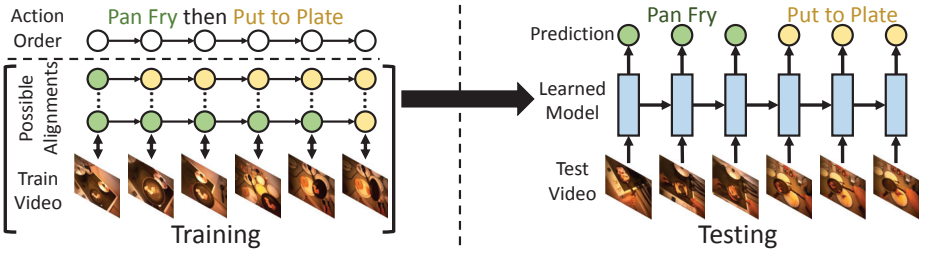
Connectionist Temporal Modeling for Weakly Supervised Action Labeling

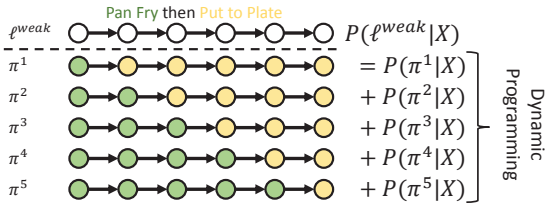
De-An Huang, Li Fei-Fei, Juan Carlos Niebles
Computer Science Department, Stanford University

1. Weakly Supervised Action Labeling

- Challenge: no temporal localization of the actions during training (only order)
- Contributions: (1) first introduce ECTC to efficiently evaluate all alignments
  (2) comparable to fully supervised methods with 1% label



2. Connectionist Temporal Modeling



- Challenge: large number of possible alignments
- CTC [2]: evaluate all possible alignments using DP

Path probability: $P(\pi|X) = \prod_{t=1}^{T} z_t^{\pi_t}, \quad z_t^k = P(k,t|X)$

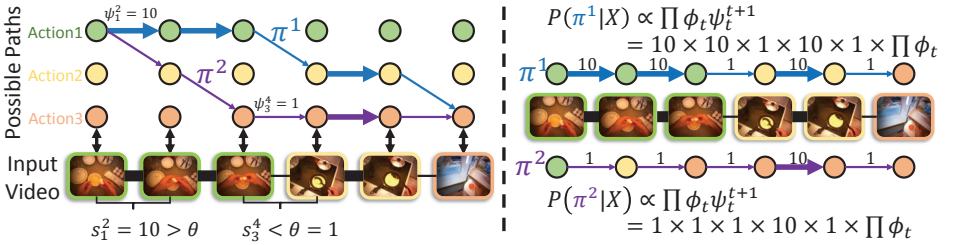
Label probability: $P(\ell|X) = \sum_{s=1}^{S} \frac{\alpha(s,t)\beta(s,t)}{z_t^{\ell_s}}$

Forward variable: $\alpha(s,t) = \sum_{\{\pi_{1:t}|B(\pi_{1:t})=\ell_{1:s}\}} P(\pi_{1:t}|X)$

$= z_t^{\pi_t}[\alpha(s,t-1) + \alpha(s-1,t-1)]$

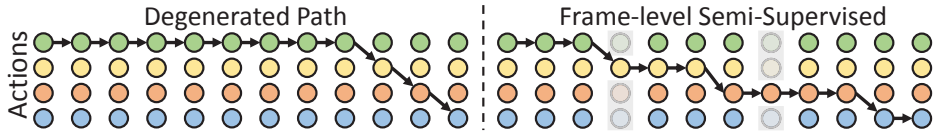
Backward variable: $\beta(s,t) = \sum_{\{\pi_{t:T}|B(\pi_{t:T})=\ell_{s:S}\}} P(\pi_{t:T}|X)$

3. Extended Connectionist Temporal Classification (ECTC)



- Challenge: much larger space of possible alignments with degenerated alignments
- ECTC: explicitly enforce consistency with the frame-to-frame visual similarities

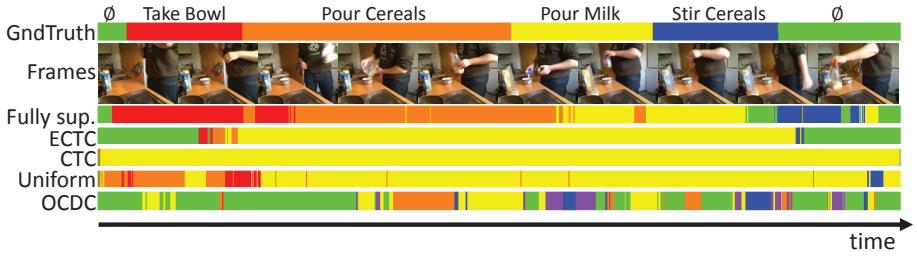
4. Frame-level Semi-supervised Learning



- Could be extracted from movie scripts or by labeling actions for a small number of frames
- Significantly reduce the alignment space and boosts the performance of our approach

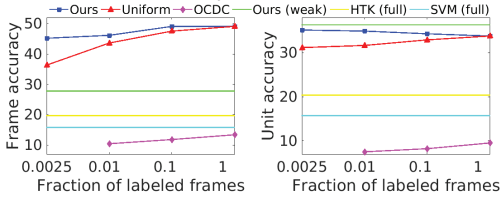
5. Evaluating Complex Activity Segmentation

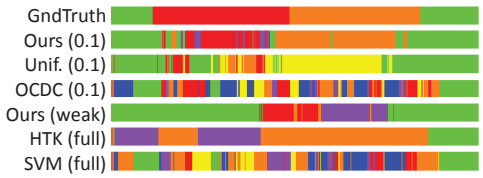
- The Breakfast Actions Dataset [3]



- OCDC: Ordering Constrained Discriminative Clustering [4]
- Uniform: uniformly distributing the occurring actions among frames
- HTK: Hidden Markov Model Toolkit [3]

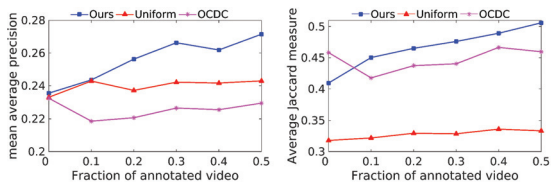
- Frame-level semi-supervised results

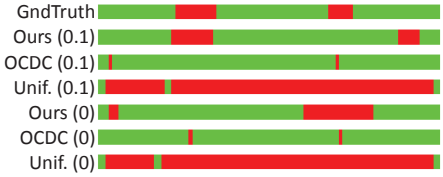




6. Evaluating Action Detection

- Clips from subset of the Hollywood2 Dataset [4]





- Weakly supervised action detection results

[1] Huang et al. "Connectionist Temporal Modeling for Weakly Supervised Action Labeling" ECCV 2016.
[2] Graves et al. "Connectionist Temporal Classification: Labelling Unsegmented. Sequence Data with Recurrent Neural Networks." ICML 2006.
[3] Kuehne et al. "The language of actions: Recovering the syntax and semantics of goal-directed human activities." CVPR 2014.
[4] Bojanowski et al. "Weakly supervised action labeling in videos under ordering constraints." ECCV 2014.