Unsupervised Semantic Parsing of Video Collections

Ozan Sener, Amir Zamir, Silvio Savarese, Ashutosh Saxena

http://robo.watch

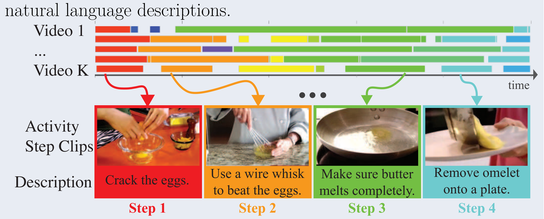
Motivation

- Instructional videos are generally structured and have a clear beginning, end, and a set of activity steps in between.
- Our goal is to discover this structure using large-scale data and unsupervised learning.

Problem Definition

- Given a collection of videos of same category

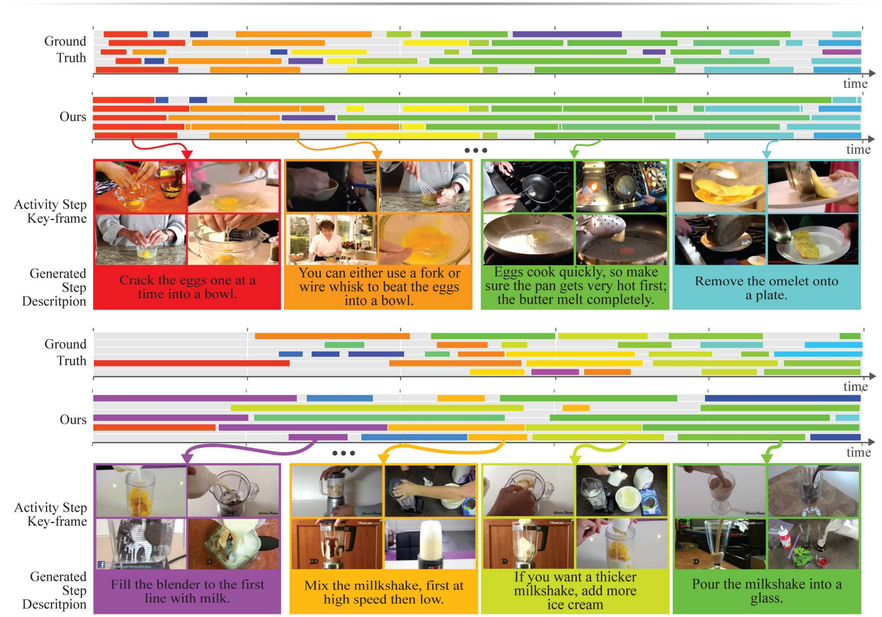
- We discover a common *semantic storyline* of activity steps with their natural language descriptions.



In order to do so; We propose a **fully unsupervised** and **multi-modal** framework;

- Learns a multi-modal dictionary of words and objects.
- Discovers a set of activity steps via *hierarchical beta process*
- Generates captions for each discovered activity via Markov language model.

Storylines



Our algorithm can generalize other domains like travel videos;

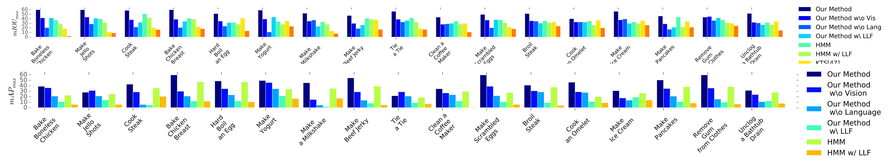
Cornell University

Stanford University

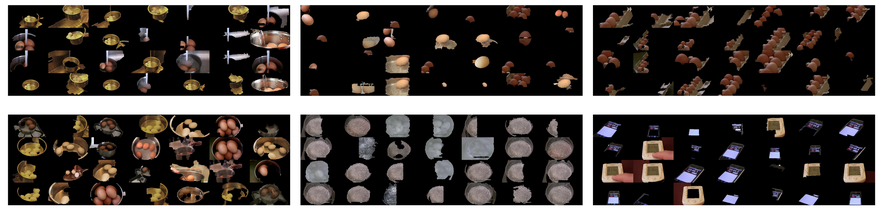
Quantitative Results

- mIOU: Mean Intersection over Union
- mAP: Mean Average Precision
- X_{cms}: Unsupervised extension of metric X through maximization over matching.

