Urban Sound & Sight: Dataset and benchmark for Audio-Visual Urban Scene Understanding

Magdalena Fuentes[1], Bea Steers[1], Pablo Zinemanas[2], Martín Rocamora[3], Luca Bondi[4], Julia Wilkins[1], Qianyi Shi[1], Yao Hou[1], Samarjit Das[4], Xavier Serra[2], Juan Pablo Bello[1]
[1] MARL, NYU [2] MTG, UPF [3] GPA, UdelaR [4] Bosch Research

Abstract

Automatic audio-visual urban traffic understanding is a growing area of research with many potential applications of value to industry, academia, and the public sector. Yet, the lack of well-curated resources for training and evaluating models to research in this area hinders their development. To address this we present a curated audio-visual dataset, Urban Sound & Sight (Urbansas), developed for investigating the detection and localization of sounding vehicles in the wild. Urbansas consists of 12 hours of unlabeled data along with 3 hours of manually annotated data, including bounding boxes with classes and unique id of vehicles, and strong audio labels featuring vehicle types and indicating off-screen sounds. We discuss the challenges presented by the dataset and how to use its annotations for the localization of vehicles in the wild through audio models.

1. Introduction

The automatic understanding of audio-visual urban scenes is a growing area of research, with many potential applications of value to industry, academia, and the public sector, with potential applications such as assistive devices for the hearing-impaired, the quantification of traffic for policy making, autonomous driving, among others. Audio-visual information is fundamental for the full understanding of real-world scenes, as visual and acoustic modals provide complementary information: images help identify sources and understand their motion, audio help understand the proximity of sources, the presence of relevant off-screen sounding objects, and help solve occlusions and improve estimations with poor lighting. Understanding an audio-visual urban scene includes estimating the class, spatial location, direction and speed of movement of beings and objects in real environments by the sounds they make and the way they look. Ideally, automatic solutions would be robust across a wide range of sound scenes and sensing conditions: noisy, sparse, with varying compositions of sources, with moving sources, with moving sensors.

While there is a large body of research in related computer vision (e.g. object detection and pedestrian counting [13, 17]), and machine listening areas (e.g. urban sound event detection and classification, [15, 11]), there is little work on audio-visual classification and localization of sounding sources in realistic urban settings. Recently the machine listening community has turned its attention to localization, seeking to apply the same deep learning techniques that have proven successful in classification before [12, 1, 8], mostly using synthetic datasets. Research on the co-occurrence of audio and video has recently received increasing attention due to the development of self-supervised models that exploit audiovisual cues for their pretext-task [2, 18, 3]. Most of this research is carried out using unlabeled videos from Youtube or Audioset [9], and models learn a representation of the data (either audio, visual or both) to later be applied to a downstream task [5]. Except for a few exceptions [7], these works have focused on audio-visual localisation mostly of sources such as musical instruments or in low-complexity settings, where objects are relatively close to the camera and central to the scene.

One of the main challenges to audio-visual urban research is the lack of labeled data. While most of the existing resources involve either audio [10] or video [4] alone, the available audio-visual datasets of urban scenes have limited annotations, restricted to audio events only [20] or clip labels intended for scene classification [19]. Moreover, since manually annotating real-world data is very arduous and time consuming, the amount of labeled data tends to be small for machine learning standards. A way to alleviate the work of manual annotation is to create synthetic audio mixtures using isolated sound events [16] or synthetic visual scenes from video games [14], but they fail to capture the diversity and complexity of naturally occurring sound scenes. Another challenge is how to annotate moving sources in such complex settings: dealing with off-screen sounds, occlusions, or objects that can be seen but not heard.

2. The Urban Sound & Sight dataset

We set four main goals for creating this dataset: 1) to compile a set of real-field audio-visual recordings; 2) the recordings should be stereo to allow exploring sound localization in the wild; 3) the compilation should be varied in terms of scenes and recording conditions to be meaningful for training and evaluation of machine learning models; 4) the labeled collection should be accompanied by a bigger unlabeled collection with similar characteristics to allow exploring self-supervised learning in urban contexts in the future. In the following we explain how we have compiled Urbansas to fulfill these goals.

