Rethink Cross-Modal Fusion in Weakly-Supervised Audio-Visual Video Parsing (Supplementary Material)

Yating Xu Conghui Hu Gim Hee Lee
Department of Computer Science, National University of Singapore
xu.yating@u.nus.edu conghui@nus.edu.sg gimhee.lee@nus.edu.sg

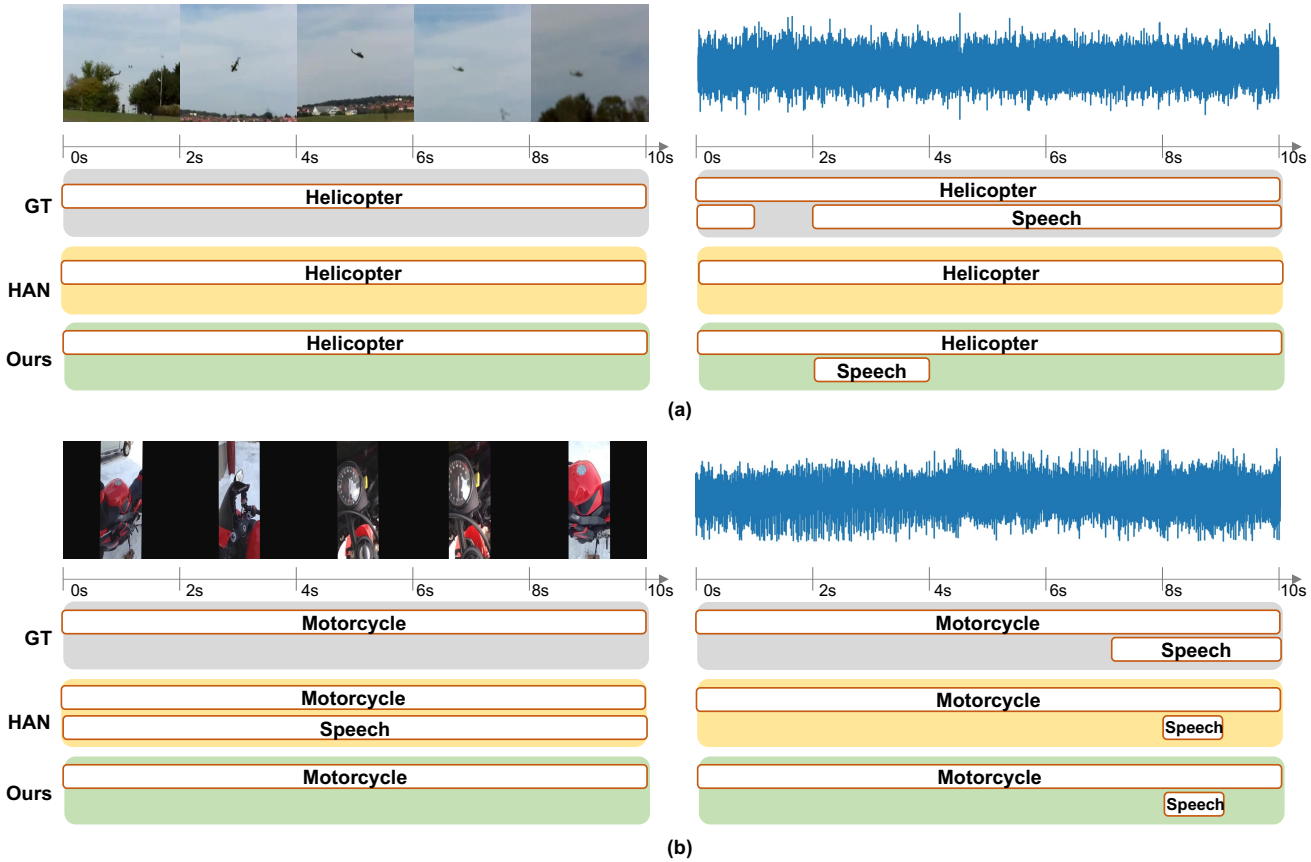

Figure 1. Qualitative comparison with HAN.

A. Additional Analysis of CAPC.

Tab. 1 presents the results of detecting visual events by different models. We split the visual events into visual-exclusive and multi-modality events. Visual-exclusive events refers to events only happening in the visual modality, while multi-modality events appear on audio and visual streams with temporal overlap. The results are averaged F-scores per event. As shown in Tab. 1, the results for the visual-exclusive events consistently improve with an increase in the value of N, suggesting the CAPC can effectively reduce the influence of unmatched audio context. However, the accuracy of multi-modality events drops when $N = 3$, suggesting that large N can impede the positive impact of audio on its correlated visual event. Only a small number of N, *i.e.* $N = 1$, improves fusion effectiveness for both events.

Model	Visual Event	
	Visual-exclusive	Multi-modality
No CAPC	35.6	63.0
N=1	38.5	**63.6**
N=3	**38.6**	62.9

Table 1. Additional Analysis of cross-audio prediction consistency (CAPC).

B. Qualitative Comparison with HAN

Fig. 1 shows the qualitative results. HAN [1] fails to identify the single-modality event ('Speech') in Fig. 1(a) or wrongly detect it on two modalities in Fig. 1(b), suggesting the information on the audio and visual streams are highly confounded. In contrast, our model can correctly detect the audio and visual events.

C. Illustration of Single-modality and Multi-modality Event

Single-modality event in Table 1 of main paper refers to events *only* happening in one modality, while audio (visual) event in Table 2 includes event both only happening in the audio (visual) and happening in the audio and visual modality.

Multi-modality event in Table 1 of main paper refers to event appearing with temporal overlap (either partial or full overlap) on audio and visual streams, while audio-visual event in Table 2 refers to event with *full* overlap on audio and visual streams.

The split of single-modality and multi-modality event in Table 1 is to illustrate that the strong entanglement with another non-fully correlated modality is harmful in detecting its own exclusive events. The evaluation of audio, visual and audio-visual event in Table 2 is the standard benchmark of audio-visual video parsing task [1].

References

[1] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 2