The 5th Advanced NLP Summer Camp

"*Grounded Language Learning from Video Described with Sentences*"

Haonan Yu and Jeffrey Mark Siskind

ACL2013

Daichi Mochihashi

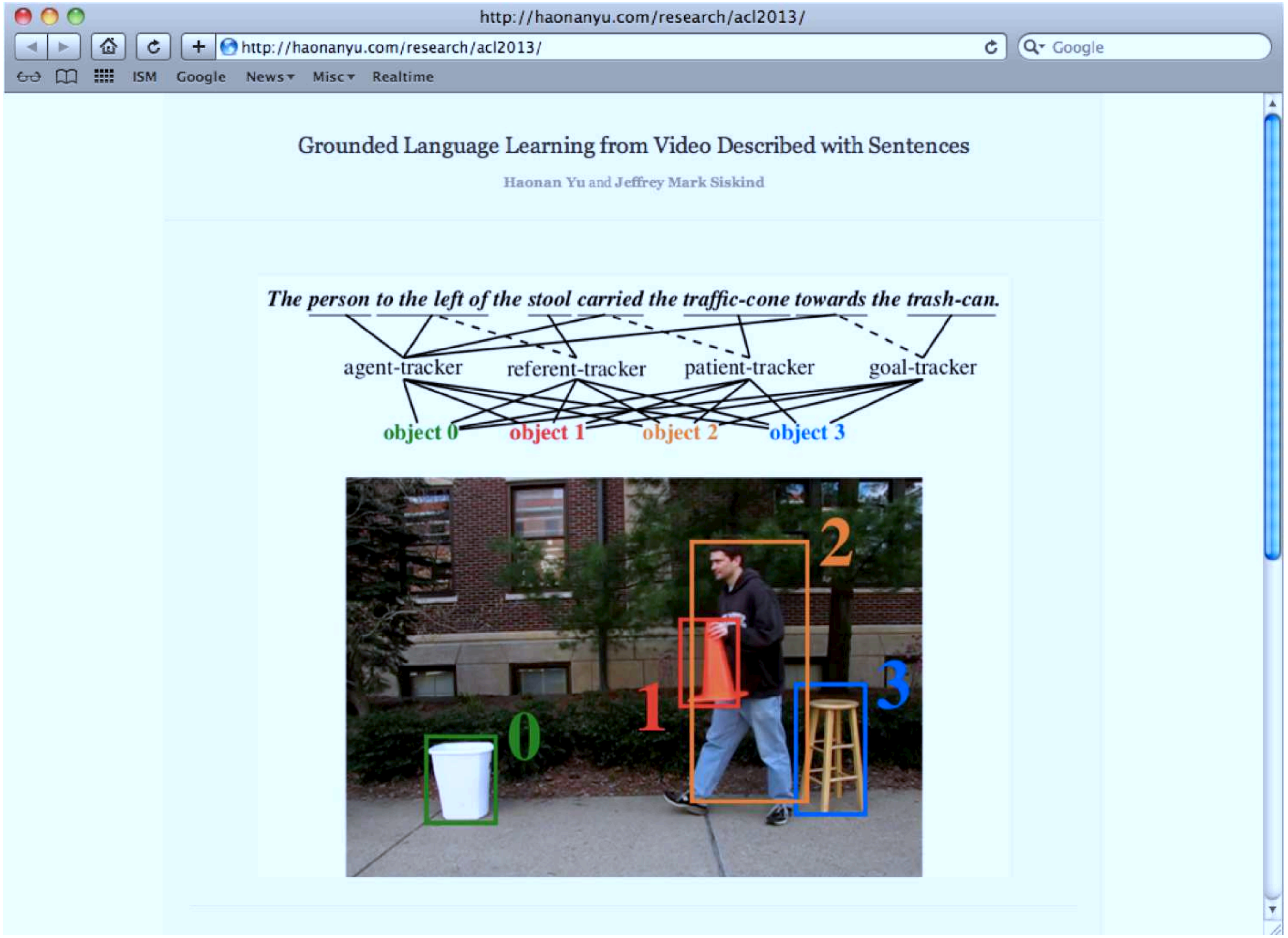
daichi@ism.ac.jp

The Institute of Statistical Mathematics

2013-8-31(Sat), Kujukuri-hama, Japan

About this paper

- ACL 2013 Best Paper Award
- http://haonanyu.com/research/acl2013/ provides a paper, slides, all codes and data



About the author

- Jeffrey Siskind: famous for his extraordinary optimized scheme compiler “Stalin”
  - https://engineering.purdue.edu/~qobi/software.html
- As a researcher, he pursues grounded language learning from 90s
- This paper is an extension to Barbu&Siskind (2012) with sentences
  - Fundamentals: “Recognizing Human Action in Time-Sequential Images using Hidden Markov Model”, Junji Yamato, Jun Otani, Kenichiro Ishii (NTT), CVPR 92 (citation 976!)

Objectives of this paper

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

Objectives of this paper

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

Objectives of this paper

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

Objectives of this paper

What Children Learn From

The person picked up the traffic-cone.
The person picked up the traffic-cone to the left of the stool.
The person put down the trash-can quickly.

The person carried the chair.
The person carried the backpack.
The chair approached the backpack.

