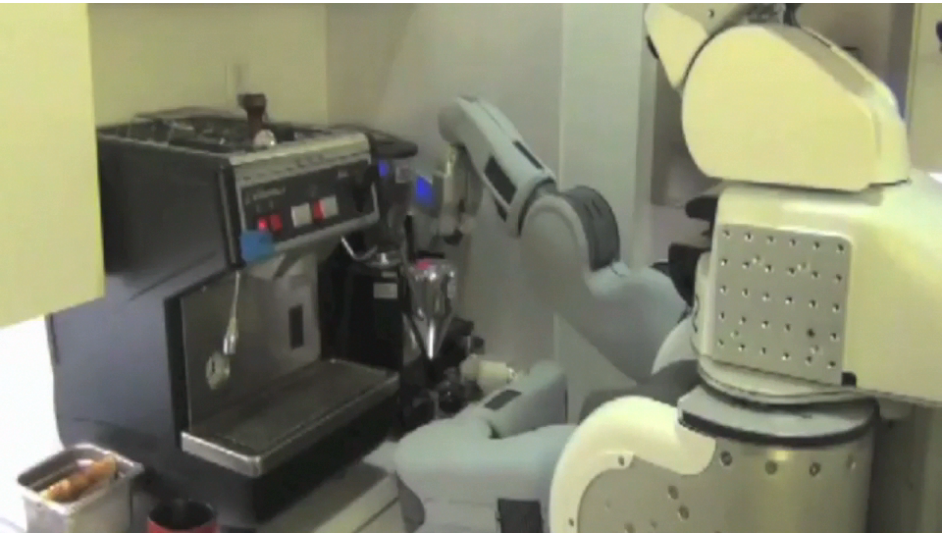
Unsupervised Discovery of Structure in Human Videos

Ozan Sener

Joint work with Ashutosh Saxena, Silvio Savarese, Ashesh Jain and Amir Zamir

Committee: Ashutosh Saxena, David Mimno, Emin Gun Sirer

We envision robots



courtesy of Sung et al.

doing human like activities

courtesy of Koppula et al.

while working with humans

It requires

understanding humans, their environments, objects and activities


What is he doing?
What is the next step?
How can I perform X activity?

Understanding Videos

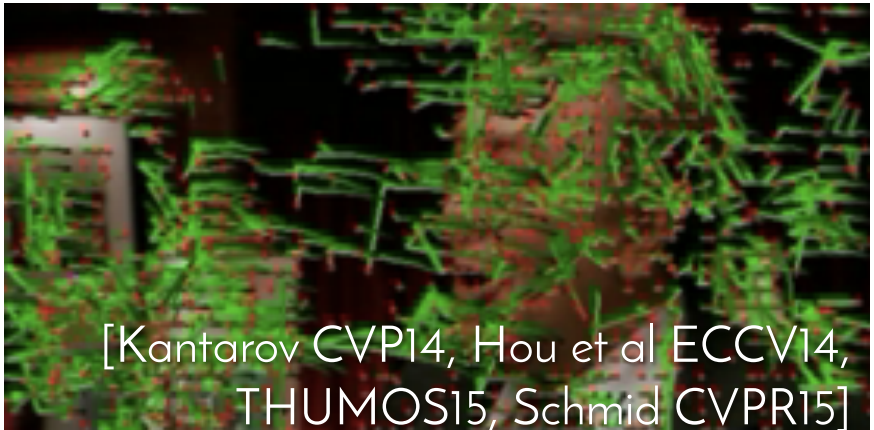
Image Centric

Video is an trivial extensions of images

[Koppula et al RSS,IJRR13; Wojek et al ECCV08; Sutton et al. JMLR08; Tian et al ICCV15...]

Video Centric

We need video specific features/models



[Kantarov CVP14, Hou et al ECCV14, THUMOS15, Schmid CVPR15]

Understanding Videos

Image Centric

Rich models like CRF

Easy to model context

Super linear in #of-frames

Hard to obtain supervision
(~10s activitiy, ~10s objects)

Video Centric

Scales linearly in #of-frames

Requires only frame labels

Does not model the context

Inefficient
(~30sec for ~1sec of vid)

Exclusively supervised

Hard to scale in #of-videos (~100s videos)

What we have?

~30sec to process 1sec of video

support ~10 activities

learning from ~100 videos

covering only indoor/sport environments

What we need?

real-time

any activity

learn from all available information

any environment

Discover, understand and share the underlying **semantic** structure of the **videos**.

Outline

Structured understanding of a single video

Large-scale understanding of video collections

Sharing knowledge to other domains and modalities

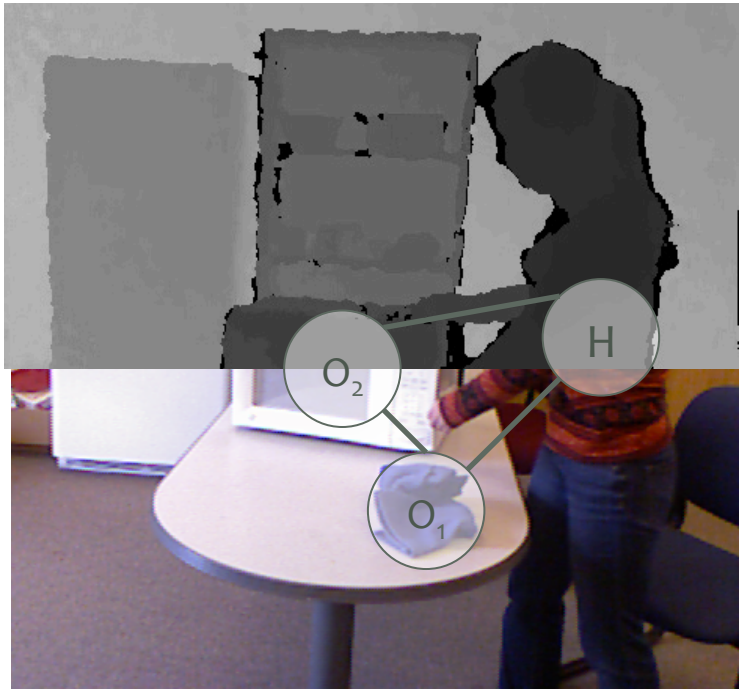
Outline

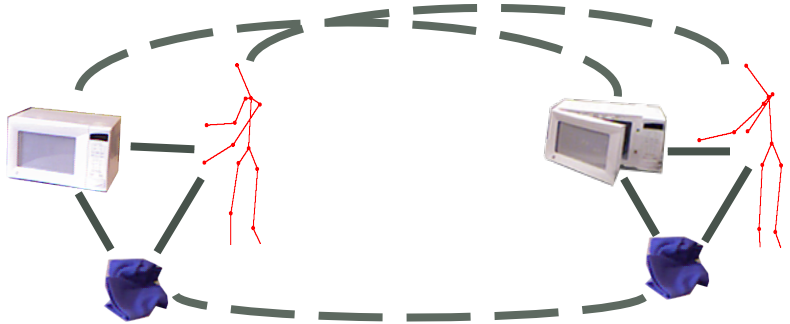
Structured understanding of a single video

Large-scaled understanding of video collections

Sharing knowledge to other domains and modalities

Revisit the Image Based Approach

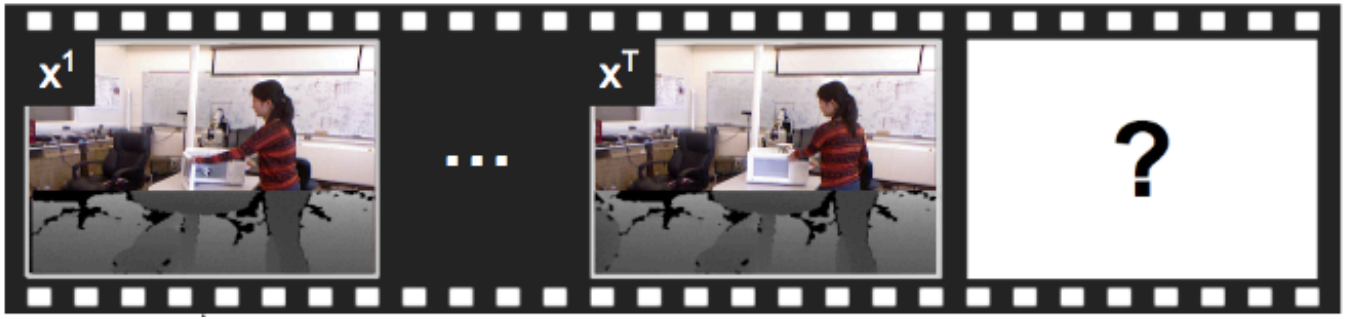




$$P(O_1^{1,\ldots,T}, O_2^{1,\ldots,T}, H^{1,\ldots,T} | \Phi_{O_1}^{1,\ldots,T}, \Phi_{O_2}^{1,\ldots,T}, \Phi_H^{1,\ldots,T}) \sim \exp\left(\sum_{v \in V} E(\Phi_v) + \sum_{v,w \in \mathcal{E}} E(\Phi_v, \Phi_w)\right)$$

Context of humans and objects are successfully modeled as CRFs

How to Find MAP [Koppula RSS 2013]



Compute features
Define the energy function
Solve the Combinatorial Optimization

Activity/Object labels for the past and future

Sample Possible Futures

Shortcomings[Koppula RSS 2013]

We also need probabilities in addition to MAP solution (future is unknown)

Compute features
Define the energy function
Solve the Combinatorial Optimization

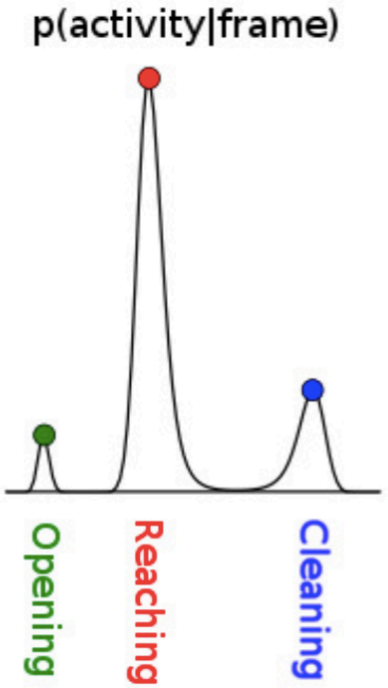
$$\mathcal{O}\left((TN_OL_OL_A)^3\right)$$

Sample Possible Futures

Activity/Object labels for the past and future

Dimension ~ $1o^{6xT}$ ~ 10^{3600}
(#ObjLabels$^{\#Objects}$x#ActLabels)^{Time}

Structured Diversity

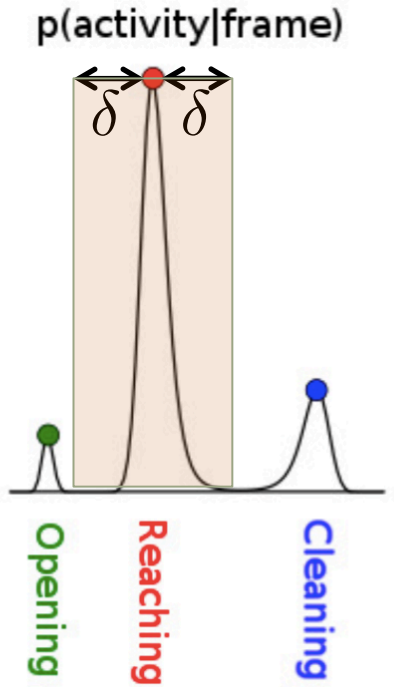


Although the state dimensionality is high, probability **concentrates** on a few **modes**

Structured Diversity

$$\mathbf{y}^{t,i} = \arg\max_{\mathbf{y}} bel^t(y)$$

$$s.t. \quad \Delta(\mathbf{y}, \mathbf{y}^{t,i}) \geq \delta \quad \forall j < i$$



Modes are likely and structurally diverse

HMM – Recursive Belief Estimation

HMM Derivation [Rabiner]

$$bel^t(\mathbf{y}) \propto \underbrace{p(\mathbf{y}^t = \mathbf{y}|\mathbf{x}^1, \ldots, \mathbf{x}^t)}_{\alpha^t(\mathbf{y})} \underbrace{p(\mathbf{x}^{t+1}, \ldots, \mathbf{x}^T|\mathbf{y}^t = \mathbf{y})}_{\beta^t(\mathbf{y})}$$

$$\alpha^t(\mathbf{y}^t) = p(\mathbf{x}^t|\mathbf{y}^t) \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) p(\mathbf{y}^t|\mathbf{y}^{t-1})$$

$$\beta^t(\mathbf{y}^t) = \sum_{\mathbf{y}^{t+1}} p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{y}^{t+1}|\mathbf{y}^t)$$

rCRF: Structured Diversity meets HMM

$$bel(\mathbf{y}^t) \propto \exp\left[\sum_{v,w\in\mathcal{E}^t}\left(E^b(\Phi_v,\Phi_w)-\tilde{E}^b(\Phi_v,\Phi_w)\right)\right.$$

