Large-Scale Unsupervised Understanding of Multi-Modal Recipes

Ozan Sener

Unsupervised Learning of How to Perform a Task

We (humans & robots) are using recipe books since 9th century.

Humans

- ✔ Expert curated solution
- ✘ Generally a single recipe is available
- ✘ Lacks visual information
- ✘ Even if the visual information exists, only a single view with some important parts are occluded

Robots [1,2]

- ✘ Assume common sense knowledge
- ✘ Ambigous language
- ✘ Requires lots of supervision for (partially) succesfull operation.

[1] M. Bollini, J. Barry, and D. Rus, BakeBot: Baking Cookies with the PR2, IROS 2011

[2] M. Beetz, U. Klank, I. Kresse, A. Maldonado, L. Mosenlechner, D. Pangercic, T. Ruhr and M. Tenorth, Robotic Roommates Making Pancakes, Humanoids 2011

Community Driven Solution

YouTube already has large-scale, multi-modal recipes (eg 281,000 video with speech for how to tie a bow tie). For a given query,

- It has multiple ways of performing the task
- It has (automtically generated) language information.
- Videos cover variety of environment conditions, camera angles, close-up etc.

However,

- A typical human will not watch 281,000 videos.
- A robot needs a structured information.

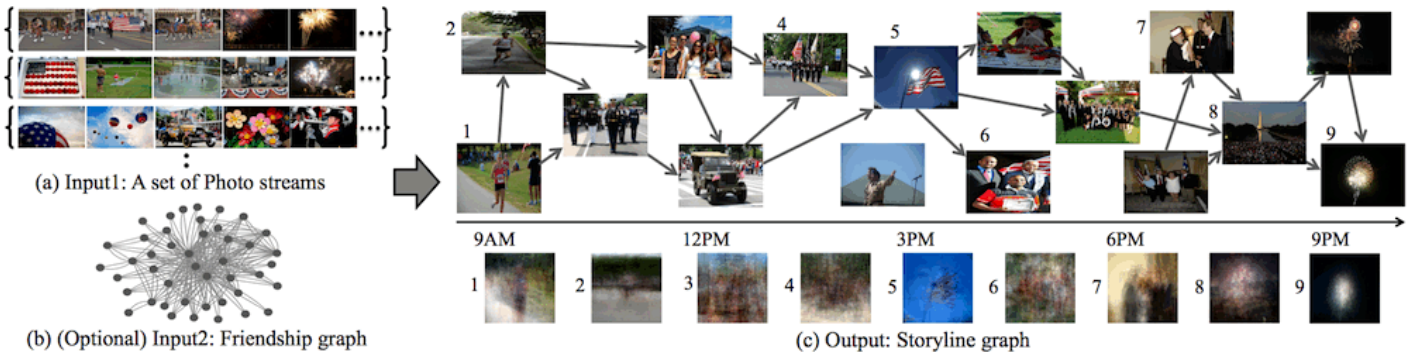
If we can automatically find the activities/steps, we can bring a structure to the videos and summarize 281,000 videos. Win-win scenario for both robots and humans.

Related Work

- Recipe Understanding
  - Cooking with Semantics [J Malmaud et al.]: Parsing text recipes and learning affordances
  - What's Cookin'? [J Malmaud et al.]: Matching a text recipe to the subtitles
  - BakeBot: Baking Cookies with the PR2 [M.Beetz et al.] Extracting a plan out of a recipe
- Multi-Modal Activity Detection/Recognition
  - Grounding Action Descriptions in Videos [M. Regneri]: Can visual data help semantic understanding of sentences?
  - Language Learning from Videos [H. Yu], YouTube2Text [S. Guadarrama], MPII Movie Description [A. Rohrbach]: Generating captions for video clips.

Related Work-2

- Video Summarization
  - Reconstructing Storyline Graphs [G. Kim]: Recovering the temporal dynamics of large photo collection.
  - Understanding Videos, Constructing Plots [A. Gupta]: Given activity labels, summarizing all videos as a graph.



Challanges-Noise

These are part of top 25 results when the query is How to make a milkshake?

Challanges-Visual Variety

Overview

Query: *How to make an ommellette?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

. . .

