BGM2Pose: Active 3D Human Pose Estimation with Non-Stationary Sounds

Yuto Shibata
Keio University
yuto071508@keio.jp

Yusuke Oumi
Keio University
youmi@keio.jp

Go Irie
Tokyo University of Science
goirie@ieee.org

Akisato Kimura
NTT Corporation
akisato@ieee.org

Yoshimitsu Aoki
Keio University
aoki@elec.keio.ac.jp

Mariko Isogawa
Keio University, JST Presto
mariko.isogawa@keio.jp

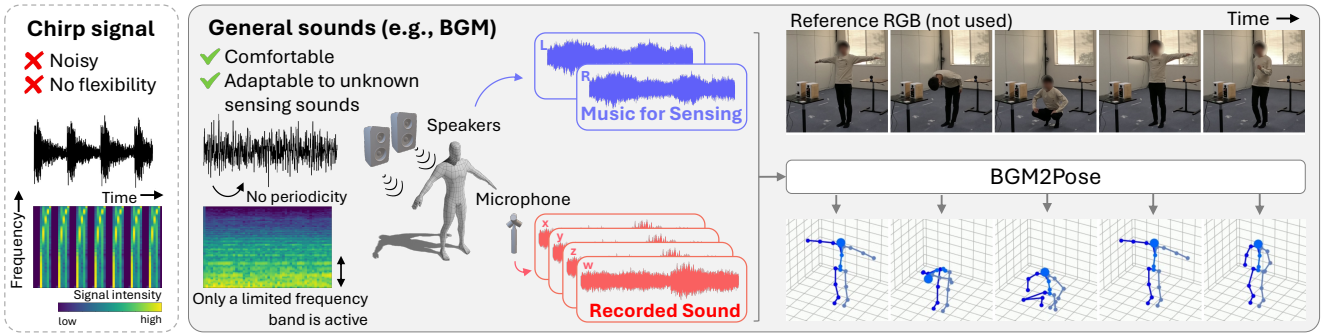

Figure 1: We propose BGM2Pose, a 3D human pose estimation that utilizes common background music.

Abstract

We propose BGM2Pose, a non-invasive 3D human pose estimation method using arbitrary music (e.g., background music) as active sensing signals. Unlike existing approaches that significantly limit practicality by employing intrusive chirp signals within the audible range, our method utilizes natural music that causes minimal discomfort to humans. Estimating human poses from standard music presents significant challenges. In contrast to sound sources specifically designed for measurement, regular music varies in both volume and pitch. These dynamic changes in signals caused by music are inevitably mixed with alterations in the sound field resulting from human motion, making it hard to extract reliable cues for pose estimation. To address these challenges, BGM2Pose introduces a Contrastive Pose Extraction Module that employs contrastive learning and hard negative sampling to eliminate musical components from the recorded data, isolating the pose information. Additionally, we propose a Frequency-wise Attention Module that enables the model to focus on subtle acoustic variations attributable to human movement by dynamically computing attention across frequency bands. Experiments suggest that our method outperforms the existing methods, demonstrating substantial potential for real-world applications.

1. Introduction

This paper addresses the task of estimating human poses using ambient background music, without relying on specialized sensing signals. While existing vision-based methods achieve strong performance, these approaches face challenges such as vulnerability to occlusion and low-light conditions, as well as privacy concerns due to the capture of identifiable features like faces.

Acoustic sensing is lighting-invariant and visual privacy-friendly. Compared to other non-visual modalities like mmWave or WiFi [3], sound is more universally applicable across environments, including those with sensitive equipment or during flights. Several methods using acoustic signals for human state estimation have been explored [1, 2]. Existing approaches utilize chirp signals where the frequency monotonically increases or decreases over time as the sensing source. However, this method faces several limitations. First, chirp signals typically span a wide frequency range, including audible frequencies, making them highly uncomfortable for human ears. Second, a general constraint in methods using chirp signals is the requirement to use the exact same sensing signal during inference as was used during training. Both conditions highly limit the method's applicability in

real-world scenarios.

