SOUND EVENT CLASSIFICATION WITH OBJECT-BASED LABELS

James Afolaranmi, Irene Martín-Morató, Annamaria Mesaros

Computing Sciences, Tampere University, Tampere, FINLAND
james.afolaranmi@tuni.fi, irene.martinmorato@tuni.fi, annamaria.mesaros@tuni.fi

ABSTRACT

Availability of audio-visual datasets and increase of computational resources have made possible the use of deep learning techniques that exploit the relationship between audio and video. In this paper, we present an approach that makes use of pretrained models for object detection to label audio clips based on objects that are expected to make sound. The study consists of performing object detection for four target classes belonging to vehicle category and training sound classifiers in supervised way using the resulting labels. We conclude that object detection is a useful alternative for labeling audio-visual material for audio classification, with substantial improvements in different datasets. Results show that even for data provided with reference audio labels, labeling through video object detection can identify additional, non-annotated acoustic events, thus improving the quality of the labels in existing datasets. This promotes exploitation of video content not only as an alternative, but also to complement the available label information.

Index Terms— sound event classification, deep neural networks, object-based labels.

1. INTRODUCTION

Audio classification tasks have increased in popularity in recent years, due to applicability of methods for acoustic monitoring [1], environment monitoring [2], or emotion recognition [3] along with others. The analysis of acoustic scenes aims at recognizing different types of information in the environment, for example vehicles in urban scenes. The diversity of acoustic information in everyday environments increases the complexity of such task.

Deep learning methods allow obtaining high performance on classification tasks. Among the challenges that the traditional supervised learning scenario must overcome, one is data availability for training robust models. Supervised methods rely on the labeled data to train models effectively. These methods require accurate labelling, meaning that when incorrect labels are present in the training data, the learning process is compromised, leading to suboptimal performance of the model and reducing its generalization capabilities [4]. Proper data curation, label verification, and quality control mechanisms are essential to mitigate the impact of incorrect labels and ensure the training of robust models.

The release of AudioSet [5] has been a milestone for the audio datasets. It contains 527 sound classes from over 5000 hours of audio recordings collected from YouTube videos, provided as annotations for 10-second long clips. Following AudioSet, more datasets providing video and audio modalities have been published, e.g. MAVD-traffic [6] and TAU Urban Audio-Visual Scenes 2021 [7] datasets. While these are the main audio-visual datasets used in audio research, there is a much larger number of such datasets that are used in image/video research. Audio-visual datasets provide a rich source of information that combines auditory and visual modalities, offering valuable insights into the correlation and complementarity between audio and visual cues.

The task of annotating sound events within such datasets is time-consuming and expensive. As a result, the majority of audio-visual datasets are primarily annotated for visual content; some datasets have information on acoustic scene, while the annotation of sound events remains limited to a smaller subset. For example in EPIC-SOUNDS [8], the authors collected a large scale dataset of audio annotations as an extension of the original EPIC-KITCHEN dataset [9], which is originally aimed at computer vision research.

In this work we propose to investigate if labels derived through object detection methods based on the video modality are suitable for audio classification. Generally, the information in the audio and visual modalities is highly correlated, and sound-producing objects may be visible in the video, even though this is not guaranteed, for example in poor light conditions or in the presence of obstructions. We investigate how well YOLO (You Only Look Once) object detector [10] can be used to provide labels for audio content to ultimately train an audio classification model. Experiments performed on three different datasets show that even though the labels inferred based on objects are not fully corresponding to the audio ground truth, they provide a sufficient supervision signal for training a sound event classification system.

The rest of the paper is organized as follows: Section 2 introduces the approach used for obtaining the object-based labels and how they are used for audio classification purposes; Section 3 presents the datasets used in the experiments and introduces the classification system; it also includes an analysis of the results and discusses the comparison of the object-based labels with the reference audio labels; finally, Section 4 presents the conclusions and future work.

