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

Edward Fish
University of Surrey
edward.fish@surrey.ac.uk

Jon Weinbren
University of Surrey
j.weinbren@surrey.ac.uk

Andrew Gilbert
University of Surrey
a.gilbert@surrey.ac.uk

0.1 Introduction

In this appendix we provide further information about the network architecture proposed in the paper, training details, feature extraction, additional results, and discussion.

0.2 Feature extraction

Visual Features: We use the features provided by existing works in TAL **???**. For EPIC-Kitchens features are extracted using a SlowFast network **?** pre-trained on EPIC-Kitchens **?**. During extraction we use a 32-frame input sequence with a stride of 16 to generate a set of 2304-D features.

Audio Features: For the audio preprocessing and feature extraction, we followed a series of well-established steps to derive meaningful representations:

1. **Resampling:** All audio data was resampled to a uniform rate of 16 kHz in mono.
2. **Spectrogram Computation:** We computed the spectrogram by extracting magnitudes from the Short-Time Fourier Transform (STFT). This utilized a window size of 25 ms, a hop size of 10 ms, and a periodic Hann window for the analysis.
3. **Mel Spectrogram Mapping:** The computed spectrogram was then mapped to a mel scale, producing a mel spectrogram with 64 mel bins that cover the frequency range from 125 Hz to 7500 Hz.
4. **Log Mel Spectrogram Stabilization:** To enhance the stability and avoid issues with the logarithm function, we calculated a stabilized log mel spectrogram as:

   $$\text{Log-Mel} = \log(\text{Mel-Spectrogram} + 0.01)$$

   Here, the offset of 0.01 prevents the computation of the logarithm of zero.
5. **Framing:** Finally, the derived features were segmented into non-overlapping examples spanning 0.96 seconds each. Every example encapsulates 64 mel bands and 96 time frames, with each frame lasting 10 ms.

Following extraction the features are projected to 128-D features via a pre-trained VGG audio encoder network **?** pretrained on AudioSet **?**. The network outputs embeddings of shape $T \times 128$ where T is the temporal input dimension as defined in the paper.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

0.3 Ablation Results

We perform initial ablation experiments to evaluate the performance of our proposed method and present the results in Tab 1. Each experiment is conducted on EPIC-Kitchens, where we edit the temporal fusion method in each temporal block. We first exchange our MRAV-FF temporal block for simple feature fusion in which we concatenate and project the audio-visual features at each temporal scale via a 1D-CNN. We notice that this actually harms network performance over unimodal features demonstrating the need for a gated approach to fusion. Similarly we also replace the block with a max-pooling layer inspired by **?** which pools channel-wise for feature fusion. Again this method has a negative impact on network performance.

0.4 Further results

Furthermore in Tab 2 we evaluate our method with other approaches to audio-visual fusion for TAL on EPIC-Kitchens. We show a large increase in performance, which can be attributed to both the effectiveness of the FPN structure for audio visual temporal pooling and also our MRAV-FF fusion module. The lack of available comparative methods for audio-visual fusion further illustrates the importance of updated baselines in this field.

Finally, we also evaluate the method on the THUMOS14 dataset which **?** contains 200 validation videos and 213 testing videos with 20 action classes. THUMOS14 presents a different challenge to ego-centric audio-visual fusion, since the videos are heavily edited and contain many actions that do not have audio-visual alignment. For example, many videos are of sporting events where there is no localized audio information, contain music, narration, or have no audio at all. Due to these challenges there are no existing TAL audio-visual fusion works that test their methods on THUMOS14.

Following previous work **?????**, we trained the model on the validation set and evaluate on the test set. Our results in Tab 3 demonstrate that our method struggles to handle this audio-visual disparity only improving on the 0.7 iou threshold.

0.5 Discussion

In **?**, the authors propose that the research direction in audio-visual fusion for TAL should be split in to edited vs non-edited video. However, there exist very few large datasets of video content that match the audio-visual quality of EPIC-Kitchens. Our future direction in this field is to develop our MRAV-FF approach to encompass audio separation modules to handle both edited and un-edited content to improve performance on the THUMOS task and develop a unified system for improved audio-visual TAL. Ultimatley, we believe that TAL methods should handle a diverse range of video content 'in the wild' and approaches should not be designed around specific datasets. We also plan to thoroughly evaluate the gating mechanism proposed to understand which actions require more or less audio-visual information and how this can be utilised for more effective fusion.