	KTS LLF	KTS Sem	HMM LL F	HMM Sem	Ours w/o Vis	Ours w/o Lng	Ours 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
mAP_{sem}	n/a	n/a	6.44	24.83	7.28	28.93	14.83	39.01

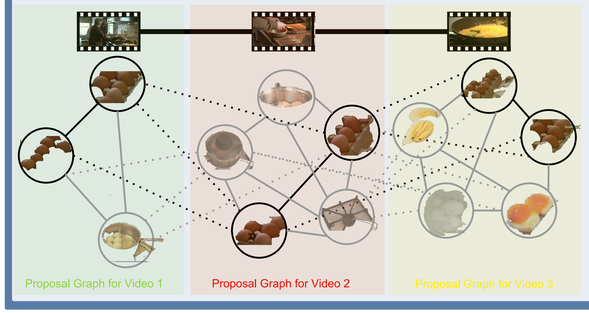


Sample Discovered Objects



Learning Visual Atoms

- Start with set of object proposals from all frames($0.1Hz$).
- Create a hierarchical graph of proposals by connecting $k-NN$ proposals within one video and across videos.
- Propose a hierarchical clustering algorithm to discover atoms.



Notation

- $\mathbf{A}^{(\mathbf{i})}$: Distance of the proposals within the video (i)
- $\mathbf{A}^{(\mathbf{i},\mathbf{j})}$: Distance of the proposals across the video (i) and (j)
- $\mathbf{x}^{(\mathbf{i})}$: Indicator vectors, $x_j^{(i)} = 1$ if j^{th} prop. is in the cluster

Optimization Problem

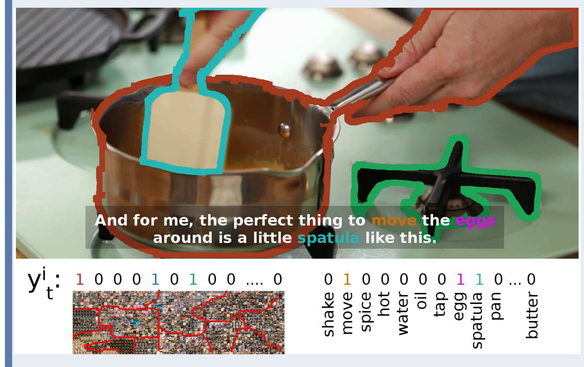
$$\arg\max \sum_{i\in N} \frac{\mathbf{x^{(i)^T} A^{(i)} x^{(i)}}}{\mathbf{x^{(i)^T} x^{(i)}}} + \sum_{i\in N}\sum_{j\in N(i)} \frac{\mathbf{x^{(i)^T} A^{(i,j)} x^{(j)}}}{\mathbf{x^{(i)^T} 1 1^T x^{(j)}}},$$

we use Stochastic Gradient Descent as the cost function is quasi-convex when relaxed as;

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

where $r^{(i)} = \frac{\mathbf{x^{(i)^T} A^{(i)} x^{(i)}}}{\mathbf{x^{(i)^T} x^{(i)}}}$ and $r^{(i,j)} = \frac{\mathbf{x^{(i)^T} A^{(i,j)} x^{(j)}}}{\mathbf{x^{(i)^T} 1 1^T x^{(j)}}}$

Representation and Learning



We represent each frame as a binary vector over the dictionary.

Notation

- Probability of seeing i^{th} atom in k^{th} activity is Bern. with Θ_k^i
- Activity id of the t^{th} frame of the i^{th} video is $z_t^{(i)}$
- $\pi^{(i)}$ sampled $\eta_{j,k}^{(i)} \sim Gam(\alpha + \kappa\delta_{j,k}, 1)$, $\pi_{\mathbf{j}}^{(\mathbf{i})} = \frac{\eta_{\mathbf{j}}^{(\mathbf{i})} \mathbf{o} \mathbf{f}^{(\mathbf{i})}}{\sum_k \eta_{j,k}^{(i)} f_k^{(i)}}$