Data Sources. We have compiled and manually annotated Urbansas from two publicly available datasets, plus

the addition of unreleased material. The public datasets are the TAU Urban Audio-Visual Scenes 2021 Development dataset [19] and the Montevideo Audio-Visual Dataset (MAVD) [20]. The TAU dataset consists of 10-second segments of audio and video from different scenes across European cities, traffic being one of the scenes. Only the subset of scenes labeled as traffic were included in Urbansas. MAVD is an audio-visual traffic dataset curated in different locations of Montevideo, Uruguay, with annotations of vehicles and vehicle components sounds (e.g. engine, brakes) for sound event detection. Besides the published datasets, we include a total of 9.5 hours of unpublished material recorded in Montevideo, with the same recording devices of MAVD but including new locations and scenes.

subset	places	clips	mins	frames	labeled mins
Montevideo	8	3978	663	955k	90
TAU	42	1387	231	333k	90

Table 1. Breakdown of Urbansas per city and location. Last column indicates the portion data in the labeled set.

3. Annotating audio-visual urban scenes

In order to understand an audio-visual urban scene, we want to estimate the class and location of each source as it moves over time. To that goal, we have annotated: 1) bounding boxes of objects with a class assignment and object id; 2) "strong" audio labels, with beginning and end timestamps and the correspondent class of the acoustic event; 3) relevant metadata about lighting and weather conditions (e.g. night vs. day). This dataset focuses on traffic since vehicles are a compelling case-study of sounding moving objects in urban settings. Consequently, our ontology focuses on the four most predominant vehicle types: *car*, *truck*, *bus*, and *motorbike*. In the following, we discuss the decisions we have made to annotate Urbansas. We used CVAT [1] for the bounding box annotations, and VIA [6] for annotating the audio with the video as reference. The video annotations were performed at 2fps to reduce redundant annotations, improve annotation quality, and allow for a larger volume of annotated clips.

Notation. For a specific file in the dataset, let us define an audio annotation as a tuple $(t_{s,i}, t_{e,i}, l_i),\ i \in [1, N_A]$, where $t_{s,i}$ and $t_{e,i}$ are the start and end time of an audio event with label l_i, and N_A is the total number of audio annotations for the file. We also define a video annotation as a tuple $(t_j, l_j, tr_j, v_j, x_j, w_j, y_j, h_j),\ i \in [1, N_V]$, where t_j is the timestamp of an object with label l_j and visibility flag v_j; track id tr_j is used to identify a single object across frames in a file; the bounding box for the object is defined in terms of horizontal (x_j) and vertical (y_j) shift between the top-left corner of the frame and bounding box, with corresponding height (h_j) and width (w_j); N_V is the total number of video annotations for the file.

Video annotations of sounding vehicles. Vehicles in the video are annotated if they are believed to contribute to the acoustic scene. Primarily, this includes vehicles that either drive past or idle near the observer, while excluding vehicles with their engines off (i.e. parked). In complex scenes, there are often multiple roads at different distances. In these scenarios, acoustic masking is taken into account - e.g., if vehicles from closer road mask sounds from the further road, then only the closer vehicles are annotated. If the closer road is less busy, then the further road may be annotated as well. If a vehicle is temporarily occluded (hidden behind something, partially or fully) it is still annotated with an estimate of its true location, with an additional flag ($v_j = 0$) identifying it as occluded.

Integrating audio and video annotations. The audio annotations can be used in combination with the video annotations to identify vehicles that are both audible and visible. In some cases, an object could have no audio events (and vice versa) if the sound occurs before or after the vehicle enters/leaves the scene (this can happen for certain camera angles). In other cases, an audio event may have no corresponding object in the video, which may happen when a vehicle passes outside of the camera's view; these are labeled as off-screen sounds. Since we have the audio annotations to disambiguate when the object is both present in the image and producing sound at the same time, we annotate vehicles when they are "close enough" to understand error types in visual-only or audio-visual models.

Scene annotations. Some scenes have many vehicles passing at the same time and it is perceptually very hard to attribute sounds to a particular vehicle, they rather produce a "constant background sound" altogether. To address this, we include a binary flag at the clip level indicating the presence of *non_identifiable_vehicle_sound*. In cases where particular vehicles are identifiable on top of this constant sound, we annotate them with strong labels as well as indicate the presence of non identifiable vehicle sounds. Additionally, we include flags indicative of the lighting: night vs. day.

