Audio–Visual Model Distillation Using Acoustic Images

Andrés F. Pérez[1], Valentina Sanguineti[1,2], Pietro Morerio[1], Vittorio Murino[1,3,4]
andres.perez@mail.polimi.it {valentina.sanguineti, pietro.morerio, vittorio.murino}@iit.it
[1]Pattern Analysis & Computer Vision - Istituto Italiano di Tecnologia, [2]Università degli Studi di Genova, Italy,
[3]Computer Science Department - Università di Verona, Italy, [4]Huawei Technologies Ltd., Ireland Research Center

Abstract

In this paper, we investigate how to learn rich and robust feature representations for audio classification from visual data and acoustic images, a novel audio data modality. Former models learn audio representations from raw signals or spectral data acquired by a single microphone, with remarkable results in classification and retrieval. However, such representations are not so robust towards variable environmental sound conditions. We tackle this drawback by exploiting a new multimodal labeled action recognition dataset acquired by a hybrid audio-visual sensor that provides RGB video, raw audio signals, and spatialized acoustic data, also known as acoustic images, where the visual and acoustic images are aligned in space and synchronized in time. Using this richer information, we train audio deep learning models in a teacher-student fashion. In particular, we distill knowledge into audio networks from both visual and acoustic image teachers. Our experiments suggest that the learned representations are more powerful and have better generalization capabilities than the features learned from models trained using just single-microphone audio data.

1. Introduction

Humans experience the world through a number of simultaneous sensory observation streams. The co-occurrence of these streams provides a useful learning signal to understand the environment surrounding us [13]. There is in fact evidence that audio-visual mirror neurons play a central role in the recognition of actions given their temporal synchronization [4]. Furthermore, it was found that many neurons with receptive fields spatially aligned across modalities show a super-additive response to coincident and co-localized multimodal stimulations [44].

In this paper, motivated by these findings, we investigate whether and how visual and acoustic data *synchronized in time* and *aligned in space* can be exploited for scene understanding. We take advantage of a recent audio-visual sensor, called DualCam, composed by an optical camera and a 2D planar array of microphones (see Figure 3), able to provide spatially localized acoustic data aligned with the corresponding optical image (see Figure 1, right) [47]. Specifically, by combining the raw signals acquired by 128 microphones (by beamforming [43]), this sensor is able to output an acoustic image where each pixel represents the imprint of the sound coming from the corresponding pixel location in the optical image. Using this sensor, we generate a new multimodal dataset depicting different subjects performing several actions in multiple scenarios. By exploiting spatialized audio information coupled to the related visual data and designing suitable multimodal deep learning models, we aim at generating more discriminant and robust features, likely resulting in a better description of the scene content for robust audio classification. Figure 1 shows the multispectral acoustic image used as input data, which has 512 frequency bins and an example of visualization of an acoustic image overlaid upon an optical image.

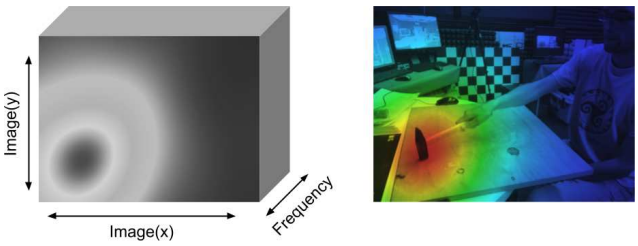


Figure 1. *Left*: multispectral acoustic image volume associated to the audio content of the sensed scene. It has two spatial dimensions (aligned with the visual image space) and a frequency axis of 512 bins that cover the sensor's audible range. Each image in the volume represents the spatial audio information associated to each frequency bin. *Right*: visualization (as heat color map) of an acoustic image formed by summing the energy of every frequency bin between $900Hz$ and $6400Hz$ for each spatial location, overlaid on the corresponding video frame.

The idea of leveraging the co-occurrence of visual and audio events as supervisory signal is not new. Former approaches in the pre deep-learning era combined visual and

auditory signals in rather simplistic ways. For instance, in [45] a neural network was trained to predict the auditory signal given the visual input. A particularly relevant earlier work is [8], which introduced a self-supervised learning algorithm for jointly training audio and visual networks by minimizing codebook disagreement. Another interesting work is [22], which presented an algorithm based on canonical correlation analysis (CCA) to detect pixels associated to the sound, while filtering out other dynamic (but silent) pixels.

Several recent works address audio-related tasks such as natural sound recognition [25], speech separation and enhancement [1, 10], audio event classification or sound source localization [34, 41], either by directly modeling raw audio signals with 1D convolutions [5, 30, 36] or, most popularly, by modeling intermediate sound representations such as spectrograms or cochleograms [2, 3, 32, 31, 37, 38, 39, 46]. Nevertheless, none of the past works tried to exploit spatially localized acoustic data to assess the potentialities of such richer information source.

