Audio–Visual Model Distillation Using Acoustic Images

Supplementary Material

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

1. Data Preparation

We implemented all of our networks and our data processing pipeline using TensorFlow. In particular we store our dataset in multiple compressed TFRecord files, each of which contains 1 second of synchronized data from the three modalities, video images, raw audio waveforms, and acoustic images. We use the `tf.data` API to retrieve this data and compose at runtime variable length sequences. We grouped contiguous TFRecord files into full audio-video sequences and then randomly sampled shorter length sequences, e.g. we compose a full audio-video sequence of 30 seconds and sample from it 10 sequences of 5 seconds.

2. Dataset Splitting

In Section 3 of the paper, we mentioned our dataset consists of 378 audio-video sequences from 30 to 60 seconds each. However we did not comment on how it was split for training purposes. Since only a few sequences were longer than 30 seconds, and in order to keep a balanced dataset, we cropped all the sequences up to 30 seconds and assign 80% of them for training, 10% for validation and 10% for test.

Splitting the dataset this way accounts for 302 training sequences, 39 validation sequences, and 37 test sequences. We then extracted sequences of the desired length. In case that the required length was 1 second we extracted 30 samples, while in case the required length was 5 seconds we extracted 6 samples. Extracting more samples would result in a high load of data repeated. Finally to keep some consistence across the experiments, we used a fixed seed for random crops extraction and the epoch number as seed for data shuffling.

3. Hyperparameter Optimization

In Section 6 of the paper, we presented the obtained experimental results and mentioned that in some cases we used a different learning rate. Basically we considered only two values, 1×10^{-3} and 1×10^{-4}. Table 1 shows the values used throughout all the experiments. For all teacher networks we used a learning rate of 1×10^{-4} except for *DualCamNet* which required a bigger value. For the student networks (*OursSoundNet* and *HearNet*) we used a mix of both considered values, and almost always the same across all scenarios settings, except for *HearNet* when trained from *DualCamNet* soft labels on first scenario which required a smaller learning rate.

Network	Learning rate
DualCamNet	1×10^{-3}
ResNet-50	1×10^{-4}
Temporal ResNet-50	1×10^{-4}
AVNet	1×10^{-4}
HearNet (G)	1×10^{-4}
HearNet (D)	1×10^{-3} and 1×10^{-4}
HearNet (R)	1×10^{-3}
OurSoundNet (G)	1×10^{-4}
OurSoundNet (D)	1×10^{-3}
OurSoundNet (R)	1×10^{-4}

Table 1. Training learning rates. Supervision is indicated as follows: (G): from ground truth hard labels, (D): from DualCamNet soft labels, (R): from ResNet-50 soft labels.

It is worth mentioning that in all cases when training our student networks with distillation, we performed hyperparameters optimization using grid search by cross-validation on the held-out validation set. We basically looked at three hyperparameters, learning rate (lr), temperature value (T), and imitation parameter (λ).

Finally, regarding the transfer learning results, also presented in Section 6 of the paper, we validated the considered number of nearest neighbors k. We computed accuracy with odd values between 7 and 15 included for validation set, choose on it best k and use that value for the testing accuracy which we report.

4. Dataset Qualitative Analysis

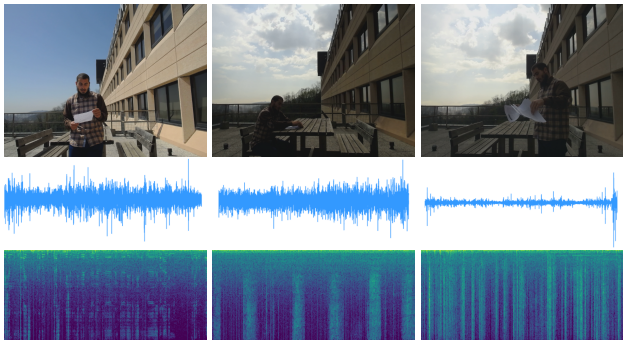
In this section we provide additional qualitative insights on the proposed dataset, which may clarify some statements made in the paper. We first illustrate the problem of visual clutter mentioned in Section 6 of the paper. Figure 1 shows three examples of actions performed over all three scenarios with varying conditions of visual clutter. Comparing scenarios 1 and 3, it can be observed that on the first case the object involved on the action execution is well visible in the foreground, making easier for the visual models to identify the corresponding action. With scenario 2 the difficulty is that often other people appear on the background or non-related objects are present on the foreground, thus making it harder to identify the action.