Therefore, this paper addresses these challenges by proposing a task of **3D Human Pose Estimation Based on Non-Stationary Sounds** (Fig. 1), which significantly expands the framework of non-invasive estimation of dynamic human pose. In this task, we utilize standard background music (BGM) as the sensing signal. With BGM, our system significantly enhances comfort and practicality compared to a noisy, chirp-based framework.

Utilizing everyday BGM as a sensing signal introduces several critical challenges: **(i)** the phase- and amplitude shifts caused by human motion are subtle, so the much larger variations in background music readily mask these pose-related cues. **(ii)** unlike chirp signals that sweep across a wide range of frequencies, BGM has a limited frequency band that changes over time (see Fig. 1). Since pose information is observed as changes occurring within the BGM, the model needs to selectively attend to the relevant portions of the spectrogram. **(iii)** since this framework does not assume specific predefined signals such as chirp, the method needs to adapt to unseen acoustic signals for inference.

To address these challenges, we propose a novel model called BGM2Pose. To overcome the first challenge, we introduce a Contrastive Pose Extraction module (CPE module). This module effectively promotes the extraction of pose components while simultaneously excluding music components by employing a contrastive loss within a shared feature space. To address the second challenge of limited frequency bands and the third challenge of generalizing to unseen acoustic signals, we incorporate a Frequency-wise Attention Module (FA module). This module uses an attention mechanism to effectively identify the frequency bands of the spectrogram containing human pose information from the recorded signal, allowing it to extract posture-estimation-relevant information even from limited frequency bands or unseen music.

Since no existing work has addressed inferring 3D human poses with BGM, we create Acoustic Music-based Pose Learning (AMPL) dataset, a large-scale original dataset for this task. Through experiments using the proposed dataset, we demonstrate that the proposed method significantly outperforms existing methods.

2. Methodology

Our goal is to estimate the 3D human pose sequence $\{\mathbf{p}_t\}_{t=1}^{T}$ of a target subject standing between a microphone and loudspeakers, given the recorded sound sequence $\{\mathbf{s}_t\}_{t=1}^{T}$ and original music sequences $\{\mathbf{m}_{i,t}\}_{t=1}^{T}$ emitted from the i-th speakers. Here, T indicates the length of input and output sequences and t means each timestep. We assume a typical consumer speaker setup with two speakers (*i.e.*, right and left). Following [1], we also use one ambisonics microphone, which captures omnidirectional (w) and x, y, and z components with four channels in B-Format.

The overview of our BGM2Poseis presented in Fig.2. In the following sections, we detail the acoustic feature extraction, the proposed Frequency-wise Attention (FA) module, and the Contrastive Pose Extraction (CPE) module.

2.1. Acoustic Feature Extraction

To generate the sequence of the audio feature vectors as the input to our framework, we generate three types of acoustic features. From the recorded sound signals $\mathbf{s}_t$, we extract the intensity vector $\mathbf{I}_{\text{intensity}} \in \mathbb{R}^{3\times b\times T}$, including three channels of (x, y, z)-directional components, and the log-mel spectrum $\mathbf{S}_{\text{logmel}} \in \mathbb{R}^{4\times b\times T}$. Here, b denotes the number of frequency bins. Given the original BGM sound signals $\mathbf{m}_i$ emitted from the left and right speakers, we generate the log-mel spectrum $\mathbf{M}_{\text{logmel}} \in \mathbb{R}^{2\times b\times T}$. These features are normalized before concatenation. In our implementation, we use $b = 128$ and $T = 12$.

To filter out the influence of BGM from our recorded audio, we explicitly subtract the original music data $\mathbf{M}_i$, emitted from the i-th speaker, from the recorded audio data $\mathbf{S}$ after standard normalization. By concatenating the intensity vector and these difference features with two speakers, we obtain the feature $\mathbf{X}$ with $4 + 4 + 3 = 11$ channels in total, which is fed into our network model.