Binary Term

$$\sum_{v\in\mathcal{V}^t}\left(E^u(\Phi_v)-\tilde{E}^u(\Phi_v)+\sum_{\mathbf{y}^{t-1}}\alpha^{t-1}(\mathbf{y}^{t-1})\log p(y_v^t|y_v^{t-1})\right.$$

Unary Term

$$\left.\left.\frac{1}{\gamma}\sum_{\mathbf{y}^{t+1}}\beta^{t+1}(\mathbf{y}^{t+1})bel(\mathbf{y}^{t+1})\log p(y_v^{t+1}|y_v^t)\right)\right]$$

Proposition: A Belief over rCRF is a CRF

rCRF: Algorithm

Compute energy function for each frame-wise CRF

Forward-Backward loop for message passing

Compute the energy function of rCRF

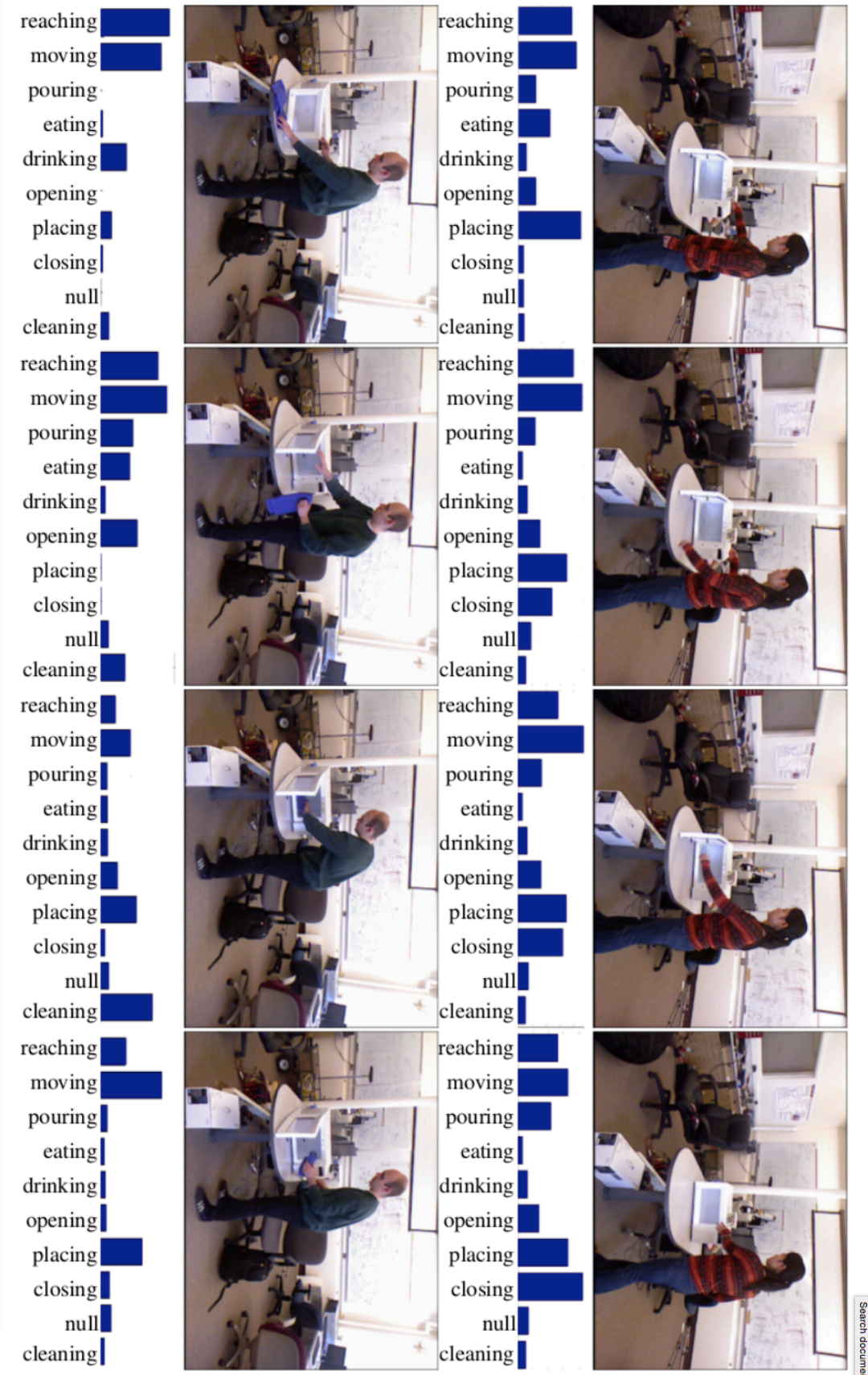
Sample by using Lagrangian relaxation [Batra et al]

rCRF: Algorithm

$$\mathcal{O}\left((TN_OL_OL_A)^3\right) \rightarrow \mathcal{O}\left(T(N_OL_OL_A)^3\right)$$

Computes **probabilities** for past/present/future states as a part of the formulation with **no random sampling**

Resulting Belief



Efficiency and Accuracy Improvement

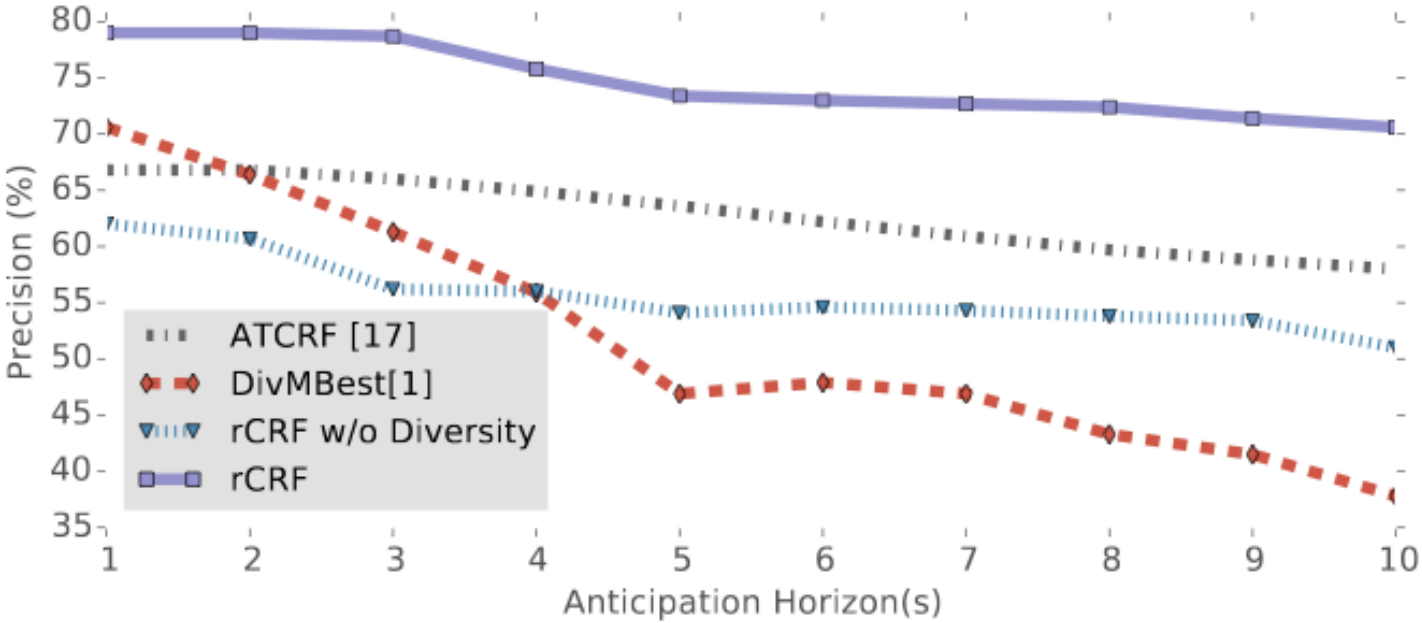
rCRF is **30x** faster than the state-of-the-art algorithms and runs in real-time

Accurate handling of uncertainty also increases accuracy

	Sub-activity			Object Affordance		
Method	micro prec(%)	macro f1-scr(%)	robot ant. metric(%)	micro prec(%)	macro f1-scr(%)	robot ant. metric(%)
Chance	10.0±0.1	10.0±0.1	30.0±0.1	8.3±0.1	8.3±0.1	24.9±0.1
GP-LCRF [17]	52.1±1.2	43.2±1.5	76.1±1.5	68.1±1.0	44.2±1.2	74.9±1.1
ATCRF [22]	47.7±1.6	37.9±2.6	69.2±2.1	66.1±1.9	36.7±2.3	71.3±1.7
DivMBest[1]	47.9±1.4	43.2±3.6	71.5±2.7	61.3±1.4	56.3±2.1	73.3±0.5
DCRF[43]	48.3±2.6	35.4±1.8	66.6±1.1	55.2±3.1	48.5±3.1	71.24±2.2
rCRF w/o div	49.6±2.1	39.7±2.6	65.1±1.1	56.2±1.9	47.4±3.1	70.8±2.5
rCRF	**54.3±3.9**	**45.8±2.7**	**76.5±2.6**	**78.7±3.4**	**74.9±3.8**	**82.1±2.9**

Efficiency and Accuracy Improvement

Resulting belief **also stays informative** trough **time**



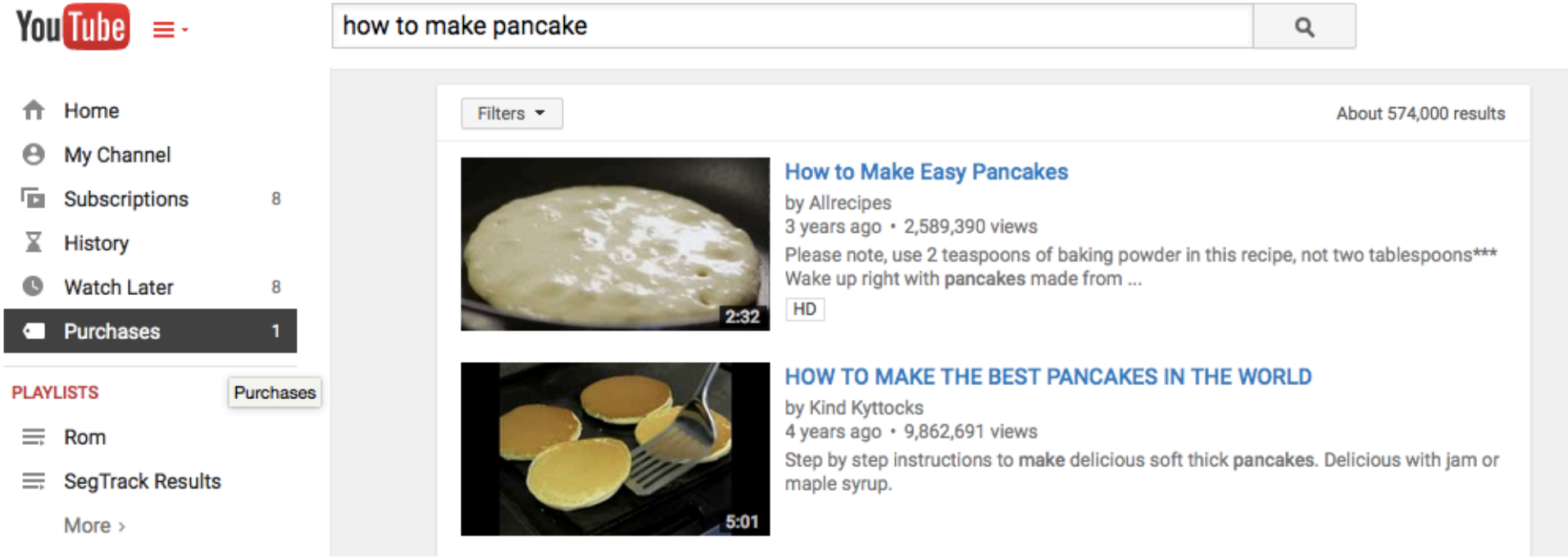
Outline

Structured understanding of a single video

Large-scaled understanding of video collection

Sharing knowledge to other domains and modalities

is Unsupervised Learning Possible



Is there an underlying **structure** in the YouTube "*How-To*" videos?

Is there an underlying **structure** in the YouTube "*How-To*" videos?

1st Result

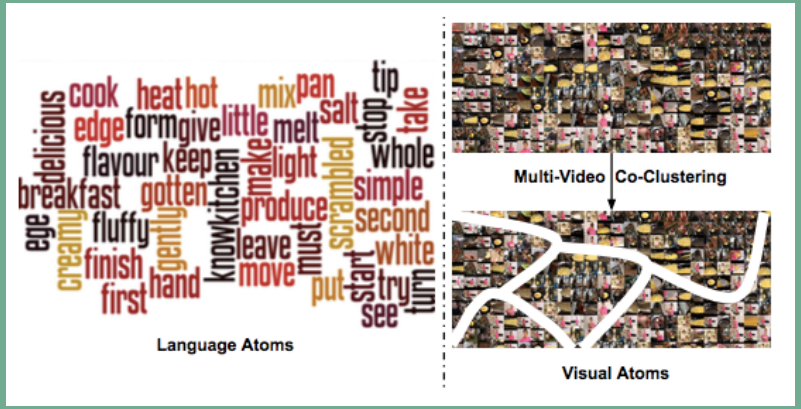
Is there an underlying **structure** in the YouTube "*How-To*" videos?

