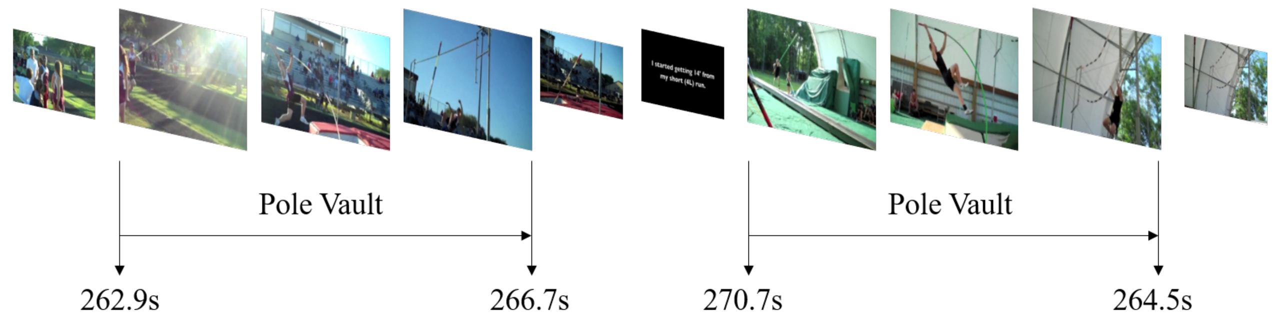
Learning Temporal Co-Attention Models for Unsupervised Video Action Localization

Guoqiang Gong, Xinghan Wang, Yadong Mu, Qi Tian
Peking University, Huawei

CVPR VIRTUAL JUNE 14-19

Introduction

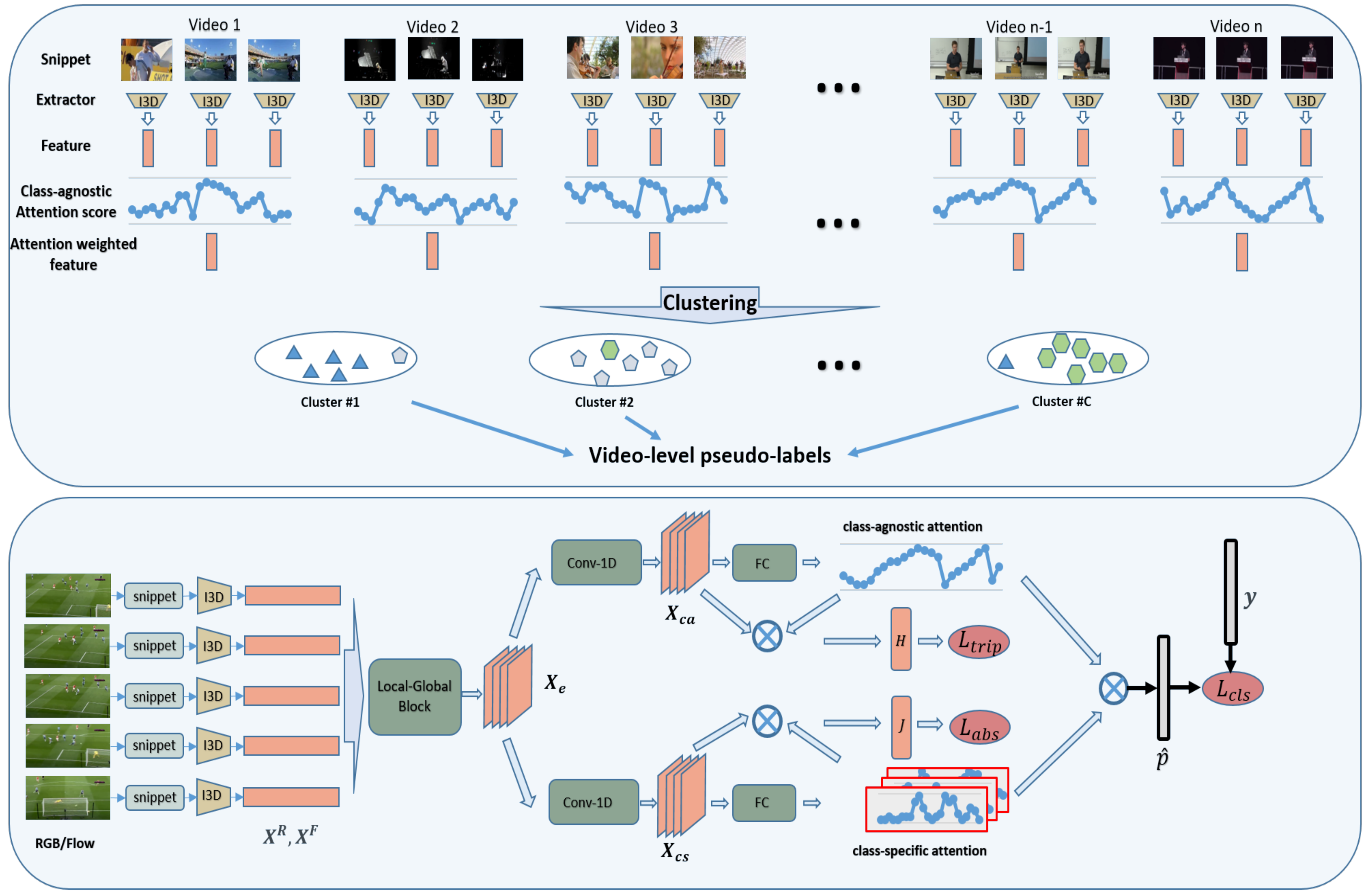
The goal of temporal action localization is to precisely find the starting and ending time for each action instance and determine its category from an untrimmed video. Most existing action localization methods are based on fully supervision, requiring manually annotated temporal boundaries and action category label for each action instance. However, delimiting the temporal boundary of an action instance is time-consuming. The scarcity of instance-level annotation has inspired recent works on weakly-supervised action localization methods. Specifically, for every training video, only a video-level action category is available. In this work, we propose an unsupervised temporal action localization task. In the unsupervised case, all we know regarding the training videos is the number of action categories in the video collection.