In our work, we claim that it is possible to train audio deep learning models to face an action recognition problem in a more robust way across different scenarios utilizing a teacher-student framework able to distill knowledge [12, 27] from state-of-the-art vision network models and from a novel architecture that operates on the spatialized acoustic data. Similarly to [29], our intuition is to learn better features for a given modality assuming the availability of other complementary modalities at training time. We leverage video and multispectral acoustic image sequences aligned in space/time as side information at training, and predict actions given only a raw audio signal acquired by a single microphone at test, in a cross-scenario setting, where the environmental noise conditions are significantly different. Current methods, even best deep learning models, lead to very low classification accuracies [14, 28] in such conditions.

Hence, in essence, in this work we try to answer the following question: *Does spatialized data allow to learn more discriminant features for single-microphone audio classification?* In this respect, our main contributions can be summarized as follows.

1. We propose a thorough study to assess whether visual and acoustic data aligned in space and synchronized in time bring advantage for single-microphone audio classification.
2. We introduce a new multimodal dataset consisting in 14 action classes, in which acoustic and visual data are spatially aligned. This type of multi-sensory data has no counterpart in the literature and may lead to further studies by the scientific community.
3. We develop a deep teacher-student model to deal with such new data, showing that it is indeed possible to extract semantically richer representations for improving audio classification from single microphone. In particular, we distill knowledge learned from spatialized audio-visual modalities to a single-microphone model.

It is worth to note that we are the first to propose an algorithm in which the transfer of knowledge involves teacher models considering 2 different modalities (2D audio and 2D visual data) and the student model is devised for a different modality (1D audio signal), when typically the student deals with the task of one of the teacher models.

We validate our approach 1) on the proposed action dataset, and 2) by transferring learned representations on a standard sound classification benchmark dataset, demonstrating remarkable capabilities and the usefulness of distillation for cross-scenario learning.

The remainder of this paper is organized as follows. We first discuss the related work in Section 2, mainly focusing on audio-visual models and benchmark datasets. In Section 3, we describe our new acquired multimodal action dataset, and in Section 4, we describe acoustic image pre-processing and we propose the network architecture to deal with acoustic images. In Section 5, we present our distillation-based approach to deal with multispectral acoustic data, and in Section 6, we extensively validate our proposed framework by devising a set of experiments in order to assess the soundness of the learned representations. Finally, we draw conclusions in Section 7.

2. Related Work

We briefly review related work in the areas of multimodal learning, video and sound self-supervision, and transfer learning. We also review already existing audio and audio-visual datasets.

Multimodal learning. Multimodal learning concerns relating information from multiple data modalities. Such data provides complementary semantic information due to correlations in between them [29]. We consider the cross-modality learning setting, in which data from multiple modalities is available only during training, while only data from a single modality is provided at testing phase. In [6, 7] the authors learn shared representations from multimodal aligned data and use them for cross-modal retrieval. [6] for instance considers three major natural modalities: vision, sound and language, while [7] considers five weakly aligned modalities: natural images, sketches, clip art, spatial text, and descriptions. Other works such as [12, 19] utilize RGB video images and depth information to learn feature representations through modality hallucination. In our work instead, we consider RGB video images, raw audio and acoustic images for training phase, and only raw audio at testing time.

Video and sound self-supervision. There has been increased interest in using deep learning models for multimodal fusion of auditory and visual signals to improve the performance of visual models or solve various speech-related problems, such as speech separation and enhancement.

First approaches trained single networks on one modality using the other one to derive some sort of supervisory signal [5, 16, 32, 31, 33]. For example [5, 16] train an audio network to correlate with visual outputs using pre-trained visual networks as a teacher. Others such as [31, 32] train a visual network to generate sounds by solving a regression problem consisting in mapping a sequence of video frames to a sequence of audio features. In [33] instead, they learn visual models using ambient sounds as scene labels.

More recent works [2, 3, 9, 30, 37] train both visual and audio networks aiming at learning multimodal representations useful for many applications, such as cross-modal retrieval, speech separation, sound source localization, action recognition, and on/off-screen audio source separation. For instance in [2, 3] they learn aligned audio-visual representations, using an audio-visual correspondence task. In [30] they train an early-fusion multisensory network to predict whether video frames and audio are temporally aligned. In [37] a two-stream network structure is trained utilizing an attention mechanism guided by sound information to localize the sound source.

They key factor in all these works is that they exploit the natural synchronization between auditory and visual signals by training in a self-supervised manner. Although we address our problem in a pseudo-supervised manner using a combination of hard and soft labels, we notice that the natural spatial alignment and time synchronization of the data produced by the DualCam sensor opens the door to also train models through self-supervision.

Transfer learning. Our work is strongly related to transfer learning which deals with sharing information from one task to another. In particular we transfer knowledge between networks operating on different data modalities (see Section 5). We perform transferring with the aid of the generalized distillation framework which proposes to use the teacher-student approach from the distillation theory to extract knowledge from a privileged information source [27], also called a teacher. In our case the privileged information leveraged at training time is represented by the additional modalities, i.e. video and acoustic images. A rather simple transfer mechanism is that of [5] which proposes a teacher-student self-supervised training procedure based on the Kullback-Leibler divergence to transfer knowledge from a vision model into sound modality using unlabeled video as a bridge. This mechanism resembles the generalized distillation framework, however they only rely on the teacher soft labels which are in general less reliable than hard labels. An interesting work is [19] which introduces a novel technique for incorporating additional information, in the form of depth images, at training time to improve test time RGB only detection models. We draw inspiration from [12] which addresses action recognition by distilling knowledge from a depth network into a vision network. They accomplish this by training a hallucination network [19] that learns to distill depth features. It is worth noticing that although [12] works with different data modalities, it is the closest to ours since they transfer knowledge with the aid of the generalized distillation framework.

