Leveraging Action Affinity and Continuity for Semi-supervised Temporal Action Segmentation

Guodong Ding and Angela Yao
National University of Singapore

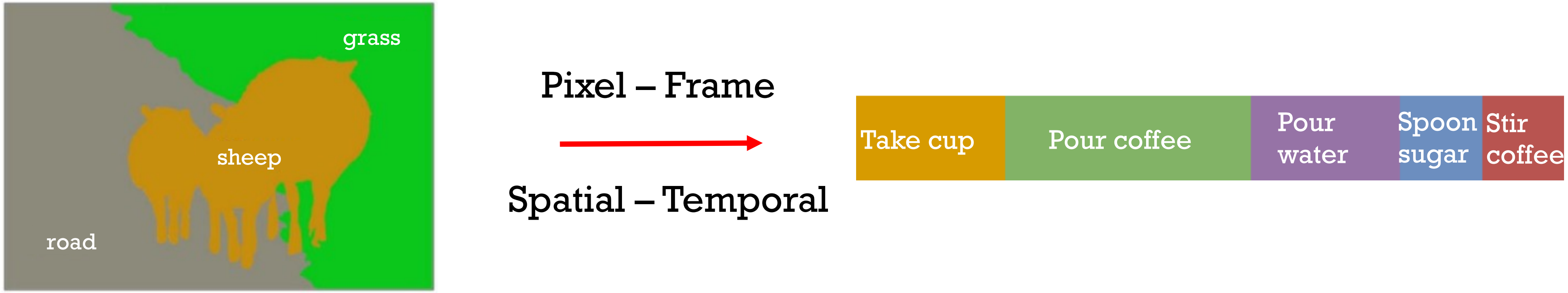
NUS
National University
of Singapore

THE TASK

Temporal Action Segmentation

- Temporally segments long-range procedural video
- Assigns semantic labels for each segment

1D Analogy of semantic segmentation



SEMI-SUPERVISED

Frame-wise annotation for procedural videos is time-consuming
- Number of videos (hundreds if not thousands)
- Temporal span of videos (minutes long)

Semi-supervised only requires
- A small portion of annotated videos (as low as 3)
- A large collection of videos unlabelled (cost free)

TRAINING UNLABELLED DATA

Incorporating unlabelled videos for training, factors to consider:

- What action compositions are likely to occur?

- What is a reasonable temporal proportion for each action to take?

- What kind of constraints should the action labels follow?

OBSERVATIONS

Action Affinity

- Videos performing the same activity will share a similar set of actions
- There exist pairs of videos sharing resembling action temporal portions

Action Continuity

- Action labels stay locally constant and only transit at the actual boundaries.
- Existing models tend to over-segment, leading to over-fragmentation problem

OBSERVATIONS

Action Affinity

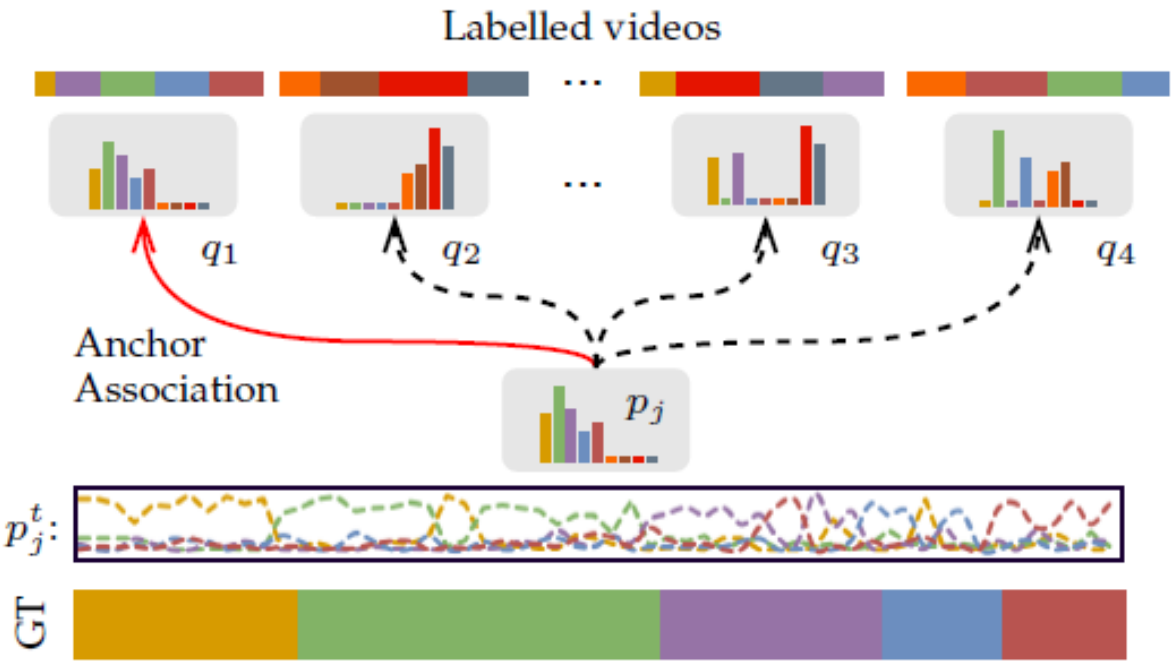
- Videos performing the same activity will share a similar set of actions
- There exist pairs of videos sharing resembling action temporal portions