2nd Result

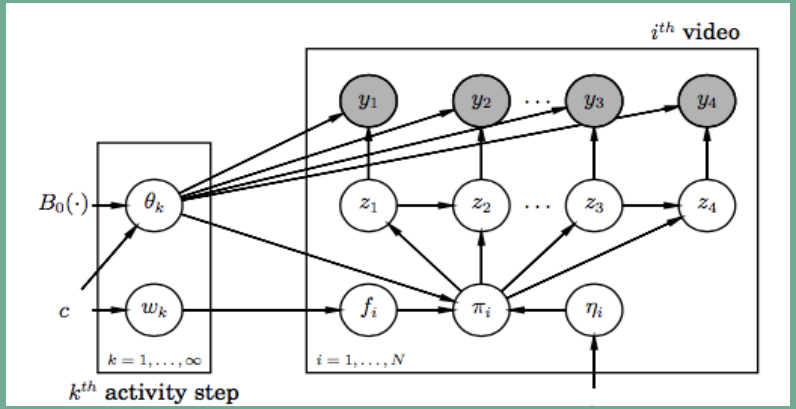
Summary of the Approach

We automatically download and filter large multi-modal activity corpus from YouTube

We learn a multi-modal dictionary



We discover activities by using a NP-Bayes approach



Dictionary Learning (Language)

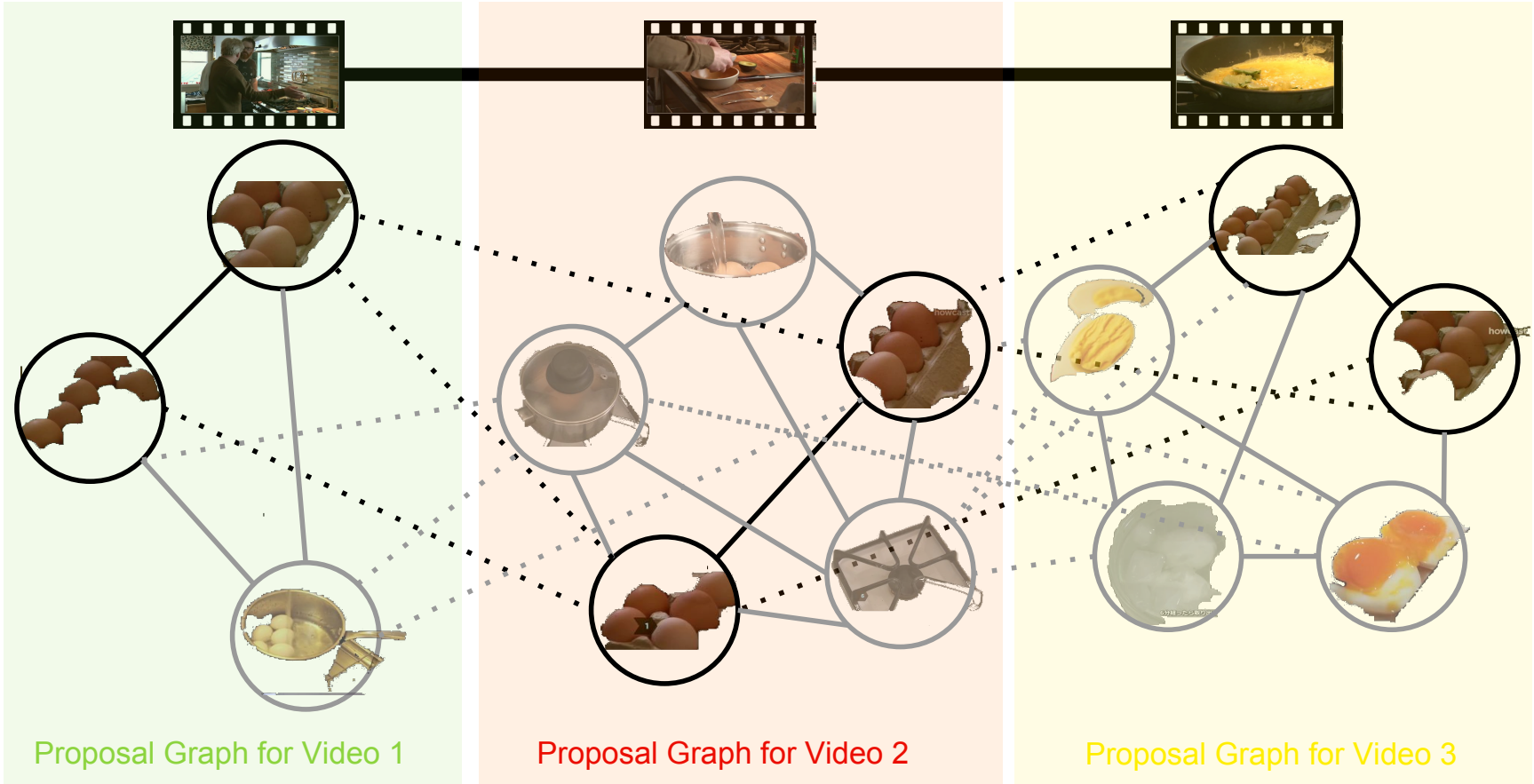
We use the tf-idf metric by considering each video as a document.

We choose the K mostfrequent words with max tf–idf

Dictionary for category "Hard Boil an Egg" with K=50

sort, place, water, egg, bottom, fresh, pot, crack, cold, cover, time, overcooking, hot, shell, stove, turn, cook, boil, break, pinch, salt, peel, lid, point, haigh, rules, perfectly, hard, smell, fast, soft, chill, ice, bowl, remove, aside, store, set, temperature, coagulates, yolk, drain, swirl, shake, white, roll, handle, surface, flat

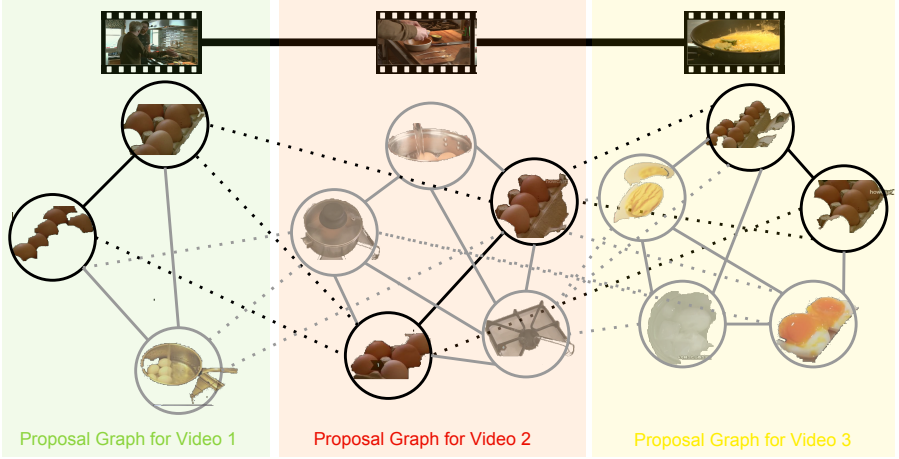
Dictionary Learning (Visual)



Dictionary Learning (Visual)

$$\arg\max \sum_{i \in V} \frac{\mathbf{x}^{(i)^T} \mathbf{A}^{(i)} \mathbf{x}^{(i)}}{\mathbf{x}^{(i)^T} \mathbf{x}^{(i)}} + \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} \frac{\mathbf{x}^{(i)^T} \mathbf{A}^{(i,j)} \mathbf{x}^{(j)}}{\mathbf{x}^{(i)^T} \mathbb{1}\mathbb{1}^T \mathbf{x}^{(j)}}$$

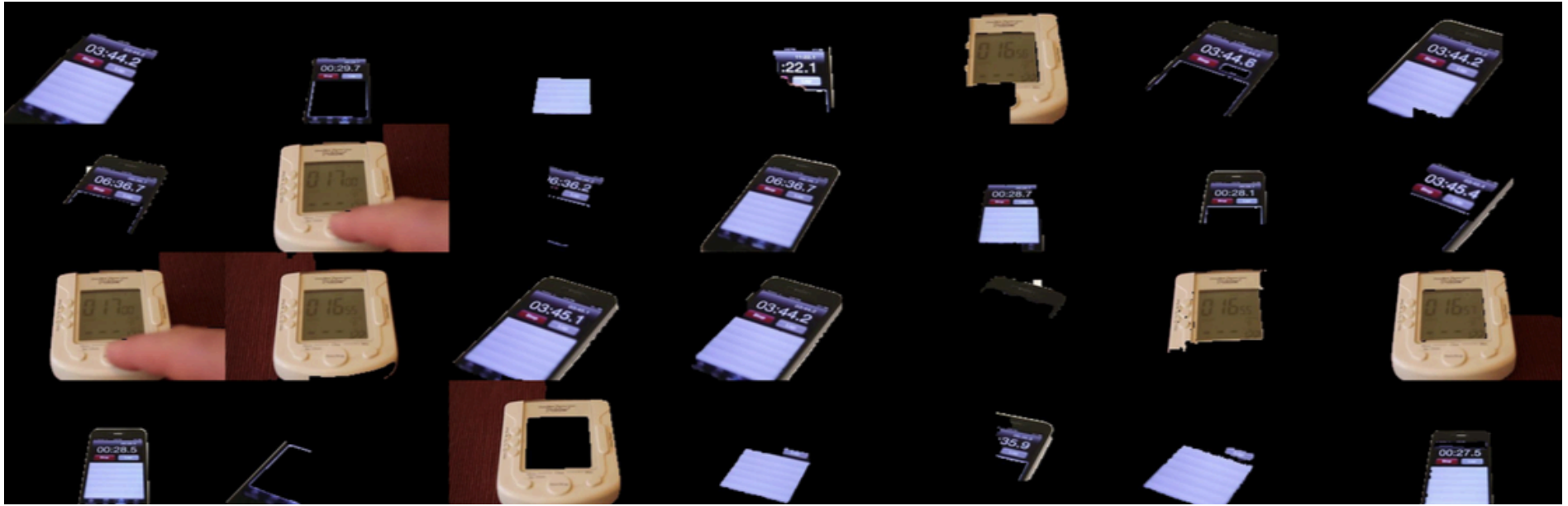
where V is set of videos,
$\mathcal{N}(i)$ is neighbour videos of i,
and $\mathbf{A}$ is similarity matrices



This function is quasi convex and can be optimized via SGD as

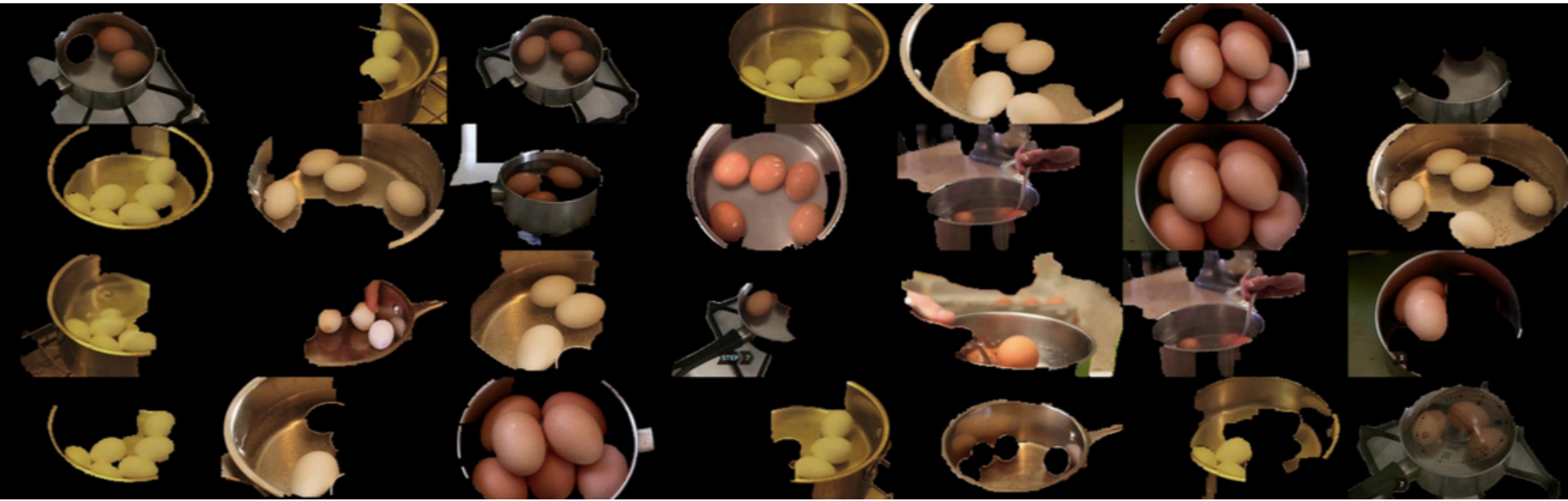
$$\nabla_{\mathbf{x}^{(i)}} = \frac{2\mathbf{A}^{(i)}\mathbf{x}^{(i)} - 2\mathbf{x}^{(i)} r^{(i)}}{\mathbf{x}^{(i)^T}\mathbf{x}^{(i)}} + \sum_{i \in N} \frac{\mathbf{A}^{i,j}\mathbf{x}^{j} - \mathbf{x}^{(j)^T}\mathbb{1} r^{(i,j)}}{\mathbf{x}^{(i)^T}\mathbb{1}\mathbb{1}^T\mathbf{x}^{(j)}}$$

