Listening Human Behavior: 3D Human Pose Estimation with Acoustic Signals
Supplementary Materials

Paper ID: 1844

Contents

1. Overview of the Supplementary Materials

This supplementary document contains additional details and discussions of our 3D human pose estimation given low-level acoustic signals. Please also refer to the supplementary video 1844_video.mp4 for additional results. This video contains our results in cross-subject settings and in both anechoic chamber and classroom. We highlight reference numbers associated with the main paper in blue, and those associated with this supplementary document in red.

2. Additional Discussion about Acoustic Sensing

2.1. The Effect of Using Stereo Speakers

We used loudspeakers consisting of two speakers (*i.e.*, right/left speaker) that were vertically aligned toward microphones to transmit acoustic signals. This stereo transmission has several advantages: when a single speaker is used as the point sound source, the signal wave propagates along a spherical surface. In this case, as shown in Fig. 1 (a), the area of the signal wave propagation at distance r from the speaker is $4\pi r^2$, resulting in signal attenuation at distance r being proportional to $1/r^2$. On the other hand, suppose we have multiple speakers distanced apart h. In this case, the signal wave propagates along a cylindrical surface and the area of the signal wave at distance r is $2\pi rh$ (see Fig. 1 (b)). Therefore, the attenuation of the

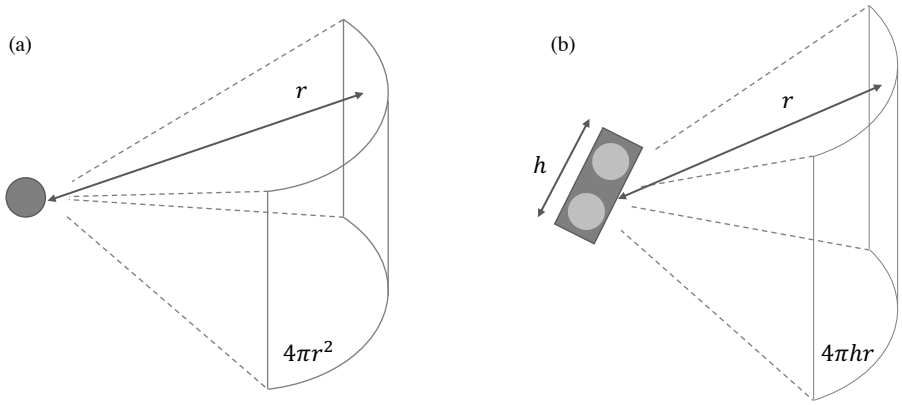

Figure 1. The area of the acoustic signal wave propagation at distance r from the speaker, with (a) a single point source sound, and (b) stereo speakers.

signal at distance r is proportional to $1/r$, much less than the use of a single speaker. This effect makes it easier to perceive the behavior of humans, standing at a distance since the acoustic signal is less attenuated and can reach farther.

2.2. Received Signal Property

The magnitude of the measured reflection relative to the emitted signal depends on the positions of loudspeakers, microphones, and a human. The transmitted time-stretched pulse (TSP) signal that we use propagates along a cylindrical wavefront from the speakers to the human, and occluded or diffracted signals are captured via microphones. Unlike RGB images in which most human body is visible unless it is occluded, at the audible audio signals frequency that we used (*i.e.*, 85 Hz $\sim$10 kHz), the human body acts as a reflector rather than a scatterer. Also, as the density of the object material increases, sound signals are reflected rather than transmitted through the object. As a result, our microphone can capture only occluded or diffracted signals. Fig. 2 explains this in more detail. When we transmit signals towards the human body as shown in Fig. 2 (a), signals that fall close to the normal to the surface are reflected back toward the loudspeakers (see Fig. 2 (b)). Signals that deviate from the normal to the surface are diffracted (see Fig. 2 (c)). Then a part of diffracted signals is absorbed into anechoic room walls and becomes invisible to our microphones, while only the rest of diffracted signals are captured. Besides, due to the longer wavelength, a part of the signals pass through the human body, although such signals are highly attenuated.

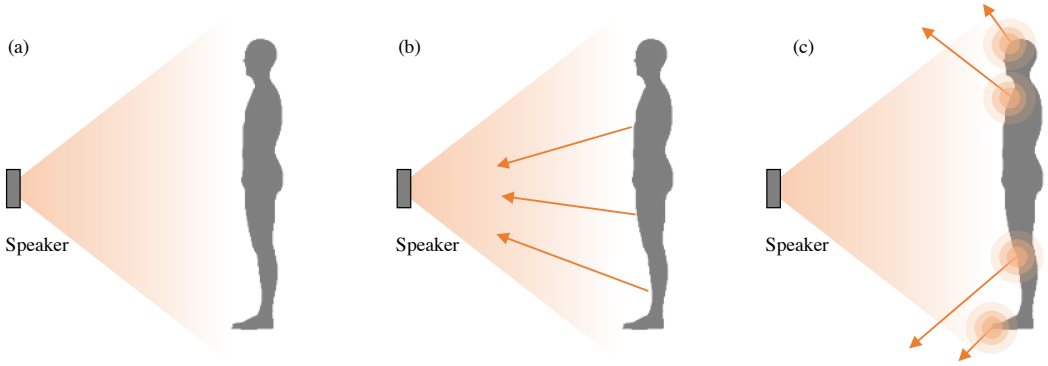

Figure 2. The property of audio signal reflection and diffraction.

