Actor-Action Semantic Segmentation with Grouping Process Models

Chenliang Xu and Jason J. Corso
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor

CVPR 2016

OBJECTIVE

We seek to label each pixel in a video with a pair of actor (e.g. adult, baby and dog) and action (e.g. eating, walking and jumping) labels.

- We propose a novel grouping process model (GPM) that adaptively adds long-ranging interactions of the supervoxel hierarchy to the labeling CRF.

- We incorporate the video-level recognition into segment-level labeling by the means of global labeling cost and the GPM.

Definition & Joint Modeling

Segment-Level:

$\mathcal{V} = \{q_1, q_2, \ldots, q_N\}$ - a video segmentation with n segments.

$\mathbf{L} = \{l_1, l_2, \ldots, l_N\}$ - a set of random variables defined on the segments taking labels from both actor space and action space, e.g. adult-eating, dog-crawling.

Supervoxel Hierarchy:

$\mathcal{T} = \{T_1, T_2, \ldots, T_S\}$ - a segmentation tree extracted from a supervoxel hierarchy with S total supervoxels.

$\mathbf{s} = \{s_1, s_2, \ldots, s_S\}$ - a set of binary random variables defined on the supervoxels denoting its active or not.

The Overall Objective Function:

$$(\mathbf{L}^*, \mathbf{s}^*) = \arg\min_{\mathbf{L},\mathbf{s}} E(\mathbf{L}, \mathbf{s}|\mathcal{V}, \mathcal{T})$$

$$E(\mathbf{L}, \mathbf{s}|\mathcal{V}, \mathcal{T}) = E^v(\mathbf{L}|\mathcal{V}) + E^h(\mathbf{s}|\mathcal{T}) + \sum_{t\in\mathcal{T}} (E^h(\mathbf{L}_t|s_t) + E^h(s_t|\mathbf{L}_t))$$

Grouping Cues from Segment Labeling. The GPM uses evidence directly from the segment-level CRF to locate supervoxels across various scales that best correspond to the actor and its action.

$$E^h(s_t|\mathbf{L}_t) = (\mathcal{H}(\mathbf{L}_t)|\mathbf{L}_t| + \theta_h)s_t$$

The Tree Slice Constraint. We seek a single labeling over the video. Each node in CRF is associated with one and only one supervoxel in the hierarchy. This constraint is the same as our previous work in Xu et al. ICCV 2013.

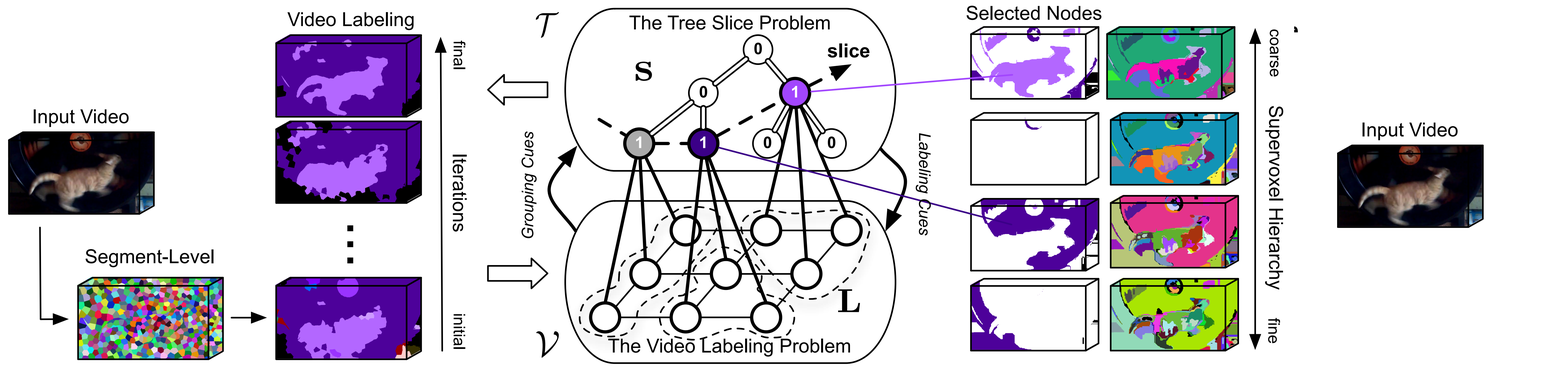
$$E^h(\mathbf{s}|\mathcal{T}) = \sum_{p=1}^{P} \delta(\mathbf{P}_p^\mathsf{T}\mathbf{s} \neq 1)\theta_\tau$$

Labeling Cues from Supervoxel Hierarchy. Once the supervoxels are selected, they provide strong labeling cues to the segment-level CRF. The CRF nodes connected to the same active supervoxel are encouraged to have the same label.

$$E^h(\mathbf{L}_t|s_t) = \begin{cases} \sum_{i\in\mathbf{L}_t}\sum_{j\neq i, j\in\mathbf{L}_t} \psi_{ij}^h(l_i, l_j) & \text{if } s_t = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_{ij}^h(l_i, l_j) = \begin{cases} \theta_t & \text{if } l_i \neq l_j \\ 0 & \text{otherwise} \end{cases}$$

Overview of the Grouping Process Model

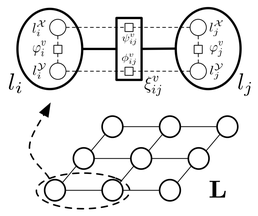


Segment-Level CRF

The segment-level CRF considers the interplay of actors and actions.