Audio-visual datasets. Due to recent interest in audio-visual and multimodal learning, several audio and audio-visual datasets have emerged. Here we summarize some of the most prominent ones.

Flickr-SoundNet [5] is a large unlabelled dataset of completely unconstrained videos from Flickr, compiled by searching for popular tags and dictionary words. It contains over 2 million videos which total for over one year of continuous natural sound and video.

Kinetics-Sounds [2] comprises a subset of the Kinetics dataset [21], which contains YouTube videos manually annotated for human actions, and cropped to 10 seconds around the action. The subset contains 19k video clips formed by filtering the Kinetics dataset for 34 human action classes, which have been chosen to be potentially manifested visually and aurally.

FAIR-Play [11] is an unlabelled video dataset with binaural audio that mimics human hearing. It consists of 1.871 short clips of 10 seconds long musical performances, totaling 5.2 hours. It depicts different combinations of people playing musical instruments including cello, guitar, drum, ukelele, harp, piano, trumpet, upright bass, and banjo, in a large music room, in solo, duet, and multiplayer performances.

Environmental Sound Classification (ESC-50) [35] is a labeled collection of 2.000 environmental audio recordings manually extracted from Freesound. It consists of 5 seconds long recordings organized into 50 semantical classes loosely arranged into five major categories: animals, natural soundscapes & water sounds, human non-speech sounds, interior/domestic sounds, and exterior/urban noise.

Detection and Classification of Acoustic Scenes and Events (DCASE) [28] is a dataset consistent of of recordings from various acoustic scenes. It was recorded in six large European cities, in different locations for each scene class. For each recording location there are 5 to 6 minutes of audio split into segments of 10 seconds.

The closest dataset to ours is *FAIR-Play* because of its size and the nature of its data, since binaural audio is a form of spatial audio. Similarly to *Kinetics-Sounds* we propose a dataset of human actions, but with data in multiple modalities which try to describe more realistic conditions.

3. Audio-Visually Indicated Action Dataset

We introduce a new multimodal dataset comprised of visual data as RGB image sequences and acoustic data as raw audio signals acquired from 128 microphones. The latter signals, opportunely combined by a beamforming algorithm, compose a multispectral acoustic image volume, which is aligned in space and time with the optical images (see Figure 1). The following 14 actions were chosen:

1. Clapping
2. Snapping fingers
3. Speaking
4. Whistling
5. Playing kendama
6. Clicking
7. Typing
8. Knocking
9. Hammering
10. Peanut breaking
11. Paper ripping
12. Plastic crumpling
13. Paper shaking
14. Stick dropping

For the acquisition, we acknowledge the participation of 9 people performing the aforementioned actions recorded in three different scenarios, with increasing and varying noise conditions, namely, an anechoic room, an indoor open space area, and a terrace outdoor. We name them scenario 1, 2, and 3, respectively. In our dataset, the same action is performed by different subjects in distinct places, so allowing to show the equivariance properties of the multispectral acoustic images across subjects, scenarios and position in the scene, which are exploited when learning audio features from an acoustic teacher model. In the end, the dataset consists of 378 audio-visual video sequences (27 per action) between 30 and 60 seconds depicting different people individually performing a set of actions producing a characteristic sound in each scenario. Figure 2 shows representative samples of our dataset for the 3 considered scenarios.

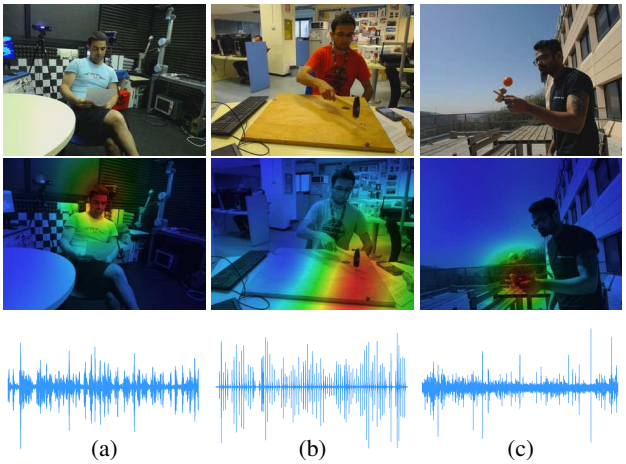


Figure 2. Three examples of Audio-Visual Indicated Actions dataset represented as video frame, acoustic image visualization overlaid on the frame, and raw waveform (from a single microphone). (a) Speaking in anechoic room. (b) Hammering in the indoor open space area. (c) Playing Kendama in the terrace.

