Weakly Supervised Representation Learning for Unsynchronized Audio-Visual Events

Sanjeel Parekh[1,2], Slim Essid[1], Alexey Ozerov[2], Ngoc Q. K. Duong[2], Patrick Pérez[2], and Gaël Richard[1]
[1]LTCI, Télécom ParisTech, Université Paris–Saclay, France [2]Technicolor, France

1. Introduction

We are surrounded by events that can be perceived via distinct audio and visual cues. Be it a ringing phone or a car passing by, we instantly identify the audio-visual (AV) components that characterize these events. This remarkable ability helps us understand and interact with our environment. For building machines with such scene understanding capabilities, it is important to design algorithms for learning audio-visual representations from real-world data. This work is a step in that direction, where we aim to learn such representations through weak supervision i.e., supervision only in the form of video-level event labels without any timing information.

To motivate our tasks and method, consider a video labeled as “train horn”. Assuming that the train is both visible and audible at some time in the video, in addition to identifying the event, we are interested in learning representations that help us answer the following questions:

- *Where is the visual object or context that distinguishes the event?* In this case it might be the train (object) or platform (context) *etc*. We are thus aiming for their spatio-temporal localization in the image sequence.
- *When does the sound event occur?* Here it is the train horn. We thus want to temporally localize the audio event.

The variety of noisy situations that one may encounter in unconstrained videos adds to the difficulty of this very challenging problem. Apart from modality-specific noise such as visual clutter or low audio SNR, in real-world scenarios, the audio and visual elements characterizing the event are often unsynchronized in time. This is to say that the train horn in the previous example may sound before or after the train is visible. In the extreme, not so rare case, the train may not appear at all. We are interested in designing a system to tackle the aforementioned questions and situations.

Prior research has utilized AV modalities for classification and localization tasks in various contexts. Fusing modality-specific hand-crafted or deep features has been a popular approach for problems such as multimedia event detection [5]. Lately, several inspiring multimodal deep networks have been proposed [7, 1], focusing primarily on unsupervised representation learning and synchronous AV training. However, to our knowledge, a unified framework for simultaneous event classification and characteristic AV cue localization, especially in asynchronous situations, has not been extensively studied in literature.

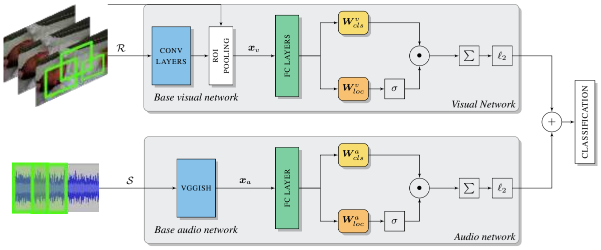

Figure 1. **Proposed approach**: Given a video, we consider the depicted pipeline for going from audio and visual proposals to localization and classification. Here $\boldsymbol{W}_{\text{cls}}$ and $\boldsymbol{W}_{\text{loc}}$ refer to the fully-connected classification and localization streams respectively; σ denotes softmax operation over proposals for each class, $\odot$ refers to element-wise multiplication; Σ to a summation over proposals and ℓ_2 to a normalization of scores.

2. Proposed Approach

Our tasks can be naturally interpreted as *multiple instance learning* (MIL) problems [3]. MIL is typically applied to cases where labels are available over bags (sets of instances) instead of individual instances. The task then amounts to jointly selecting appropriate instances and estimating classifier parameters.

In our case, a video, V can be decomposed as a bag of M selected image regions, $\mathcal{R} = \{r_1, r_2, \ldots, r_M\}$, obtained from sub-sampled frames and T audio segments, $\mathcal{S} = \{s_1, s_2, \ldots, s_T\}$. Each region/segment *proposal* is dealt with in separate visual & audio sub-modules. The key idea is to extract features from generated proposals and transform them for: (1) scoring each according to their relevance for class labels; (2) aggregating these scores in each modality & fusing them for video-level classification. This allows us to train both the sub-modules together through weak-supervision and learn representations for event classification and localization. Moreover, use of both the modalities makes the system robust against noisy scenarios.