Action Continuity

- Action labels stay locally constant and only transit at the actual boundaries.
- Existing models tend to over-segment, leading to over-fragmentation problem

AFFINITY LOSS

Impose the action prior induced from labelled videos to guide the learning of unlabelled samples.



Action frequency (labelled):

$$q_i(k) = \frac{1}{T_i}\sum_{t}^{T_i} \mathbb{1}(y_i^t == k); \quad k \in [1, \dots, K]$$

Action frequency (unlabelled):

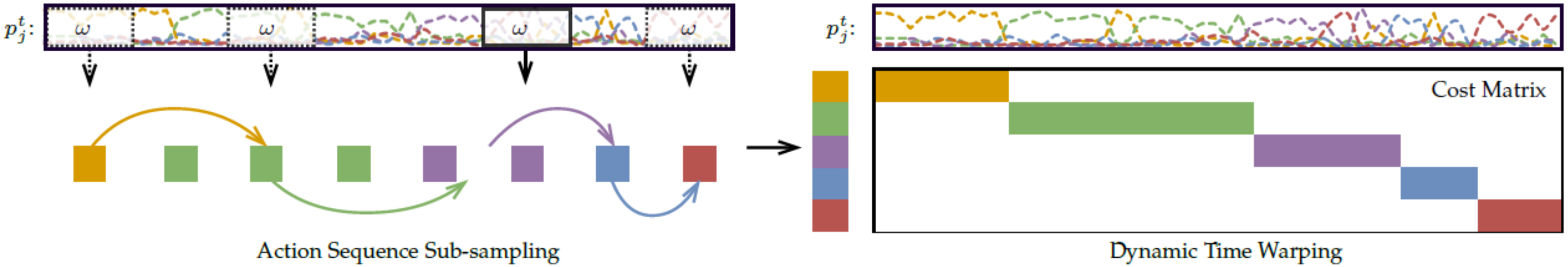
$$p_j(k) = \frac{1}{T_j}\sum_{t}^{T_j} p_j^t(k); \quad k \in [1, \dots, K].$$

Affinity loss w/ KL-Divergence:

$$\mathcal{L}_{\text{aff}} = \min_i \sum_k q_i(k) \log\left(\frac{q_i(k)}{p_j(k)}\right)$$

CONTINUITY LOSS

Mitigate the fragmentation problem in network predictions.



1. Sub-sample actions in time

$$o = \arg\max_k \frac{1}{\omega} \sum_{t=t'}^{t'+\omega} p^t(k),$$

2. Remove adjacent repetitive actions in sampled list

3. Using the KL-Divergence for cost calculation

$$d(l,t) = KL(o^l || p^t) = \sum_k o^l(k) \log \left(\frac{o^l(k)}{p^t(k)} \right).$$

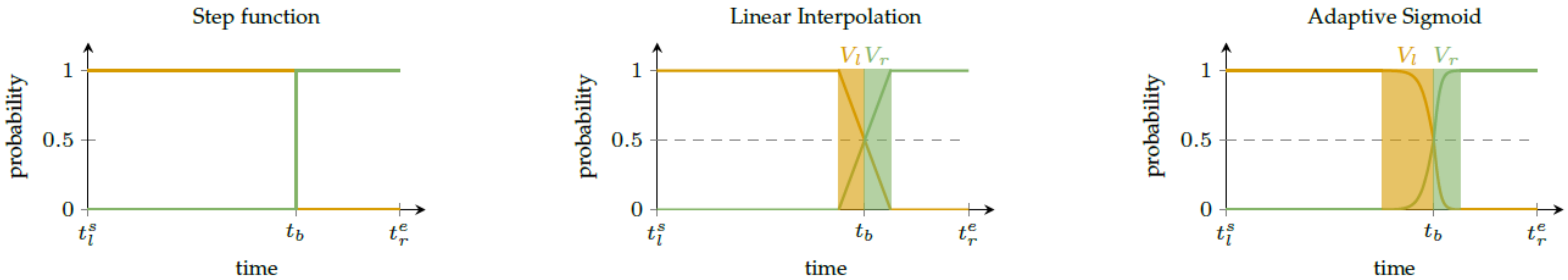
4. Optimize the cost along the optimal path.

$$\mathcal{L}_{\text{cont}} = \frac{1}{T} \min_Y \langle Y, \Delta \rangle = \frac{1}{T} \sum_t -\log(p^t(\tilde{y}^t)).$$

Identical to classification loss!