We acquired the dataset using the DualCam acoustic-optical camera described in [47]. The sensor captures both audio and video data using a $0.45m \times 0.45m$ planar array of 128 low-cost digital MEMS microphones located according to an optimized aperiodic layout, and a video camera placed at the device center as depicted in Figure 3.

Figure 3. DualCam acoustic-optical camera.

The device is capable of acquiring audio data in the range $200\,\mathrm{Hz}$ – $10\,\mathrm{kHz}$ and audio-video sequences at a frame rate of 12 fps. In our acquisition setup the camera was static looking at the scene, while the subjects moved around within its field of view at a minimum distance of 2 meters from the device.

After collecting the dataset, audio and video data had to be synchronized since they were acquired in an interleaved way at different frame rates.

The data provided by the sensor consists in RGB video frames of 640×480 pixels, raw audio data from 128 microphones acquired at a frequency of 12 kHz, and $36 \times 48 \times 512$ multispectral acoustic images obtained from the raw audio signals of all the microphones using beamforming, which summarize the per-direction audio information in the frequency domain. This means that each acoustic pixel corresponds to 13,3 visual pixels, in fact acoustic resolution is lower than optical one. Among the raw audio waveforms, we choose the one of just one microphone for testing single microphone audio networks.

4. Learning with Acoustic Images

In this section, we describe acoustic images representation, their pre-processing and the network architecture we proposed for modelling this novel type of data.

Acoustic Images Pre-processing. Multispectral acoustic images are generated with the frequency implementation of the filter-and-sum beamforming algorithm [43], aimed at producing a volume of size 36×48, with 512 channels corresponding to the frequency bins which represent the frequency information. Full details of the algorithm can be found in [47].

Handling input acoustic images with 512 channels is a computationally expensive task and typically the majority of information in our dataset is contained in the low

frequencies. Consequently, we decided to compress the acoustic images using Mel-Frequency Cepstral Coefficients (MFCC), which consider audio human perception characteristics [40]. Therefore, we compute 12 MFCC, going from from $36 \times 48 \times 512$-D volumes to $36 \times 48 \times 12$-D volumes, retaining the most important information and reducing consistently the computational complexity and the memory footprint.

DualCamNet Architecture. Acoustic images provide a small temporal support which is generally not enough for discriminating information over time intervals of several seconds. For this reason, we feed to our network a set of 12 consecutive $36 \times 48 \times 12$ acoustic images corresponding to 1 second of audio data. We deem that 1 second of acoustic images is a reasonable trade-off between sound information content and processing cost.

In order to train a model able to discriminate information from acoustic images, we explicitly model both the spatial and the temporal relationships among them. To this end, we propose the architecture structure shown in Figure 4a which utilizes 3D convolutions as commonly done in visual action recognition [42], where the spatial and temporal convolutions are decoupled.

We follow the LeNet [24] design style, with 5×5 convolutional filters, and 2×2 max-pooling layers with stride 1 and zero-padding to keep the spatial resolution. The network includes 3 blocks of convolutional layers plus a block of 3 fully convolutional layers which produces the output prediction.

The first block consists of a single 1D convolutional layer over time followed by a ReLU nonlinearity. The aim of this layer is modeling the temporal relationship of consecutive acoustic images by aggregating them. In particular, we apply a filter of size 7 with stride 1 and zero-padding to keep the temporal resolution. We experimented with several filters sizes finding 7 to be the best one.

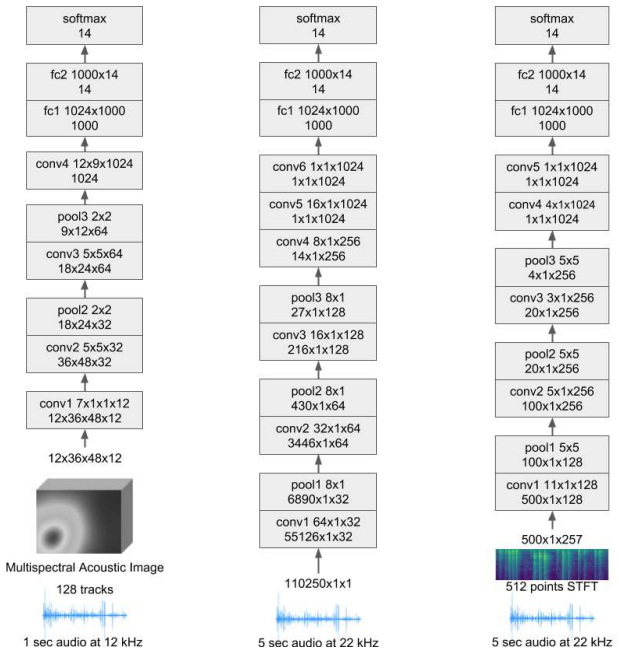
The second and third blocks model the spatial equivariance of the acoustic images and consist of a 2D convolutional layer followed by max-pooling. We go from the 12 channels of the input to 32 channels and then double it to 64. Each convolutional layer is followed by batch normalization [20] and ReLU nonlinearity.

The final block comprises 3 fully convolutional layers with ReLU in between. It converts the input feature map into a 14-D classification vector as output, namely the predicted class probabilities, using intermediate features size of 1024-D and 1000-D.