Objectives of this paper

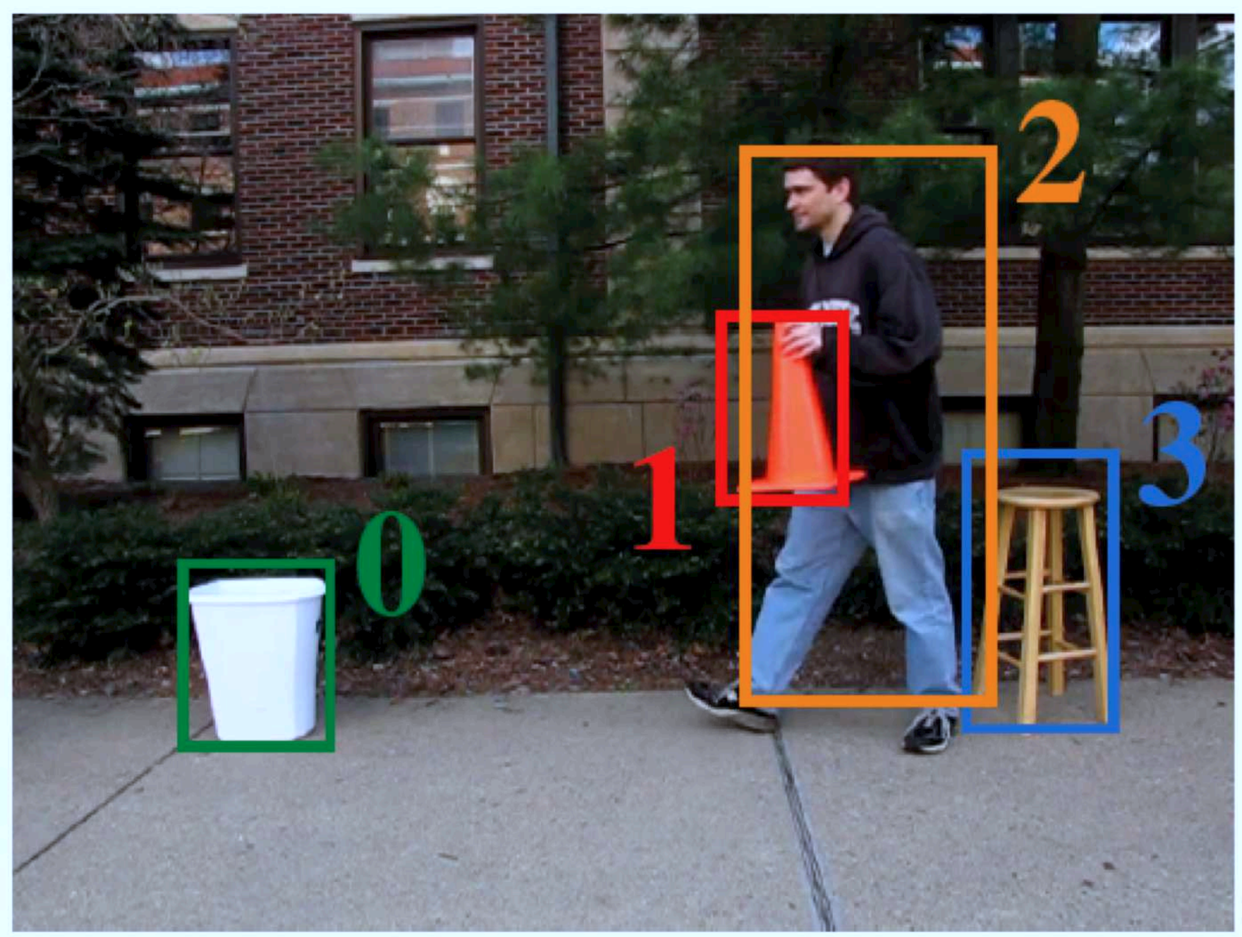
- From the set of {video,sentence} pairs, we will learn
  - HMM for the “meaning” of each word
    - Actual state trajectory
    - Emission distribution, State transition matrix
- .. almost automatically.
- Foundation of this model:
  Factorial HMM (Ghahramani and Jordan 1995)

Notice

- Original paper is very difficult to understand.
  - Unintuitive notations
  - No intuitive figure of the model
- See this slides, and draw a picture by yourself!

Meanings as HMM

- This paper uses a fixed vocabulary:
  person, backpack, trash-can, chair, traffic-cone, stool
- Imagine we already track regions in a video as objects:



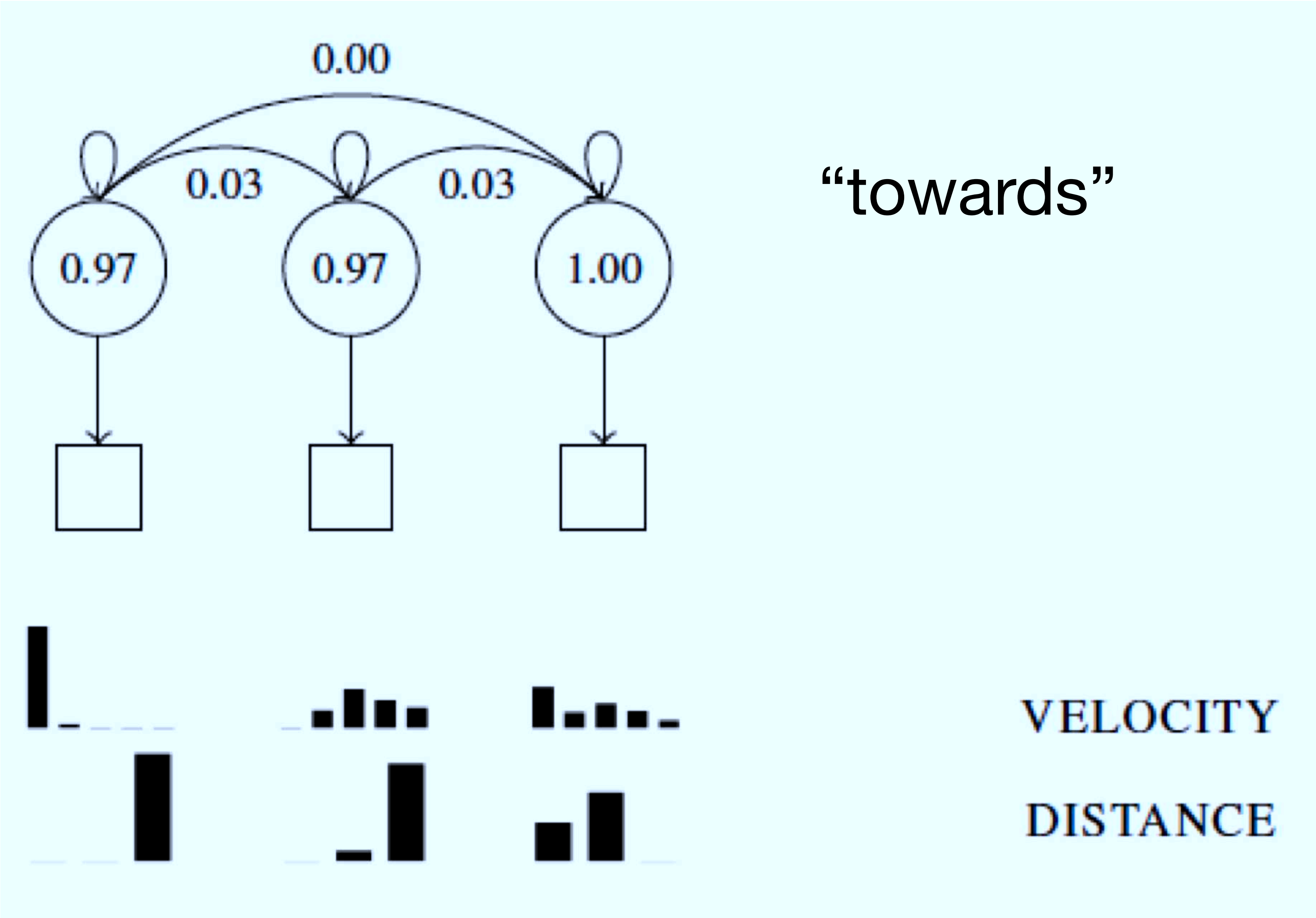
Meanings as HMM

- Each region has features (=outputs) like
  - Velocity
  - Movement direction
  - Color
  - X-coordinate, Y-coordinate
  - Size
- Then,

Meanings as HMM

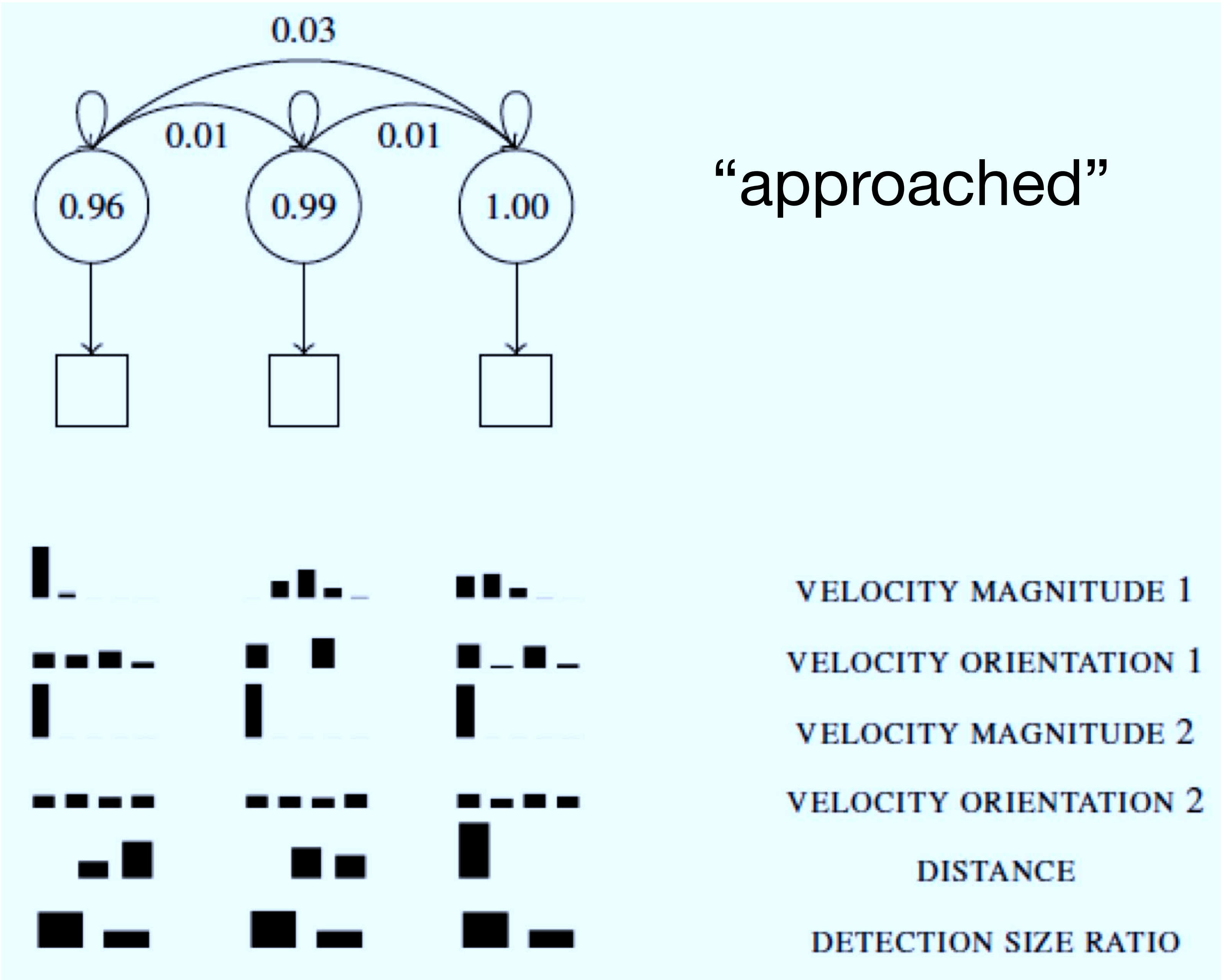
- "*jump*" is a 2-state HMM over velocity-direction
- "*pick up*" is a 2-state HMM over two objects features, like distance and y-coordinates
- "*person*" is a 1-state HMM emitting image features (like some specific colors or textures)
- "*quickly*" is a 1-state HMM over the velocity of its argument

Meanings as HMM



"towards"