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Semantic Multi-Modal Representation

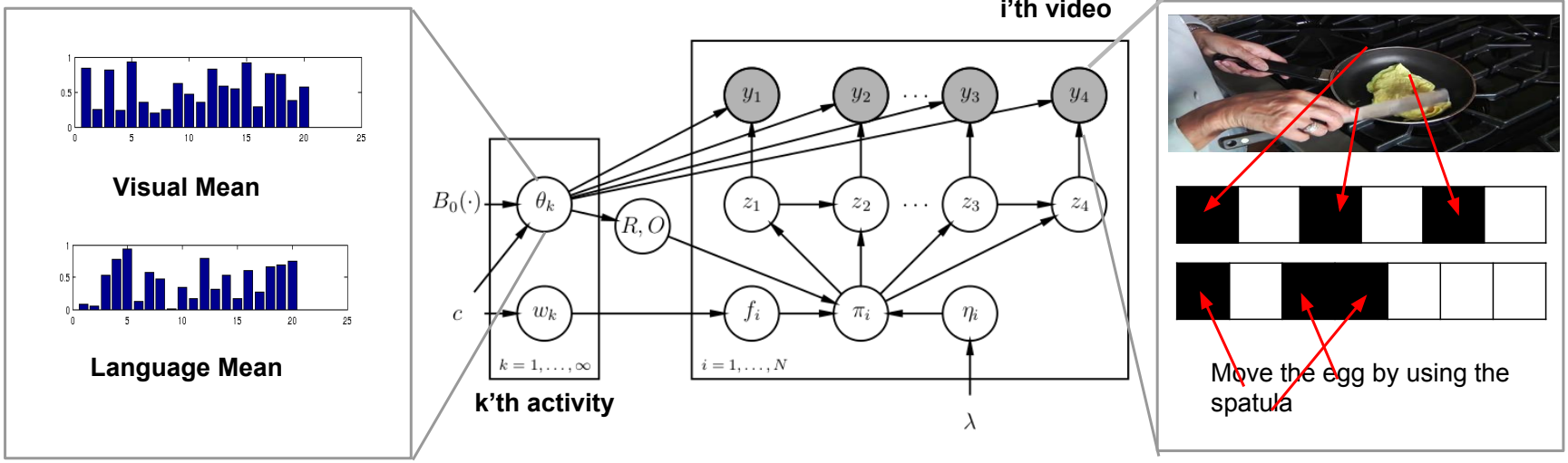
Salient Action Verbs/Object Names
(Language Atoms)

Object Proposals

Multi-Video Co-Clustering

Object Clusters
(Visual Atoms)

Unsupervised Activity Representation



Collecting Data

- For given a query, we download top 100 videos from YouTube.
- Each video has a language description (D_i), and we compute pairwise semantic distances by using Dice coefficients of n-grams ($n = 2$)
- We compute the normalized cut as thresholding the result of; ($A_{i,j} = d(D_i, D_j)$)

$$\arg\max_{x \in [0,1]^{100}} \frac{x^T A x}{x^T x}$$

- Resulting dominant cluster is used for the rest of the video.