This model will be used as teacher network in our validation experiments.

5. Model Distillation

In this section, we describe the utilized network architectures and the knowledge transfer procedure.



(a) DualCamNet (b) OursSoundNet (c) HearNet

Figure 4. Our proposed networks. (a) DualCamNet architecture, used as teacher model. (b) OursSoundNet architecture, used as student model. (c) HearNet architecture, used as student model.

5.1. Architectures

Similarly to [12], we utilize data from multiple modalities at training phase, and only data from a single modality at testing phase. We leverage either RGB video images or multispectral acoustic images in training as side information, and we test only on audio data from a single microphone.

We want to emphasize here that, to the best of our knowledge, this is the first time that model distillation is performed from modalities different from those utilized in testing. Specifically, we train on 2-dimensional spatialized audio and video data, to improve accuracy on a model working on mono-dimensional audio signals only as input. As further original aspect, [12] trains one ResNet-50 [17] network per stream, while we use different network architectures for each stream of our model.

Teacher Networks. For the visual stream, we experimented with two models, ResNet-50 [17] and its variation including 3D temporal convolutions introduced in [12], here called Temporal ResNet-50. We choose ResNet-50 respect to Temporal ResNet-50 as it provides a good compromise between network size and accuracy. On the other hand, Temporal ResNet-50 stands as a strong action recognition model dealing with action dynamics with the aid of temporal connections between residual units. It has also been

selected since it constitutes a powerful baseline model to compare with. DualCamNet, explained earlier in Section 4, will be used as teacher model as well in the following.

Student Networks. Regarding the raw audio waveform stream, we experimented two models that capture different characteristics of audio data. The first one is SoundNet [5], which operates over time domain signals. We preferred the 5-layer version over the 8-layer one, as our dataset is not big enough to allow SoundNet to grasp the underlying data patterns. We used the exact same architecture described in [6], adding 3 fully convolutional layers at the bottom of the network with 1024, 1000 and 14 filters, respectively. To avoid further confusion, we named our version *OurSoundNet*.

The second model is a network based on the sound subnetwork presented in [6], called from here on, *HearNet*. Its architecture is shown in Figure 4c. This network operates on amplitude spectrograms obtained from an audio waveform of 5 seconds, upsampled to 22 kHz. Such spectrogram was produced by computing the STFT [1] considering a window length of 20 ms with half-window overlap. This produces 500 windows with 257 frequency bands. The resulting $500 \times 1 \times 257$ spectrogram is interpreted as a 257-dimensional signal over 500 time steps.

HearNet processes spectrograms with 3 1D convolutions using kernel sizes 11, 5, 3, and 128, 256, 256 filters, respectively, with stride 1. The last convolutional layers are fully convolutional and use 1024, 1024, 1000 and 14 filters to obtain the class predictions. We applied zero-padding in all layers except conv4 in order to keep the spatial resolution. The chosen activation function is ReLU. After each of the first 3 convolutional layers, we downsampled with one-dimensional max-pooling by a factor of 5.

5.2. Training procedure

Following the *generalized distillation* framework [27], we first learn a teacher function $f_t \in \mathcal{F}_t$ by solving a classification problem and, second, we compute the teacher soft labels s_i. As third step, we distill $f_t \in \mathcal{F}_t$ into $f_s \in \mathcal{F}_s$ by using both the hard and soft labels. The knowledge transfer procedure is graphically illustrated in Figure 5.

In particular we transfer knowledge between multiple modality network streams by using Hinton's distillation loss [18] to extract knowledge from privileged representations. More formally, we distill the teacher learned representation $f_t \in \mathcal{F}_t$ into $f_s \in \mathcal{F}_s$ as follows:

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} (1-\lambda)\ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))), \tag{1}$$

where $s_i = \sigma(f_t(x_i^*/T)) \in \delta^c$ are the soft labels derived from the teacher about the training data, $\mathcal{F}_t$ and $\mathcal{F}_s$ are classes of functions described by the teacher and student models [27], respectively, σ is the softmax operator, and y_i are the ground truth hard labels. The imitation parameter $\lambda \in [0, 1]$ allows to balance the weight of soft labels with respect to the true hard labels y_i. The temperature parameter $T > 0$ allows to smoothen the probability vector predicted by the teacher network f_t.

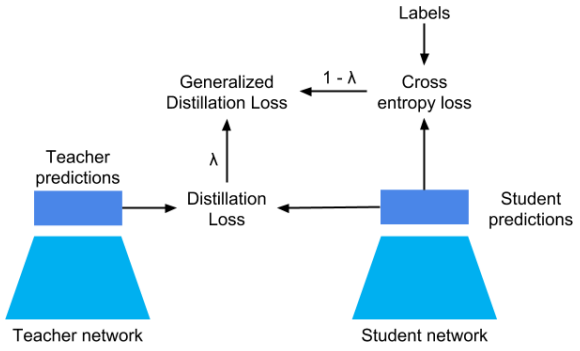


Figure 5. Teacher-student training procedure

6. Experimental Results

Our goal is to learn feature representations for raw audio data by transferring knowledge across networks operating on different data modalities. To evaluate how well our method addresses this problem we perform two sets of experiments with the objectives of 1) showing the improvement brought by distilling knowledge from different data modality networks and 2) assessing the quality of the distilled representations on a standard sound classification benchmark.

