Self-supervised Learning of Audio Representations from Audio-Visual Data using Spatial Alignment

Shanshan Wang, Archontis Politis, Annamaria Mesaros, and Tuomas Virtanen

Tampere University, Finland,
firstname.lastname@tuni.fi

Abstract. In this work, we present a method for self-supervised representation learning based on audio-visual spatial alignment (AVSA), a more sophisticated alignment task than the audio-visual correspondence (AVC). In addition to the correspondence, AVSA also learns from the spatial location of acoustic and visual content. Based on 360° video and Ambisonics audio, we propose selection of visual objects using object detection, and beamforming of the audio signal towards the detected objects, attempting to learn the spatial alignment between objects and the sound they produce. We also investigate the use of spatial audio features to represent the audio input. Experimental results show a 10% improvement on AVSA for the first order ambisonics intensity vector (FOA-IV) in comparison with log-mel spectrogram features; the addition of object-oriented crops also brings significant performance increases for the human action recognition downstream task. A number of audio-only downstream tasks obtain performance comparable to state-of-the-art methods on acoustic scene classification from ambisonic and binaural audio.

Keywords: Self-supervised learning, Feature learning, Audio-Visual Correspondence, Audio-Visual Spatial Alignment, Audio classification

1 Introduction

Self-supervised learning is a common approach for learning representations based on unlabeled data, using proxy learning tasks. A popular strategy is the use of audio-visual correspondence as proxy task [1,2,5,8], shown to be useful in various downstream audio-visual or audio-only classification or recognition tasks [1, 15].

The spatial information in multichannel recordings provides more information about the recorded scene. Yang et.al. [16] showed that using a proxy task that learns whether the left and right channels in a video are in the correct order or flipped produces representations that outperform non-spatial versions in further audiovisual tasks. Morgado et al. [7] extended this idea to 360° videos with 4-channel ambisonic audio, and trained a network to distinguish whether crops in a 360° video frame are spatially aligned with the corresponding ambisonic audio using contrastive learning. They employ both AVC and AVSA as proxy task.

The method was shown to learn useful embeddings for a number of downstream video classification task.

In this work, we propose a system that learns from 360° audio and video data through spatial alignment. We build on the method proposed by Morgado et al. [7] by proposing a number of specific audio processing steps which focus the method towards downstream audio tasks. Specifically: (1) We combine audio beamforming with visual object detection to create a strong spatial correspondence between the audio and video modalities. (2) We use spatial audio features, to provide an explicit representation of the spatial content. We test our approach on both audio and video downstream tasks, including the in-domain AVC and AVSA, human action recognition, and acoustic scene classification with ambisonic audio using the Eigenscape dataset [4].

2 Approach

A simplified block diagram of the proposed method is illustrated in Fig. 1. The AVC learning process aims to learn feature representations based on audio-visual correspondence while AVSA aims to learn feature representations by using multiple crops and rotated audio signals of the same clip, and their spatial correspondence. The AVC and AVSA networks follow the same structure as [7].

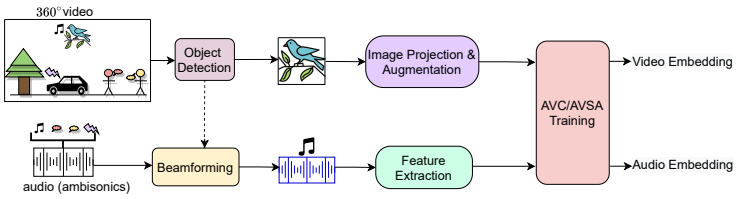

Fig. 1. Simplified block diagram of the proposed system. Object-oriented crops are selected from the video using an object detection method, and the ambisonic audio is rotated towards the center of the crop to form positive pairs.

Beamforming with ambisonic signals is typically done using a weighted version of the encoding basis as beamforming weights [12]. In the simplest case, the beamformed signal $y(n)$ for a beamforming direction (θ_0, ϕ_0) is $y(n) = \mathbf{u}^{\mathrm{T}}(\theta_0, \phi_0)\mathbf{x}(n)$ for the ambisonic signals $\mathbf{x}(n)$ corresponding to a source signal $s(n)$:

$$\mathbf{x}(n) = \begin{bmatrix} w(n) \\ y(n) \\ z(n) \\ x(n) \end{bmatrix} = \mathbf{u}(\theta, \phi)s(n). \tag{1}$$

Another common operation in Ambisonics is the capability to rotate the sound field. In the case of FOA only, and contrary to higher-order Ambisonics, the

rotation can be simply performed with a standard rotation matrix [10]. Following e.g., the yaw-pitch-roll convention corresponding to angles (α, β, γ) such a rotation $\mathbf{Q}$ is applied to the ambisonic signals as $\mathbf{x}_{\text{rot}}(n) = \mathbf{Q}(\alpha, \beta, \gamma)\mathbf{x}(n)$.

Since in [7] the crops are taken randomly from a video frame, beamforming to the crops may be focusing on regions where there are no sound sources present. To minimize such audiovisual mismatches, we implement the beamforming combined with object-oriented crops obtained from YOLO detector [13]. We apply YOLO on the equirectangular frames (360° image) of the video, to obtain objects with their bounding boxes and center points. The YOLO-based selection process is illustrated in Fig.2.

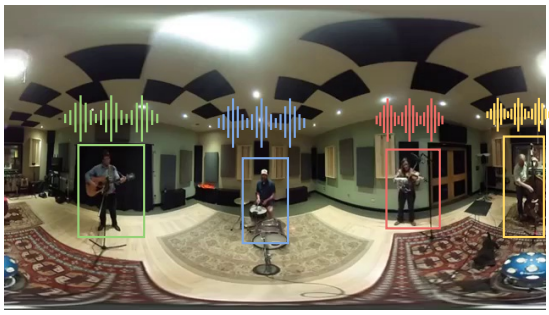

Fig. 2. The AVSA learning procedure uses four crops of the same clip and corresponding audio, presented as green, blue, red and yellow pairs.