Learned Dictionaries



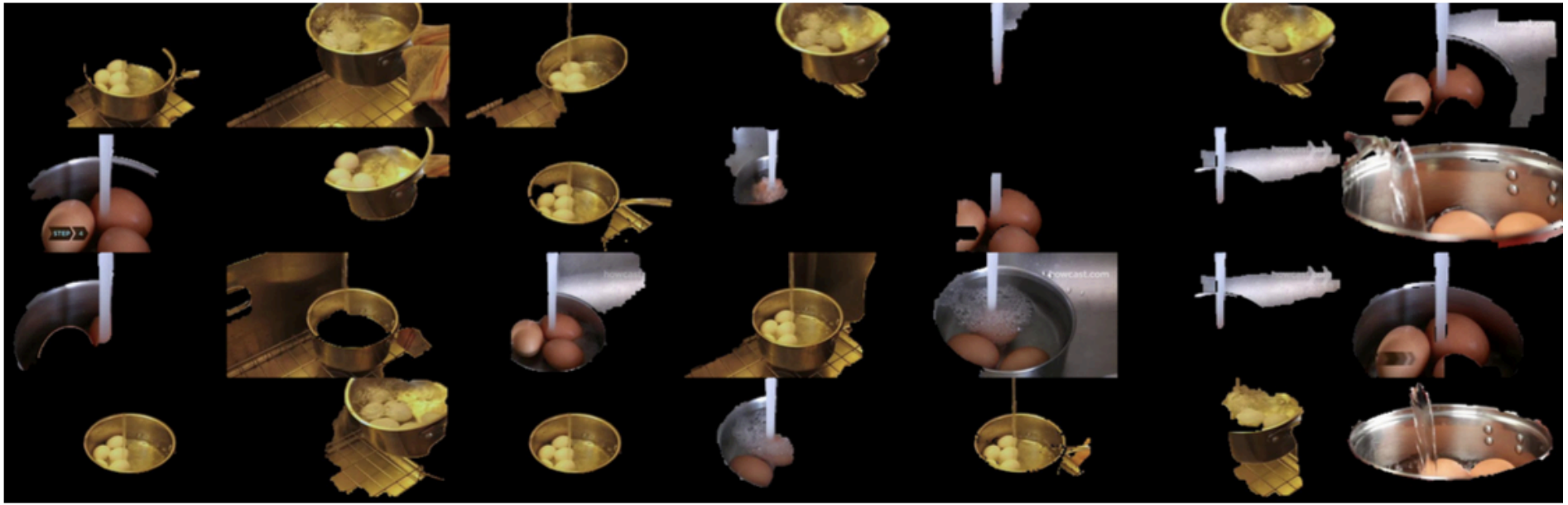
Semantically Correct

Learned Dictionaries



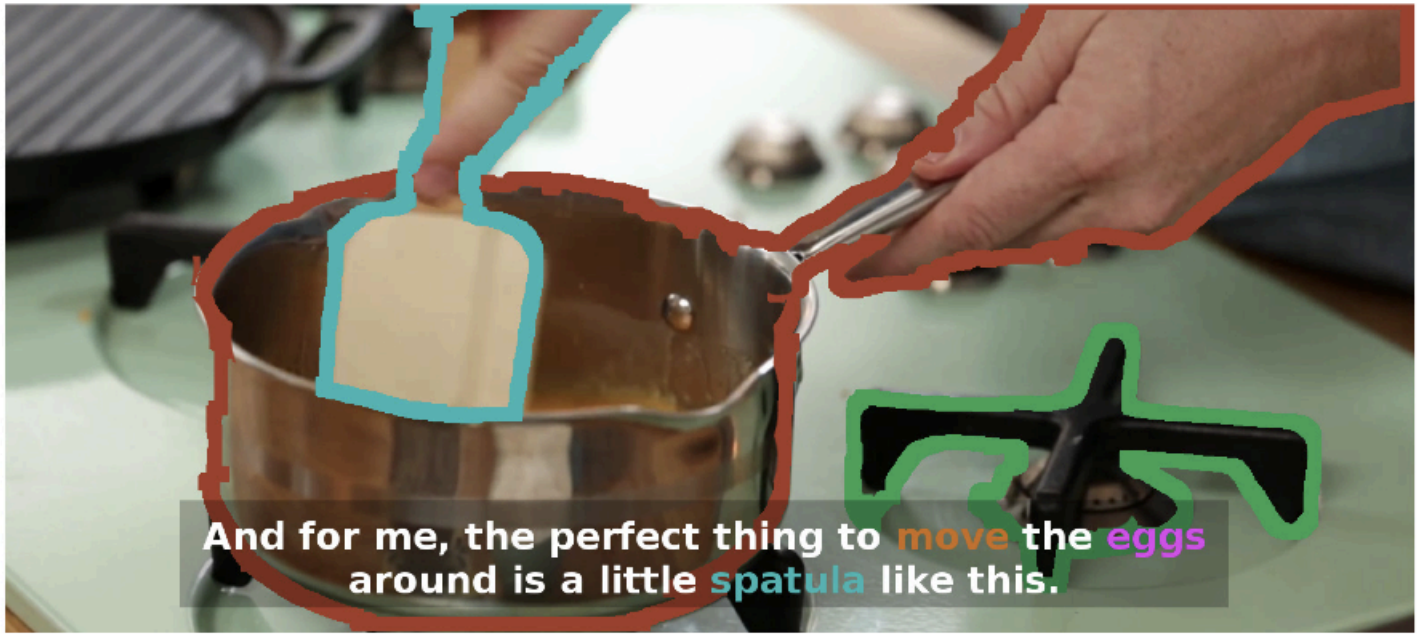
Semantically Correct

Learned Dictionaries

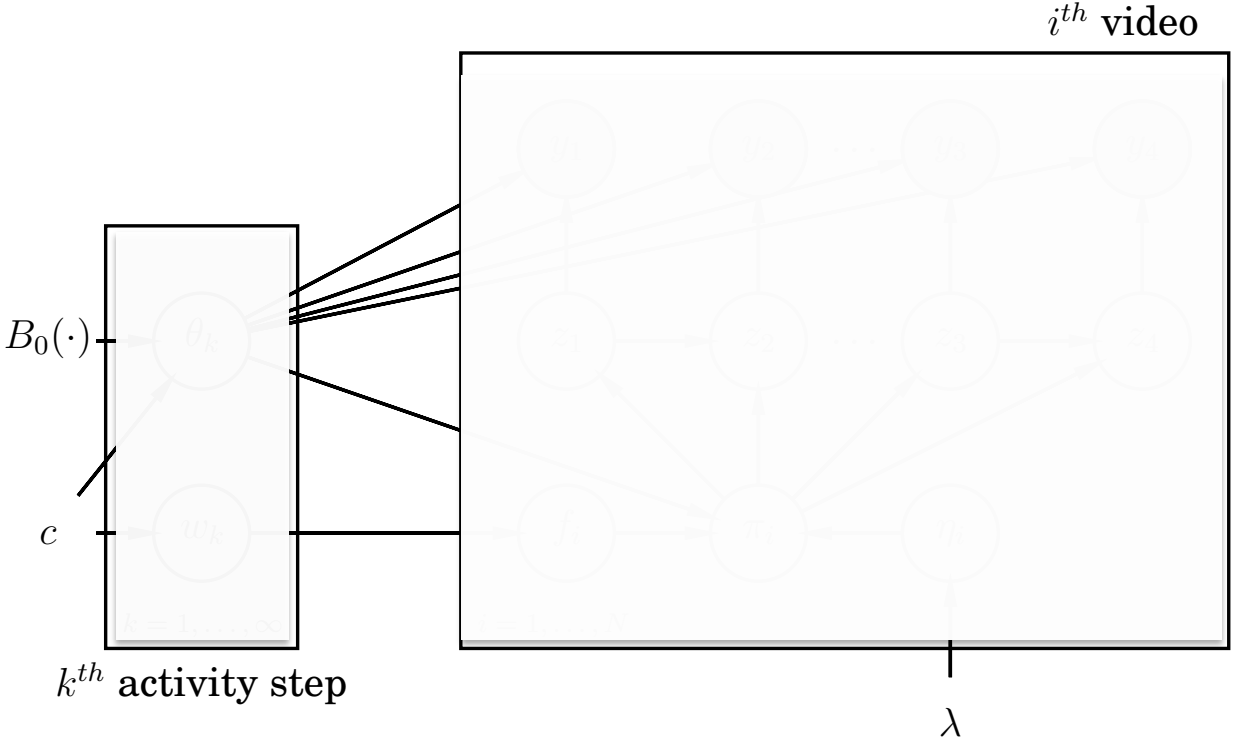


Accuracy vs Semantic Meaning

Representing Each Frame

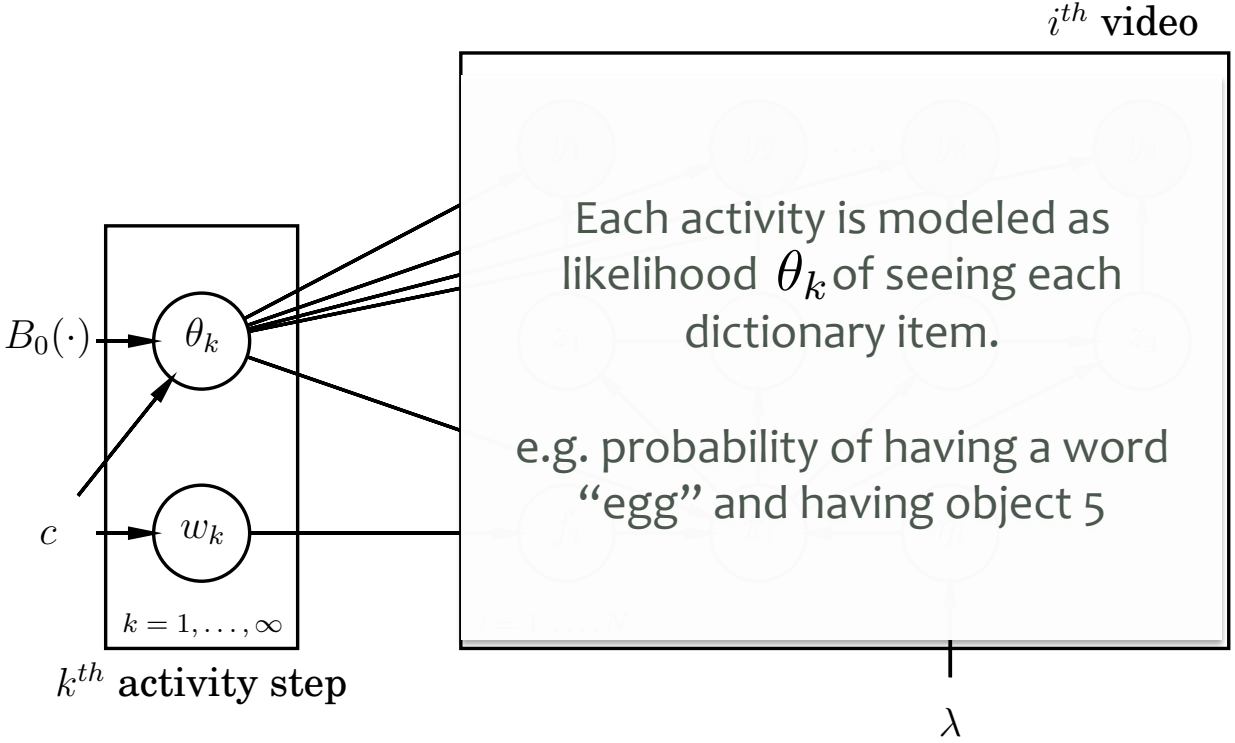


Unsupervised Discovery via NP-Bayes

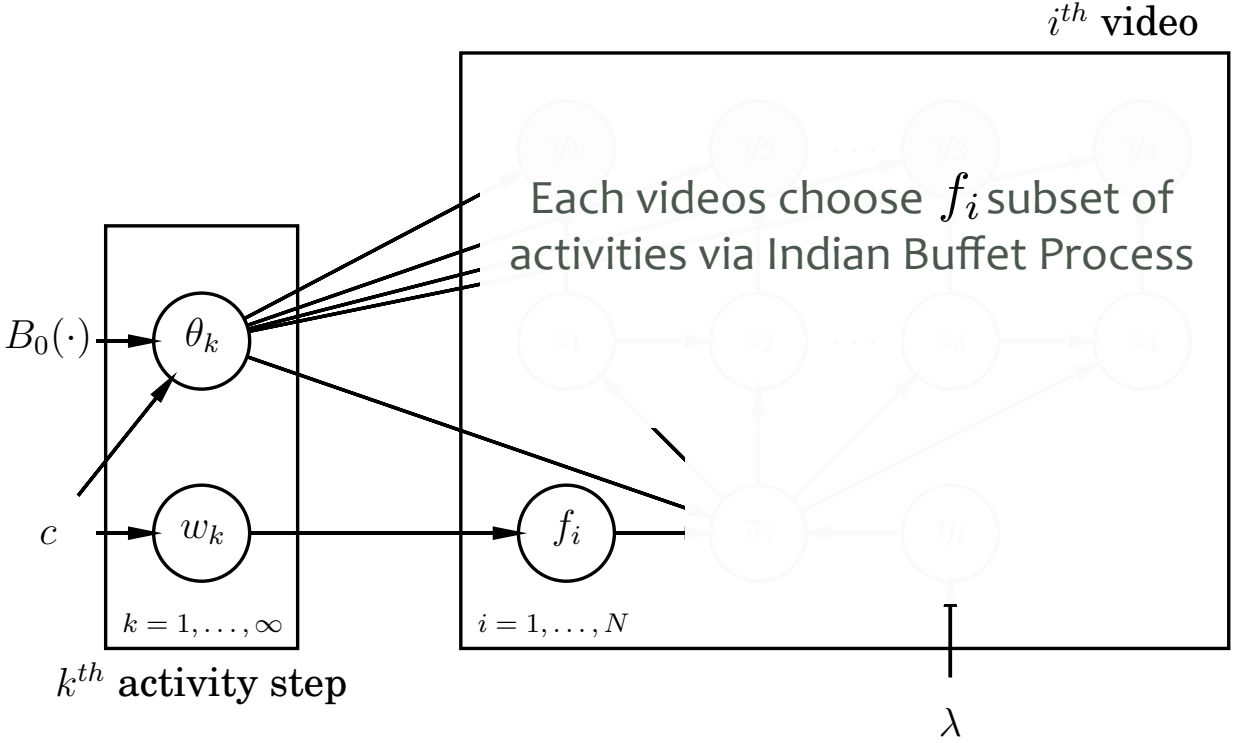


We jointly model activities and videos

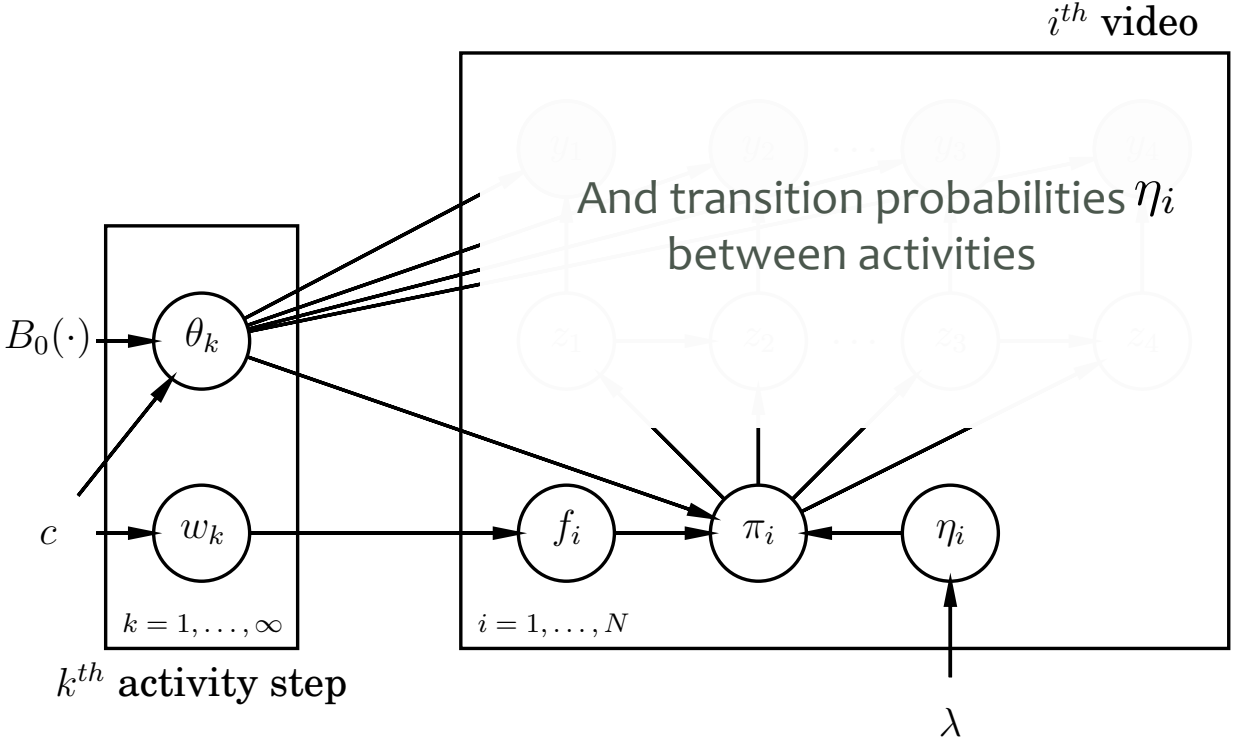
Unsupervised Discovery via NP-Bayes



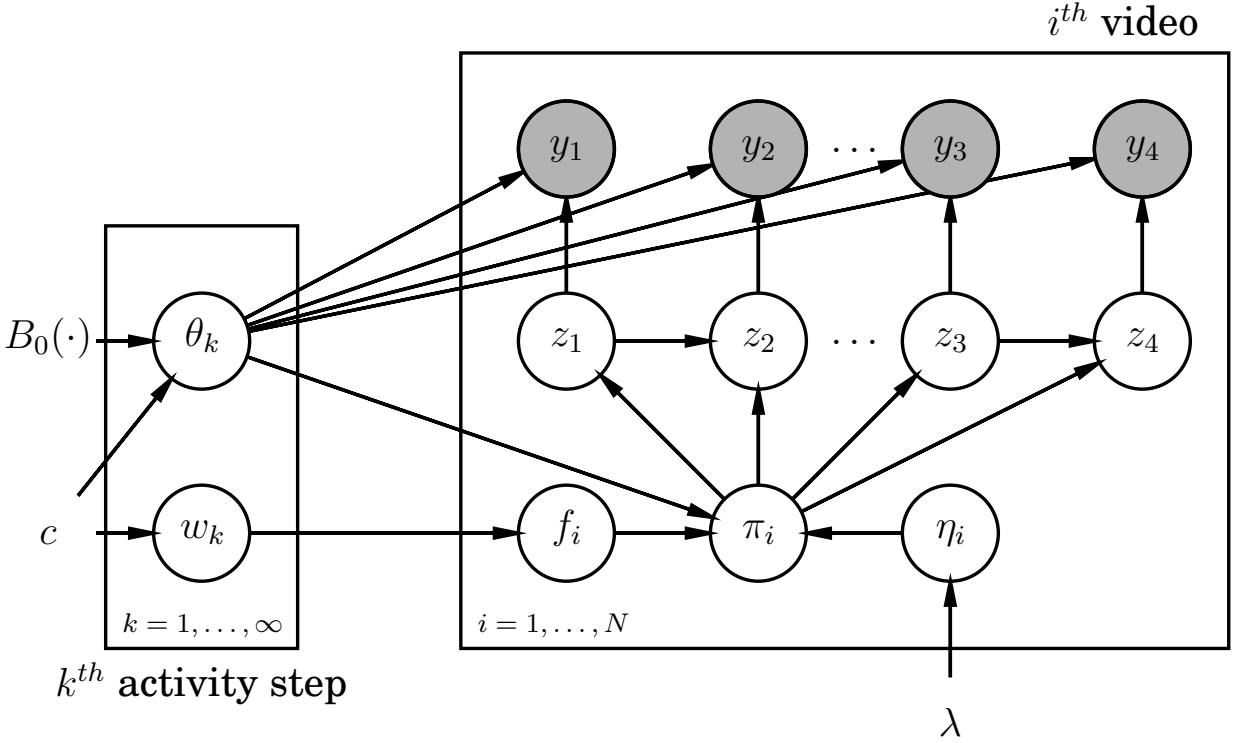
Unsupervised Discovery via NP-Bayes



Unsupervised Discovery via NP-Bayes

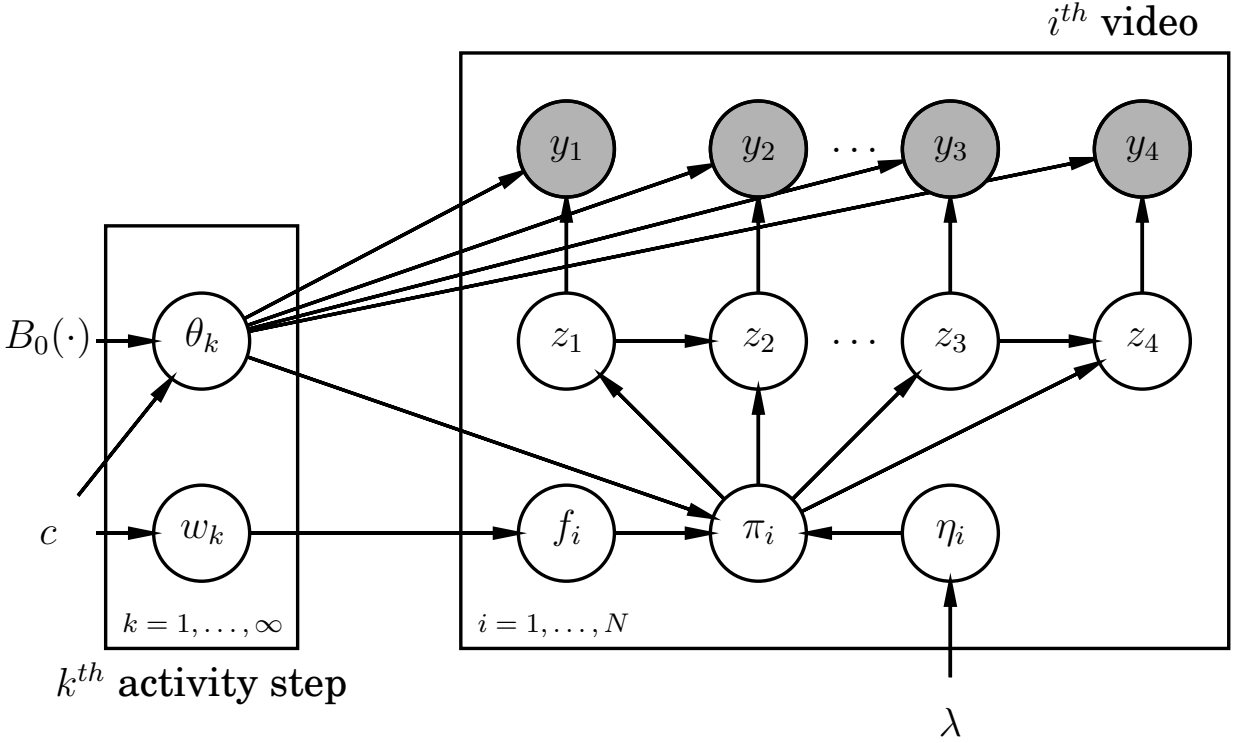


Unsupervised Discovery via NP-Bayes



Given activities/transition probabilities, it is HMM

Unsupervised Discovery via NP-Bayes



We learn by Gibbs Sampling