2. OBJECT-BASED AUDIO CLASSIFICATION

Figure 1 illustrates the workflow followed in this study. To obtain labels for the audio content, object detection using YOLO [10] is performed on video frames from the video clip. The pretrained model OpenL3 [11] is used to perform feature extraction and to obtain the embeddings for the corresponding aucio clip. The labels and the embeddings are used as input to the audio classification model. The target labels are the labels obtained from the object-detection model, and the input data are the embeddings from the pretrained OpenL3 model. The acoustic model is then trained using this information for classifying the selected target sounds.

This work was supported by Academy of Finland grant 332063 “Teaching machines to listen”. The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources.

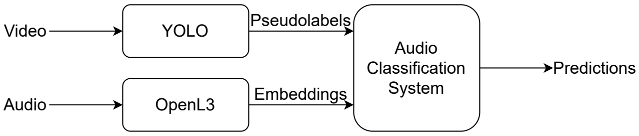

Figure 1: Proposed approach; Object detection is applied to frames of the video, and the resulting output labels are used in training the audio classification system.

2.1. Object detection framework

Object detection is a popular research task in computer vision. It involves localization of target objects into bounding boxes and classification of those objects. Object detectors can be classified into two categories: single-stage or two-stage object detectors, depending on the method used to locate and classify objects. YOLO falls under the category of single-stage detector that carries out object localization and classification in the same run [12].

YOLO architecture is based on multiple CNN layers followed by fully connected layers. It predicts bounding boxes and class probabilities simultaneously, making it efficient for real-time detection. In [10], the authors showed that YOLO was able to score 57.9% mean average precision on the PASCAL VOC 2012 test set on 20 labelled classes, and generalized better than other detectors when tested for person detection in two artwork datasets. In our study, we used a pretrained version of YOLOV5[13] capable of recognizing 80 classes to perform detection of four vehicle-related classes within the video data.

2.2. Audio classification framework

Because the scope of this work is to investigate feasibility of labeling audio through video, for the audio classification model employed in our study uses an existing architectures rather than designing and optimizing one for the task. We use embeddings from the pretrained L3-Net [14] implemented in OpenL3 [11] as a backbone, and three dense linear layers of 512, 128 and 4 neurons stacked upon each other; the network uses ReLU as an activation function for the first two dense layers and sigmoid activation function for the output layer to perform multi-label classification.

2.3. Datasets and baseline system

We use three different audio-visual datasets, namely: AudioSet [5], the MAVD dataset in Urban environments [15] and a subset of TAU Urban Audio-Visual Scenes 2021 Development Dataset [7]. In this work we use four target sound classes: *Bus*, *Car*, *Motorcycle* and *Truck*, which can all be found in these three datasets.

From AudioSet, a subset of 121.8 hours of data was selected based on the target classes. A 70/30 ratio is used to partition this subset into training and test set. The labels provided in AudioSet are used as ground truth in our comparative experiments. As documented in [16], some clips in AudioSet may have incorrect or missing labels. This is due to the annotation process which included a verification step for the candidate labels [5]; in this process the labels were manually verified, but no new labels were added. In general, AudioSet has a highly imbalanced class distribution which is prominent also in the subset used in our experiments, with the majority of the data examples belonging to the *Car* class.

Dataset	Training (hh:mm)	Test (hh:mm)
AudioSet	85:15	36:32
MAVD	01:03	00:27
TAU UrbanASC	02:00	01:30

Table 1: Amount of data available per dataset.

As a second source of annotated audio-visual data we use the Urbansas dataset [17], which consists of 3 hours of manually annotated data, compiled from two different datasets: MAVD [6] and TAU Urban Audio-Visual Scenes 2021 dataset (TAU UrbanASC) [7]. MAVD is an audio-visual dataset created to monitor urban noise in Montevideo, Uruguay, and consists of 1.5 hours of manually annotated data divided into train and test set. The TAU Urban Audio-Visual Scenes 2021 dataset (TAU UrbanASC) [7] consists of synchronized audio and video segments with a length of 10 seconds recorded in 12 different European cities. Of these, 1.5 h of the *street traffic* clips was annotated within the Urbansas dataset. We treat MAVD and TAU UrbanASC separately in our experiments. The total amount of data available in the training and test subsets used in our classification experiments is presented in Table 1.