2.2. Frequency-wise Attention Module

To effectively extract acoustic changes caused by human posture in response to dynamic sensing signals, our method proposes a Frequency-wise Attention Module (FA module). Unlike TSP (chirp) signals designed to emit signals with constant intensity across all frequencies within a specific time frame, the frequency in BGM does not have periodicity and varies over time (see Fig. 1 (left)). To extract pose-related information captured by dynamically changing sensing signals, we compute frequency-wise attention over the input features at each time step.

The FA module takes acoustic feature $\mathbf{X}$ and the music feature $\mathbf{M}$ as input. To aggregate local information of the inputs, we independently apply 2D CNNs to both $\mathbf{X}$ and $\mathbf{M}$. Then, we obtain the recorded sound feature $\mathbf{X'}_t \in \mathbb{R}^{b\times d}$ and the music feature $\mathbf{M'}_t \in \mathbb{R}^{b\times d}$ for each time index t, where d is the latent dimension.

To capture frequency characteristics, we also add learnable shared frequency embedding $\mathbf{F} \in \mathbb{R}^{b\times d}$ to the aforementioned two features, $\hat{\mathbf{X}}_t = \mathbf{X'}_t + \mathbf{F}$ and $\hat{\mathbf{M}}_t = \mathbf{M'}_t + \mathbf{F}$. Then, the attention mechanism is calculated as follows:

$$\text{Attention}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t) = \text{softmax}(\frac{\mathbf{Q}_t \cdot \mathbf{K}_t^T}{\sqrt{d}})\mathbf{V}_t, \quad (1)$$

where $\mathbf{K}_t = \hat{\mathbf{M}}_t\mathbf{W}^K$, $\mathbf{Q}_t = \hat{\mathbf{X}}_t\mathbf{W}^Q$, and $\mathbf{V}_t = \hat{\mathbf{X}}_t\mathbf{W}^V$. Here, $\mathbf{W}^K$, $\mathbf{W}^Q$, and $\mathbf{W}^V \in \mathbb{R}^{d\times d}$ are the key, query,

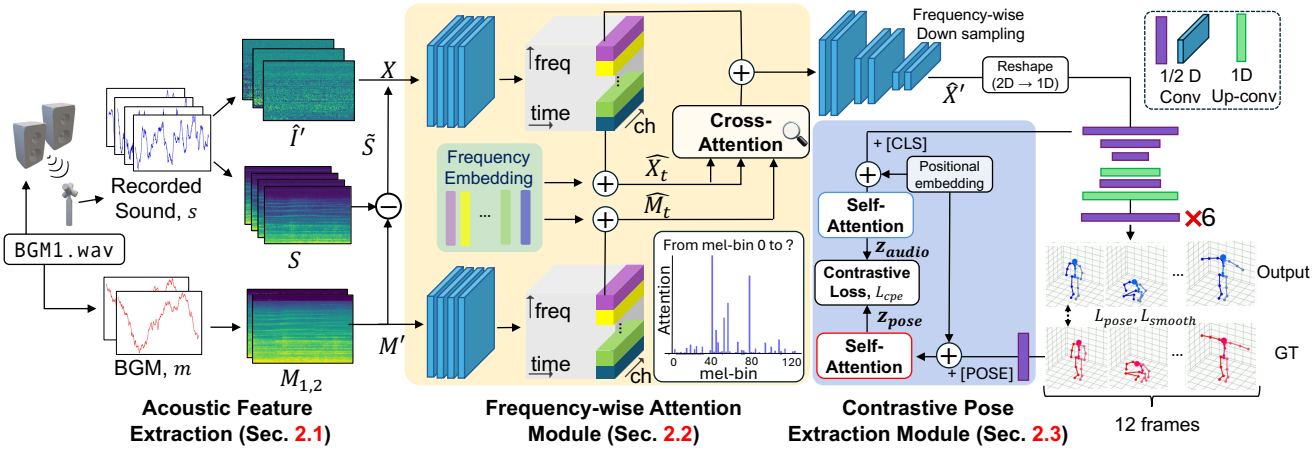

Figure 2: The overview of our framework for BGM–based 3D human pose estimation.

and value projection matrices, respectively. Following prior works [1, 4], we also apply 2D CNNs, a reshaping layer, and 1D CNNs to output pose $\mathbf{p}$ with a size of 12×63, which represents 3D $\times 21$ joints for the 12 frames.

