On Audio-Visual Sound Source Segmentation: Towards Better Multimodal Perception

William T. Doan, Siva Sai Nagender Vasireddy, and Yapeng Tian

THE UNIVERSITY OF TEXAS AT DALLAS
Undergraduate Research Apprenticeship Program

Introduction

Abstract

The audio-visual segmentation (AVS) problem seeks output a correct, pixel-wise mapping of sound producing object(s). Unfortunately, current AVS solutions are weak as their collective cross-modal alignment is off. We propose forth a furtherance of the AVS task: Audio-Visual Sound Source Segmentation (AVSSS).

Fig. 1. From left to right: The original video, the isolated sounding object, and the segmented polygon.

Figure 1 is a canonical demonstration of the AVS task and is typical of most datasets in this field. This leaves models susceptible to annotating the most salient object in a video as opposed to actual sounding objects.

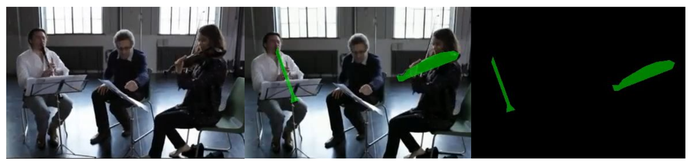

Fig. 2. From left to right: The original video, the isolated sounding objects, and the segmented polygons.

Our work, utilizing the LLP dataset, fills the dearth of hard samples which exists. See Figure 2. An AVS model's zero-shot performance in like cases would be especially illuminating and bring the field closer to genuine audio-visual perception.

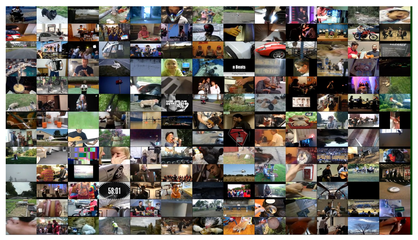

Fig. 3. The LLP Dataset.

Methods and Analysis

Annotating the Dataset

To build our AVSSS dataset, we used videos from the LLP dataset and used CVAT and the VGG Image Annotator (VIA) to annotate key frame ground-truth annotations for: single audible objects, multiple audible objects, silent objects, and mixed audible objects.

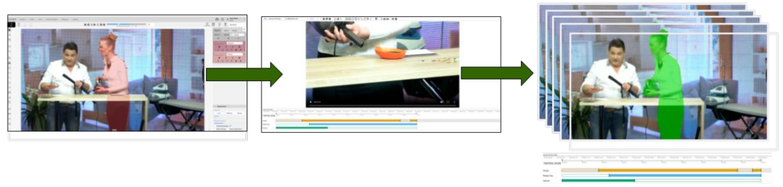

Fig. 4. The Annotation Process: CVAT for video, VIA for audio, and finally exported.

Findings

Presently, the AVSSS dataset has over 200 annotations of randomly selected videos from the LLP dataset. We tested AVSBench on five fully annotated videos, and our results in Table 1. are promising.

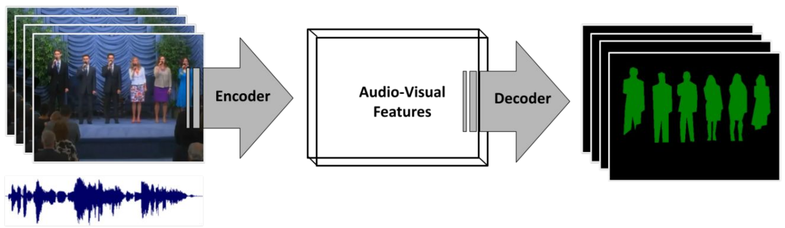

Fig. 5. AVSBench Network Architecture: Audio and Visual Inputs are given, then processed through the Encoder before the Decoder enlarges and outputs the feature maps.

	mIoU	F-Score
AVSBench on original dataset	0.787	0.879
AVSBench on our validation set	0.786	0.877

Table 1.

AVSBench performed well on our preliminary validation set, demonstrating relative parity across all classes of our dataset with a percent difference of 0.127% and 0.228% for the mean value and F-scores respectively.

Conclusion

Impact

The work of this project is ongoing and promising, especially in the realm of developing a model which can surmount segmenting the most salient object in a video—instead segmenting genuine sounding objects—using pixel-wise spatiotemporal annotations.

A phenomenon [1] found is most models are fusing audio and visual features and then try to predict segmentation masks. These are not meaningful to the broader multimodal understanding of the sample. The reason for that is the segmentations produced are more likely to classify the most prominent objects in a video, regardless of whether or not sound is actually being produced.

AVSSS is a solution to this problem. By deploying a dataset having both audio and visual annotations which include multiple-source and silent-source objects, AVSSS can bring the field closer to more robust sound-source localization.

References

[1] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In Proceedings of the 31st ACM International Conference on Multimedia, MM '23, page 7590–7598, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In European Conference on Computer Vision (ECCV), 2020.

[3] Jinxing Zhou, Jianyuan Wang, Jiayi Eris Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. ArXiv, abs/2207.05042, 2022.4

References

William T. Doan
w.doan@utdallas.edu
https://wtd-1.github.io/web/

Acknowledgments:
I would like to thank Dr. Yapeng Tian for his patience and guidance during this project, as well as Mr. Siva Sai Nagender Vasireddy for his sustained commitment to this work. AVSSS was brought about in part by the generous support of the Undergraduate Research Apprenticeship Program at the University of Texas at Dallas.