Multi-Resolution Audio-Visual Feature Fusion for Temporal Action Localization

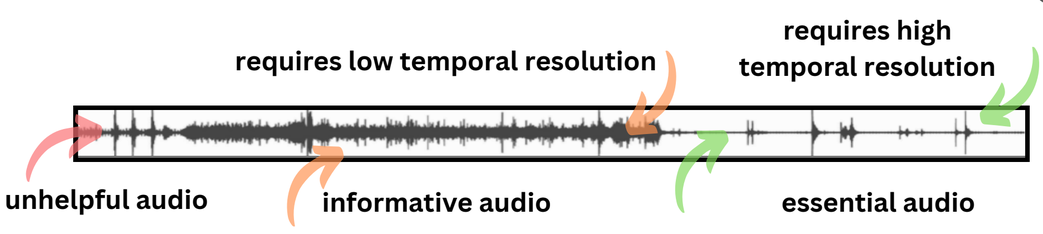
Edward Fish, Jon Weinbren, Andrew Gilbert | University of Surrey | edward.fish@surrey.ac.uk

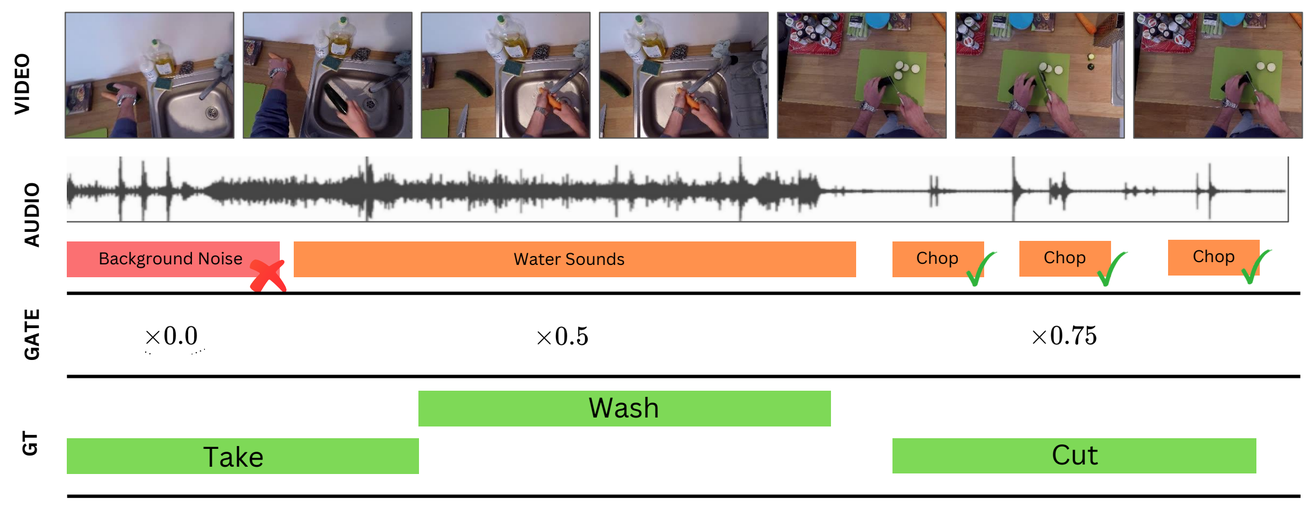
Audio is not always useful for temporal action localization in videos. Gated audio-visual feature fusion at different temporal resolutions improves performance.

Read the paper

In Temporal Action Localization (TAL) we are searching for the start and end of an action in a video as well as its label.

Adding audio features to Temporal Action Localization degrades performance of some classes because audio is not always useful for the localization task.





Feature Pyramid Network (FPN): Encodes audio-visual features along different temporal resolutions, selectively gating audio features based on their relevance to action classification and regression.

Multi-Resolution Fusion: The framework applies a cross-attention mechanism to audio and visual features, followed by a gated fusion that refines the contribution of each modality over each temporal resolution.