2.3. Contrastive Pose Extraction Module

To mitigate the influence of sensing music on the model's representations and to enable the extraction of pose information only, we propose a Contrastive Pose Extraction module (CPE module) based on contrastive learning. This module is designed to incur a high loss when the model's output is influenced by the sensing BGM. This module performs multi-modal contrastive learning that maps human poses and recorded audio into a common feature space. When samples recorded with similar sensing music are included in the same mini-batch, the learning process encourages separating them, thereby promoting the extraction of only pose information. Additionally, we propose a novel sampling method called "BGM-based hard negative sampling." In this approach, similar recordings (and corresponding pose data) obtained using the same sensing BGM are included in the same mini-batch, imposing a more challenging setting for contrastive learning to further enhance the model's discriminative ability.

3. Human Pose Dataset with Music

For our task, we created a large-scale dataset, **AMPL** (**A**coustic **M**usic-based **P**ose **L**earning dataset), which links musical sensing signals with 3D human poses. These data are captured in a classroom environment using motion capture (Mocap) and 16 cameras. This dataset consists of approximately 1.4 million pose-annotated frames, collected from eight subjects and four BGM tracks (three ambient and one jazz music). Additionally, for each of the music tracks, casual clothing data from two individuals without pose annotations is also included to evaluate system practicality.

Table 1: Quantitative experimental results in the (a) single-music and the (b) cross-music settings.

Method	(a) Single-Music			(b) Cross-Music		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Jiang *et al.* [3]	1.338	0.768	0.251	1.417	0.800	0.272
Ginosar *et al.* [4]	1.223	0.666	0.379	1.274	0.682	0.375
Shibata *et al.* [1]	1.090	0.574	0.468	1.110	0.556	0.499
Ours	**0.923**	**0.453**	**0.573**	**1.036**	**0.494**	**0.570**

Table 2: Quantitative experimental results with (c) cross-genre setting (train=ambient, test=jazz).

Method	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Jiang *et al.* [3]	1.418	0.849	0.221
Ginosar *et al.* [4]	1.326	0.727	0.360
Shibata *et al.* [1]	1.112	0.595	0.376
Ours	**1.065**	**0.543**	**0.463**

4. Experimental Settings

Baselines To our knowledge, no prior work has performed 3D human pose estimation using both music and recorded data as inputs. We therefore compare our model to the following related approaches: (i) **Jiang *et al.* [3]** uses WiFi signals and, like us, processes low-dimensional actively sensed input; (ii) **Ginosar *et al.* [4]** uses acoustic signals, but for gesture generation from speech-like audio; (iii) **Shibata *et al.* [1]** is most relevant, using acoustic signals for 3D pose estimation, but with chirp-based active sensing. Please note that we trained these baselines on AMPL dataset with matched I/O layers for fair comparison.

Evaluation Metrics. Our quantitative metrics are as follows: root mean square error (RMSE), mean absolute error (MAE), and percentage of correct key point with a threshold of 50% for the head–neck bone link (PCKh@0.5).

Table 3: Ablation study in the setting (a), (b).

Method	(a) Single Music			(b) Cross Music		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Ours	**0.923**	**0.453**	**0.573**	1.036	**0.494**	**0.570**
w/o CPE module	0.945	0.481	0.531	**1.024**	0.501	0.547
w/o BGM Hard Negative	0.928	0.457	0.566	1.065	0.506	0.559
w/o FA module	1.025	0.509	0.536	1.210	0.593	0.512
w/o BGM conditioning	0.970	0.484	0.537	1.075	0.522	0.543

5. Experimental Results

5.1. Comparison with Other Baselines

We evaluated our proposed method against the baseline models within the following three scenarios: (a) a single-music setting, wherein the same ambient BGM was used as the acoustic source for both training and testing; (b) a cross-music setting, in which two ambient and one jazz BGM were used for training and the remaining ambient music was used for testing only; and (c) a cross-genre setting, in which all ambient tracks were used for training and one jazz BGM was used for tesing. These accuracies were obtained by computing the average performance on unseen subjects using K-fold cross-validation and 3-seed averaging.

