EXPLORING METHODS FOR MEASURING AUDIENCE BODY MOTION IN CONCERT HALLS

Hugh A. VON ARNIM (hughav@uio.no) (0009-0005-7511-9019)[1,2], **Finn UPHAM** (finnu@uio.no) (0000-0002-9948-2024)[1,2], and **Alexander Refsum JENSENIUS** (alexanje@uio.no) (0000-0001-6171-8743)[1,2]

[1]*RITMO Centre for Interdisciplinary Studies in Rhythm, Time, and Motion*, **University of Oslo**, Oslo, Norway
[2]*Department of Musicology*, **University of Oslo**, Oslo, Norway

ABSTRACT

Measurement of audience members' body motion in ecological concert settings must navigate environmental constraints, access to audience members, privacy considerations, and behavioural norms of the performance event. This study compares five distinct technologies for capturing motion from "audience member" participants under simulated symphonic concert conditions: visible-range imaging (RGB), near-infrared imaging (IR), thermal imaging, thermal data, and torso-mounted accelerometry. Seven quantity of motion (QoM) time series were generated from these measurements, each cleaned and normalised to reveal similarities and differences in the information they provide regarding four concert-typical behaviours: Large Actions such as posture changes, Applause, Stillness, and Respiration. Results show comparable, though not identical, quantity of motion measurements across devices in this simulated concert setting. We discuss the advantages and disadvantages of the various signals, share recommendations on key measurement parameters, and discuss scaling these motion capture methods for larger audiences in real concert halls.

1. INTRODUCTION

In Western classical concerts, when and how audience members move during the performance has been shown to reflect moment-to-moment musical understanding and experience [1, 2], as well as correlate with reported engagement, connectedness, enjoyment, and mutual awareness [3, 4]. As part of our ongoing work in live concert research [5], we have collaborated with several Northern European symphonic orchestras, testing new technologies and strategies for collecting data on audience members' body motion. This allows for the empirical study of both music performance and perception in real-world settings, further exploration of theories of embodied music cognition [6], and *musicking* [7].

This paper compares methods for measuring audience motion typical of the Western classical symphonic concert, where strict behavioural norms demand that audience members remain unobtrusive during the performance [1, 8]. In such events, direct access to audience members for sensor distribution can also be challenging due to the size of the audience and the intrusiveness of some devices, making distance-based motion tracking substantially more practical.



Figure 1. The simulated concert hall environment as shown by thermal imaging.

Data collection is much more challenging in ecological settings—such as real-world concert venues—than in controlled laboratory settings. This typically rules out the use of state-of-the-art marker- or sensor-based motion capture systems and instead relies on camera-based systems [9]. However, regular cameras often struggle with low or changing lighting conditions in concerts.

In this study, we compare and explore strategies for capturing and analysing audience motion in a simulated symphonic concert hall environment (Fig. 1) by calculating the quantity of motion (QoM) using five measurement types: visible-range imaging (RGB), near-infrared imaging (IR), thermal imaging, thermal data (temperature measurements), and torso-mounted accelerometry. [1] We aim to answer the following research questions:

RQ1 How do different motion capture strategies make different aspects of audience motion available?

RQ2 What are the advantages, disadvantages, and contingencies of each motion capture strategy?

RQ3 How can these motion capture strategies be scaled to measure larger audiences in ecological settings?

[1] Supplementary materials: https://doi.org/10.5281/zenodo.21373858

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2. BACKGROUND AND RELATED WORK

Studying audience experience during live performance increasingly involves measuring body motion, although integrating sensors while maintaining ecological validity remains challenging. Methods include individual measurements via body-worn accelerometers [2] or mobile phone inertial measurement units (IMU) [3, 4], as well as group-level imaging such as IR [10]. Furthermore, commercially available thermographic sensors now offer spatial and temporal resolutions that enable investigations into thermal imaging for audience tracking.

While wearable sensors can provide direct motion data from individuals, they are highly intrusive and impractical to scale for an entire concert audience. Measurements also tend towards a local description of the motion of the body part to which they are attached. Moreover, measurement at the scale of an audience will yield a large number of time series that need to be synchronised and aligned.

For imaging sensors, the measured spectral band and the radiation source are key factors. RGB and IR cameras measure *reflected* radiation and therefore require active illumination of the scene. Conversely, thermal cameras measure *emitted* mid- and long-wave infrared radiation and can therefore bypass illumination issues. However, the microbolometer arrays standard in thermal cameras possess lower spatial and temporal resolutions than RGB and IR sensor arrays. Despite thermal imaging's growing use in biological motion analysis (e.g. [11]), rapid motion can cause visual "smearing" due to the microbolometer's thermal time constant—its response time to changes in thermal radiation intensity [12, p.112]. Commercially available thermal cameras are often intended for security and building inspection rather than for millisecond-level motion measurement.

Thermal images are generated through a two-step mapping process. First, radiation intensity is mapped to a high-precision temperature range based on camera calibration and assumptions about surface emissivity. Second, these temperatures are mapped to an 8-bit image using either fixed or frame-adaptive temperature ranges. Consequently, while diverse sensors may all provide information about the audience's motion, it is not a given that this information is easily comparable across sensor types.