(a) Stick dropping (b) Clicking (c) Plastic crumpling

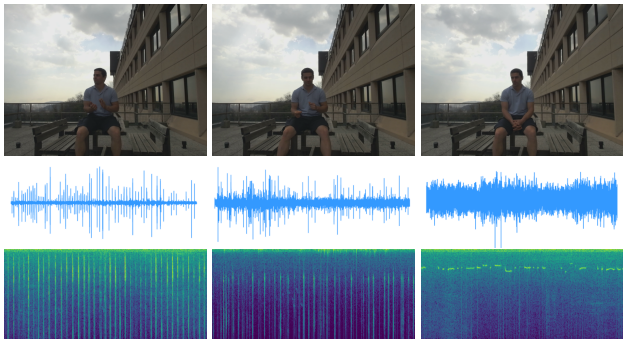
Figure 1. Comparison of three actions performed on all scenarios. From top to bottom, scenario 1 on the first row, scenario 2 on the second row, and scenario 3 on the third row.

A key finding on the paper was that models based on acoustic data achieved better classification results than models based on visual data. Here we illustrate the difficulty of identifying actions from visual data in contrast to identifying actions from audio data. Figure 2 shows two subjects on the third scenario performing three different actions each. It can be seen that some actions involving the same subject are visually similar although they depict completely different actions, but they are distinguishable by their acoustic signature.

Looking more closely at Figure 2, it can be seen that some actions have a visually distinguishable pattern. For instance, "clapping" and "snapping fingers" have a periodic pattern and concentrate on the low frequencies rather than on the high ones. Such patterns are more difficult to grasp from raw waveform. This lead us to think that spectrograms are better audio representations since they summarize the



(a) Speaking (b) Paper ripping (c) Paper shaking

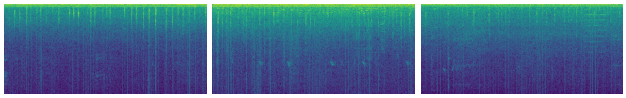


(d) Clapping (e) Snapping fingers (f) Whistling

Figure 2. Comparison of six actions visually similar but distinguishable from audio. All six actions where performed on the third scenario corresponding to the terrace.

scene acoustic content in a better way when compared to raw waveform. This observation gives some more clues into why *HearNet* performs better than *OurSoundNet* in many cases.

Figure 3 shows the spectrograms for the same action performed by three different subjects on the third location. There can be seen that the same pattern of multiple events spaced at short time intervals with the energy concentrated on the low frequencies, repeats across different subject executions.

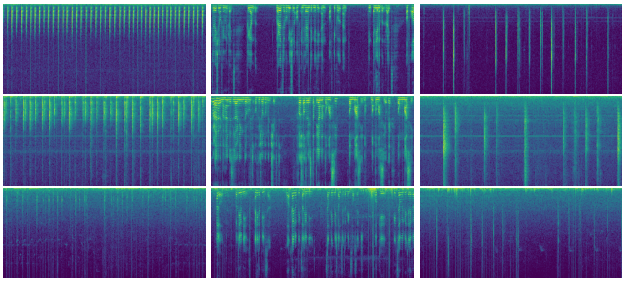


(a) Subject 3 (b) Subject 4 (c) Subject 5

Figure 3. Comparison of the spectrograms for the "knocking" action performed by three distinct subjects on the third scenario.

Figure 4 compares the spectrograms of the audios of three different actions performed by the same subject on the three considered scenarios. Here we also see that the audios for the same actions share a visual pattern when visualized

as a spectrogram, even when performed across locations. Interestingly, the cleanest spectrograms are those from actions performed at first scenario, while for second and third scenarios there are two different kinds of noise. In second scenario the noise is mainly due to indoor echoes, while for third scenario it is due to ambient noise.



(a) Knocking (b) Speaking (c) Playing kendama

Figure 4. Comparison of the spectrograms of three actions performed by the same subject at the three considered scenarios. From top to bottom, scenario 1 on the first row, scenario 2 on the second row, and scenario 3 on the third row.

5. Dataset Quantitative Analysis

We report here the confusion matrices for all the student and teacher models, in order to get a deeper understanding of the dataset's challenges.