3. Difference from "underwater sonar" or "geophysical sub-surface mapping"

Our task seems similar to "underwater sonar" or "geophysical sub-surface mapping" tasks. However, we would like to reiterate that our problem is rather different, as is evident from the following aspects regarding the physical nature of air and liquids/solids. First, sounds attenuate more quickly when traveling through the air than they do underwater/underground, as attenuation depends on the density of the medium sounds travel [5]. Second, unlike underwater/underground with fewer

occluders like walls, in indoor environments, a sound bounces multiple times and reaches the receiver via multiple pathways, thereby making it much more difficult or even impossible to solve analytically. These remarkable differences make air sonar much less popular in the real world than underwater/underground-based sonars. Therefore, we believe our success in predicting human poses with acoustic signals in the air indicated the possibility of solving this challenging task.

4. Differences from scene structure estimation

Additionally, our task is similar to estimating scene structure problems. We acknowledge that this is true with respect to the fact that we obtain 3D scene information by identifying the room impulse response. However, our work is rather different from the scene structure estimation method with respect to identifying **dynamic** human poses that have high time-wise correlation, unlike the work in which the scene is static and requires the sensor and target to remain stationary after emitting an active measurement signal and then wait for a response. To tackle this rather challenging task, we used techniques in windowed-based signal processing fashion — that is longer time scale inputs ($\approx 0.6s$) and processed these inputs in a windowing manner. Moreover, with regard to the network part, we leveraged time-wise CNN to capture dynamic motions.

5. Evaluation Metric

In our main paper, we evaluate our proposed model and previous work based on root mean square error (**RMSE**), mean absolute error (**MAE**), and percentage of correct key points (**PCK**). Particularly, RMSE and MAE are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{TJ}\Sigma_{t=1}^{T}\Sigma_{j=1}^{J}(x_t^j - \hat{x}_t^j)^2}, \tag{1}$$

$$\text{MAE} = \frac{1}{TJ}\Sigma_{t=1}^{T}\Sigma_{j=1}^{J}|x_t^j - \hat{x}_t^j|, \tag{2}$$

where x_t^j is the j^{th} joint position of the estimated pose, and $\hat{x}_t^j$ is the ground truth. The length of the data and the number of total joints are denoted as T and J, respectively.

6. Network Architecture

We show in Fig. 3 a more detailed version of our audio to 3D human pose network f that was described in Sec. 3.3 and Fig. 2 in the main paper. As discussed in the main paper, our model takes time-series audio features which have a size of $7 \times 128 \times 12$, where 128 indicates a number of frequency resolution bins, 7 is the number of channels and 12 is the sequence length. As shown in Fig. 3 our model has 4 2D CNN blocks, a 1D U-Net module, a subject discriminator module, and 4 1D CNN layers followed by an output layer. The output shape (12×63) means 3-dimensional positions of 21 human joints in 12 frames. Numbers in each layer in Fig. 3 indicate the number of output channels. Each 2D CNN block has 1 max pooling layer at last to implement frequency-wise downsampling. Additionally, every time-wise downsampling in 1D U-Net was done with a max pooling layer to reduce sequence length to half. We use activation function (*i.e.*, Leaky ReLU [6]) after each convolutional layer. Please refer to Fig. 3 for each layer's kernel size.

7. Qualitative analysis in single-subject setting

In addition to the cross-subject setting, here we showed our qualitative analysis in a single-subject setting in Fig 4. As Table 2 in our paper indicates, our model has close outputs to Ginosar *et al.*'s model and outperformed Jiang *et al.*'s model as the green arrow in the third line shows.

8. Predictions in no-light setting

As mentioned in Sec. 5 in our main paper, we tested our model's accuracy in a no-light setting. Table 1 shows the quantitative results from two subjects with our best model in a classroom and cross-subject setting (= Ours w/o Intensity Vector), which indicates even in the condition in which it was extremely difficult to estimate human poses with a visible light signal-based method, the proposed method successfully estimated finer 3D human poses. We also demonstrated our qualitative results in Figure 5 to show our model was working properly although we could see almost nothing in RGB images.