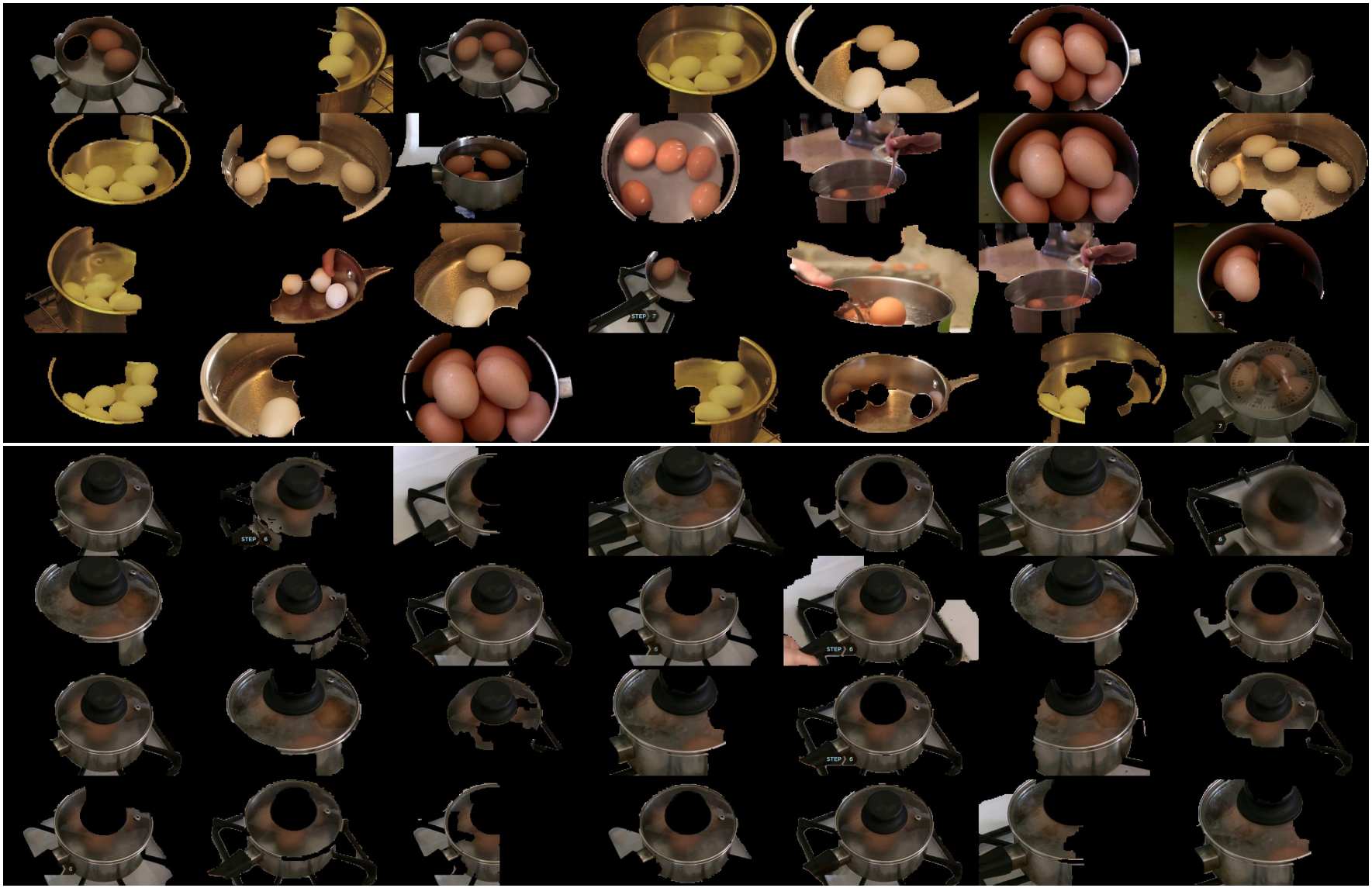
Finding Objects

- We generate object proposals $(P_i^1, \ldots, P_i^R)$ for each frame (t) of each video (i).
- We cluster the proposals in order to find objects; let's say $x_{i,r}^p$ is 1 if p^{th} cluster, has r^{th} proposal of i^{th} video.
- We define the clustering problem as (A is pairwise distance matrix with fc7 AlexNet features)