4. Localizing sources in the wild

Indexing of video annotations for audio localization. We approximate the vehicles position using linearly spaced regions corresponding to the angles within the camera's field of view (FoV). For each video annotation, we approximate the position (θ_j) of the object based on the coordinates of the bounding box, and then we quantize θ_j to the closest region. We explore two ways of computing θ_j:

[1] https://github.com/openvinotoolkit/cvat

1) We consider the vehicles as point sources. For this we used the center point of the bounding box as the position indicator. Formally:

$$\theta_j(x_j) = \left(\frac{x_j + \frac{w_j}{2}}{W} - \frac{1}{2} \right) f_{ov}, \quad (1)$$

where W is the width of the frame and f_{ov} is the FoV of the camera. Working with θ_j approximated this way allow us to combine data with different FoVs and resolutions (W) in the future. 2) We relax the point-wise estimation and instead approximate the vehicle location to be the left ($\theta_{j,L}$) and right ($\theta_{j,R}$) bounds of the bounding box.

$$\theta_{j,L} = \theta_j \left(x_j - \frac{w_j}{2} \right), \text{ and } \theta_{j,R} = \theta_j \left(x_j + \frac{w_j}{2} \right). \quad (2)$$

Finally, we map θ_j to a specific region r_j as $r_j = \text{argmin}_i |\theta_j - r_i|$ for $i \in \{1, .., R_N\}$, where r_i denotes the region i and R_N is the total number of regions the FoV (120° from camera specification) is divided, which we set to 5.

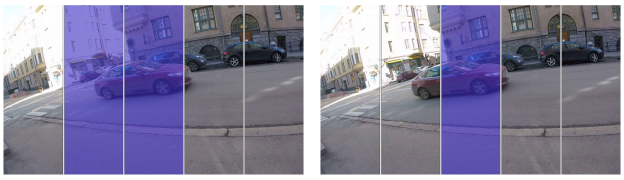

Figure 1. Left: Regions activated using the full bounding box, Right: regions activated using the bounding box center. Each region conveys 24°. Note: black vehicle is parked so it is not annotated.

Audio annotations as filter to video objects. At training time, we only consider video events that are confirmed by audio annotations. This condition is met if the timestamps for a given video object overlap with the start and end time of an audio annotation with the same label. Note that this is not always the case since there are scenes where vehicles are visible but not audible and the other way around. Formally, given a video object $\mathcal{V}^k$ characterized by label l^k and a set of timestamps $\left\{t_p^k \; p \in [1, N_{\mathcal{V}_k}]\right\}$, we consider the video object as valid for training if it exists at least one audio annotation $(t_{s,i}, t_{e,i}, l_i)$, $i \in [1, N_A]$ such that $l_i = l^k$ and $\sum_{p=1}^{N_{\mathcal{V}^k}} \left[(t_p^k \geq t_{s,i}) \wedge (t_p^k \leq t_{e,i}) \right] >= 0$, i.e. audio and video overlap and their labels coincide.

5. Data challenges and baseline

To learn about the challenges of the data and usefulness of the metric, we ran a set of experiments with simple baselines. We are not searching for an optimal model that maximizes accuracy but rather we are interested in understanding the characteristics of the dataset and metric themselves, and identifying venues for future research.

Baselines. We adapt the architecture of [1] to use stereo audio, and to be multi-class and multi-direction model, i.e. to predict overlapping sources of the same class and with different positions. To do so, our model predicts a tensor $T(i, c, j) = (t_i, c, r_k)$ for each time t_i, $i \in [1, N_f]$ with N_f the number of frames, vehicle class $c \in \{C_1, .., C_4\}$, and region $r_j \in [R_1, R_5]$. We use a sigmoid layer to allow for multiple activations at once. We train and evaluate the box-wise model using the regions covering the entire bounding box, and the point-wise using the regions activated at the center of the bounding box (see Figure 1). We also include two random baselines: a point-wise baseline that can predict up to two active regions at a time, and a box-wise baseline that estimates up to five regions. Each one is compared to the matching ground-truth (point- and box-wise). We split the labeled set into 5 folds stratified by location and we perform cross-fold (4-1) training and validation. We train using 4 second chunks as in [1]. We used a weighted binary cross-entropy loss for training.