Meanings as HMM



The Problem

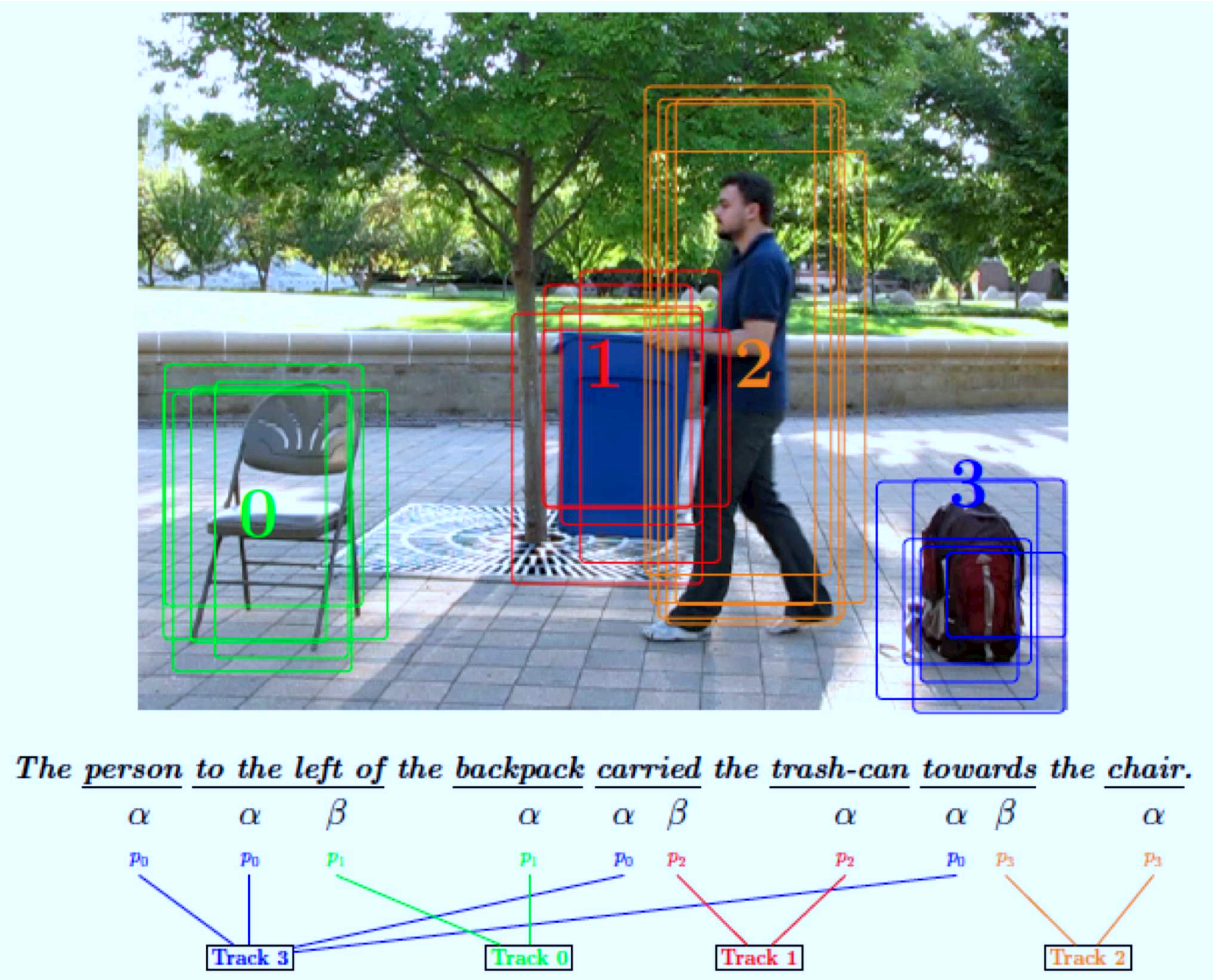
- Image regions are not aligned with words
  - And we do not know which region to select
- However, same words will appear in multiple videos

- Similar regions will be aligned to the same word
  - Word “dim” will be aligned to dark color region
  - Word “run” will be aligned to regions with high velocity
- How to optimize the correspondences?

Assumptions

- Each sentence caption is generated from a known CFG, thus we can parse the sentence
- We know the arity of each word (eg. carry(α, β))

- We can do a shallow “semantic parsing” of a sentence
- However, we don’t know what the “object” corresponds to!
- The number of objects in a video is known (4 for the next slide)


0
1
2
3
The person to the left of the backpack carried the trash-can towards the chair.
α α β α α β α α β α
p0 p0 p1 p1 p0 p2 p2 p0 p3 p3
Track 3
Track 0
Track 1
Track 2

Assumed Grammar

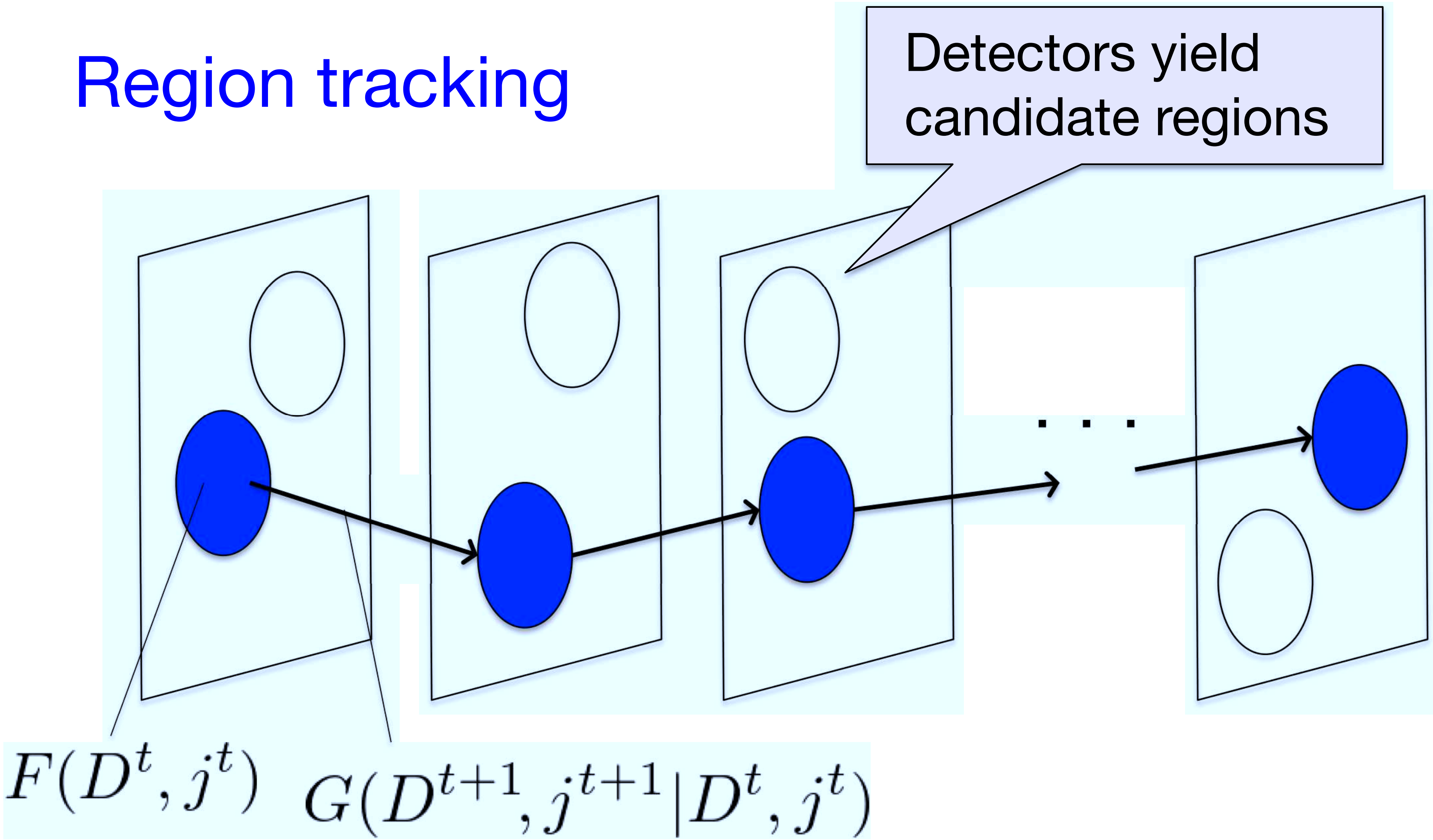
$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D N [PP]} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{VP} &\rightarrow \text{V NP [ADV] [PPM]} \\
\text{PPM} &\rightarrow \text{PM NP} \\
\text{D} &\rightarrow \textit{the} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{trash-can} \mid \textit{chair} \mid \textit{traffic-cone} \mid \textit{stool} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{V} &\rightarrow \textit{picked up} \mid \textit{put down} \mid \textit{carried} \mid \textit{approached} \\
\text{ADV} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PM} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