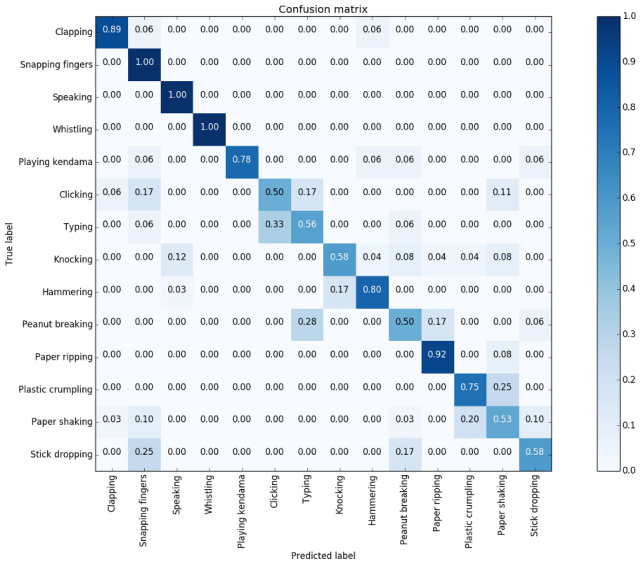


Figure 5. Hearnet trained on all scenarios confusion matrix.

For HearNet (Figure 5) we notice that Hammering is often confused with Knocking, Clicking with Typing, Paper shaking with Plastic crumpling. All the three pairs of classes, in fact, are very similar aurally.

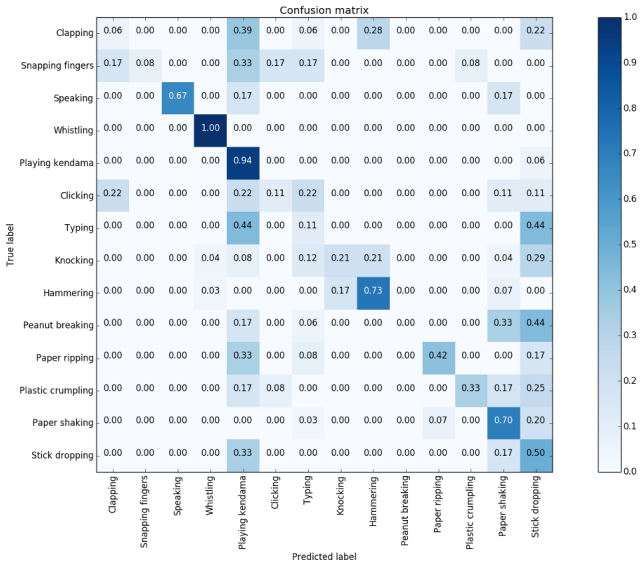


Figure 6. OurSoundNet trained on all scenarios confusion matrix.

Regarding OurSoundNet (Figure 6) many classes are confused with Playing kendama and Stick dropping. Hammering and Knocking, Paper shaking and Stick dropping are confused with each other, Peanut breaking is always misclassified, probably because of its feeble audio pattern. As stated before, HearNet superior performance may be ascribed to its more powerful input representation (spectrogram).

We now consider the teachers confusion matrices. DualCamNet (Figure 7) and AVNet (Figure 10) confusion matrices have diagonal elements with very high values, indicating high accuracy (they are good teachers indeed). Temporal ResNet-50 in Figure 9 and ResNet-50 in Figure 8 confuse many classes with Clapping and Clicking. Whistling is always misclassified. As already certified by higher accuracy, we can conclude that are DualCamNet and AVNet are better teacher.

Finally we can see in detail ResNet-50 confusion matrices when trained and tested on scenario 1 in Figure 11, scenario 2 in Figure 12 and in scenario 3 in Figure 13. We notice that when trained and tested on scenario 1, ResNet-50 presents higher accuracies for all classes. In scenario 2 many classes are confused with Clapping, in scenario 3 with Knocking. In particular, we see in scenario 1 that Snapping fingers, Speaking and Plastic Crumpling are the more difficult to recognize. In scenario 2 Speaking, Snapping fingers, Playing kendama and Paper shaking have low accuracies. In scenario 3 many classes have low results, for e.g. Clapping and Snapping fingers. As a matter of fact these classes are visually similar to other ones or sometimes the visual part of the images to recognize the action are occluded or there

