Unified Multisensory Perception: Weakly-Supervised Audio-Visual Video Parsing

Yapeng Tian[1], Dingzeyu Li[2], and Chenliang Xu[1]

[1] University of Rochester
[2] Adobe Research
{yapengtian,chenliang.xu}@rochester.edu; dinli@adobe.com

Appendix

In this appendix, we first describe implementation details of our models in Sec. 1. Then, we compare our LLP dataset with several related existing datasets in Sec 2. Finally, we illustrate different multimodal temporal contexts within in a single video in Sec. 3.

1 Implementation Details

For a 10-second long video, we first sample video frames at $8fps$, and each video is divided into non-overlapping snippets of the same length with 8 frames in 1 second. Given a visual snippet, we extract a 2048-D feature for each frame using ResNet152 [5] pretrained on ImageNet [3] and obtain a 512-D spatio-temporal visual feature from 8 frames using a 3D ResNet [16] pre-trained on Kinetics [2]. To obtain a snippet-level visual feature, we first reduce frame-level feature dimension to 512 with a fully-connected (FC) layer and then temporally averaging pool the 8 frame-level features. A 512-D snippet-level visual feature is predicted using an additional FC layer to process the concatenated the temporally pooled spatial and spatio-temporal features. Audio representation is first extracted via a pre-trained VGGish network [6] on AudioSet [4], which extracts a 128-D feature for each 1s audio snippet and then projects it to 512-D with a FC layer. In addition, $d = 512$, $T = 10$, $C = 25$, $\epsilon_a = 1.0$, $\epsilon_v = 0.9$, and $K = 2$ in our experiments.

2 Differences between Existing Datasets and Our LLP

We compare our LLP dataset with some related existing datasets: UrbanSound [11], DCASE2018 [12], THUMOS14 [7], Charades [13], ActivityNet [1], and AVE [14] in Table 1. The UrbanSound and DCASE2018 are sound detection datasets from urban and domestic domains, respectively. THUMOS14, Charades, and ActivityNet are video datasets containing different human activities. AVE is collected for audio-visual event localization. From Tab. 1, we can see that only our LLP dataset has different modality types of event annotations, which are required for learning audio-visual video parsing.

Table 1: Comparison with different datasets. We can see that only our LLP dataset contains different modality types of event annotations, which are required for learning audio-visual video parsing.

Datasets	Context	#Video	Audio Event Label	Visual Event Label	Audio-Visual Event Label
UrbanSound [11]	Urban	8,732	✓	✗	✗
DCASE2018 [12]	Domestic	1,866	✓	✗	✗
THUMOS14 [7]	Human Activity	413	✗	✓	✗
Charades [13]	Human Activity	9,848	✗	✓	✗
ActivityNet [1]	Human Activity	19,994	✗	✓	✗
AVE [14]	Open	4,143	✗	✗	✓
LLP	Open	11,849	✓	✓	✓

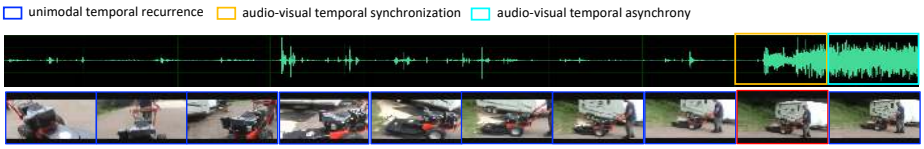

Fig. 1: Different multimodal temporal contexts for the visual event in the red box. In the video, *Lawn mower* makes sounds only at the last two seconds.

Since AVE dataset [15] is also an audio-visual video dataset, we would like to further clarify the differences between the AVE dataset and our LLP dataset. AVE dataset is a video dataset containing 4143 videos with audio-visual events, in which sound sources are visible and their sounds are audible. There are two strong assumptions inside the AVE : (1) each video contains at least one 2s long audio-visual event; and (2) there is only one audio-visual event inside a 10-second video. However, a video may only have audio or visual events without any audio-visual events, and multiple overlapping events can simultaneously occur in one video. So, numerous real-world unconstrained videos cannot hold the assumptions. Our LLP breaks the assumptions and can help us develop audio-visual video parsing algorithms for parsing videos into different audio, visual, and audio-visual temporal events towards a unified multisensory perception.

3 Different Multimodal Temporal Contexts

As discussed in our main paper, we know that audio or visual events in a video usually redundantly recur many times inside the video, both within the same modality (unimodal temporal recurrence [9,10]), as well as across different modalities (audio-visual temporal synchronization [8] and asynchrony [17]). Figure 1 illustrates different multimodal temporal contexts in a video. From the figure, we can find unimodal temporal recurrent *Lawn mower* event for the visual content in the red box (see blue boxes), audio-visual temporal synchronized event in the yellow box, and audio-visual temporal asynchronous event in the cyan boxes.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015) 1, 2
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 1
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1
4. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017) 1
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
6. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 ieee international conference on acoustics, speech and signal processing (icassp). pp. 131–135. IEEE (2017) 1
7. Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The thumos challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding **155**, 1–23 (2017) 1, 2
8. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Advances in Neural Information Processing Systems. pp. 7763–7774 (2018) 2
9. Naphade, M.R., Huang, T.S.: Discovering recurrent events in video using unsupervised methods. In: Proceedings. International Conference on Image Processing. vol. 2, pp. II–II. IEEE (2002) 2
10. Roma, G., Nogueira, W., Herrera, P., de Boronat, R.: Recurrence quantification analysis features for auditory scene classification. IEEE AASP challenge on detection and classification of acoustic scenes and events **2** (2013) 2
11. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1041–1044 (2014) 1, 2
12. Serizel, R., Turpault, N., Eghbal-Zadeh, H., Shah, A.P.: Large-scale weakly labeled semi-supervised sound event detection in domestic environments. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018). pp. 19–23 (November 2018), https://hal.inria.fr/hal-01850270 1, 2
13. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision. pp. 510–526. Springer (2016) 1, 2
14. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV. pp. 247–263 (2018) 1, 2
15. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in the wild. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops (2019) 2

16. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018) 1
17. Vroomen, J., Keetels, M., De Gelder, B., Bertelson, P.: Recalibration of temporal order perception by exposure to audio-visual asynchrony. Cognitive brain research **22**(1), 32–35 (2004) 2