- model other words also as HMMs

  *The **jump** was fast.* (Some nouns are dynamic.)

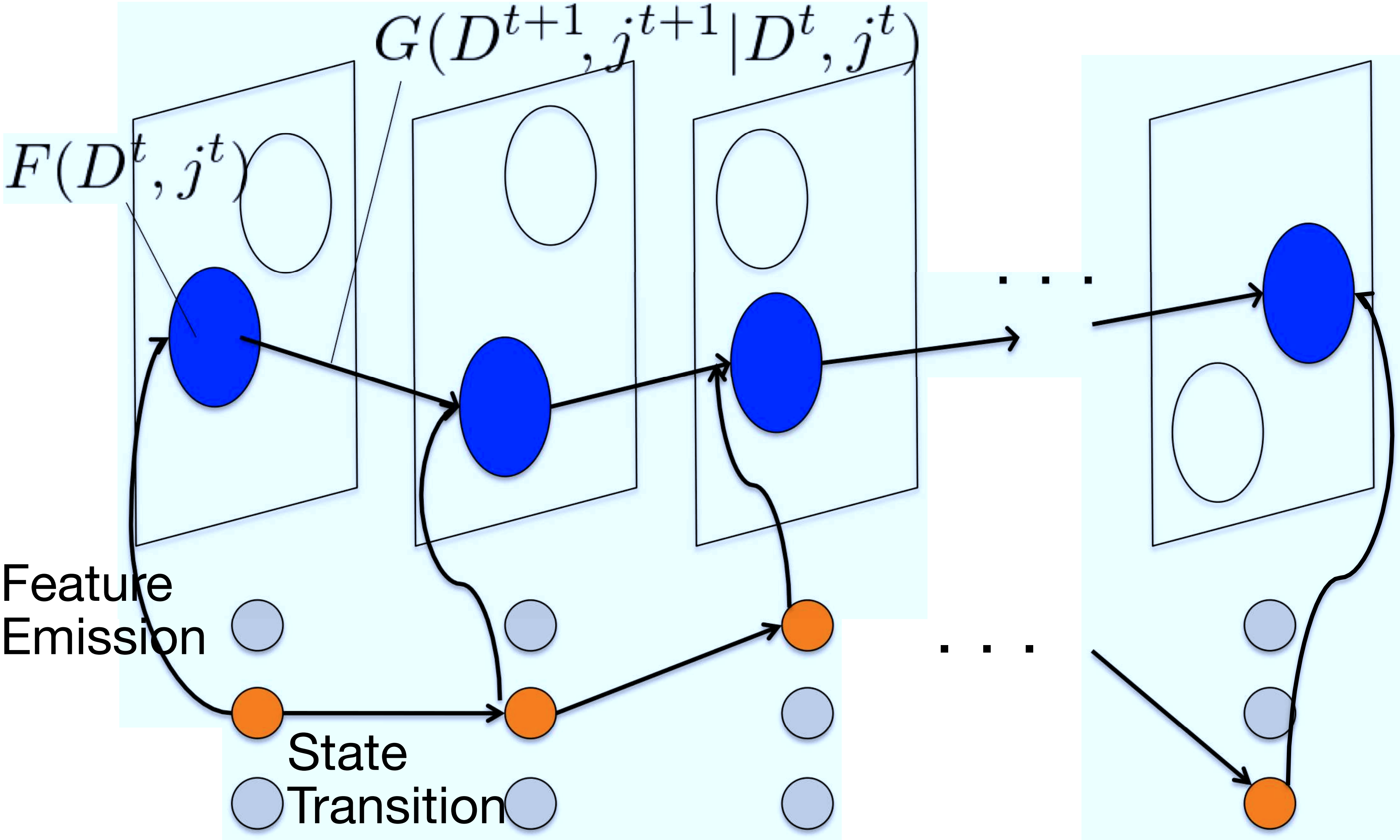
  *The person **held** the backpack.* (Some verbs are static.)