We perform the classification experiments using the object-based labels and, for comparison, the reference audio labels, when available. Since audio reference labels are only available for two of the three datasets, AudioSet and MAVD, the comparative experiment is performed only for these two datasets.

3. EXPERIMENTAL RESULTS

We performed object detection on five image frames of the video clip (one frame every two seconds) using the pretrained YOLO model. To extract frames from the video we used the OpenCV library [1] in Python. For each of these five frames, YOLO returns labels corresponding to the four target classes, and coordinates for the bounding box of each object. The predicted labels include multiple instances of different classes for each 10 s video clip. To avoid losing any information about the detected objects, we create the set of labels inferred based on the video as the union set of the predicted object labels. The audio clip is then assigned the resulting set of labels for training a model as a multilabel classifier.

3.1. Comparison of inferred labels with audio reference labels

First of all, we verify to what extent the object-based inferred labels match the audio reference labels. To this end, we compare the obtained labels with the reference labels for all the data in each dataset (including training and test set, when available). The results are presented in Table 2. The object-based labels are most similar with the reference labels for the *Car* class in all three datasets, having a significantly higher F-score than any other class. We also observe that the *Truck* class has very low precision values for all three datasets. The discrepancy between the object-based and the audio reference labels is quite large for many cases. For example in the case of AudioSet the *Bus* class the precision is 0.32, meaning that only one third of clips labeled as *Bus* by YOLO are also annotated based on audio as containing the sound.

[1] https://github.com/opencv/opencv

Class	AudioSet			MAVD			TAU UrbanASC		
	P	R	F	P	R	F	P	R	F
Bus	0.32	0.73	0.45	0.86	0.55	0.67	0.43	0.90	0.59
Car	0.72	0.89	0.79	0.67	0.97	0.79	0.65	0.99	0.78
Motorcycle	0.46	0.90	0.61	0.50	0.35	0.41	0.71	0.42	0.53
Truck	0.36	0.87	0.51	0.12	0.80	0.21	0.13	0.84	0.23
Average	0.47	0.85	0.59	0.54	0.67	0.52	0.48	0.79	0.53

Table 2: Comparison of the object-based labels and reference audio labels for the three datasets.

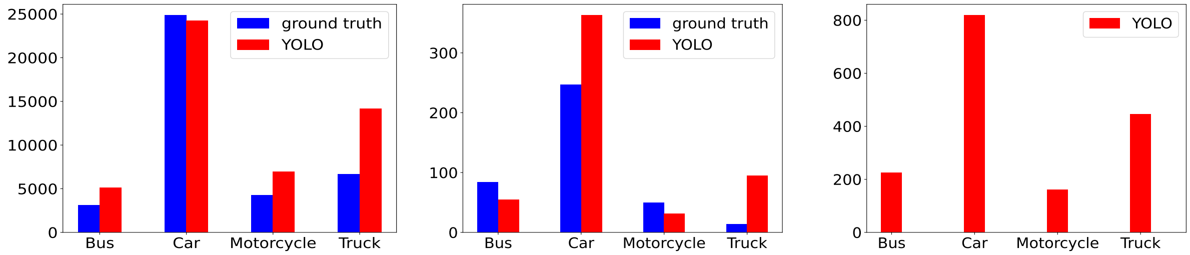

Figure 2: Training instances available for AudioSet, MAVD and TAU UrbanASC using the reference audio labels (blue) and object-based labels (red). For TAU UrbanASC, only the testing set is annotated, therefore we have no reference for comparison.