For representing ambisonic signals, a feature more suitable than the usual energy-based representations is the active intensity vector (AIV), an acoustical quantity indicating the mean flow of sound energy. In this work we are using a normalized version of the AIV as in [9], which bounds the magnitude of the vector between [0,1], similar to the *diffuseness* feature [11], with unity length in the presence of a single source, and less than unity in the presence of multiple sources or noise and ambience. We include these features in the training by stacking them as additional channels along with the mel-band energies (denoted as FOA-IV in the results).

3 Experimental results

We first evaluated the proposed extensions on the in-domain tasks of AVC and AVSA that are the same as used for training the system. The results showed a significant improvement brought by the FOA-IV features compared to [7]. In particular, the spatial alignment benefits of the use of spatial features, with an increase in performance from 71% to 81%. The combination of beamforming and YOLO did not bring any advantage for the in-domain tasks.

We tested the method on the human action recognition similar to [7] using the UCF dataset [14] that contains 101 classes, and the HMDB dataset [6] that

contains 51 classes. The results are presented in Table 1. Without fine tuning, our proposed additions outperform the baseline in [7] in most cases on the UCF dataset, and only for some combination on HMDB. When using fine tuning, the advantage brought by the different extensions is diminished, even though the performance obtained is, in some cases, higher. On UCF, the method using FOA-IV performs the best, with 1.7% advantages over baseline; on HMDB, the method using beamforming with YOLO achieves the highest with 2.2% advantages over baseline. In both cases their performance lies in the 95% confidence interval of the baseline.

	without fine tuning		with fine tuning	
Dataset	UCF	HMDB	UCF	HMDB
Baseline [7]	36.3	23.1	65.9	32.0
FOA + YOLO	35.7	23.0	65.0	32.9
FOA + FOA-IV	36.9	22.7	**67.6**	33.3
FOA + FOA-IV + YOLO	37.5	24.0	66.2	32.4
Beamform	37.1	22.1	65.9	31.8
Beamform +YOLO	**40.4**	**24.2**	66.9	**34.2**

Table 1. Action recognition accuracy (%) without and with fine tuning.

We use the audio embeddings on acoustic scene classification downstream task using ambisonic audio, using the EigenScape dataset [4], consisting of audio from eight classes. The spatial classification system provided with the data [4] had a 69% performance, later outperformed by a CNN approach, with 82% performance [3]. Our results without any fine-tuning are presented in Table 2, and show the superiority of learning through spatial correspondence, with AVSA having 89.4% accuracy compared to 82.5% for AVC, and outperforming significantly the previous state-of-the-art on this task.

	AVC training	AVSA training
Baseline [7]	82.5	**89.4**
FOA-IV	82.5	74.1

Table 2. Acoustic scene classification accuracy (%) on EigenScape data, no fine tuning

4 Conclusions

In this work, we proposed a self-supervised learning method for learning audio representations based on spatial alignment between audio and video information. To create a strong correspondence between the audio and video content, we proposed a new method for sampling crops by detecting the objects in the video frame using YOLO. Additionally, to use the spatial audio information from Ambisonics to its full extent, we proposed use of acoustic intensity vector as feature representation for the audio input.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 609–617 (2017)
2. Cramer, J., Wu, H.H., Salamon, J., Bello, J.P.: 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). pp. 3852–3856 (2019). https://doi.org/10.1109/ICASSP.2019.8682475
3. Green, M.C., Adavanne, S., Murphy, D., Virtanen, T.: Acoustic scene classification using higher-order ambisonic features. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 328–332. IEEE (2019)
4. Green, M.C., Murphy, D.: Eigenscape: A database of spatial acoustic scene recordings. Applied Sciences **7**(11) (2017). https://doi.org/10.3390/app7111204
5. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: Proc. of the European conference on computer vision (ECCV). pp. 649–665 (2018)
6. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: 2011 International Conference on Computer Vision. pp. 2556–2563 (2011). https://doi.org/10.1109/ICCV.2011.6126543
7. Morgado, P., Li, Y., Nvasconcelos, N.: Learning representations from audio-visual spatial alignment. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 4733–4744. Curran Associates, Inc. (2020)
8. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proc. of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018)
9. Perotin, L., Serizel, R., Vincent, E., Guerin, A.: Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings. IEEE Journal of Selected Topics in Signal Processing **13**(1), 22–33 (2019)
10. Politis, A., Poirier-Quinot, D.: Jsambisonics: A web audio library for interactive spatial sound processing on the web. In: Interactive Audio Systems Symposium (2016)
11. Pulkki, V., Politis, A., Laitinen, M.V., Vilkamo, J., Ahonen, J.: First-order directional audio coding (dirac). Parametric Time-Frequency Domain Spatial Audio **10**, 89–140 (2017)
12. Rafaely, B.: Fundamentals of spherical array processing (2021)
13. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
14. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. CoRR **abs/1212.0402** (2012)
15. Wang, S., Mesaros, A., Heittola, T., Virtanen, T.: Audio-visual scene classification: Analysis of DCASE 2021 Challenge submissions. In: Proc. of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021). pp. 45–49. Barcelona, Spain (November 2021)
16. Yang, K., Russell, B., Salamon, J.: Telling left from right: Learning spatial correspondence of sight and sound. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9932–9941 (2020)