Discovered Activities

Name of the category

Clips from multiple videos detected to be the same step

Automatically generated caption for the discovered activity

Name of the category

Clips from multiple videos detected to be the same step

Automatically generated caption for the discovered activity

Evaluation

Both modalities are complementary and joint modeling is necessary!
Multi-video mid-level descriptions are critical for the accuracy

Table 1: Average of IOU_{cms} and mAP_{cms} over recipes.

	KTS [47] w/ LLF	KTS[47] w/ Sem	HMM w/ LLF	HMM w/Sem	Ours w/ LLF	Ours w/o Vis	Ours w/o Lang	Our full
IOU_{cms}	16.80	28.01	30.84	37.69	33.16	36.50	29.91	52.36
mAP_{cms}	n/a	n/a	9.35	32.30	11.33	30.50	19.50	44.09

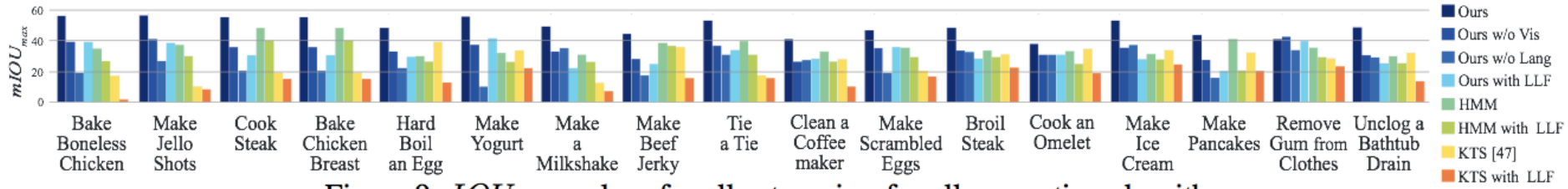


Figure 9: IOU_{max} values for all categories, for all competing algorithms.