Results. Results are depicted in Table 2. We compute the IoU score for non-empty frames (i.e. frames containing at least one bounding box that overlaps with the audio). The first observation is that both models perform better than random, the box-wise model being the best. This is to expect since the bounding box conveys more regions than the point-wise case and thus is an easier problem. We see a considerable drop in performance for the least frequent class (truck) whose sound resembles to cars and buses.

model	*IoU* ($\tau = 0.05$)			
	bus	*car*	*motorbike*	*truck*
point-wise (pw)	0.415	0.359	0.322	0.361
box-wise (bw)	0.567	0.492	0.356	0.477
pw-random	0.045	0.045	0.048	0.037
bw-random	0.102	0.100	0.089	0.115

Table 2. IoU per-class on non-empty frames.

We also compute the IoU for all frames, including inactive frames, to assess whether the baseline can determine the presence (and absence) of vehicles in a clip. For those empty frames, we compare the prediction mask of the model with an empty ground truth, obtaining a score of 1 if the model did not predict the class at any direction. We obtained better scores overall in this setting: cars ($IoU = 0.372$), buses ($IoU = 0.734$), motorbikes ($IoU = 0.758$) and trucks ($IoU = 0.877$) for the box-wise model. A counter-intuitive result is that the highest scores correspond to the least represented classes in the dataset, potentially due to the low frequency of such vehicles in the scenes and the fact that the baseline models have low confidence values in general, favoring empty predictions and scoring high in empty frames. This indicates that the joint

detection and localization of vehicles is a highly imbalanced and hard learning problem. Regarding the usefulness of the IoU metric for localization of sources in the wild, we believe that the formulation of the problem as detection and localization makes it hard to judge with this metric how good the models are at localizing and detecting respectively, and we plan to explore them separately in the future.

6. Conclusions and future work

We present Urbansas, an audio-visual dataset of traffic scenes, containing 12 hours of unlabeled data, suitable for unsupervised and self-supervised research in visual sound source detection and localization, and 3 hours of human-annotated data, containing bounding boxes, classes, and tracking information to be used for supervised research and validation of self-supervised models as a downstream task. To the best of our knowledge, Urbansas is the first audio-visual urban traffic dataset with human-annotated labels both in audio and video. We believe the dataset will open the path to new research on audio and audio-visual sound source localization, vehicle tracking, self-supervised audio-visual representation for real world applications, among others. We present first experiments on vehicle localization and detection, including a baseline. The data and code are open to the research community.[2]

References

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018. 1, 3

[2] R. Arandjelovic and A. Zisserman. Objects that sound. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[3] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 1

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[5] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE, 2019. 1

[6] A. Dutta and A. Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. 2

[7] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7062, 2019. 1

[8] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud. Tracking the active speaker based on a joint audio-visual observation model. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 702–708, 2015. 1

[9] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[10] A. Mesaros, T. Heittola, and T. Virtanen. TUT database for acoustic scene classification and sound event detection. In *In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016), Budapest, Hungary*, 2016. 1

[11] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley. Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–83, 2021. 1

[12] P. Pertilä, E. Cakir, A. Hakala, E. Fagerlund, T. Virtanen, A. Politis, and A. Eronen. Mobile microphone array speech detection and localization in diverse everyday environments. *arXiv preprint arXiv:2106.14787*, 2021. 1

[13] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[14] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 1

[15] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283, 2017. 1

[16] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello. Scaper: A library for soundscape synthesis and augmentation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, october 2017. 1

[17] V. A. Sindagi and V. M. Patel. A survey of recent advances in CNN-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, 2018. 1

[18] E. Tzinis, S. Wisdom, A. Jansen, S. Hershey, T. Remez, D. P. Ellis, and J. R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 1

[19] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen. A curated dataset of urban scenes for audio-visual scene analysis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2021. 1, 2

[20] P. Zinemanas, P. Cancela, and M. Rocamora. MAVD: A dataset for sound event detection in urban environments. *Detection and Classification of Acoustic Scenes and Events, DCASE 2019, New York, NY, USA, 25–26 oct, page 263–267*, 2019. 1, 2

[2] https://github.com/magdalenafuentes/urbansas