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 1: Results for an ablation experiment on EPIC-Kitchens 100 **?** TAL task, where we replace the MRAV-FF module with existing approached 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.

References

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	Damen ?	10.83	9.84	8.43	7.11	5.58	8.36
	AGT ?	12.01	10.25	8.15	7.12	6.14	8.73
	OWL ?	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 ?	10.31	8.33	6.17	4.47	3.35	6.53
	AGT ?	11.63	9.33	7.05	6.57	3.89	7.70
	OWL ?	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 2: The performance of our proposed method on the EPIC-Kitchens 100 dataset ? compared to existing approaches for audio-visual feature fusion on TAL. Our method demonstrates a large increase performance jointly attributed to the addition of feature pyramid architecture and our fusion strategy.

Type	Model	Feature	tIoU↑						time(ms) ↓
			0.3	0.4	0.5	0.6	0.7	Avg.	
Two-Stage	BMN ?	TSN ?	56.0	47.4	38.8	29.7	20.5	38.5	483*
	DBG ?	TSN ?	57.8	49.4	39.8	30.2	21.7	39.8	—
	G-TAD ?	TSN ?	54.5	47.6	40.3	30.8	23.4	39.3	4440*
	BC-GNN ?	TSN ?	57.1	49.1	40.4	31.2	23.1	40.2	—
	TAL-MR ?	I3D ?	53.9	50.7	45.4	38.0	28.5	43.3	>644*
	P-GCN ?	I3D ?	63.6	57.8	49.1	—	—	—	7298*
	P-GCN ? +TSP ?	R(2+1)1 D ?	69.1	63.3	53.5	40.4	26.0	50.5	—
	TSA-Net ?	P3D ?	61.2	55.9	46.9	36.1	25.2	45.1	—
	MUSES ?	I3D ?	68.9	64.0	56.9	46.3	31.0	53.4	2101*
	TCANet ?	TSN ?	60.6	53.2	44.6	36.8	26.7	44.3	—
	BMN-CSA ?	TSN ?	64.4	58.0	49.2	38.2	27.8	47.7	—
	ContextLoc ?	I3D ?	68.3	63.8	54.3	41.8	26.2	50.9	—
	VSGN ?	TSN ?	66.7	60.4	52.4	41.0	30.4	50.2	—
	RTD-Net ?	I3D ?	68.3	62.3	51.9	38.8	23.7	49.0	>211*
	Disentangle ?	I3D ?	72.1	65.9	57.0	44.2	28.5	53.5	—
	SAC ?	I3D ?	69.3	64.8	57.6	47.0	31.5	54.0	—
Single-Stage	A^{2}Net ?	I3D ?	58.6	54.1	45.5	32.5	17.2	41.6	1554*
	GTAN ?	P3D ?	57.8	47.2	38.8	—	—	—	—
	PBRNet ?	I3D ?	58.5	54.6	51.3	41.8	29.5	—	—
	AFSD ?	I3D ?	67.3	62.4	55.5	43.7	31.1	52.0	3245*
	TAGS ?	I3D ?	68.6	63.8	57.0	46.3	31.8	52.8	—
	HTNet ?	I3D ?	71.2	67.2	61.5	51.0	39.3	58.0	—
	TadTR ?	I3D ?	74.8	69.1	60.1	46.6	32.8	56.7	195*
	GLFormer ?	I3D ?	75.9	72.6	67.2	57.2	41.8	62.9	—
	AMNet ?	I3D ?	76.7	73.1	66.8	57.2	42.7	63.3	—
	ActionFormer ?	I3D ?	82.1	77.8	71.0	59.4	43.9	66.8	80
	ActionFormer ? + GAP ?	I3D ?	82.3	—	71.4	—	44.2	66.9	>80
	TemporalMaxer	I3D ?	**82.8**	**78.9**	**71.8**	**60.5**	44.7	**67.7**	**50**
	TemporalMaxer + MRAVFF	I3D ? + Audio ?	82.2	78.2	71.5	59.9	**45.3**	67.4	60

Table 3: Performance of our method on the THUMOS dataset for TAL. We observe that audio-visual fusion on edited videos is much more challenging that the raw-video setting due to the addition of background music, narration, and audio-visual misalignment.