Table 1 summarizes the quantitative results in both the single- and cross-music settings. We can see that our proposed method outperformed the previous models in all metrics under both settings, highlighting that it has a strong capacity to capture 3D human poses based on music sounds. Note that the values in the table are normalized, with the hip-spine distance set to 1. Table 2 shows the results when three ambient music tracks are used for training while jazz music is used for testing. Jazz exhibits more drastic changes in pitch and volume compared to ambient music, and it includes silent intervals during which the sensing signal does not work. Consequently, we observe a decrease in accuracy in PCKh@0.5 compared to ambient music evaluation (Table 1). However, even so, our method records significantly higher accuracy than existing methods, suggesting that our approach is robust across different music genres.

5.2. Ablative Analysis

We investigated the effects of our main technical contributions, *i.e.*, the CPE module and the FA module. For the CPE module, we also prepared a setting in which mini-batches are created randomly without using proposed hard negative sampling. Furthermore, to demonstrate the effectiveness of using playback music information, we conducted evaluations while removing the playback BGM channels from the inputs and calculated FA module based on self-attention. Table 3 shows that both our proposed modules contributed to improving the estimation accuracy.

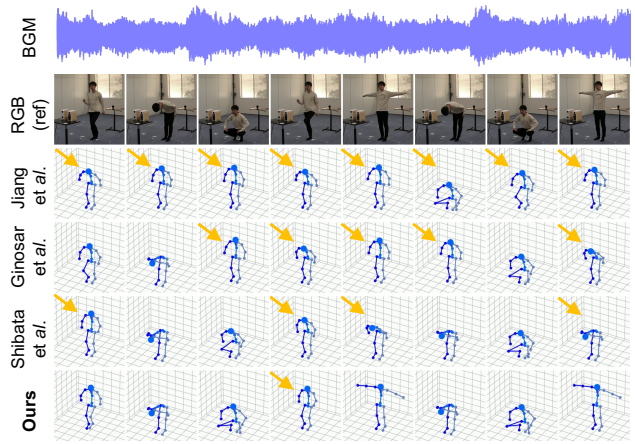

Figure 3: The qualitative results in the cross-music setting with the subjects wearing plain clothes.

5.3. Evaluation with the In Plain Clothes Dataset

Fig. 3 shows the qualitative results under the cross-music setting with the subject wearing plain clothes. As the yellow arrows highlight, we can see that our method significantly reduced false predictions, compared to the baselines.

6. Conclusion

This paper proposes BGM2Pose, a human 3D pose estimation that uses BGM as active sensing signals. Unlike existing methods that utilize chirp signals which is uncomfortable for humans, our approach uses everyday music, offering greater flexibility and practicality. This task is challenging because the amplitude and pitch of the signals vary over time, and the human pose clues in recordings are masked by music changes. Moreover, the effective frequency range of BGM is limited. Therefore, we proposed a novel model incorporating the Frequency-wise Attention Module and the Contrastive Pose Extraction Module to focus on the changes in the measurement sounds caused by human pose variations at each moment. Our method achieves accurate 3D pose estimation under diverse conditions, including unseen music and subjects in casual clothing, highlighting the potential of more practical sound-based human sensing.

Acknowledgements. This work was partially supported by JST Presto JPMJPR22C1, JSPS KAKENHI Grant Number 24K22296, Keio-BOOST, and Keio University Academic Development Funds.

References

[1] Y.Shibata et al., "Listening Human Behavior: 3D Human Pose Estimation With Acoustic Signals", in CVPR, pp.13323–13332, 2023. 1, 2, 3

[2] Z.Yang et al., "PoseKernelLifter: Metric Lifting of 3D Human Pose using Sound", in CVPR, pp. 13179–13189, 2022 1

[3] W.Jiang et al., "Towards 3D human pose construction using WiFi", in MobiCom, pp.1–14, 2020. 1, 3

[4] S.Ginosar et al., "Learning Individual Styles of Conversational Gesture", in CVPR, pp. 3497–3506, 2019. 3