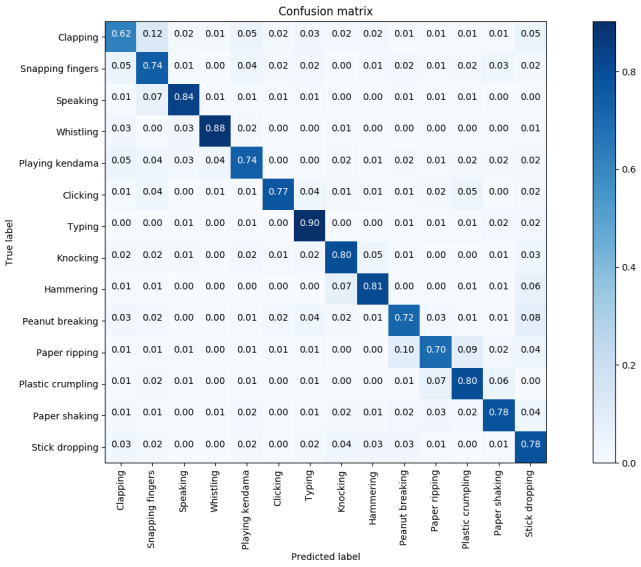


Figure 7. DualCamNet trained on all scenarios confusion matrix.

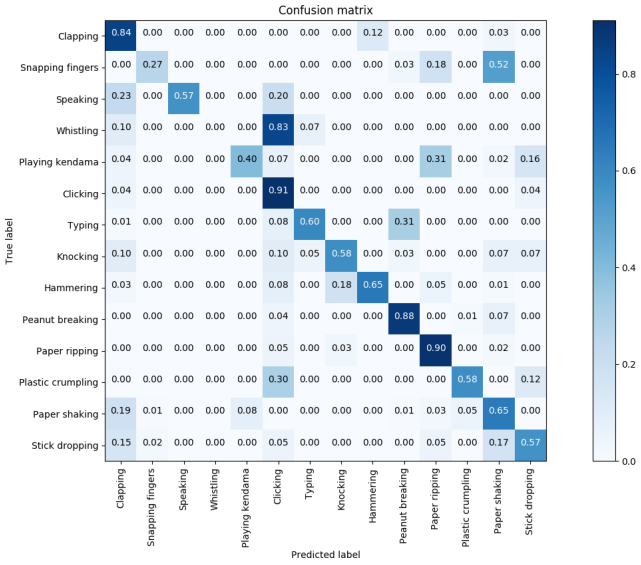


Figure 8. ResNet-50 trained on all scenarios confusion matrix.

are other objects and they can be misunderstood. This confirms the hypothesis made before in Section 4 in Figure 2 are true.

6. Reproducibility

To enable reproducibility of our results and to motivate further research on deep learning for acoustic images, our code[1], data, and models are publicly available.

[1] https://github.com/afperezm/acoustic-images-distillation

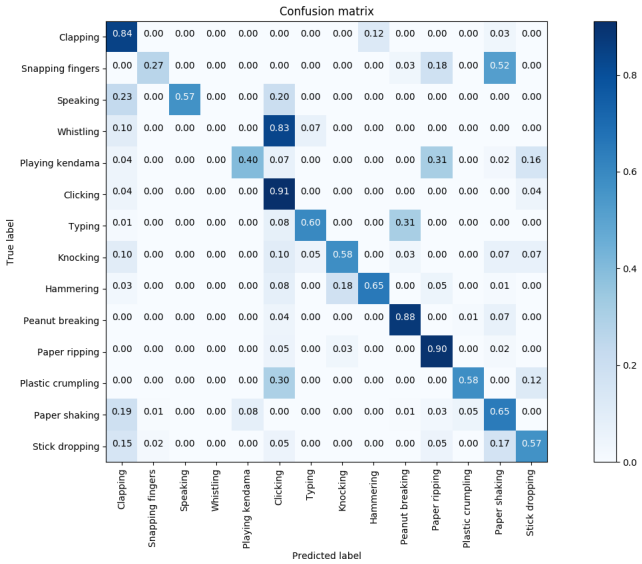


Figure 9. Temporal ResNet-50 trained on all scenarios confusion matrix.

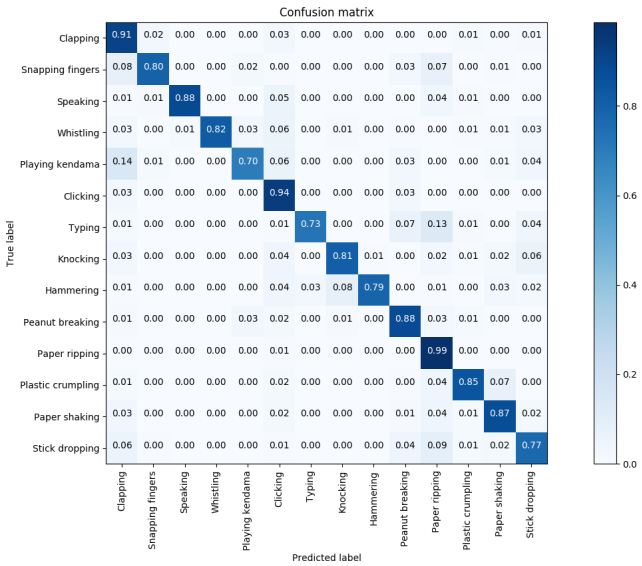


Figure 10. AVNet trained on all scenarios confusion matrix.