Outline

Structured understanding of a single video

Large-scaled unsupervised understanding of human activities

Sharing knowledge to other domains and modalities

Graph Perspective of Large-Scaled Activities

How to make pancakes

Graph Perspective of Large-Scaled Activities

How to make pancakes

Graph Perspective of Large-Scaled Activities

Pan

Egg

Heat

How to make pancakes

Flip

Beat

Flour

Graph Perspective of Large-Scaled Activities

Egg

Heat

Pan

Flip

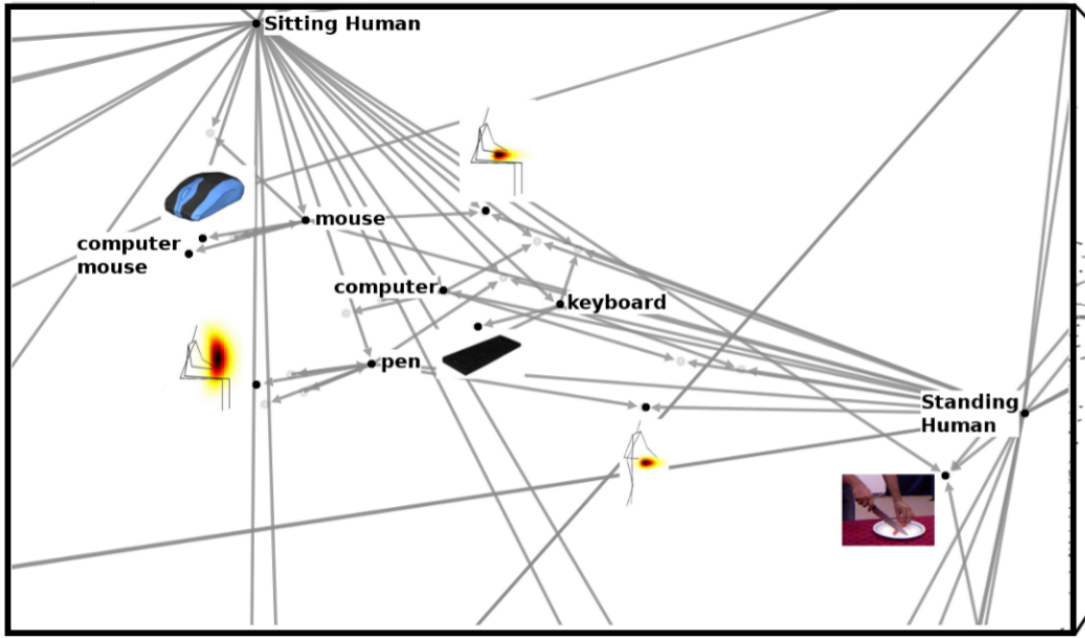
How to make pancakes

Beat

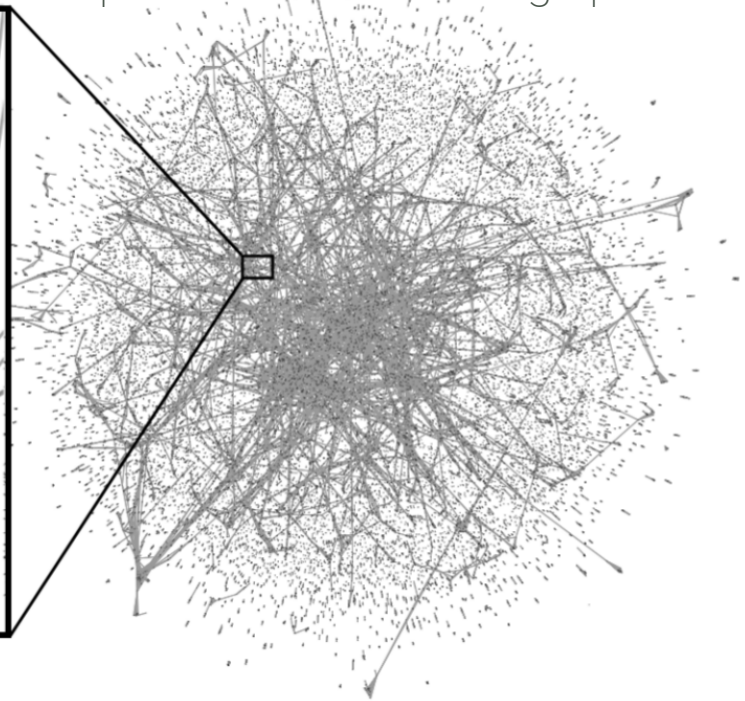
Flour

Can we go further?

RoboBrain



Snapshot of the RoboBrain graph

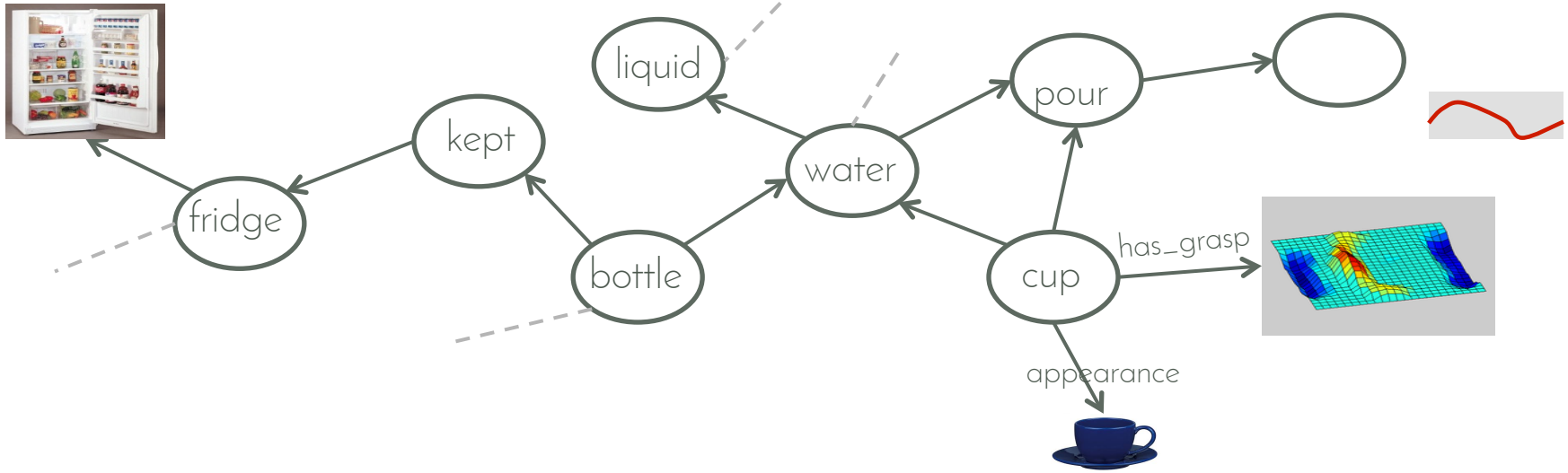


45,000 concepts (nodes)

98,000 relations (edges)

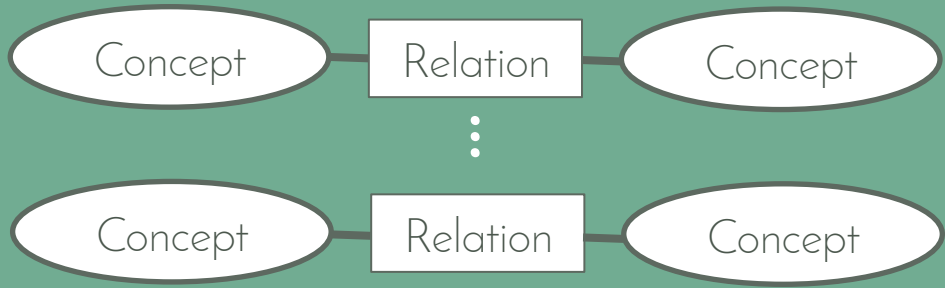
Connecting knowledge from Internet sources and manyprojects

How to scale the knowledge

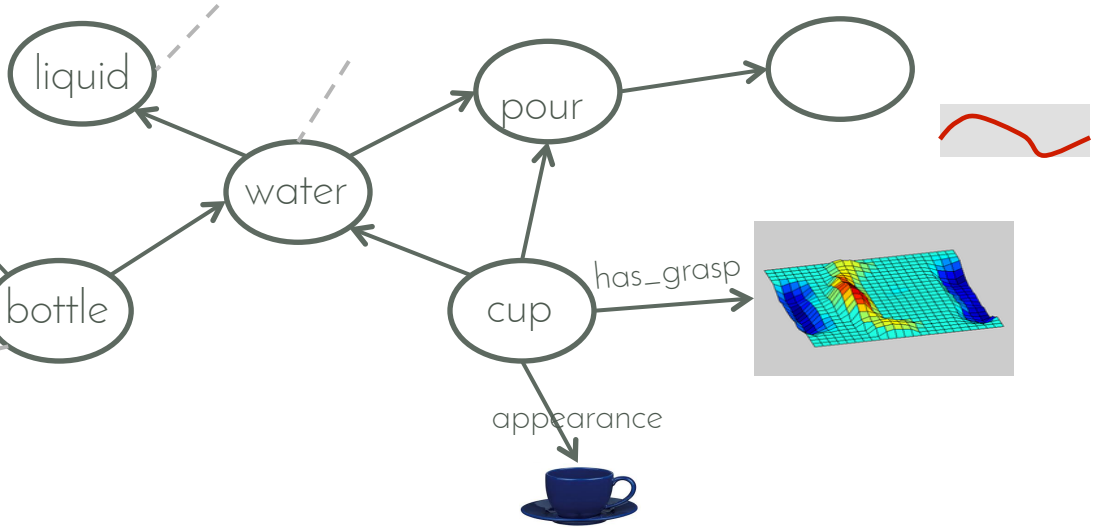


Input is in the form of "**feeds**"