6.1. Acoustic Features Transfer

In this first set of experiments, we evaluate the performance of the teacher and student networks on the task of action recognition on our dataset. We train both the teacher and student networks in a fully supervised manner using action labels as ground truth, and only the student networks following the distillation procedure described in Section 5. In all cases we trained for 100 epochs[2] with batches of 32 elements using the Adam optimizer [23] with learning rates of 1×10^{-3} and 1×10^{-4} (see details in Supplementary Material). In order to measure the generalization capabilities of the learned representations, we evaluate the accuracy of our trained models on a cross-scenario setting, i.e. when the model is trained on certain scenario, it is tested on the other two scenarios using all the available data. In all cases the data was split by assigning 80% of them for training, 10% for validation and 10% for test.

Teachers Networks. First, we train our DualCamNet model and the two proposed visual networks, ResNet-50

[1] Short-Time Fourier Transform

[2] The number of iterations varies with the size of the training set.

and Temporal ResNet-50, as they constitute our baselines. Table 1 shows their performance. We observe that our DualCamNet convincingly outperforms the visual networks in all combination of scenarios. This indicates that most of the actions in our dataset are better distinguishable aurally than visually. One possible explanation for this, is that in the majority of the cases the "object" involved in the action execution, e.g. mouth, mouse or hammer, is not easily visible but has a characteristic sound signature.

A comparison of the two visual networks reveals that they achieve similar results throughout all configurations, indicating that motion is not a key factor to model the actions performed in our dataset. Consequently, we choose ResNet-50 over Temporal ResNet-50 as visual teacher for the rest of the experiments since the former one has a simpler structure.

Additionally, we have designed a hybrid network which combines the output of the DualCamNet and ResNet-50, to check whether modality fusion brings any performance improvement. We do so by concatenating the 1024 feature volumes of the two networks and processing them with two fully convolutional layers of 1000 and 14 filters, respectively. This network achieves a 7.1% improvement in accuracy with respect to DualCamNet when trained over all scenarios. It is important to note that it also consistently improves the testing accuracy in all cross-scenario configurations (see Table 1, AV column). These findings indicate some benefits brought by modality fusion that can be further explored in future research.

Train set	Test set	D	R	T	AV
Scenario 1	Scenario 1	0.8470	0.6965	0.7117	0.8775
	Scenario 2	0.2938	0.2955	0.2616	0.3490
	Scenario 3	0.1471	0.1355	0.1410	0.1528
Scenario 2	Scenario 1	0.2986	0.1918	0.1844	0.3060
	Scenario 2	0.7600	0.5838	0.4987	0.7418
	Scenario 3	0.1504	0.1486	0.1243	0.2049
Scenario 3	Scenario 1	0.2309	0.1479	0.1571	0.2767
	Scenario 2	0.2032	0.1229	0.1063	0.2182
	Scenario 3	0.6736	0.2240	0.3013	0.5708
All scenarios	All scenarios	0.7702	0.6335	0.6393	0.8412

Table 1. Test accuracy for teacher models. D: DualCamNet. R: ResNet-50 [17]. T: Temporal ResNet-50 [12]. AV: AVNet.

Student Networks. In order to measure the improvement brought by distillation, we need to look first at the performance of the two proposed student networks when trained only from hard labels only. Column G from Tables 2 and 3 show the accuracy results for OurSoundNet and HearNet, respectively. It can be observed that both networks perform well, with HearNet achieving a higher accuracy in all scenarios settings.

This result is impressive considering that OurSoundNet was fine-tuned from SoundNet-5 which was trained on the Flickr-SoundNet dataset, while HearNet instead was trained from scratch on our dataset. A reasonable explanation for this is that shallow networks such as HearNet perform better under small data regimes.

Train set	Test set	G	D	R
Scenario 1	Scenario 1	0.4881	0.6071	0.5238
	Scenario 2	0.4114	0.4669	0.4378
	Scenario 3	0.1958	0.2844	0.1958
Scenario 2	Scenario 1	0.4339	0.3598	0.4220
	Scenario 2	0.3333	0.3810	0.2619
	Scenario 3	0.1931	0.1799	0.1786
Scenario 3	Scenario 1	0.3796	0.4352	0.3955
	Scenario 2	0.2513	0.3386	0.2725
	Scenario 3	0.3690	0.3452	0.2619
All scenarios	All scenarios	0.4102	0.5299	0.4145

Table 2. Test accuracy for OurSoundNet trained with distinct supervisory information. G: Ground truth hard labels. D: DualCamNet soft labels. R: ResNet-50 soft labels.