Contributions

- ✓ To our best knowledge, it is the first work that explores unsupervised temporal action co-localization (ACL) in the literature;
- ✓ This paper presents a novel two-step "clustering + localization" solution to the task of unsupervised ACL. In particular, we devise class-agnostic and class-specific temporal co-attentions, which are iteratively reinforced to gradually elevate the accuracy.
- ✓ Our comprehensive experiments on 20-action THUMOS14 and 100-action ActivityNet-1.2 have established first baselines and evaluation protocol for ACL. Surprisingly, the proposed model for ACL exhibits competitive performances to state-of-the-art weakly-supervised methods on both benchmarks.

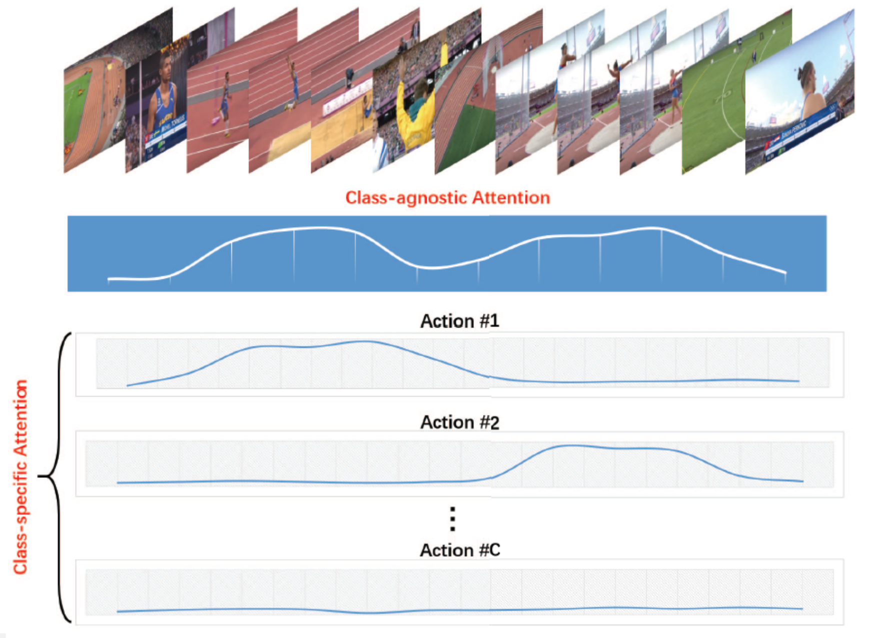
The Proposed Method



➢ Pipeline

We propose a two-step "clustering + localization" iterative procedure to solve unsupervised action localization. In the unsupervised case, true semantic annotations are missing, so we use clustering algorithm to group videos into C clusters, each of which defines a pseudo-action. Each unlabeled untrimmed video is assigned with a pseudo action class label based on clustering results. Then, two co-attention models will be learned based on these noisy video-level pseudo-labels, which is capable of detecting action instances and predicting their pseudo-labels.