An overview of our approach is provided in Fig. 1. For the visual network, we use class-agnostic bounding box proposals and for audio, short overlapping temporal segments. Following [2], we employ a parallel two-stream architec-

This is an extended abstract. The paper is available at https://arxiv.org/abs/1804.07345

	System	F1	Precision	Recall
(a)	Proposed AV Two Stream	**64.2**	59.7	**69.4**
(b)	TS Audio-Only	57.3	53.2	62.0
(c)	TS Video-Only	47.3	48.5	46.1
(d)	TS Video-Only WSDDN-Type [2]	48.8	47.6	50.1
(e)	AV One Stream	55.3	50.4	61.2
(f)	CVSSP - Fusion system [8]	55.6	**61.4**	50.8
(g)	CVSSP - Gated-CRNN-logMel [8]	54.2	58.9	50.2

Table 1. Results on DCASE smart cars task test set. We report here the micro-averaged F1 score, precision and recall values and compare with state-of-the-art. TS is an acronym for two-stream.

ture for the scoring network in each modality. As in earlier studies, we find that the two-stream architecture gives better results than a straightforward one-stream `log-sum-exp` operation implementation (soft approximation to `max`).

3. Experiments

We validate the system's performance, both quantitatively and qualitatively over DCASE challenge smart cars data [6], a large-scale multi-label weakly labeled dataset for audio events consisting of 17 classes, spread over approximately 50K YouTube videos from AudioSet [4].

Baselines. In Table 1, systems (b) and (c) only utilize the audio and visual sub-modules, respectively; (d) refers to the system proposed in [2]; (e) uses a single stream with `log-sum-exp` operator and (f)-(g) are systems from DCASE 2017 smart cars challenge event classification task winners [8], using only audio and no external data. All the systems are evaluated on the micro-averaged F1 score. This was the official metric used by the task organizers for ranking systems. For further insight, we also compute the class-wise F1 scores.

Table 1 shows event classification results for all the methods. Our system achieves state-of-the-art performance. We also find that the audio-visual complementarity is clearly reflected by the class-wise F1 scores (reported in the full paper). Briefly, the data can be categorized into two sets: (i) classes with well defined AV objects like car, motorcycle etc. and (ii) *warning sounds*, like, siren, car alarm. Well-defined visual cues enhance the performance of the proposed multimodal system over audio-only approaches. On the other hand, for warning sounds, frames alone are insufficient. For such cases, audio assists in improving system performance. Fig. 2 displays results for object localization in video frames and Fig. 3 illustrates an example of asynchronous AV cues. For more examples, please refer to additional material at https://youtu.be/C-jrZ9SDMDY

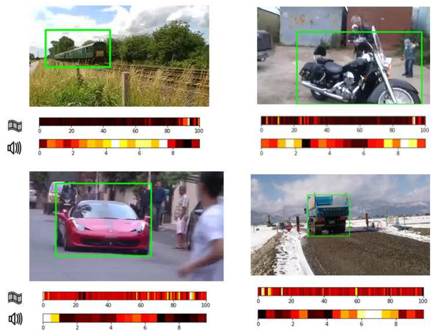

Figure 2. Examples of localization on video frames for a few categories from the test data. The localization results are shown in green. Below each image we display the scaled region proposal and audio segment scores for labels under consideration. The visual heatmap is a concatenation of proposals from all the subsampled frames, arranged in temporal order.

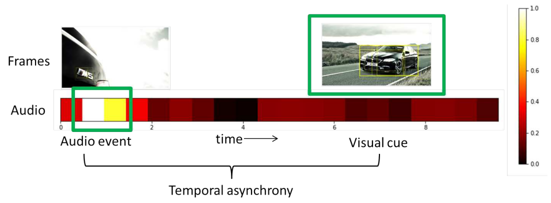

Figure 3. Qualitative result for unsynchronized AV events. The heatmap at the bottom denotes audio localization over segments. The top row depicts video frames roughly aligned to the audio temporal axis. This is a 'car' video from the validation split, where the visual object of interest appears after the audio event. The video frames show bounding boxes where edge opacity is controlled by the box's detection score. Higher score implies better visibility.

References

[1] R. Arandjelović and A. Zisserman. Look, listen and learn. In *Proc. ICCV*, 2017. 1

[2] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proc. CVPR*, 2016. 1, 2

[3] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 1

[4] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. ICASSP*, 2017. 2

[5] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2013. 1

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In *Proc. DCASE Workshop*, 2017. 2

[7] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *Proc. ECCV*, 2016. 1

[8] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley. Surrey-CVSSP system for DCASE 2017 Challenge Task 4. Technical report, DCASE Challenge, 2017. 2