Region tracking



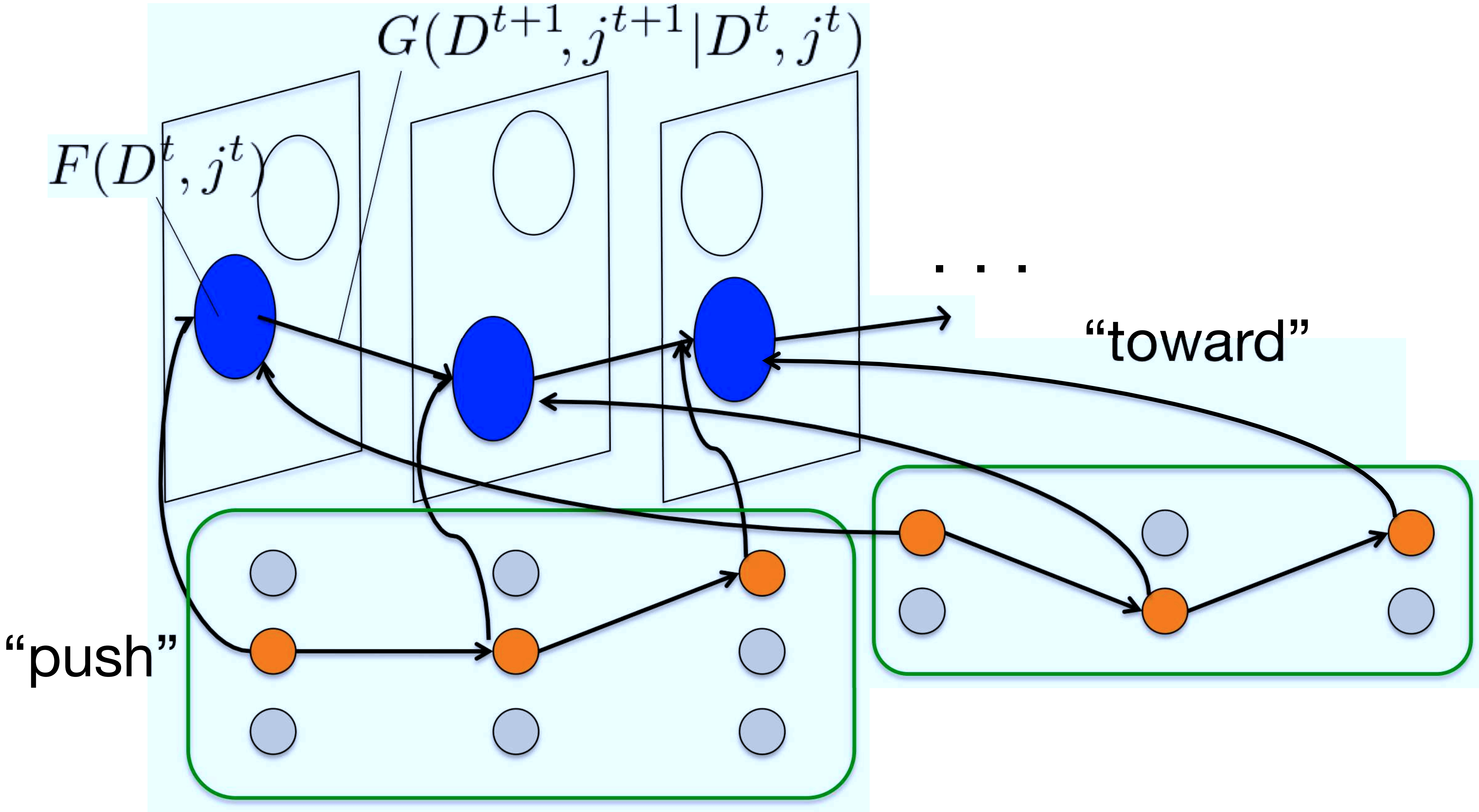
- Regions evolve like a HMM
- Each frame has a “correct” region for object #i

Region tracking+Word HMM



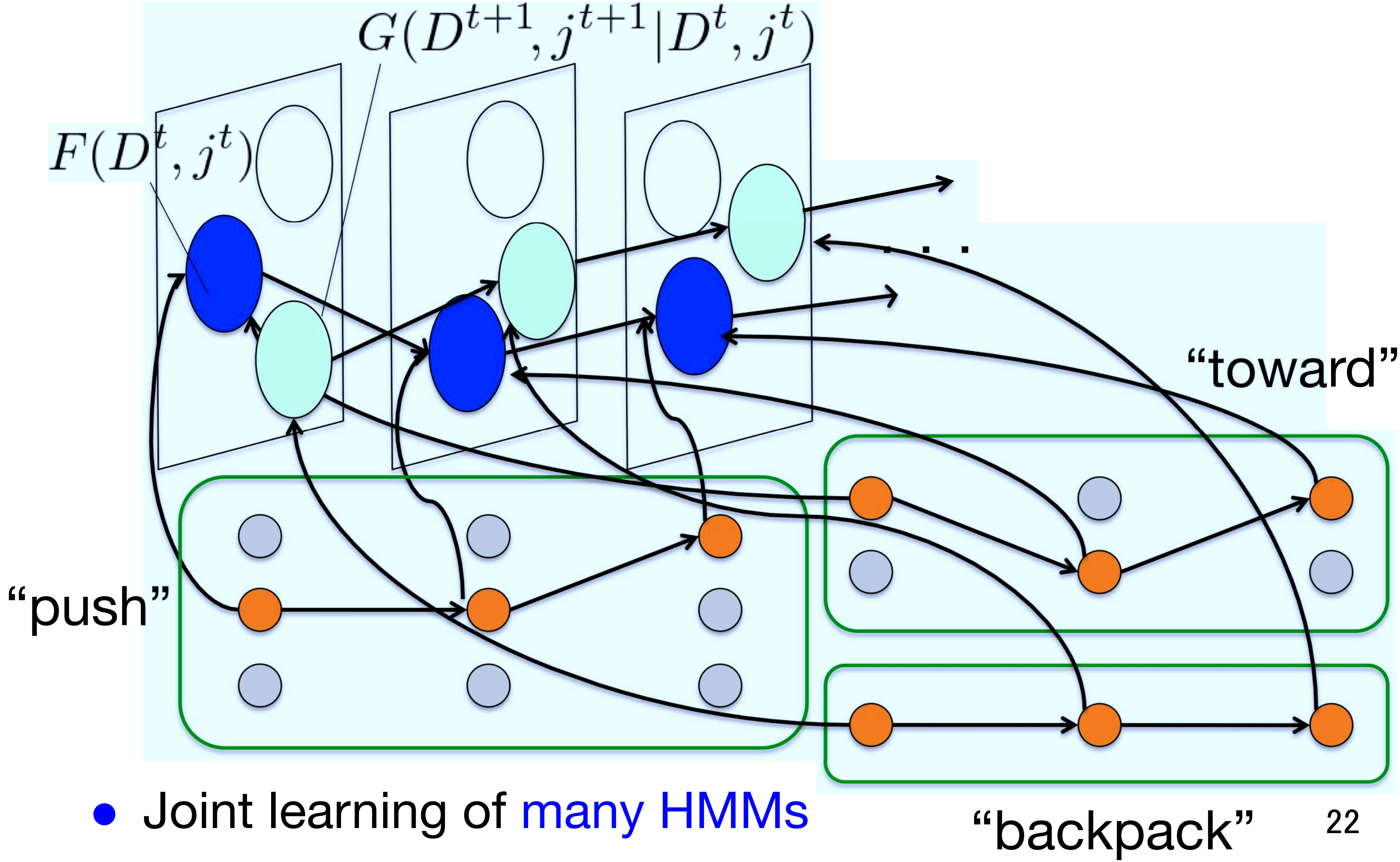
- Joint learning of two HMMs (product of HMMs)