In this example background noise is unhelpful for finding the task "Take". Water sounds are informative for the label "Wash" but do not inform when the object is washed. Chopping sounds are highly informative for both action boundaries and labels.

Feature Extraction

Visual and audio features are extracted from pretrained I3D and VGG networks. We use 1D convolution followed by local self-attention.

Downsampling

Feature Pyramid Network with max pooling downsamples features at different temporal resolutions.

Multi-Resolution Gating and Cross-Attention

Cross Attention between visual and audio features at each temporal resolution is applied. We include a sigmoid gate from the visual pathway to scale the magnitude of the audio and visual features during fusion.

Regression and Classification

Fused temporal features are processed via separate convolutional regression and classification heads to obtain start and end regression scores and classification label.

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	BMN [27, 11]	10.8	9.8	8.4	7.1	5.6	8.4
	G-TAD [57]	12.1	11.0	9.4	8.1	6.5	9.4
	ActionFormer [63]	26.6	25.4	24.2	22.3	19.1	23.5
	TemporalMaxer [47]	27.8	26.6	25.3	23.1	19.9	24.5
	ActionFormer + MRAV-FF	27.6	26.8	25.3	23.4	19.8	24.6
	TemporalMaxer + MRAV-FF	**28.5**	**27.4**	**26.0**	**23.7**	**20.12**	**25.1**
Noun	BMN [27, 11]	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [57]	11.0	10.0	8.6	7.0	5.4	8.4
	ActionFormer [63]	25.2	24.1	22.7	20.5	17.0	21.9
	TemporalMaxer [47]	26.3	25.2	23.5	21.3	17.6	22.8
	ActionFormer + MRAV-FF	26.4	25.4	23.6	21.2	17.4	22.8
	TemporalMaxer + MRAV-FF	**27.4**	**26.2**	**24.4**	**21.8**	**17.9**	**23.5**

Table 1: The performance of our proposed method on the EPIC-Kitchens 100 dataset. [11]

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	Concatenation	28.02	26.96	25.5	23.48	19.87	23.89
	Channel Pooling	25.63	24.59	23.09	21.14	17.95	23.06
	MRAV-FF	**28.5**	**27.4**	**26.0**	**23.7**	**20.12**	**25.1**
Noun	Concatenation	26.39	25.42	23.57	21.19	17.42	22.8
	Channel Pooling	25.7	24.53	22.95	20.52	17.04	22.21
	MRAV-FF	**27.4**	**26.2**	**24.4**	**21.8**	**17.9**	**23.5**

Table 2: Results for an ablation experiment on EPIC-Kitchens 100 [11] TAL task, where we replace the MRAV-FF module with existing approaches to feature fusion including concatenated projection and channel pooling. We observe that simple fusion methods hinder performance when compared with uni-modal FPN networks demonstrating the need for a more nuanced fusion strategy.

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	Damen [12]	10.83	9.84	8.43	7.11	5.58	8.36
	AGT [38]	12.01	10.25	8.15	7.12	6.14	8.73
	OWL [41]	14.48	13.05	11.82	10.25	8.73	11.67
	MRAV-FF	**28.5**	**27.4**	**26.0**	**23.7**	**20.12**	**25.1**
Noun	Damen [12]	10.31	8.33	6.17	4.47	3.35	6.53
	AGT [38]	11.63	9.33	7.05	6.57	3.89	7.70
	OWL [41]	17.94	15.81	14.14	12.13	9.80	13.96
	MRAV-FF	**27.4**	**26.2**	**24.4**	**21.8**	**17.9**	**23.5**

Table 3: The performance of our proposed method on the EPIC-Kitchens 100 dataset [11] compared to existing approaches for audio-visual feature fusion on TAL. Our method demonstrates a large increase in performance jointly attributed to the addition of feature pyramid architecture and our fusion strategy.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911