➢ Illustration of two co-attention models



➢ Class-Specific Attention

- Model the temporal distribution of different actions.
- Generate and rank action-specific proposals.
- Mainly supervised by action-background separation loss(L_{abs}).

$$\mathcal{L}_{inter,z} = \sum_{m=1}^{K} \sum_{n=1,n\neq m}^{K} max\{d(J_m, J_n) - \tau_1, 0\}$$

$$\mathcal{L}_{intra,z} = \sum_{m=1}^{K} \sum_{n=1,n\neq m}^{K} max\{d(J_m, J_n) - d(J_m, B_m) + \tau_2, 0\}$$

$$\mathcal{L}_{abs} = \sum_{z=1}^{Z} (L_{inter,z} + \theta \cdot L_{intra,z})$$

➢ Class-Agnostic Attention

- Distinguish background and foreground frames.
- Modulate the video clustering step.
- Mainly supervised by cluster-based triplet loss (L_{trip}).

$$\mathcal{L}_{trip} = \sum_{z=1}^{Z} \sum_{a=1}^{K} max\{d(H_a, H_p) - d(H_a, H_n) + m, 0\}$$

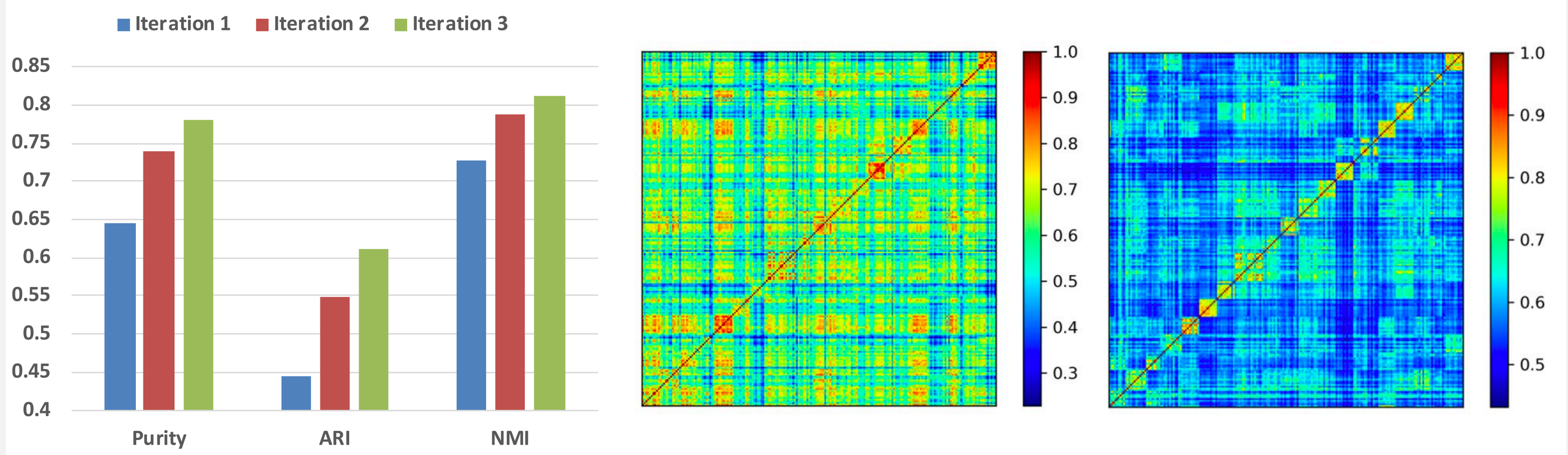
Experimental Results

	Methods	mAP@IoU (%)				
		0.3	0.4	0.5	0.6	0.7
FS	SLM-mgram [33]	30.0	23.2	15.2	-	-
	Glimpse [50]	36.0	26.4	17.1	-	-
	PSDF [54]	33.6	26.1	18.8	-	-
	S-CNN [39]	36.3	28.7	19.0	10.3	5.3
	SSAD [17]	43.0	35.0	24.6	-	-
	CDC [37]	40.1	29.4	23.3	13.1	7.9
	R-C3D [48]	44.8	35.6	28.9	-	-
	SSN [57]	51.9	41.0	29.8	-	-
	TAL-Net [4]	53.2	48.5	42.8	33.8	20.8
WS	Hide-and-seek [41]	19.5	12.7	6.8	-	-
	UntrimmedNet [46]	28.2	21.1	13.7	-	-
	STPN [27]	35.5	25.8	16.9	9.9	4.3
	Autoloc [38]	35.8	29.0	21.2	13.4	5.8
	W-TALC [29]	40.1	31.1	22.8	-	7.6
	MAAN [55]	41.1	30.6	20.3	12.0	6.9
	CMCS [20]	41.2	32.1	23.1	15.0	7.0
	3C-Net [26]	44.2	34.1	26.6	-	8.1
	BM [28]	46.6	37.5	26.8	17.6	9.0
	TSM [53]	39.5	-	24.5	-	7.1
	CleanNet [14]	37.0	30.9	23.9	13.9	7.1
	Ours	**46.9**	**38.9**	**30.1**	**19.8**	**10.4**
US	Ours	39.6	32.9	25.0	16.7	8.9