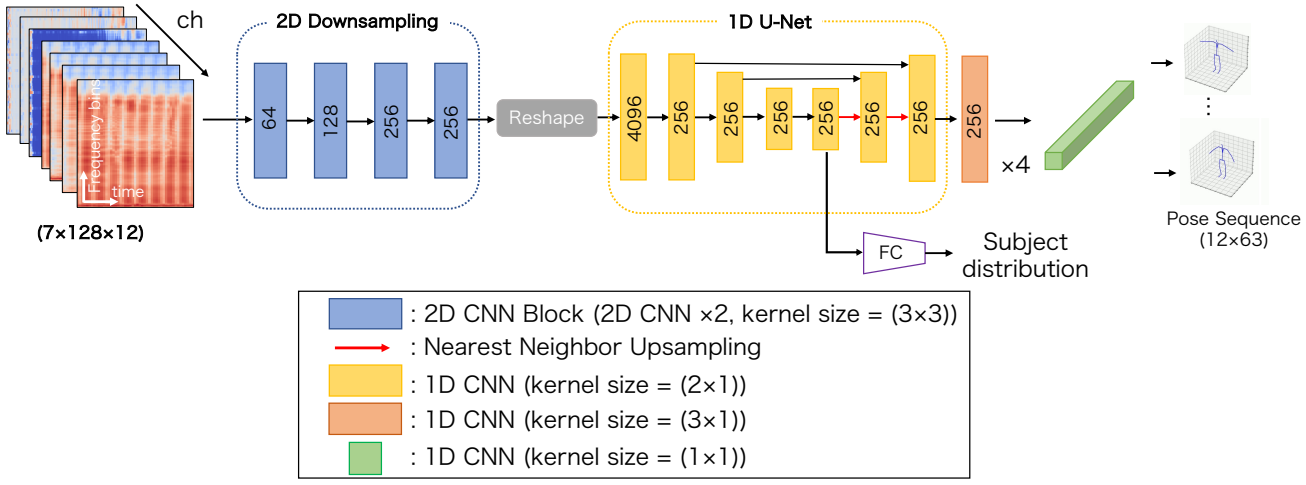

Figure 3. The network architecture.

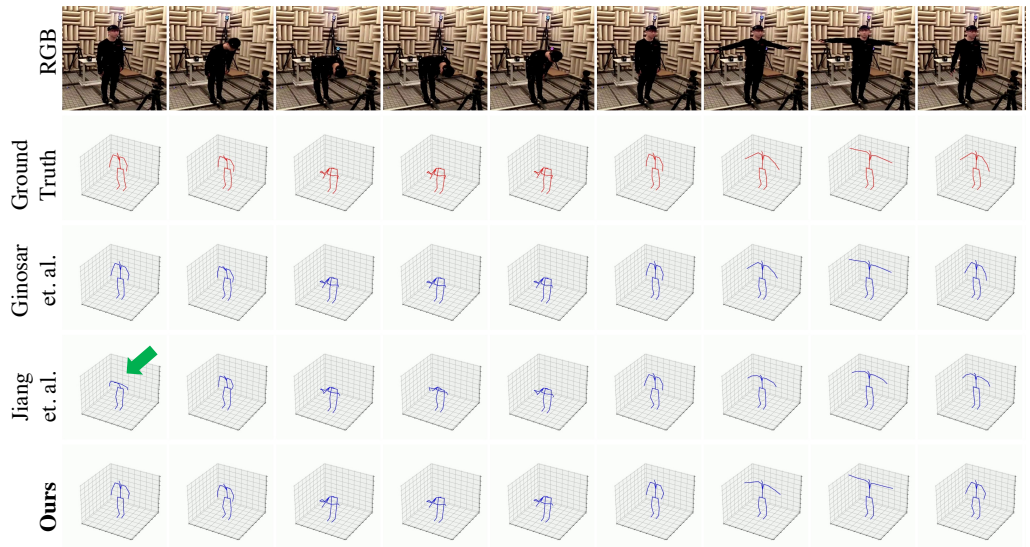

Figure 4. Qualitative results in single-subject setting

Table 1. The prediction accuracy in no light setting

Method	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Ours best	**0.91**	**0.53**	**0.72**

9. Subject Discriminator Comparison

As we did in Sec. 5 in the main maper, we also compared our subject discriminator loss L_{std} with L_d which was used in the previous work [4] in a classroom with various noise. Here, we set $w_\gamma = 0.5$ for L_{std}, and $w_\gamma = 0.1$ for L_d. The results are shown in Table 2. While these two methods had close scores on both RMSE and MAE, we can see our method outperformed the previous method on PCK@0.5, which indicates the model's prediction uncertainty is more helpful than traditional discriminator loss to force the model to use subject-invariant features only.