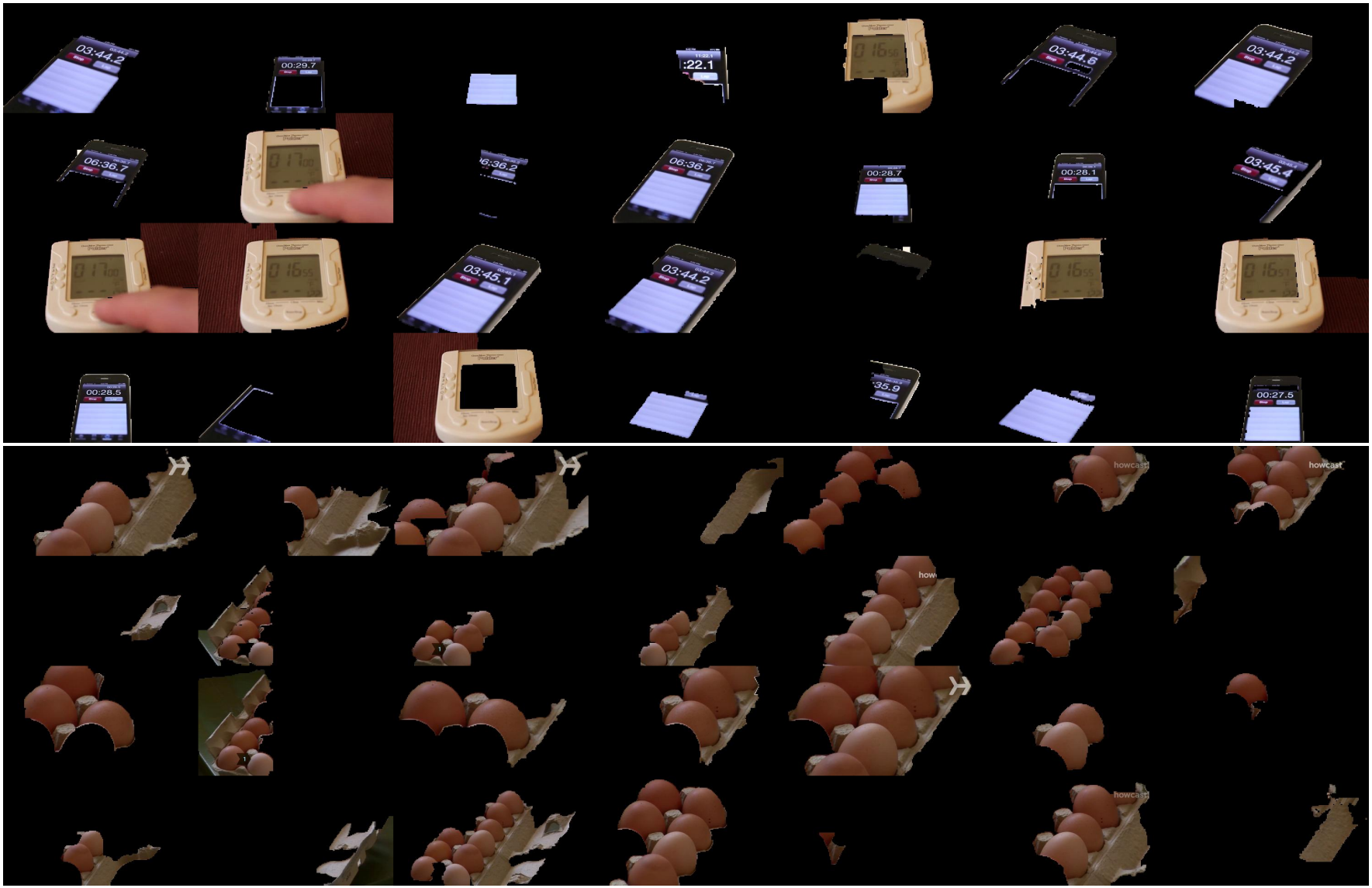
$$\arg\max_x \sum_i \frac{x_{i,.}^T A_i x_{i,.}}{x_{i,.}^T x_{i,.}} + \sum_i \sum_{j \in \mathcal{N}_k(i)} \frac{x_{i,.}^T A_{ij} x_{j,.}}{x_{i,.}^T \mathbb{1}\mathbb{1}^T x_{j,.}}$$

- This maximum of this function is a cut over proposals which all videos agree on and has maximum normalized total similarity within cluster.
- This function is quasi-convex and can be maximized by using sub-gradient method.

