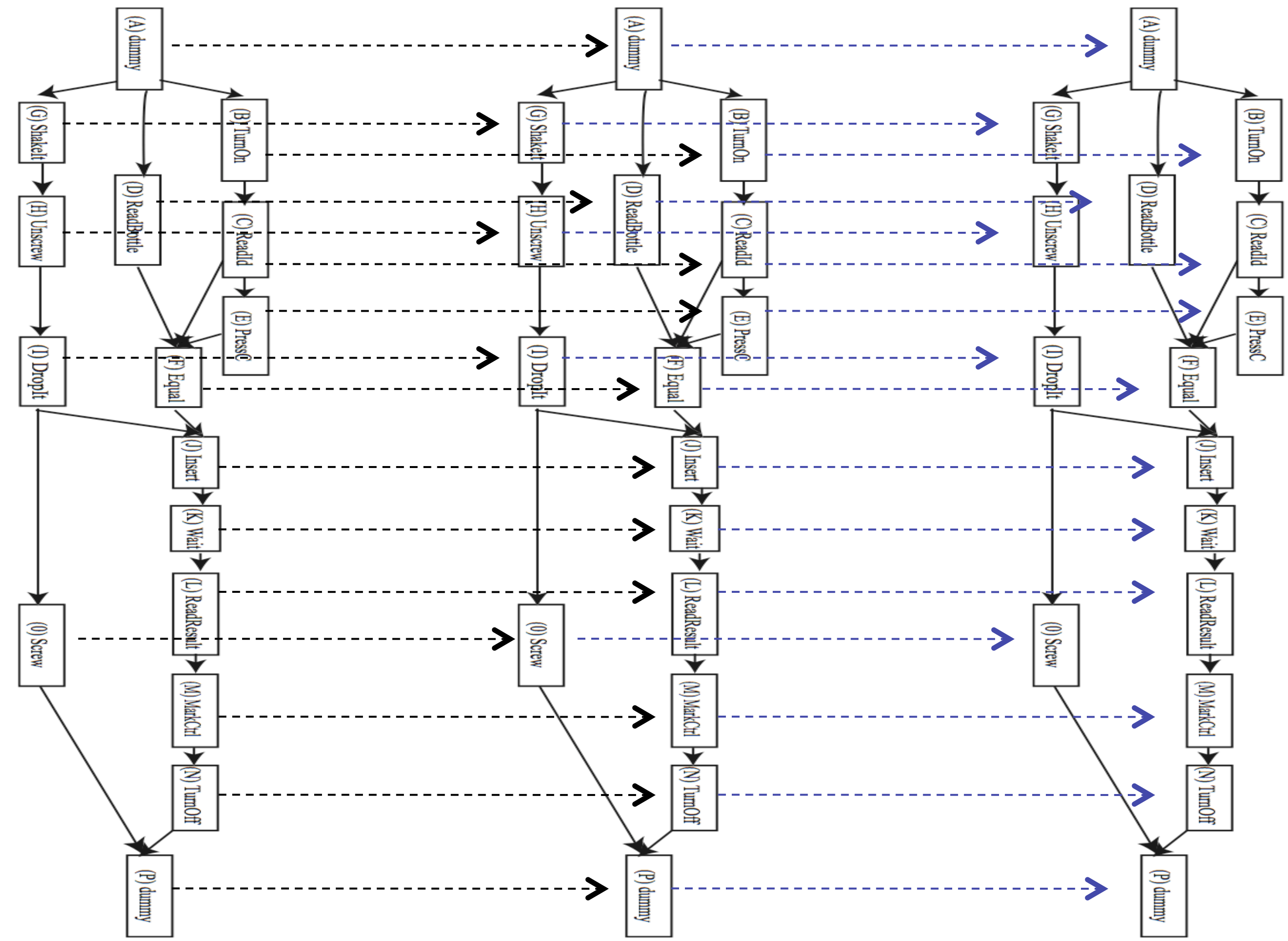
Propagation Networks for Recognition of Partially Ordered Sequential Action. Shi et al 2004





Propagation Networks for Recognition of Partially Ordered Sequential Action. Shi et al 2004