A feed is collection of **binary relations**



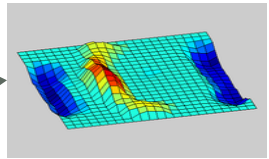
How to scale the knowledge

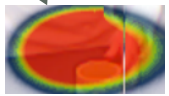


How to scale the knowledge

Bottle	has appearance	
Table	has appearance	

How to scale the knowledge





liquid

pour

kept

water

fridge

bottle

table

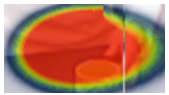
cup

has_grasp

appearance

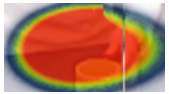
Cup

spatially distributed

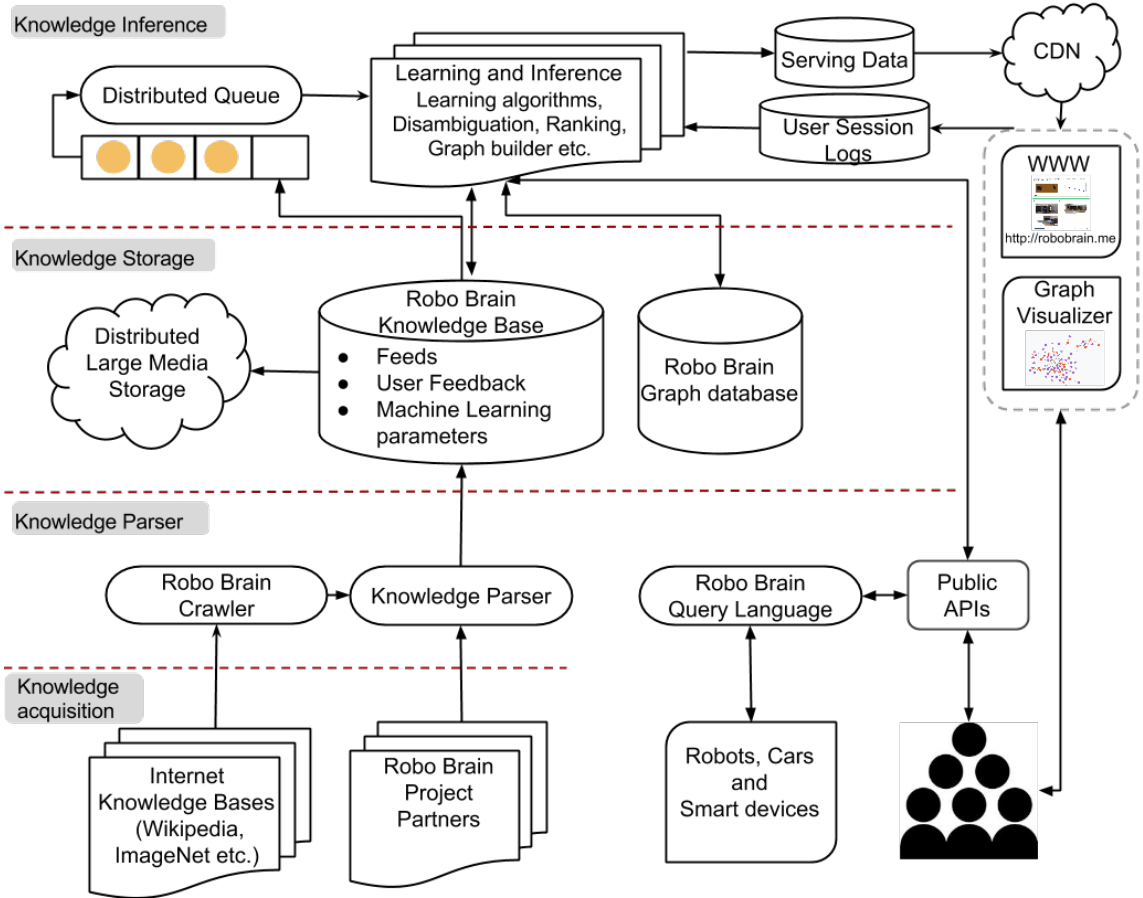


Table

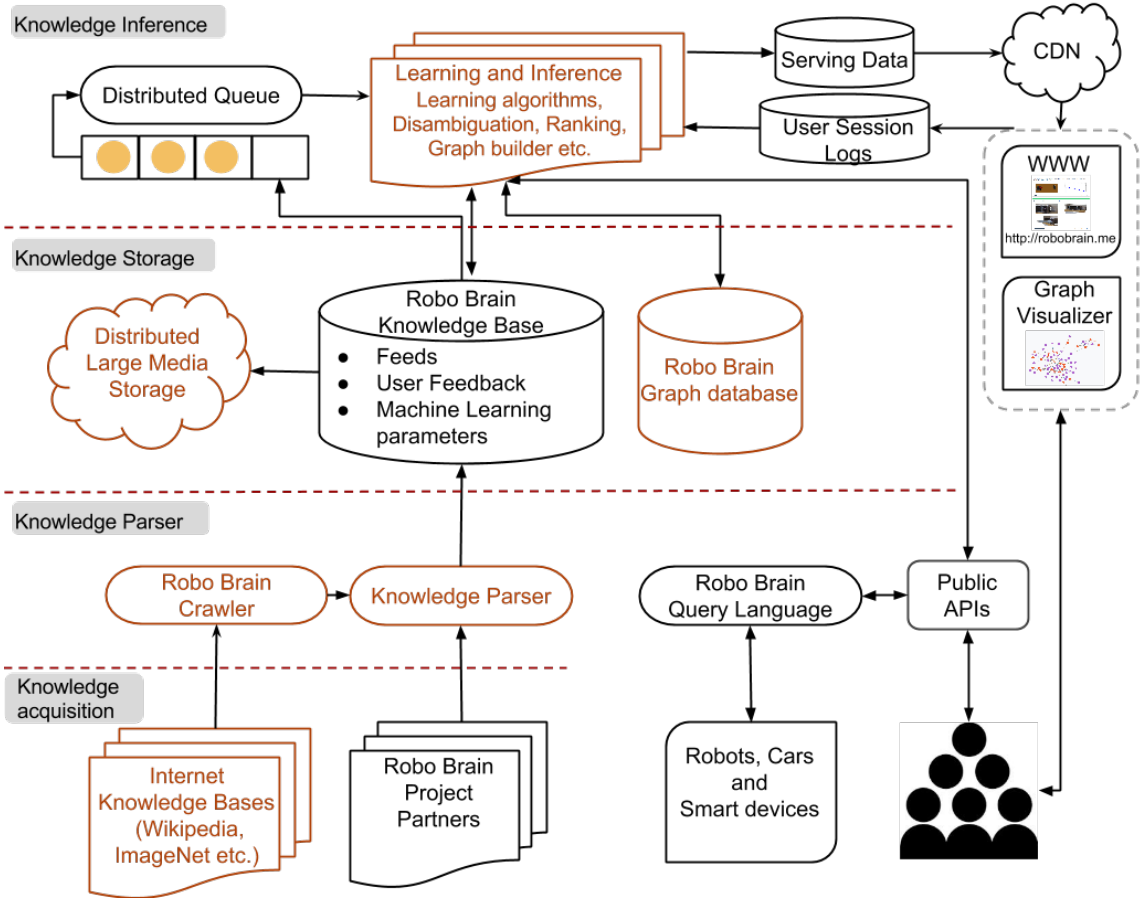
spatially distributed



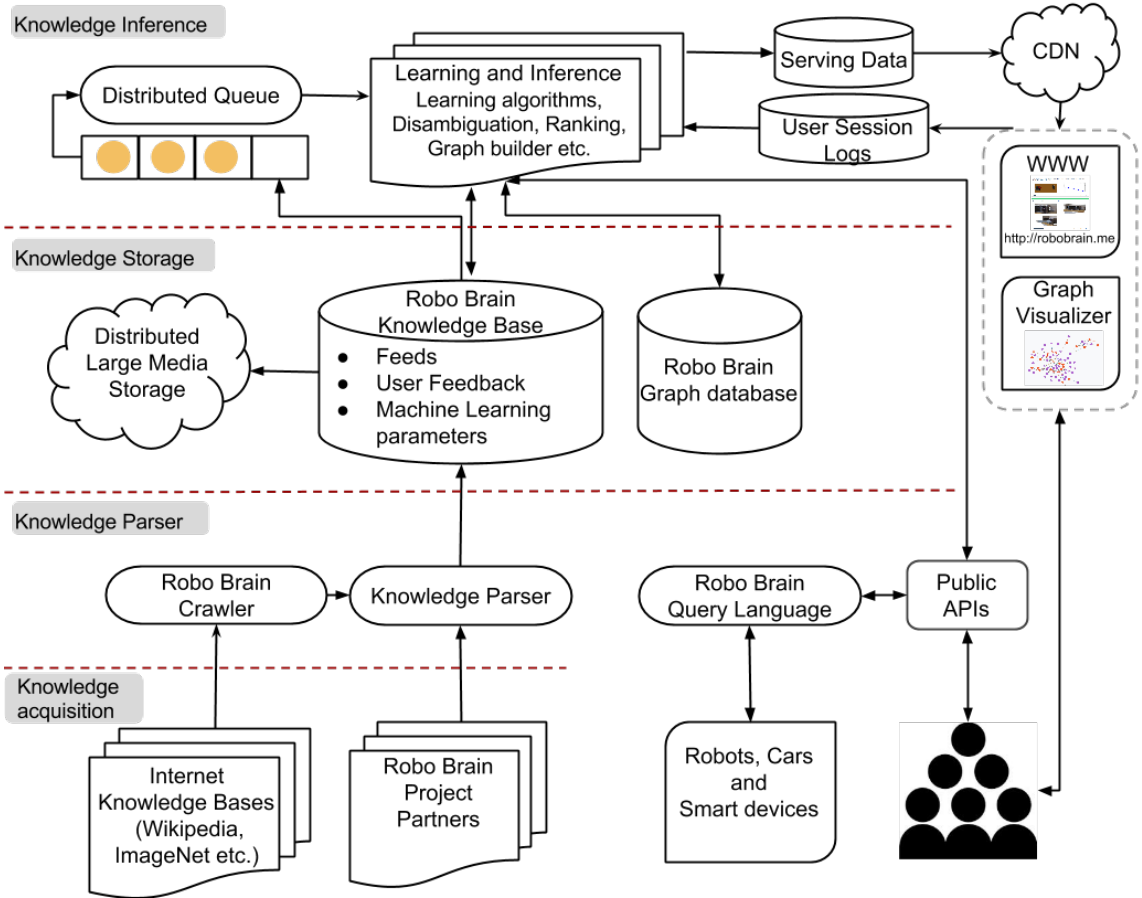
System Architecture



How we processed 25k videos in a day

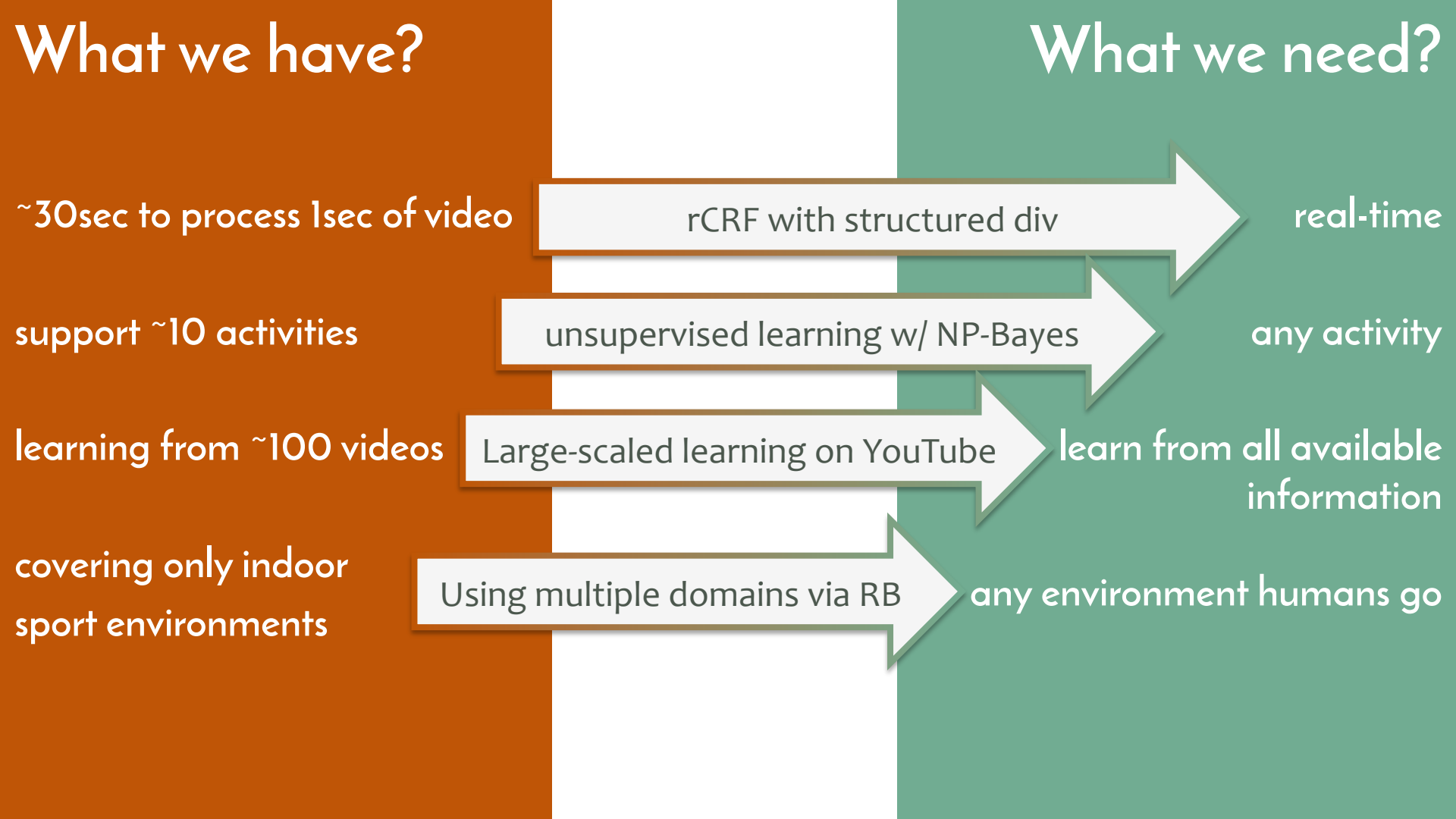


Robotics-as-a-Service



We shared a reliable interface to multiple universities.