Train set	Test set	G	D	R
Scenario 1	Scenario 1	0.6548	0.7857	0.7262
	Scenario 2	0.4286	0.4325	0.4960
	Scenario 3	0.1627	0.1825	0.2989
Scenario 2	Scenario 1	0.4100	0.5542	0.5106
	Scenario 2	0.3214	0.2619	0.4524
	Scenario 3	0.1627	0.1825	0.1799
Scenario 3	Scenario 1	0.3307	0.3770	0.4405
	Scenario 2	0.2976	0.3056	0.2765
	Scenario 3	0.5000	0.6190	0.6071
All scenarios	All scenarios	0.6966	0.7009	0.6282

Table 3. Test accuracy for HearNet [6] trained with distinct supervisory information. G: Ground truth hard labels. D: DualCamNet soft labels. R: ResNet-50 soft labels.

Teacher-Student Networks. Finally, we compare the performance of the student networks when trained by distilling knowledge from the teacher networks. These results are shown in columns D and R from Tables 2 and 3.

We observe that whenever we perform training by transferring either from DualCamNet or ResNet-50 using data from scenario 1, we obtain better results and good generalization. When transferring from DualCamNet, the improvement can be ascribed to the fact that data acquired in the anechoic room is cleaner than in other scenarios. Similarly, when transferring from ResNet-50, there is little clutter in the scene, allowing the network to easily capture the objects involved in the action execution.

Such ideal conditions do not occur in scenarios 3 and (especially) in scenario 2, which are considerably more acoustically noisy and visually cluttered. In particular the worst case is scenario 2, where echoes are even more disturbing than background noise of scenario 3. In fact, in scenario 3, distilling from DualCamNet improves accuracy in all cases

except for the OurSoundNet student tested on scenario 3. We are not able to improve results in all testing scenarios when training on scenario 2 as echoes introduce too much noise. ResNet-50 soft labels, on the contrary, do not always guarantee an increase in accuracy neither in scenario 2 nor in scenario 3.

Overall we thus notice that distillation from DualCamNet provides higher improvement over distillation from ResNet-50, especially when training on all scenarios. In some exceptional cases, the teacher is able to help the student even though it could not achieve a good accuracy. For instance, ResNet-50 trained on the scenario 3 achieves a 22.4% test accuracy and HearNet on the same setting reaches 50.0% but, when transferring from the ResNet-50 to HearNet the accuracy improves up to 60.71%.

We validated the chosen hyper-parameters, and found $T = 1$ and $\lambda = 0.5$ to be the best temperature value and imitation parameter, respectively. This means that we keep the teacher predictions unchanged and give them equal importance than to the hard labels. Interestingly our finding about λ is consistent with that of [12].

In summary, these results show that knowledge distillation allows learning more robust features given there is not much noise corrupting the data.

6.2. Acoustic Features Quality Assessment

Finally, we tested our student networks trained through distillation on a simple classification task on a standard sound benchmark, the DCASE-2018 dataset [28]. Specifically, we performed both k-NN and SVM classification on the features extracted with our distilled student networks to verify whether the learned representations were general enough to perform well in a different audio domain.

Table 4 reports our results in comparison with those obtained from established baselines [28, 26, 15] which were trained on DCASE-2018 and that of SoundNet-5 pre-trained on Flickr-SoundNet.

For OurSoundNet, we employed features both from the fc1 and conv4 layers, in order to be comparable with the conv5 and conv4 of the original SoundNet-5. We notice that OurSoundNet/conv4 outperforms OurSoundNet/fc1 by around 14%. This might be because fc1 has learned feature which are very specific for our dataset. On the other hand, conv4 layer features for both models perform similarly, because they captures less class-specific information, so also OurSoundNet has more general features.

Finally, for HearNet we considered the features from fc1 and some convolutional layers. We observe that lower layer learn features which are more general and thus transfer better to DCASE-2018. This is reasonable since higher layer encode more label-specific information, and this network was trained from scratch on our dataset. Performances of Hearnet lower layers are similar to those of OurSoundNet/fc1, which was pre-trained on Flickr-SoundNet, but was then adapted to our dataset. In conclusion, features learned with our dataset, which comprises 3 hours of videos, transfer reasonably well to DCASE if compared to features learned from the huge Flickr-SoundNet (2M videos).

Features	Training Dataset	Test accuracy	
Mesaros *et al.* [28]	DCASE-2018	0.597	
Liping *et al.* [26]	DCASE-2018	0.798	
Golubkov *et al.* [15]	DCASE-2018	0.801	
HearNet/fc1	Ours	0.2419	0.2609
HearNet/conv5	Ours	0.2488	0.2740
HearNet/conv4	Ours	0.2631	0.2967
HearNet/conv3	Ours	0.2754	0.3100
HearNet/conv2	Ours	0.2810	0.3403
OurSoundNet/fc1	Flickr-SoundNet+Ours	0.2746	0.3014
OurSoundNet/conv4	Flickr-SoundNet+Ours	0.4067	0.4420
SoundNet-5/conv5	Flickr-SoundNet	0.4180	0.4643
SoundNet-5/conv4	Flickr-SoundNet	0.4184	0.4275

Table 4. Dataset transfer results for DCASE-2018 [28]. Feature extracted by the models distilled from DualCamNet presented in Section 5 are fed into k-NN (*left*) and SVM (*right*) classifiers. The number of nearest neighbours is validated on the validation set.