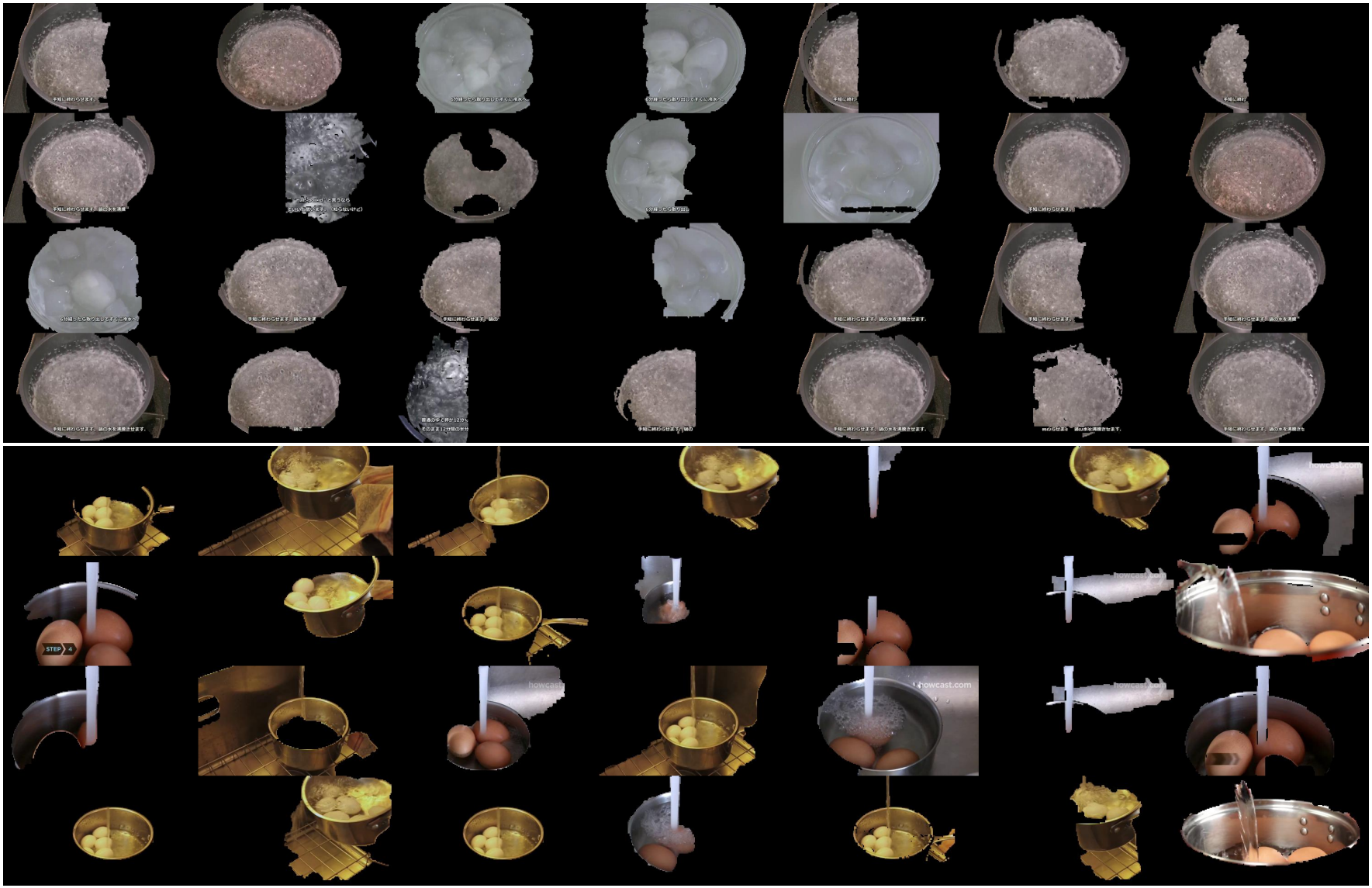
Some Visual Objects



Some Visual Objects-2



Some Visual Objects-3



Finding Salient Words

- We concatanate all subtitles and compute the frequency of each word (f_t).
- For each selected word, we compute the frequency of each word in NYTimes corpus (f_d).
- We choose the K most frequent words with property $f_t > f_d$.