ADAPTIVE BOUNDARY SMOOTHING

NUS National University of Singapore | Computing



(a) Standard One-hot Labels (b) Fixed-Duration Linear [18] (c) ABS (Ours)

The adaptive boundary:

- Adopts a sigmoid shape for mixed action probability assignment
  - Faster probability descending speed when approaching the boundary
- Is proportional to the action duration
  - Smoothing in a longer boundary for long actions provides more training samples for adjacent shorter segments
  - Smoothing in a shorter boundary for short actions preserves more high confident frames for shorter segments

ABLATIONS & RESULTS

Pseudo-labeling

$\mathcal{L}_{cls}$	$\mathcal{L}_{pse}$	$\mathcal{L}_{aff}$	$\mathcal{L}_{cont}$	ABS	F1@10	F1@25	F1@50	Edit	Acc
✓					47.9	40.6	28.6	51.8	36.8
✓	✓				49.3	44.8	33.9	49.7	40.2
✓		✓			52.0	46.5	34.3	53.4	44.0
✓	✓	✓			54.1	46.7	34.9	54.1	47.8
✓		✓	✓		53.8	50.1	37.6	56.6	49.2
✓		✓	✓	✓	**56.9**	**51.3**	**39.0**	**57.7**	**49.5**

Action prior by the affinity loss is effective:

- Stand-alone outperforms Pseudo
- Avoid overfitting to the incorrect pseudo labels esp. when data annotation is rather limited.

$\%D_L$	Method	Breakfast					50Salads					GTEA				
		F1@10	F1@25	F1@50	Edit	Acc	F1@10	F1@25	F1@50	Edit	Acc	F1@10	F1@25	F1@50	Edit	Acc
5	Base	36.7	28.4	19.5	37.5	28.2	26.8	19.7	11.5	26.1	28.1	29.9	25.8	14.8	31.0	37.2
	Pseudo	40.2	28.5	20.1	41.3	20.9	22.6	17.0	12.1	22.0	24.0	48.4	42.3	30.2	45.4	48.1
	Ours	44.5	35.3	26.5	45.9	38.1	37.4	32.3	25.5	32.9	52.3	59.8	53.6	39.0	55.7	55.8
	Gain	**7.8**	**6.9**	**7.0**	**8.4**	**9.9**	**10.6**	**12.6**	**14.0**	**6.8**	**24.2**	**29.9**	**27.8**	**24.2**	**24.7**	**18.6**
10	Base	46.8	41.1	29.2	50.9	37.1	27.6	24.3	16.0	27.4	32.0	38.1	29.6	15.3	39.6	41.1
	Pseudo	49.3	44.8	33.9	49.7	40.2	36.2	32.4	24.5	33.5	41.1	65.5	60.7	45.8	59.9	57.9
	Ours	56.9	51.3	39.0	57.7	49.5	47.3	42.7	31.8	43.6	58.0	71.5	66.0	52.9	67.2	62.6
	Gain	**10.1**	**10.2**	**9.8**	**6.8**	**12.4**	**19.7**	**18.4**	**15.8**	**16.2**	**26.0**	**33.4**	**36.4**	**37.6**	**27.6**	**21.5**

ADAPTIVE BOUNDARY SMOOTHING

ABS is generic and applied to the fully supervised setting:

	Breakfast					50Salads					GTEA				
	F1@{10,25,50}			Edit	Acc	F1@{10,25,50}			Edit	Acc	F1@{10,25,50}			Edit	Acc
Base	63.2	57.7	45.6	65.5	65.1	66.8	63.7	55.2	59.8	78.2	84.9	82.4	67.6	79.7	76.6
+ABS	71.3	65.9	52.2	71.8	68.9	72.5	70.1	61.8	66.8	79.8	87.6	85.4	71.7	82.8	77.4
Gain	**8.1**	**8.2**	**6.6**	**6.3**	**3.8**	**5.7**	**6.4**	**6.6**	**7.0**	**1.6**	**2.7**	**3.0**	**4.1**	**3.1**	**0.8**

TAKEAWAYS

- Two novel loss functions are proposed specifically for the semi-supervised learning of temporal action segmentation task.

- The densely labelled videos do not only provide frame-wise semantic action labels, when put together at a video level, they also serve as action priors for a specific procedural task.

- The action boundary itself and the human annotations are ambiguous in pinpointing exact transiting timestamps. Transitional action boundaries can be helpful.

THANKS!