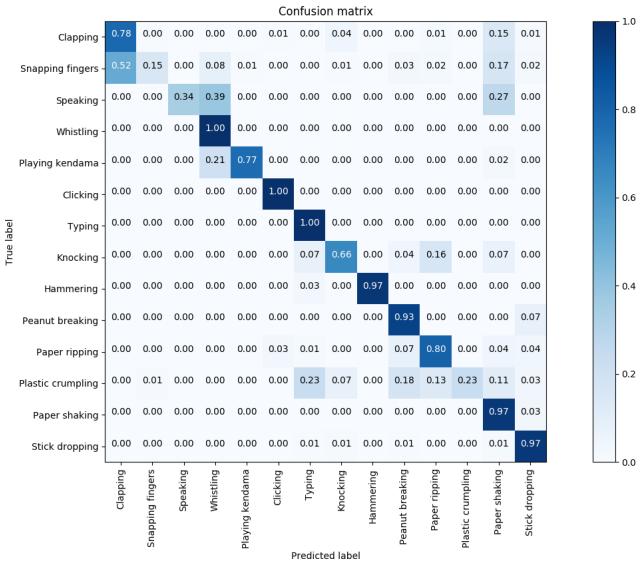


Figure 11. ResNet-50 trained and tested on scenario 1 confusion matrix.

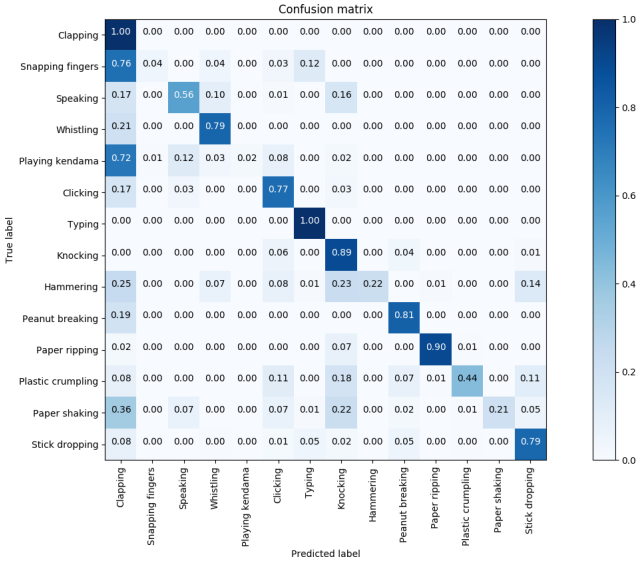


Figure 12. ResNet-50 trained and tested on scenario 2 confusion matrix.

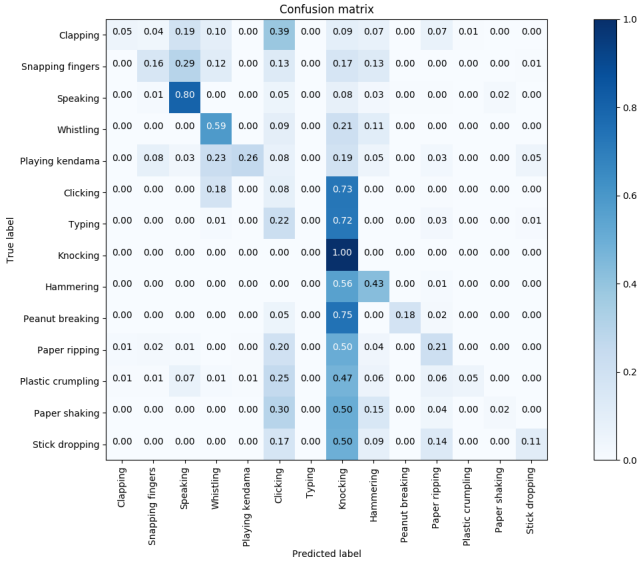


Figure 13. ResNet-50 trained and tested on scenario 3 confusion matrix.