Region tracking+Word HMMs



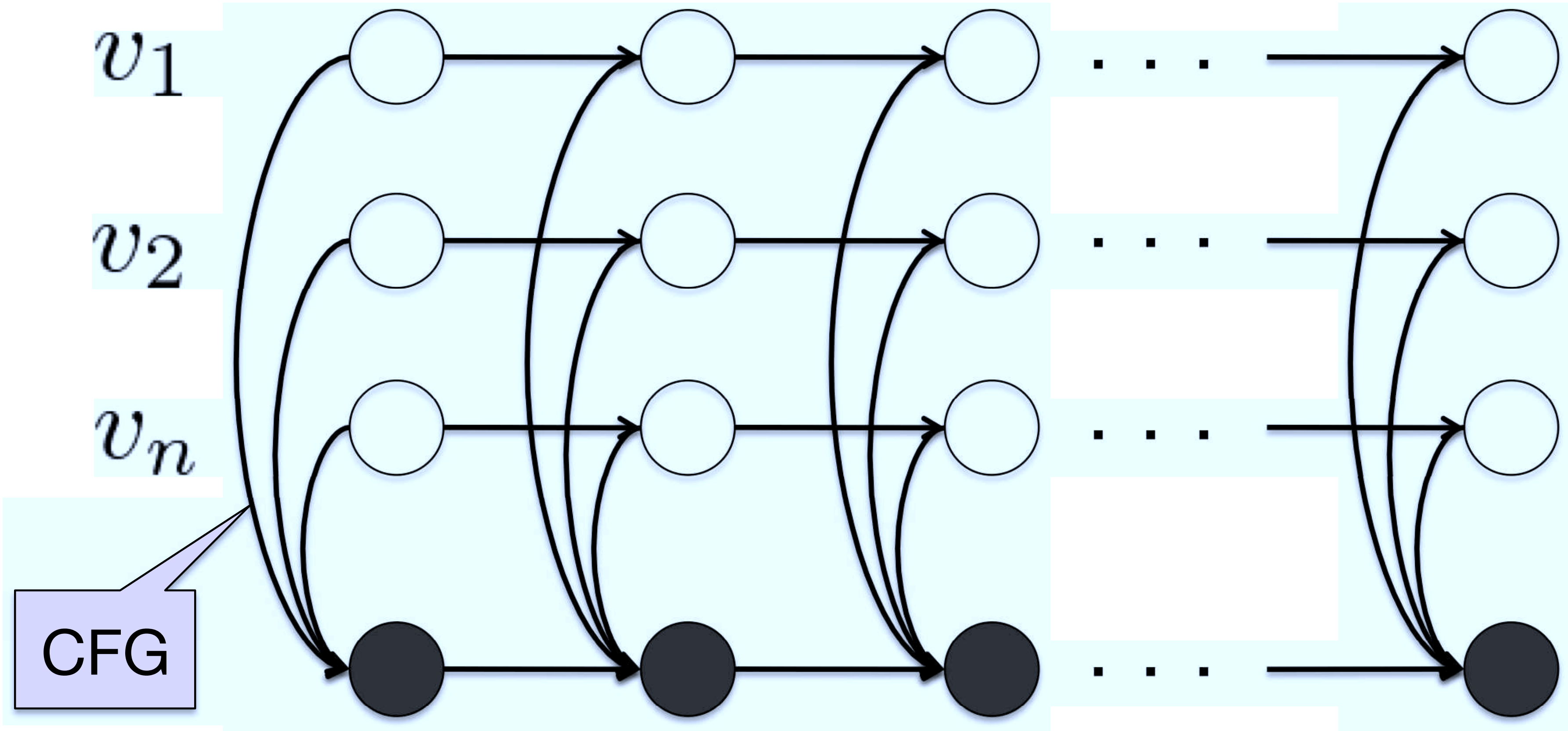
- Joint learning of many HMMs

Region trackings+Word HMMs



- Joint learning of many HMMs

Graphical Model as FHMM



- Sentence tracker can be described as a Factorial HMM (FHMM) (autoregressive FHMM)
  - CFG partly determines the portion of the output

EM in the Sentence Tracker

$$\log \sum_{\substack{j_1^1,\ldots,j_1^T \\ j_L^1,\ldots,j_L^T}} \sum_{\substack{k_1^1,\ldots,k_1^T \\ k_W^1,\ldots,k_W^T}} \exp \left[\begin{array}{l} \sum_{l=1}^{L} \sum_{t=1}^{T} f(b_{j_l^t}^t) + \sum_{t=2}^{T} g(b_{j_l^{t-1}}^{t-1}, b_{j_l^t}^t) + \\ \sum_{w=1}^{W} \sum_{t=1}^{T} h_w(k_w^t, b_{j_{\theta_w^1}^t}^t, b_{j_{\theta_w^2}^t}^t) + \sum_{t=2}^{T} a_w(k_w^{t-1}, k_w^t) \end{array} \right]$$

- Wrap the sum of log likelihoods of all video-sentence pairs in EM.
- In the E-step, compute probability for tracks, HMM states, and outputs.
- In the M-step, the transition matrix $a_w(k_w^{t-1}, k_w^t)$ and output distribution $h_w(k_w^t, b_{j_{\theta^1}^t}^t, b_{j_{\theta^2}^t}^t)$ are re-estimated.