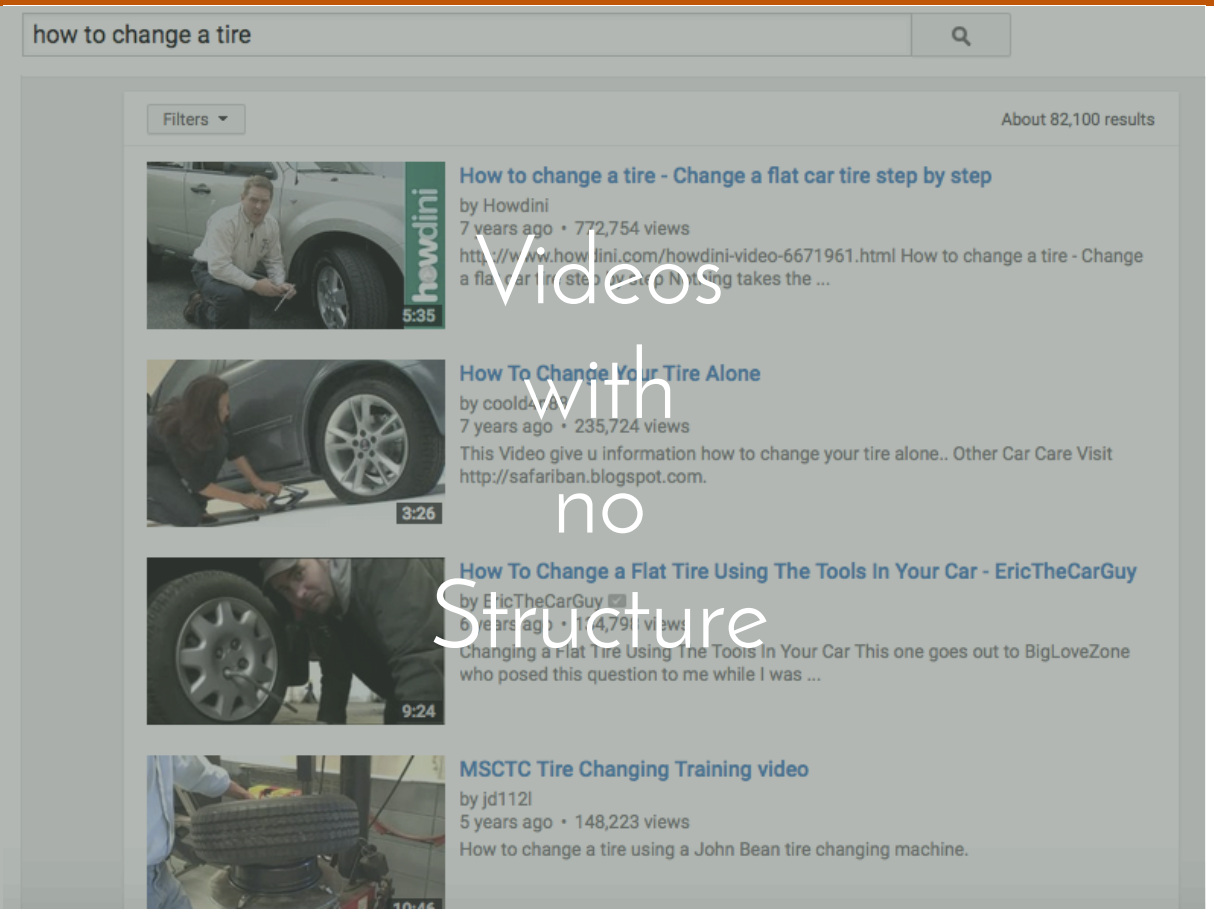
Humans to Robots Lab @ Brown University successfully used **RoboBrain-as-a-Service**


What we have?
What we need?
~30sec to process 1sec of video
rCRF with structured div
real-time
support ~10 activities
unsupervised learning w/ NP-Bayes
any activity
learning from ~100 videos
Large-scaled learning on YouTube
learn from all available information
covering only indoor sport environments
Using multiple domains via RB
any environment humans go

How can we scale further?

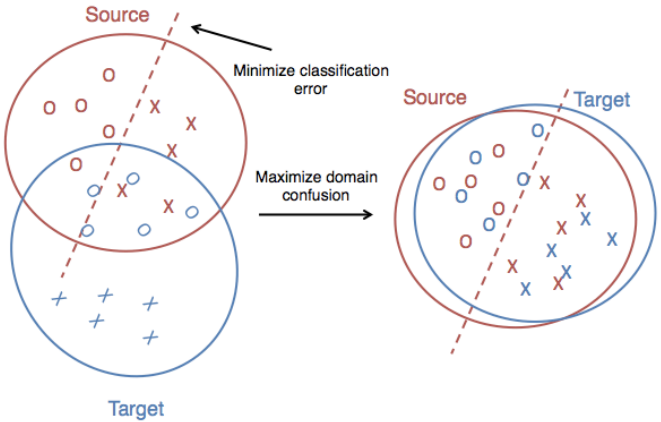
Linking more and more **modalities** and **domains** with **efficient** and theoretically sound models

Cross-Domain Information

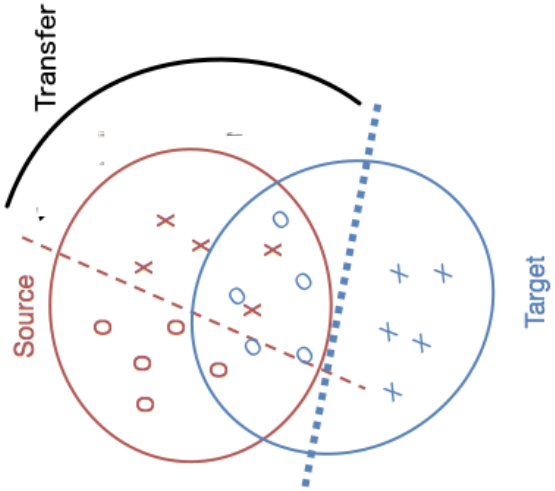


How to transfer knowledge?

Transductive Approach (ongoing/future work)



Domain invariant feature Learning followed by Induction

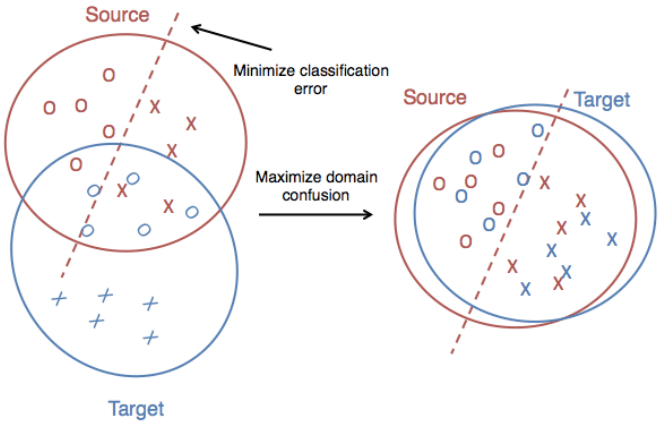


Induction followed by transformation

How to handle domain shift?

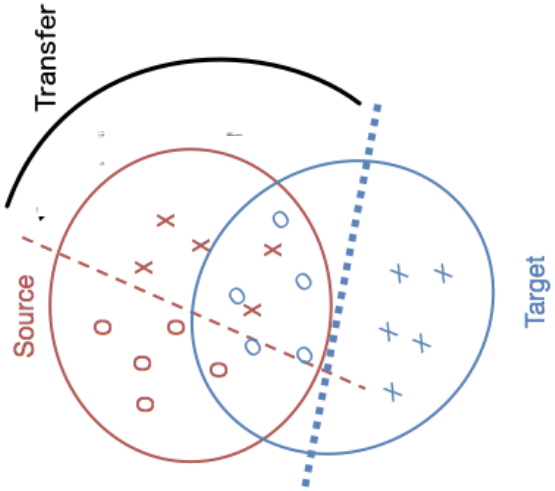
Domain adaptation vs Domain invariance

Transductive Approach (ongoing/future work)



Domain invariant feature Learning followed by Induction

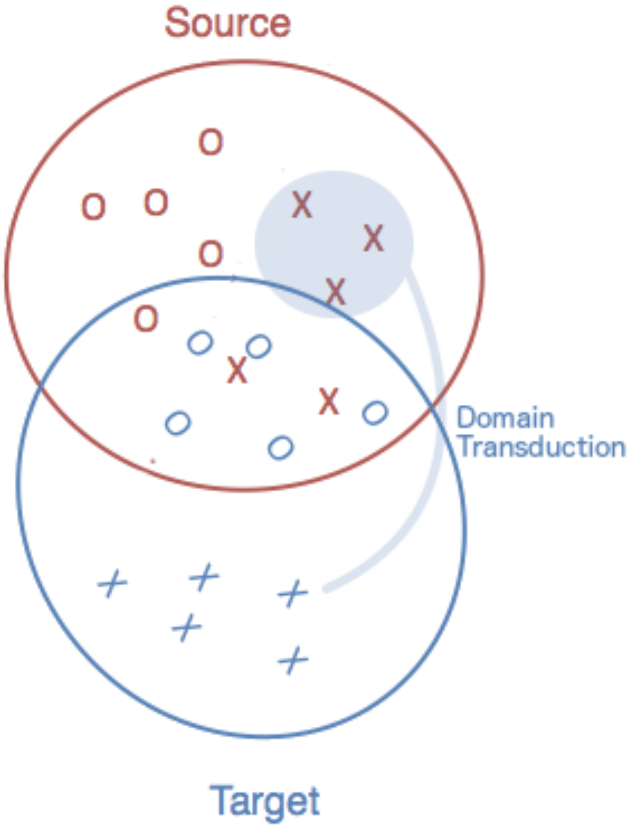
What if such feature does not exist



Induction followed by transformation

It is generally hard
Sometimes impossible
[Vapnik]

Domain Transduction [Ongoing work]

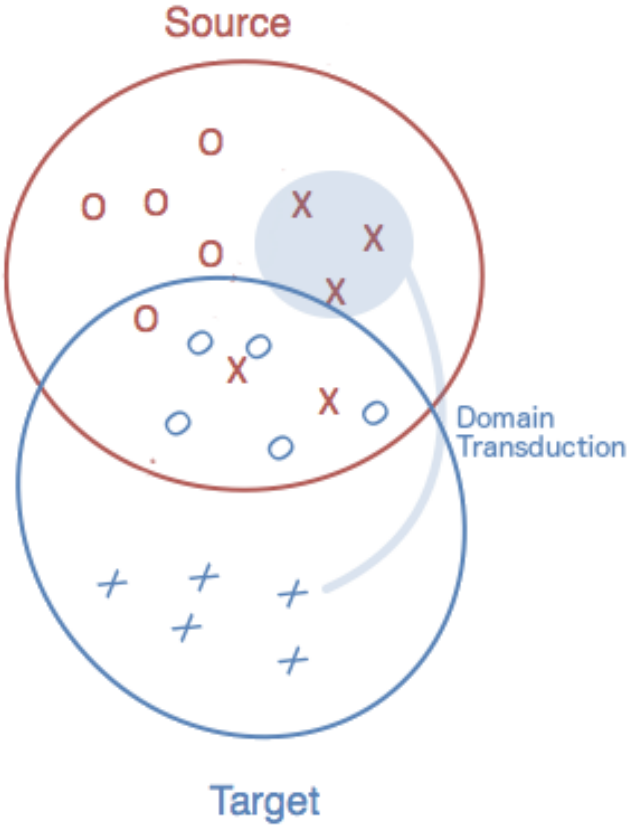


It might be possible to solve
Domain transfer with no induction

We are developing a max-margin framework based on coordinate-ascent of transduction and domain adaptation

[ongoing work]

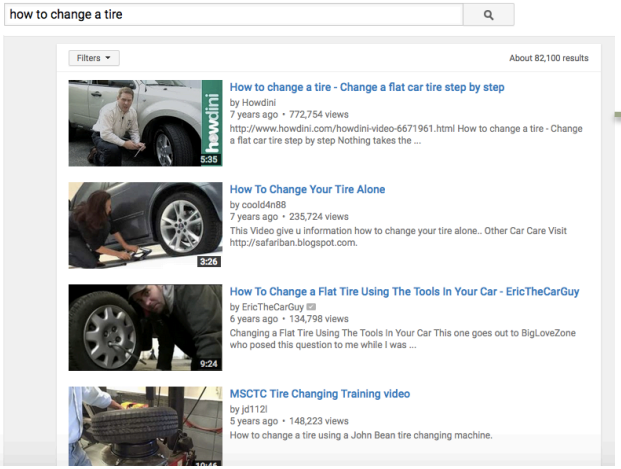
Adaptive Transduction [Future Work]



Some domains are
Intractably large like YouTube etc.

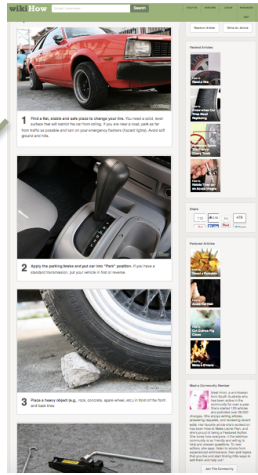
Can we solve the problem adaptively

Adaptive Transduction – Robot in the Loop



Adaptive Sampling

Sampled Videos



Domain Transduction

Unsupervised Discovery of Spatio-Temporal Semantic Descriptors
Under preperation for TPAMI submission

3D Semantic Parsing of Large-Scale Buildings
Iro Armeni, Ozan Sener et al. (In submission to CVPR 2016)

Unsupervised Semantic Parsing of Video Collections
Ozan Sener, Amir Zamir, Silvio Savarese, Ashutosh Saxena. In ICCV 2015

rCRF: Recursive Belief Estimation over CRFs in RGB-D Activity Videos
Ozan Sener, Ashutosh Saxena. In RSS 2015

RoboBrain: Large-Scale Knowledge Engine for Robots
Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra K. Misra, Hema S. Koppula. In ISRR 2015

Joint Work With

Ashutosh Saxena

Silvio Savarese

Amir R. Zamir

Ashesh Jain

Aditya Jami

Dipendra K. Misra

Hema Koppula

Jay Hack

Thank You