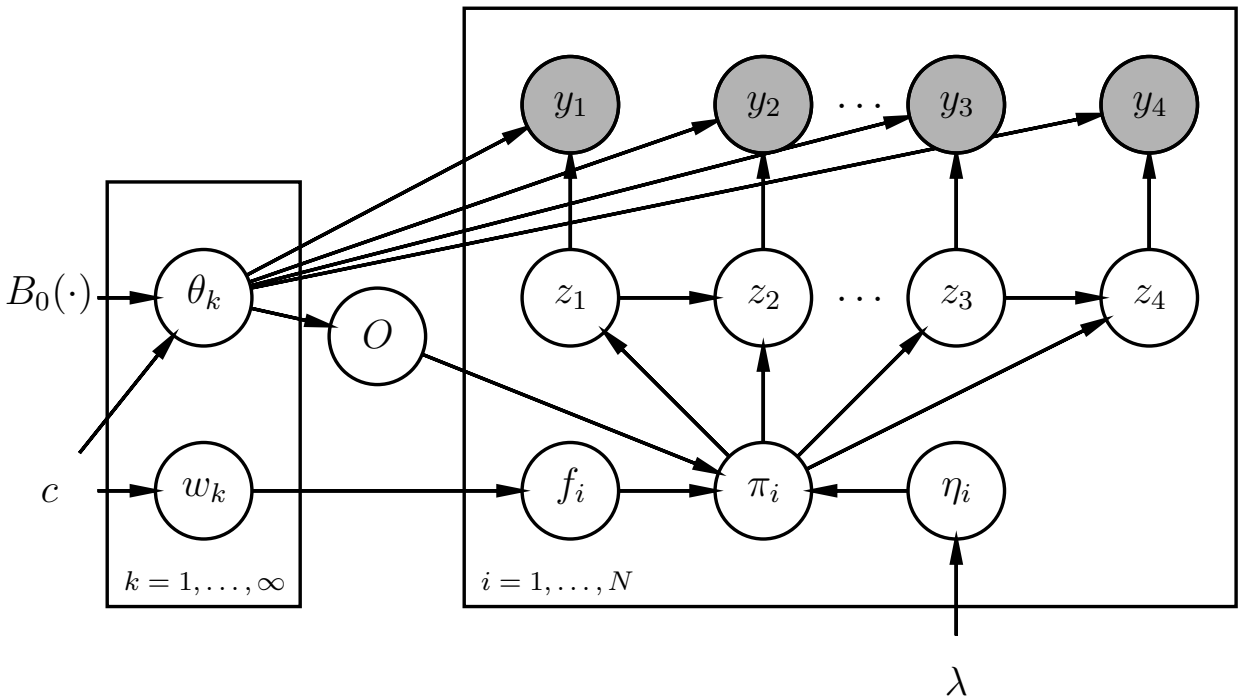
Salient Words

First 50 salient words:

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

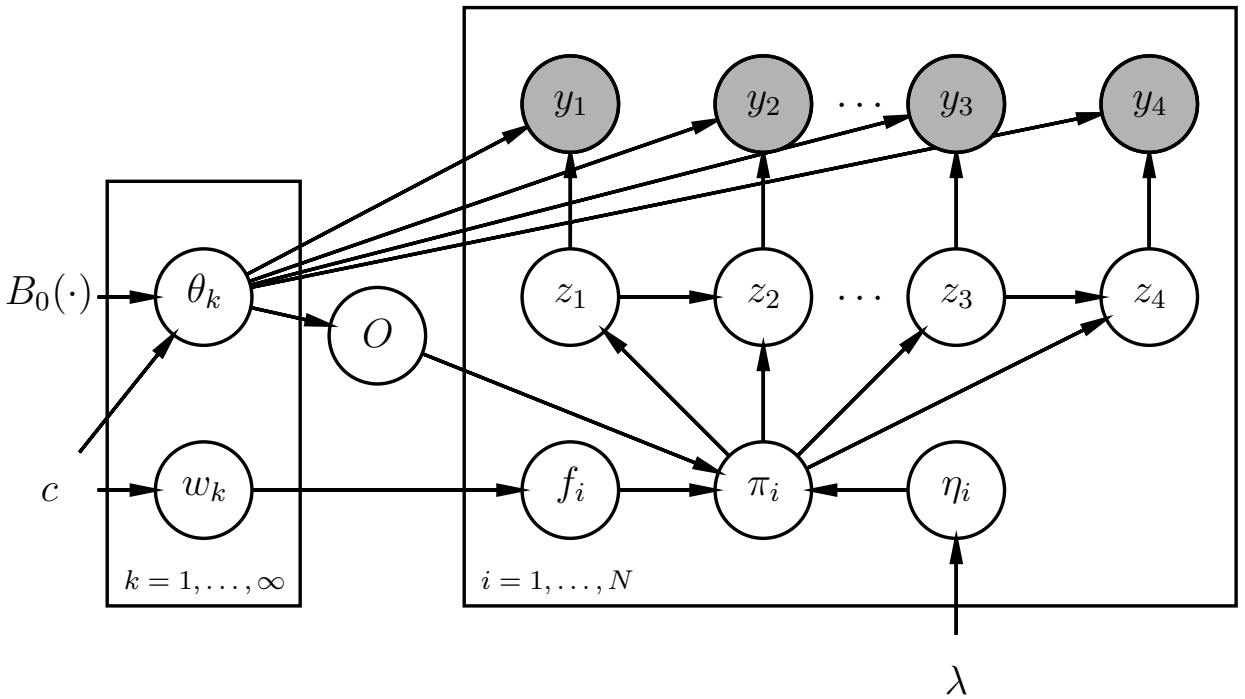
Learning Recipes

- Each activity (θ_k) is represented as Binomial r.v. over visual objects and Drichlet r.v. over language words.
- Each video choose set of activities (f_i) with likelihoods (w_k) following the Indian Buffet Process.