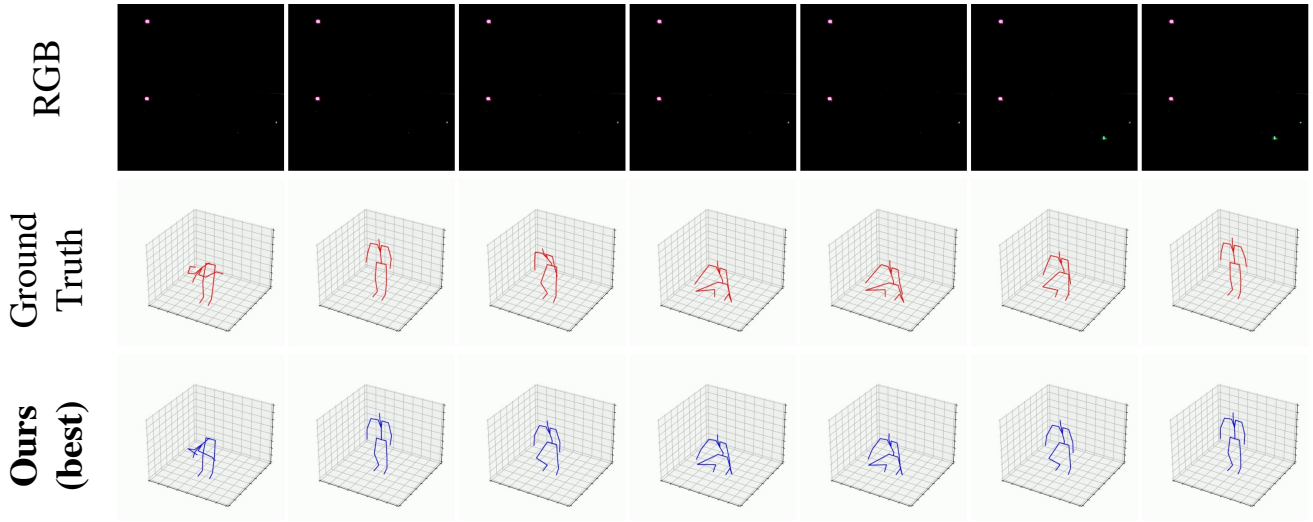

Figure 5. The results in no-light setting

Table 2. Comparison between STD and ordinal discriminator loss in a classroom

Method	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Ours (L_d [4])	**0.93**	**0.54**	0.63
Ours (L_{std})	**0.93**	0.55	**0.67**

10. Memory Size

One of the practical advantages of our approach that uses acoustic signals as input is the smaller memory size of each sample frame. To perceive human 3D poses, most approaches make use of RGB images [1, 2]. Suppose we use 20 fps of RGB video samples with (x, y)= 640×480 spatial resolution with 3 channels. In this case, 1 minute of samples will have $640 \times 480 \times 3 \times 1 \times 60 \times 20$ = 1105920000 byte ($\fallingdotseq$ 1.11 GB). Another approach [3] utilizes a lower-dimensional input called transient image. This is the 3D measurement volume containing a scene's spatio-temporal response to the laser pulse. Following [3], suppose the transient images have $(x, y, t) = 32 \times 32 \times 64$ spatio-temporal resolution with 4 fps. This case we still have $32 \times 32 \times 64 \times 1 \times 60 \times 4$ = 15728640 byte ($\fallingdotseq$ 15.73 MB). Here, our approach uses 2-dimensional multi-channel acoustic signals as input. Suppose we sample signals with 20 fps, we will cost only $128 \times 12 \times 7 \times 1 \times 60 \times 20$ = 12902400 byte ($\fallingdotseq$ 12.9 MB) of memory, that is 1.16% of RGB-based and 82% of transient-based (Please note that the transient-based method has far fewer time resolution than our model). We believe this memory-efficient approach has many practical applications for next-generation edge computing such as drone automation with human interaction, or super-fast people detection and tracking in autonomous systems.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 5

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2334–2343, 2017. 5

[3] Mariko Isogawa, Ye Yuan, Matthew O'Toole, and Kris M. Kitani. Optical non-line-of-sight physics-based 3D human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7013–7022, 2020. 5

[4] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. Towards environment independent device free human activity recognition. In *International Conference on Mobile Computing and Networking (MobiCom)*, page 289–304, 2018. 4, 5

[5] Rohan Kapoor, Subramanian Ramasamy, Alessandro Gardi, Ron Van Schyndel, and Roberto Sabatini. Acoustic sensors for air and surface navigation applications. *Sensors*, 18(2):499, 2018. 2

[6] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, page 3. Atlanta, Georgia, USA, 2013. 3