Class	AudioSet						MAVD						TAU UrbanASC		
	Reference labels			object-based			Reference labels			object-based			object-based		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Bus	0.00	0.00	0.00	0.18	0.08	0.11	0.73	0.46	**0.57**	0.82	0.22	0.35	0.23	0.27	0.25
Car	**0.73**	0.79	0.76	0.67	**0.96**	**0.79**	0.81	0.87	**0.84**	0.67	1.00	0.80	0.63	0.99	0.77
Motorcycle	0.51	0.32	0.39	0.54	0.43	**0.48**	0.50	0.32	**0.39**	0.67	0.09	0.16	0.00	0.00	0.00
Truck	0.00	0.00	0.00	0.28	0.68	**0.40**	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.34	0.11
Average	0.31	0.28	0.29	0.42	0.54	0.45	0.51	0.41	0.45	0.54	0.33	0.33	0.23	0.40	0.28

Table 3: Classification results for the three datasets, for the classifier trained with the reference labels and with the object-based labels.

Figure 2 illustrates the number of instances per class available for training the audio classification system for each case. It can be clearly seen that the *Car* class is the one with highest number of example instances for both label sets (reference and object-based labels) and all three datasets, while *Truck* has considerably less instances in the reference labels compared to the object-based labels.

3.2. Audio classification with object-based labels

We train a classifier using YOLO object-based labels. For comparison, we also train the same classifier structure using the audio reference labels. These models are then tested on the same test set and their performances are compared in terms of precision, recall, and F-score. The results are presented in Table 3.

For AudioSet, we observe that the classifier trained with object-based labels obtains a significantly higher recall for all the classes, and a higher F-score, despite the classifier output being evaluated against the reference annotations of the dataset itself. The system trained with object-based labels is able to recognize a higher number of event instances in the test data than if trained with the official reference labels provided in the dataset. However, this does not happen for the MAVD dataset. We hypothesize that the annotation process of MAVD was more efficient, and the quality of its reference labels is high. As seen from the results in Table 2, YOLO produces a lot of false positives, which in the case of MAVD are detrimental to the training process and consequently to the classification performance. Only for the *Car* class the classification performance is similar between the two training scenarios, but while recall for the model trained with the object-based labels reaches 100%, its precision suffers due to false positives. For the TAU UrbanASC we do not have audio reference labels, therefore we can only analyze the training with the object-based labels. The results in Table 3 show a high recall value for the *Car* class, which seems to be the dominant class among all datasets and label sets. At the same time. the system does not classify correctly any *Motorcycle* instances, which is the least represented class in the training data. Overall, the results on TAU UrbanASC are similar to those on MAVD.

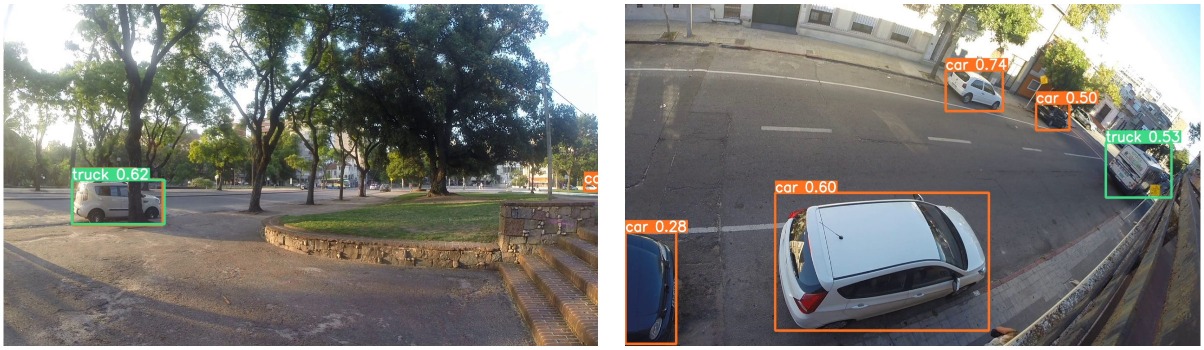

Figure 3: Example of mislabelled vehicles by YOLO.