Learning Recipes - 2

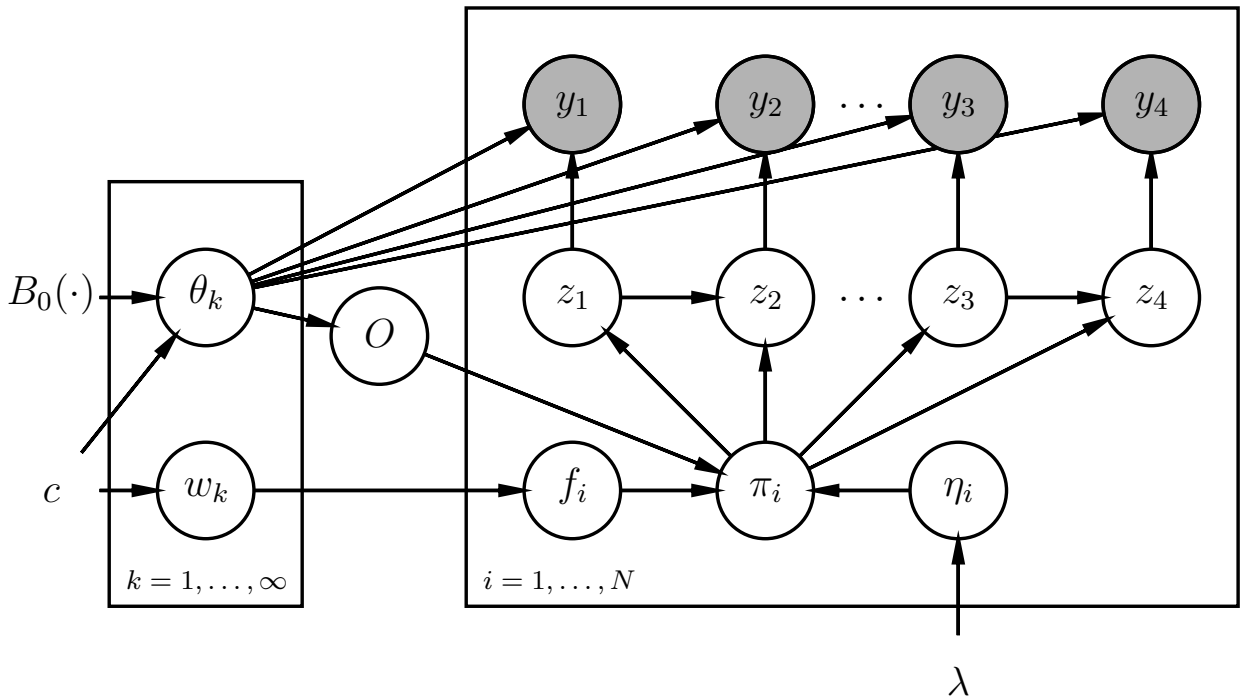
- Some activities impose time ordering via O e.g. activity 1 is always after activity 2 if $O_{1,2} = 1$.
- Activities have prior transition probabilities ($\eta_i^{m,n}$ is the transition prob from activity m to n, $\eta_i^{m,\cdot}$ follows a Drichlet r.v.