	Methods	mAP@IoU (%)			
		0.5	0.75	0.95	Average
WS	FC-CRF [58]	27.3	14.7	2.9	15.6
	AutoLoc [38]	27.3	15.1	3.3	16.0
	W-TALC [29]	37.0	-	-	18.0
	CMCS [20]	36.8	22.0	**5.6**	22.4
	3C-Net [26]	37.2	-	-	21.7
	TSM [53]	28.3	17.0	3.5	-
	CleanNet [14]	37.1	20.3	5.0	21.6
	Ours	**40.0**	**25.0**	4.6	**24.6**
US	Ours	35.2	21.4	3.1	21.1

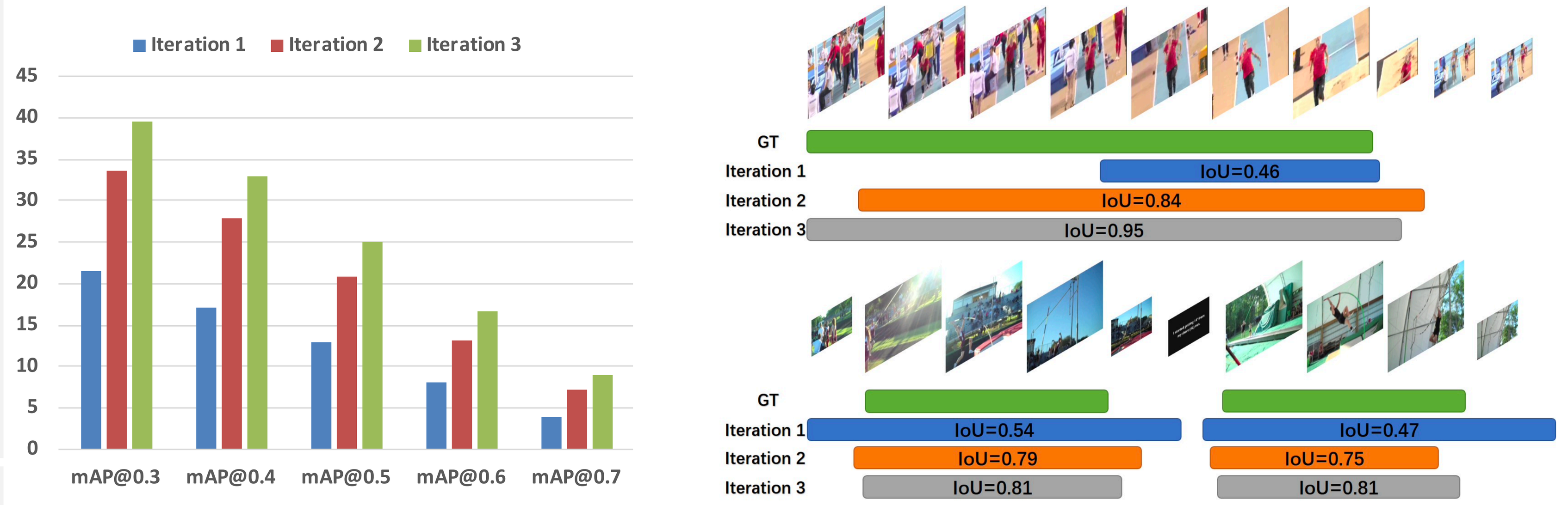
Comparisons on the THUMOS14 and ActivityNet-1.2. We denote fully-supervised, weakly-supervised and un-supervised as FS, WS and US respectively.

Clustering Results of Different Iterations



Visualize video clustering results and affinity matrices used for spectral clustering of different iterations on THUMOS14 validation set.

Localization Results of Different Iterations



Action localization results and qualitative examples of localization results on THUMOS14 testing set.