3.3. Discussion

The *Truck* class is a very difficult case for all datasets, even though it is somewhat detected in AudioSet and TAU UrbanASC by the system trained with the object-based labels. In particular, the performance in AudioSet is very high, considering that the system trained with reference labels does not find any instance of this class. To understand this significant improvement in performance for AudioSet, we checked the clips for which YOLO indicated label *Truck* but the audio reference label did not contain it. We listened to 50 randomly selected clips and observed that 14% of them indeed contain truck sounds. In these cases, YOLO indicated a correct sound label based on the image, which were missing labels in the audio reference. There were also many false positives which add noise to the training process; nevertheless, the overall effect on the system performance was positive.

We visually inspected also the MAVD dataset *Truck* class, to understand the difference between the datasets. Looking at the predictions from the object detector, we observed that different types of vehicles (cars and buses) were mislabeled as trucks, which creates confusion between the categories. Two such examples are shown in Figure 3. In addition, in MAVD there were many scenes with parked vehicles which were visible and detected by the object detector, but did not produce any sound, therefore creating misleading information for the audio classifier during training.

This investigation revealed a very obvious drawback of using this method - objects in the image that do not produce sound (in this case parked vehicles) appear as false positives for the audio modality, and may be detrimental to performance. However, even with all these drawbacks and possible failure scenarios, the approach was shown to produce reasonable labels and in some cases lead to performance improvements. While this does not solve the problem of labeling audio content in audio-visual datasets, it can serve as a tool in more advanced training approaches; for example the object-based labels can be used as suggestions for methods that use active learning, or with a human-in-the-loop for verification; or can be treated as labels with some level of uncertainty to complement data that has been manually labeled by human annotators.

4. CONCLUSIONS

This work presented a novel approach of labelling audio data utilizing video information, to investigate the suitability of the method for creating reference labels for audio. The obtained labels were used afterwards in audio classification task. The method is based on an object detector model that takes as input a few frames of video corresponding to the audio clip, and predicting the target classes. Experiments performed on three different datasets showed the feasibility of using the audio-visual connection in the data to label audio content. However, the approach is unsuitable for situations when the sound sources are obscured/absent in the video frames, as they are not found by the object detector. In addition, some target sounding objects in the scene may actually not produce a sound in specific instances, leading to false positive labels. Despite these drawbacks, the method proves to be faster and lower-cost compared to the traditional annotation methods. Results from the experiment show that the method may outperform models trained with the provided reference audio labels, if they contain noisy or possibly incorrect information. We conclude that object-based labeling provides a suitable supervision signal for training and may be a useful tool in learning about audio content if handled as complementary information or to reinforce existing information about the data. Future work will focus on exploring more datasets for including a larger number of classes, and approaches for alleviating the effect of errors introduced by the object-based detector.

5. REFERENCES

[1] M. Ohlenbusch, A. Ahrens, C. Rollwage, and J. Bitzer, “Robust drone detection for acoustic monitoring applications,” in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 6–10.

[2] E. B. Çoban, A. R. Syed, D. Pir, and M. I. Mandel, “Towards large scale ecoacoustic monitoring with small amounts of labeled data,” in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 181–185.

[3] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, “Speech emotion classification using attention-based lstm,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, 2019.

[4] C. Northcutt, A. Athalye, and J. Mueller, “Pervasive label errors in test sets destabilize machine learning benchmarks,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, December 2021.

[5] J. F. Gemmeke, D. P. W. 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. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[6] P. Zinemanas, P. Cancela, and M. Rocamora, "MAVD-traffic dataset," July 2019.

[7] 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)*, 2021, pp. 626–630.

[8] J. Huh, J. Chalk, E. Kazakos, D. Damen, and A. Zisserman, "EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[9] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[11] A. L. 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)*, 2019, pp. 3852–3856.

[12] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[13] G. Jocher, "Yolov5 by ultralytics," 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[14] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 609–617.

[15] P. Zinemanas, P. Cancela, and M. Rocamora, "Mavd: A dataset for sound event detection in urban environments," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE*, 2019.

[16] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 21–25.

[17] M. Fuentes, B. Steers, P. Zinemanas, M. Rocamora, L. Bondi, J. Wilkins, Q. Shi, Y. Hou, S. Das, X. Serra, and J. P. Bello, "Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 141–145.