Learning Recipes - 3

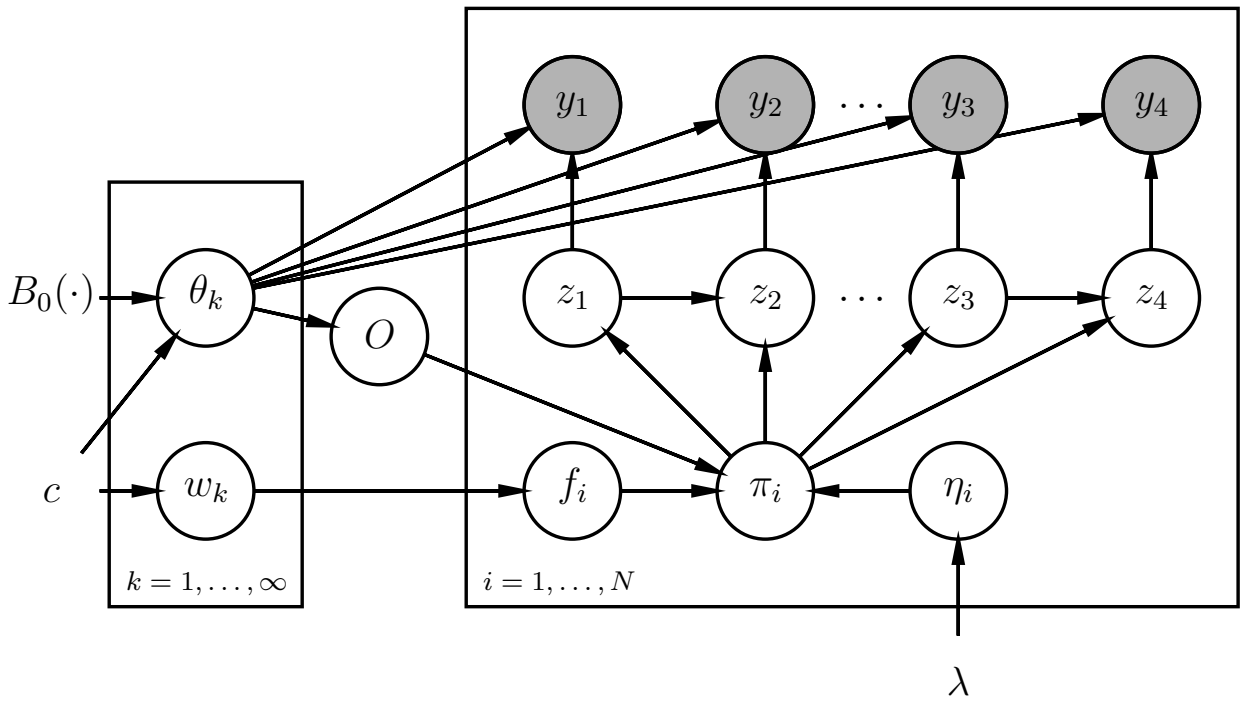
- Transition probabilities(π_i) is normalized version of prior transition probabilities obeying selected activities and time orderings.

$$\pi_i = \frac{f_i \otimes \eta_i \otimes O_{\cdot,i}}{\| f_i \otimes \eta_i \otimes O_{\cdot,i} \|}, \qquad \otimes \text{is elementwise product}$$



Learning Recipes - 4

- We represent each frame (y_i) as binary vector of occurance of visual objects and bag of language words.
- z_t is the activity of frame t and it follows an HMM.



Learning

- We use Metropolis–Hastings sampler for binary random variables and Gibbs sampler for the rest.
- Time ordering matrix and IBP is sampled via data driven manner (details omitted).

Preliminary Results

- We crawl the top 50 queries from the WikiHow.com and manually choose 25 out of them.
- For each query, we download 100 videos (total of 2500 videos, average length 7min, 80% have ASR subtitles)
- For each recipe, we have 5 evaluation videos with temporal activity labels as well as the activity names.
- We run the full model for 95 videos with no label and test in the remaining 5. For the test set, we do not sample the activities.
- We compute average intersection-over-union for the temporal segmentation (best over all matchings).

KMeans with Correct Number of States	20.74%
HMM with Correct Number of States	24.25%
Our Method	54.68%

Discussion

- Preliminary results suggest that the weak signal in the subtitles is powerful enough to recover activities.This can be used to scale-up activity detection algorithms.
- Activities are computed by using a generative model in other words we can have multiple/none activity in the same time.
- Integration of a recipe book is still an open problem and possibly a future work.

Dataset

(How to)+Bake Boneless Skinless Chicken, Cook Steak in a Frying Pan, Make Jello Shots, Tell if Gold Is Real, Bake Chicken Breast, Hard Boil an Egg, Make Pancakes, Tie a Bow Tie Broil Steak, Make a Grilled Cheese Sandwich, Make Scrambled Eggs, Tie a Tie, Clean a Coffee Maker, Make a Milkshake, Make Yogurt, Unclog a Bathtub Drain, Cook an Omelette, Make a Smoothie, Poach an Egg, Cook Lobster Tails, Make Beef Jerky, Remove Gum from Clothes, Cook Ribs in the Oven, Make Ice Cream, Tell if an Egg is Bad