Seeing What You're Told: Sentence-Guided Activity Recognition In Video

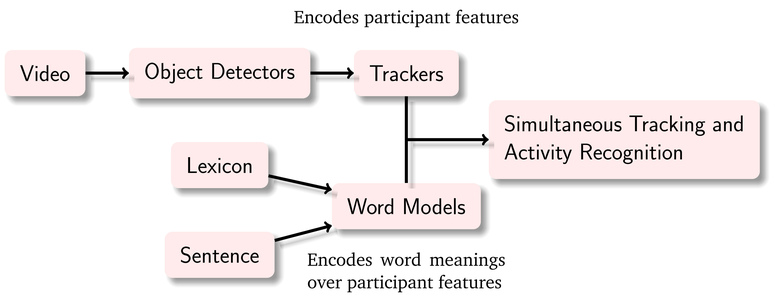
N. Siddharth, Andrei Barbu, & Jeffrey Mark Siskind

Stanford University, MIT, Purdue University

What?

The person to the left of the chair put down the blue object.

Characteristics



- Joint evaluation of both Tracking and Activity Recognition
  - A Tracker for each participant in the activity
  - A Word Model for each lexical entry in the lexicon
- Integration of top-down sentential information *and* bottom-up tracker information
- Recognize different parts of speech

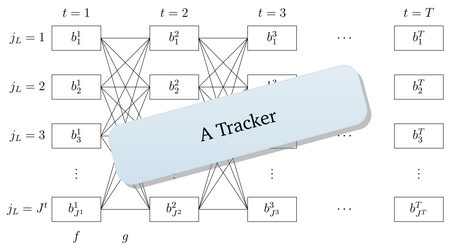
Verbs	Adverbs
Nouns	Adjectives
Determiners	Prepositions

- Sensitive to sentence structure

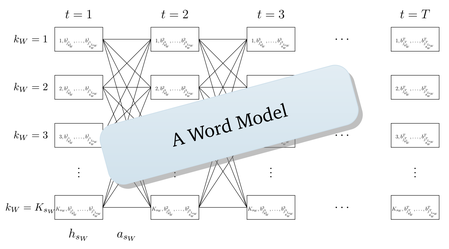
*The **person** approached the **object**.* $\not\equiv$ *The **object** approached the **person**.*

This research was supported, in part, by ARL, under Cooperative Agreement Number W911NF-10-2-0060, and the Center for Brains, Minds and Machines, funded by NSF STC award CCF-1231216. The views and conclusions contained in this document are those of the authors and do not represent the official policies, either express or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

How?



$$\sum_{t=1}^{T} \mathbf{f}(b^t_{j_l}) + \sum_{t=2}^{T} \mathbf{g}(b^{t-1}_{j_l}, b^t_{j_l})$$



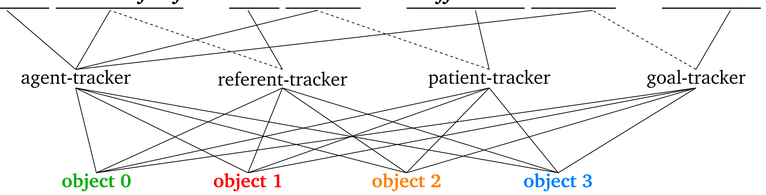
$$\sum_{t=1}^{T} \mathbf{h_w}(k^t_w, b^t_{j_{\theta^1_w}}, b^t_{j_{\theta^2_w}}) + \sum_{t=2}^{T} \mathbf{a_w}(k^{t-1}_w, k^t_w)$$

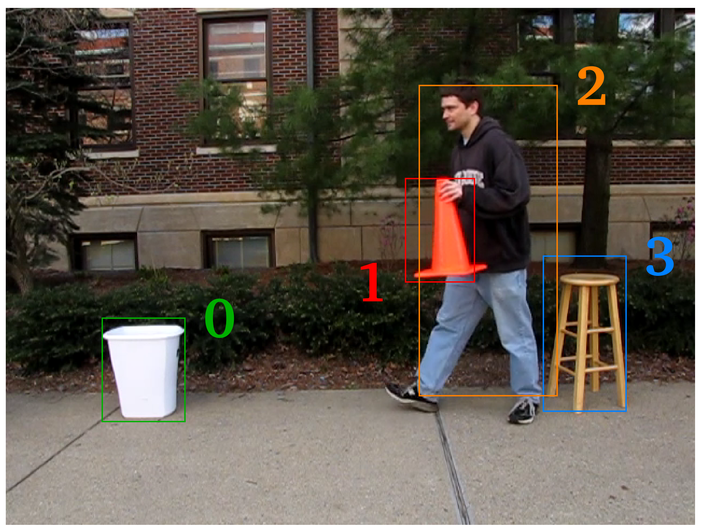
Joint

$$\max_{\substack{j^1_1,\ldots,j^T_1 \\ j^1_L,\ldots,j^T_L}} \max_{\substack{k^1_1,\ldots,k^T_1 \\ k^1_W,\ldots,k^T_W}} \sum_{l=1}^{L}\sum_{t=1}^{T} \mathbf{f}(b^t_{j_l}) + \sum_{t=2}^{T} \mathbf{g}(b^{t-1}_{j_l}, b^t_{j_l}) + \sum_{w=1}^{W}\sum_{t=1}^{T} \mathbf{h_w}(k^t_w, b^t_{j_{\theta^1_w}}, b^t_{j_{\theta^2_w}}) + \sum_{t=2}^{T} \mathbf{a_w}(k^{t-1}_w, k^t_w)$$

L tracks — W words

The person to the left of the stool carried the traffic-cone towards the trash-can.





$$\mathcal{S} : (\mathbf{B}, \mathbf{s}, \Lambda) \rightarrow (\tau, \mathbf{J})$$

$\mathbf{B}$ – detections, $\mathbf{s}$ – sentence, Λ – lexicon

τ – score, $\mathbf{J}$ – tracks

Look!

Performing disparate tasks simply by leveraging the framework differently

Sentential Focus of Attention

Same video | *Different* sentences → ***Different* tracks**

The person carried an object towards the trash can. *The person carried an object away from the trash can.*

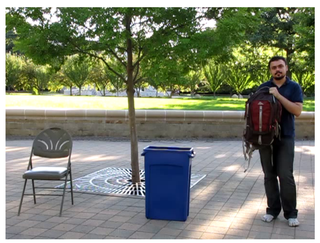
Semantic Video Retrieval

Many videos | A sentence → **That video**

The person carried the backpack away from the trash can.

Generation of Sentential Descriptions

Many sentences† | A video → **That sentence**



†

S → NP VP
NP → D [A] N [PP]
D → *an* | *the*
A → *blue* | *red*
N → *person* | *backpack* | *trash-can* | *chair* | *object*
PP → P NP
P → *to the left of* | *to the right of*
VP → V NP [ADV] [PPM]
V → *picked up* | *put down* | *carried* | *approached*
ADV → *quickly* | *slowly*
PPM → PM NP
PM → *towards* | *away from*

The person to the right of the trash-can picked up the backpack.

http://upplysingaoflun.ecn.purdue.edu/~qobi/cccp/

siddharth@iffsid.com
andrei@0xab.com
qobi@purdue.edu