$\mathcal{X}$ - denotes the set of actor labels (e.g. adult, baby and dog).

$\mathcal{Y}$ - denotes the set of action labels (e.g. eating, running and crawling).



$$E^v(\mathbf{L}|\mathcal{V}) = \sum_{i\in\mathcal{V}} \xi_i^v(l_i) + \sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{E}(i)} \xi_{ij}^v(l_i, l_j)$$

$$\xi_i^v(l_i) = \psi_i^v(l_i^\mathcal{X}) + \phi_i^v(l_i^\mathcal{Y}) + \varphi_i^v(l_i^\mathcal{X}, l_j^\mathcal{Y})$$

$$\xi_{ij}^v(l_i, l_j) = \begin{cases} \psi_{ij}^v(l_i^\mathcal{X}, l_j^\mathcal{X}) & \text{if } l_i^\mathcal{X} \neq l_j^\mathcal{X} \wedge l_i^\mathcal{Y} = l_j^\mathcal{Y} \\ \phi_{ij}^v(l_i^\mathcal{Y}, l_j^\mathcal{Y}) & \text{if } l_i^\mathcal{X} = l_j^\mathcal{X} \wedge l_i^\mathcal{Y} \neq l_j^\mathcal{Y} \\ \psi_{ij}^v(l_i^\mathcal{X}, l_j^\mathcal{X}) + \phi_{ij}^v(l_i^\mathcal{Y}, l_j^\mathcal{Y}) & \text{if } l_i^\mathcal{X} \neq l_j^\mathcal{X} \wedge l_i^\mathcal{Y} \neq l_j^\mathcal{Y} \\ 0 & \text{if } l_i^\mathcal{X} = l_j^\mathcal{X} \wedge l_i^\mathcal{Y} = l_j^\mathcal{Y}. \end{cases}$$

Iterative Inference

Directly solving the overall objective function is hard. We use an iterative inference schema to efficiently solve it.

The Video Labeling Problem. Given a tree slice, we find the best labeling.

$$\mathbf{L}^* = \arg\min_{\mathbf{L}} E(\mathbf{L}|\mathbf{s}, \mathcal{V}, \mathcal{T}) = \arg\min_{\mathbf{L}} E^v(\mathbf{L}|\mathcal{V}) + \sum_{t\in\mathcal{T}} E^h(\mathbf{L}_t|s_t)$$

- Optimization depends on
- Solvable by graph-cuts multi-label inference.

The Tree Slice Problem. Given a labeling, we find the best tree slice.

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} E(\mathbf{s}|\mathbf{L}, \mathcal{V}, \mathcal{T}) = \arg\min_{\mathbf{s}} E^h(\mathbf{s}|\mathcal{T}) + \sum_{t\in\mathcal{T}} E^h(s_t|\mathbf{L}_t)$$

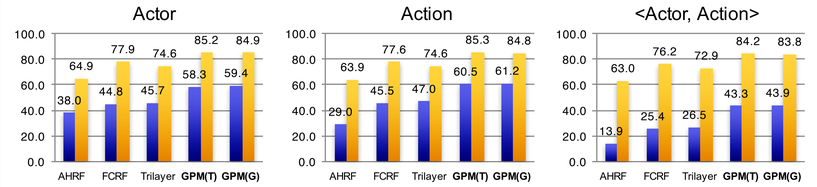
- Rewrite as a binary linear program.

$$\min \sum_{t\in\mathcal{T}} \alpha_t s_t \quad \text{s.t. } \mathcal{P}\mathbf{s} = \mathbf{1}_P \text{ and } \mathbf{s} \in \{0,1\}^\mathcal{S}$$

$$\alpha_t = \mathcal{H}(\mathbf{L}_t)|\mathbf{L}_t| + \theta_h$$

Experiments: The Actor-Action Semantic Segmentation

- Dataset: the A2D large-scale video labeling dataset.
  It consists of 3782 YouTube videos with an average length of 136 frames. One-third of videos have more than one actor performing different actions.
- Two different hierarchies: TSP and GBH.
- Video-level recognition is added through both global labeling cost and the GPM.



Blue: **Average Per-Class Accuracy**; Yellow: Global Pixel Accuracy

Visual example of the actor-action video labelings for all methods. (a) - (c) are videos where most methods get correct labelings; (d) - (e) are videos where GPM models outperform; (h) - (i) are different videos with partially correct labelings.

Acknowledgements. This work has been supported in part by Google, Samsung, DARPA W32P4Q-15-C-0070 and ARO W911NF-15-1-0354.