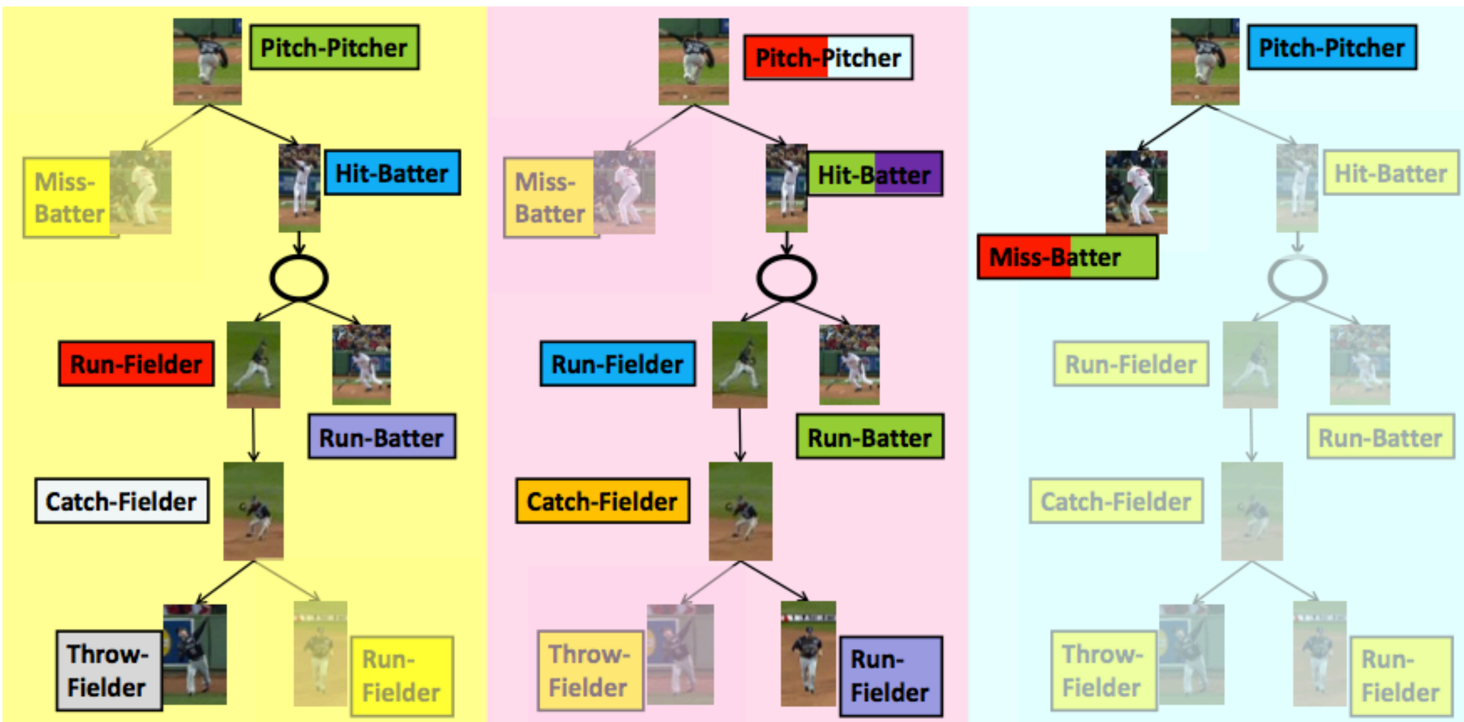
temporal inference



Inference by standing the state transition model on its side

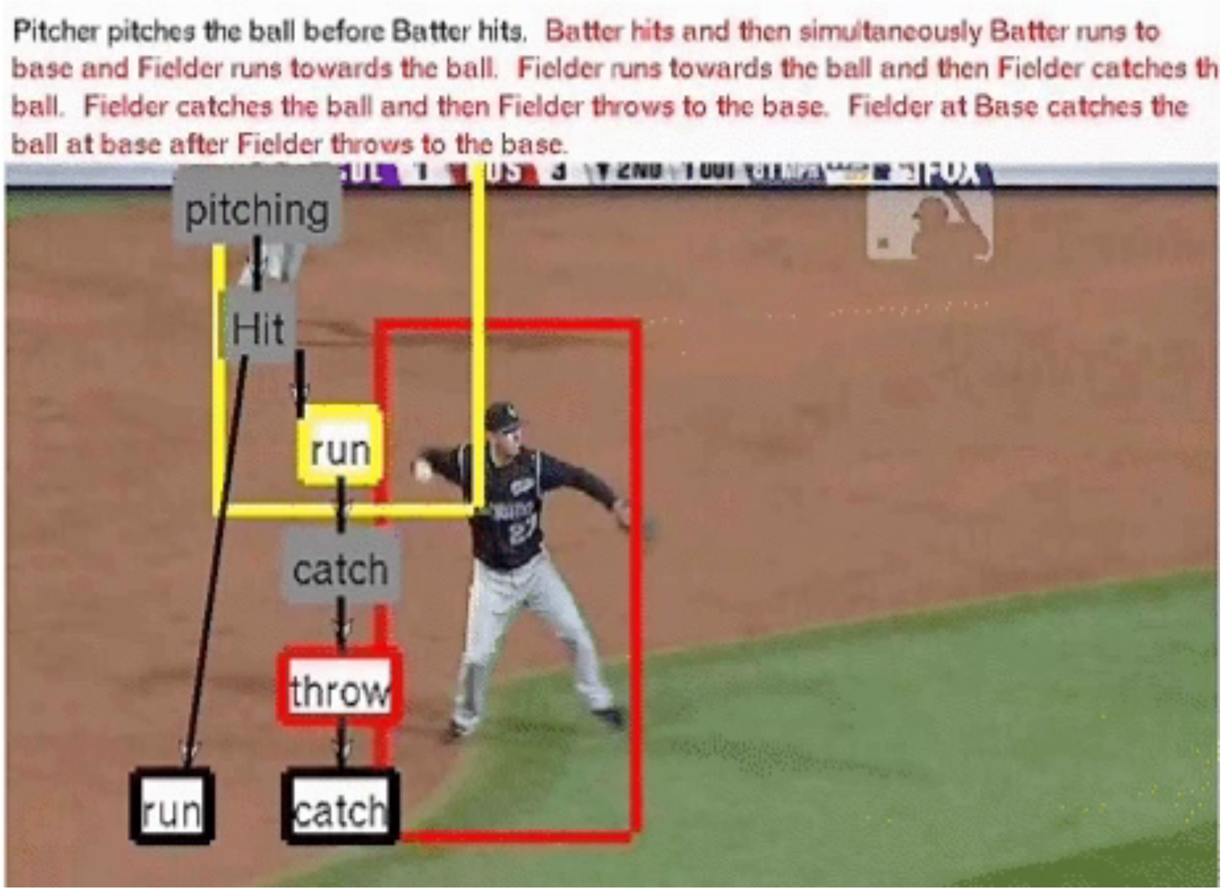
Inferring structure (storylines)

Understanding Videos, Constructing Plots –
Learning a Visually Grounded Storyline Model from Annotated Videos
Gupta, Srinivasan, Shi and Davis CVPR 2009



Learn AND-OR graphs from weakly labeled data

Scripts from structure



Understanding Videos, Constructing Plots - Learning a Visually Grounded Storyline Model from Annotated Videos.
Gupta, Srinivasan, Shi and Davis CVPR 2009

Take home message

Hierarchical statistical model

- Use when
  - Low-level action detectors are noisy
  - Structure of activity is sequential
  - Integrating dynamics
- Not for
  - Activities with deep hierarchical structure
  - Activities with complex temporal structure

Contrasting hierarchical approaches

	Actions as:	Activities as:	Model	Characteristic
Statistic	probabilistic states	paths	DBN	Robust to uncertainty
Syntactic	discrete symbols	strings	CFG	Describes deep hierarchy
Descriptive	logical relationships	sets	CFG, MLN	Encodes complex logic

References

(not included in ACM survey paper)

- W. Tsai and K.S. Fu. Attributed Grammar-A Tool for Combining Syntactic and Statistical Approaches to Pattern Recognition. SMC1980.
- M. Brand. The "Inverse Hollywood Problem": From video to scripts and storyboards via causal analysis. AAAI 1997.
- T. Wang, H. Shum, Y. Xu, N. Zheng. Unsupervised Analysis of Human Gestures. PRCM 2001.
- C.G. Nevill-Manning, I.H. Witten. On-Line and Off-Line Heuristics for Inferring Hierarchies of Repetitions in Sequences. IEEE 2000.
- K. Heller, Y.W. Teh and D. Gorur. Infinite Hierarchical Hidden Markov Model s. AISTATS 2009.
- P. Liang, S. Petrov, M. Jordan, D. Klein. The Infinite PCFG using Hierarchical Dirichlet Processes. EMNLP 2007.
- A. Gupta, N. Srinivasan, J.Shi and L.Davis. Understanding Videos, Constructing Plots - Learning a Visually Grounded Storyline Model from Annotated Videos. CVPR 2009.