HMM reestimation formula

$a_{kl}^{(v)}$: k→l transition probability of HMM of word v

$b_{kj}^{(v)}$: k→j feature emisson probability of HMM of word v

$$\begin{cases} a_{kl}^{(v)} \propto \sum_i \sum_n \sum_t \dfrac{p(q_{in}^{(t)} = k, q_{in}^{(t-1)} = l, D_i | \mathbf{w}_i, \lambda)}{p(D_i | \mathbf{w}_i, \lambda)} \mathbb{I}(w_{in} = v) \\ b_{kj}^{(v)} \propto \sum_i \sum_n \sum_t \dfrac{p(q_{in}^{(t)} = k, x_{in}^{(j)} = h, D_i | \mathbf{w}_i, \lambda)}{p(D_i | \mathbf{w}_i, \lambda)} \mathbb{I}(w_{in} = v) \end{cases}$$

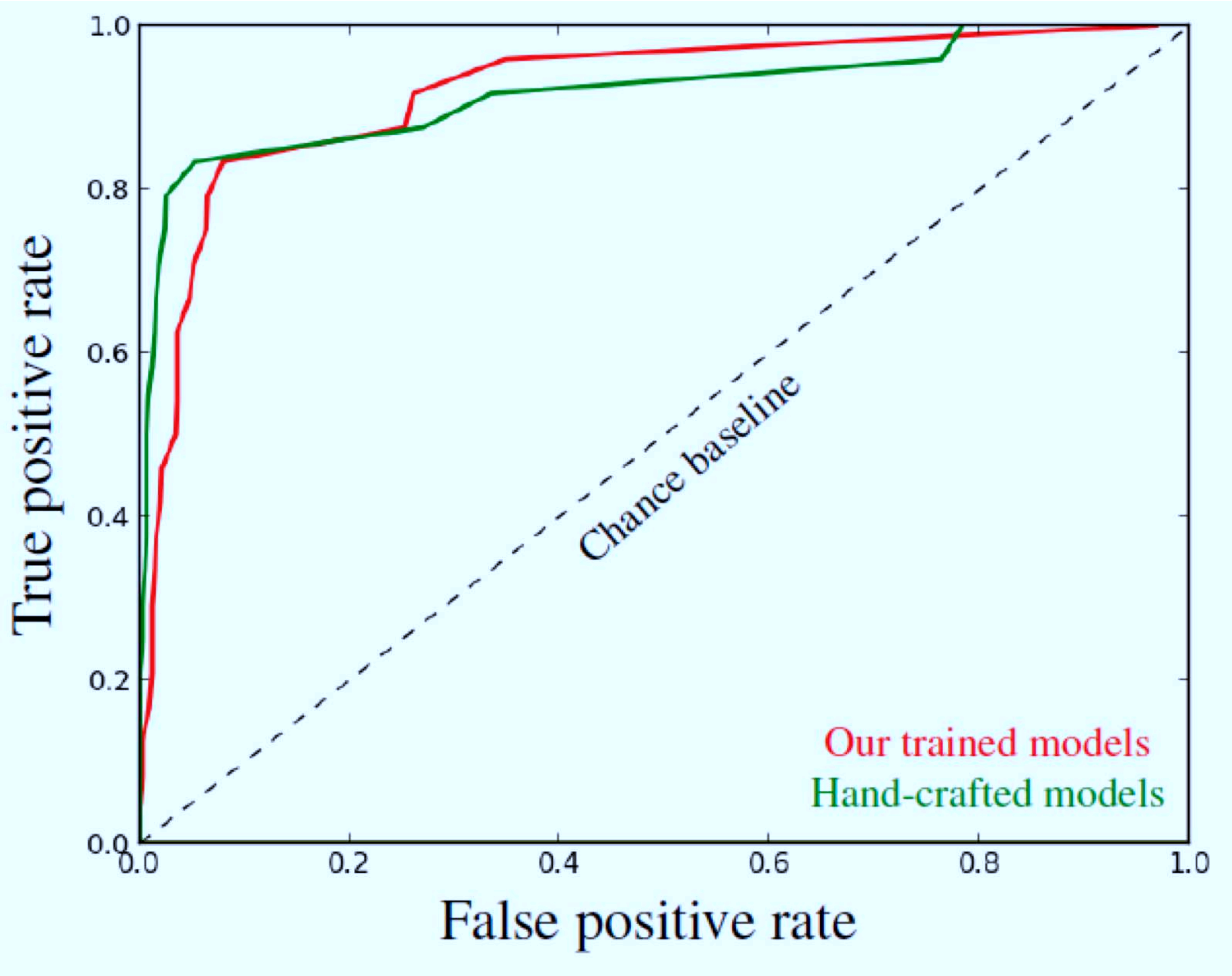
- Each term in the numerator is calculated from a standard forward-backward in HMM (each HMM in turn)

Experiments

- learn all content words in the lexicon
- 95 video clips, each video clip contains 1 person + 2 or 3 objects
- about 200 training video-sentence pairs + 240 test video-sentence pairs
- test on videos/sentences *never seen* in training set

The person to the left of the stool picked up the chair. *The person carried the backpack towards the stool.*

Experimental results



- Yielded same performance as hand-crafted models with no supervision
- Very similar model to the hand-crafted one is obtained

Summary

- Modeling “meaning” of a word by a HMM
  - Representing time series of features associated with that word
  - Strong representational power
    - Dynamic noun (eg. “jump”)
    - Static verb (eg. “hold”)
- Sentence Tracker=Factorial HMM
  - Choosing appropriate image subregions
  - Deeply nested EM, forward-backward
- Equivalent performance with a hand-crafted model
- Very complicated but interesting!