7. Conclusions

In this work, we investigate whether and how it is possible to transfer knowledge from visual data and spatialized sound, namely, acoustic images, in order to improve audio classification from single microphone. To this end, we take advantage of a special sensor, DualCam, an acoustic-optical camera that provides in output audio-visual data synchronized in time and spatially aligned. Using this sensor, we acquired a novel audio-visually indicated action dataset in 3 different scenarios, from which we aim at extracting information useful for audio classification.

The peculiar nature of the generated acoustic images synchronized with optical frames, never studied before, led to the design of deep learning models in the context of the teacher-student paradigm, in order to assess if this information was transferable and indeed useful for single-channel audio classification. We highlight here that the proposed teacher-student framework is the first able to distill from 2D visual data and acoustic images to a model taking as input a 1D modality, namely, audio signals.

On a set of experiments, in which we learnt from visual data and acoustic images separately, we found out that the distilled models are effective in the audio classification task. Future work aims at further exploring the capabilities of this sensor for detection, recognition, self-supervised learning, sound source localization and cross-modal retrieval.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. 2018.

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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

[4] R. Arrighi, F. Marini, and D. Burr. Meaningful auditory information enhances perception of visual biological motion. *Journal of Vision*, 9(4):25–25, 2009.

[5] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 892–900, USA, 2016. Curran Associates Inc.

[6] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *CoRR*, abs/1706.00932, 2017.

[7] L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2949, June 2016.

[8] V. R. DeSa. Learning classification with unlabeled data. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 112–119, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *CoRR*, abs/1804.03619, 2018.

[10] A. Gabbay, A. Shamir, and S. Peleg. Visual speech enhancement. 2018.

[11] R. Gao and K. Grauman. 2.5d visual sound. *arXiv e-prints*, page arXiv:1812.04204, Dec. 2018.

[12] N. C. Garcia, P. Morerio, and V. Murino. Modality distillation with multiple stream networks for action recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[13] W. W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[14] S. Gharib, K. Drossos, E. Çakir, D. Serdyuk, and T. Virtanen. Unsupervised adversarial domain adaptation for acoustic scene classification. *ArXiv e-prints*, Aug. 2018.

[15] A. Golubkov and A. Lavrentyev. Acoustic scene classification using convolutional neural networks and different channels representations and its fusion. Technical report, DCASE2018 Challenge, September 2018.

[16] D. Harwath, A. Torralba, and J. Glass. Unsupervised learning of spoken language with visual context. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1858–1866. Curran Associates, Inc., 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[18] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *NIPS 2014 Deep Learning Workshop*, abs/1503.02531, 2015.

[19] J. Hoffman, S. Gupta, and T. Darrell. Learning with side information through modality hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, June 2016.

[20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[22] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 88–95 vol. 1, June 2005.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR 2015*, abs/1412.6980, 2015.

[24] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. *Object Recognition with Gradient-Based Learning*, pages 319–345. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[25] X. Li, V. Chebiyyam, and K. Kirchhoff. Multi-stream network with temporal attention for environmental sound classification. *CoRR*, abs/1901.08608, 2019.

[26] Y. Liping, C. Xinxing, and T. Lianjie. Acoustic scene classification using multi-scale features. Technical report, DCASE2018 Challenge, September 2018.

[27] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. *ICLR 2016*, abs/1511.03643, 2016.

[28] A. Mesaros, T. Heittola, and T. Virtanen. A multi-device dataset for urban acoustic scene classification. *ArXiv e-prints*, July 2018.

[29] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 689–696, USA, 2011. Omnipress.

[30] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[31] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016.

[32] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 801–816, Cham, 2016. Springer International Publishing.

[33] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *International Journal of Computer Vision*, 126(10):1120–1137, Oct 2018.

[34] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard. Weakly supervised representation learning for unsynchronized audio-visual events. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[35] K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1015–1018, New York, NY, USA, 2015. ACM.

[36] M. Ravanelli and Y. Bengio. Speaker recognition from raw waveform with sincnet. *ArXiv e-prints*, July 2018.

[37] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. On learning association of sound source and visual scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[39] N. Takahashi, M. Gygli, and L. V. Gool. Aenet: Learning deep audio features for video analysis. *IEEE Transactions on Multimedia*, 20(3):513–524, March 2017.

[40] H. Terasawa, M. Slaney, and J. Berger. A statistical model of timbre perception. In *SAPA@INTERSPEECH*, 2006.

[41] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[42] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[43] H. Van Trees. *Detection, Estimation, and Modulation Theory, Optimum Array Processing*. Wiley, 2002.

[44] M. T. Wallace, M. A. Meredith, and B. E. Stein. Converging influences from visual, auditory, and somatosensory cortices onto output neurons of the superior colliculus. *Journal of Neurophysiology*, 69(6):1797–1809, 1993. PMID: 8350124.

[45] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11):65–71, Nov 1989.

[46] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[47] A. Zunino, M. Crocco, S. Martelli, A. Trucco, A. D. Bue, and V. Murino. Seeing the sound: A new multimodal imaging device for computer vision. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 693–701, Dec 2015.