2.1 Quantity of Motion in Music Research

While human motion can be quantified with great specificity, generic motion descriptors are much more practical for handling hour-long recordings of crowds in uncontrolled conditions. A popular descriptor in music-related research is the *quantity of motion* (QoM) over time. QoM has been derived from diverse measurement systems (e.g., cameras, accelerometers, inertial measurement units, optical motion capture); however, its definition, formalisation, and relationship to the physical displacement of mass vary widely across sensor apparatus.

Camurri and colleagues introduced the concept of QoM as a feature computed from whole-body videos of dancers (e.g., [13]). They tracked the alternation between active gestures and pause phases by summing the number of pixels that changed from one frame to the next. Jensenius and colleagues introduced an alternative frame-differencing calculation, summing the magnitudes of pixel-wise changes (Eq. 1) (e.g., [9]). This can be formalised as:

$$QoM(t) = \sum_{x=1}^{W} \sum_{y=1}^{H} |I(x, y, t) - I(x, y, t-1)| \qquad (1)$$

where t is the frame index, I is the greyscale pixel intensity, W is the frame width, and H is the frame height. This definition of quantity of motion is used for all imaging data types in this study.

Accelerometry on objects with a stable mass supports relatively direct measurements of physical displacement. Accelerometer-based QoM has previously been defined as the instantaneous jerk magnitude [4]. This can be formalised as:

$$QoM(t) = \sqrt{\sum_{k \in \{x,y,z\}} (k(t) - k(t-1))^2} \qquad (2)$$

where $x(t)$, $y(t)$, and $z(t)$ are the tri-axial instantaneous acceleration measurements at constant interval discrete time step t.

For the purposes of comparison, we also consider the instantaneous acceleration magnitude from the same accelerometry measurements:

$$QoM(t) = \sqrt{\sum_{k \in \{x,y,z\}} k(t)^2} \qquad (3)$$

All of these definitions of QoM assume a constant, equal sampling interval; the sample rates between measurements differ across devices, and some are less constant than expected. An essential stage of data collection is assessing and compensating for temporal inconsistencies in measurements.

While these can all be called "quantity of motion", it is not expected that each form of the measure quantifies audience motion in the same way. However, the contours of these time series indicate the types of information that are made visible by certain sensors and signals. It is principally the differences in information that will be assessed for relevance to the specific goals of the audience motion analysis, rather than absolute numerical values.

2.2 Video Encoding and Quantity of Motion

The high data rate of video makes storage at the spatial and temporal resolutions used in motion analysis impractical, so image sequences are almost always compressed with an encoder/decoder pairing (a *codec*). Because the codec–QoM relationship remains understudied, we briefly outline the *H.264* and successor *H.265* codecs used across our imaging data (following Richardson [14, 15]), which inform our subsequent methodological, analytical, and interpretive decisions.

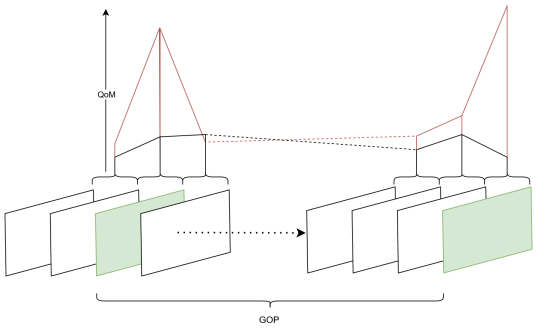


Figure 2. A simplified representation of frame-wise QoM calculated from encoded video (*H.264*/*H.265*) showing divergence across a Group of Pictures (GOP) caused by propagated distortion. The encoded QoM (red line) slowly diverges from the uncompressed QoM (black line) across the GOP, while preserving the macroscale contour. This divergence resets with a large spike at the transition to the subsequent GOP's I-frame (green frames).

Both codecs reduce data size through block-based spatial (intra-) and temporal (inter-) prediction. The residual between a predicted and an actual frame undergoes a block-based integer transform—approximating the Discrete Cosine Transform (DCT)—before the coefficients are quantised and entropy-encoded. Two aspects are important for video-based QoM. First, quantisation is governed by a variable quantisation parameter (QP) that varies both block- and frame-wise. As higher-spatial-frequency components have lower magnitudes, higher QPs more readily round them to zero, removing fine detail and introducing distortion that accumulates through intra-predicted frames. Second, the encoder periodically inserts keyframes (I-frames) relying purely on intra-prediction, resetting the prediction cycle and defining a Group of Pictures (GOP) between successive I-frames. As distortion accumulates across a GOP, decoded frames increasingly diverge from their uncompressed counterparts. When calculating QoM as in Eq. 1, this can produce a spike or discontinuity at the I-frame transition beginning the next GOP when higher QPs are used (Fig. 2).

3. METHODS

3.1 Conditions

Participants viewed a recording of Harald Sæverud's *Kjempeviseslåten* performed by the Aarhus Symphony Orchestra during one of our on-site data collections. The video was recorded through the house AV system during a research concert at the Aarhus Concert Hall. The video (*H.264*, 1920×1080, $25\,\mathrm{Hz}$, *yuv420p*) was recorded from the rear second balcony with embedded stereo audio (*AAC*, $48\,\mathrm{kHz}$) from suspended stage microphones. The stimulus was trimmed to span from pre- to post-performance applause, prepended with $30\,\mathrm{s}$ of silent black frames to allow participants time to prepare.

The study was conducted in a multipurpose hall at our institution ($17.5\,\mathrm{m} \times 6.1\,\mathrm{m} \times 5\,\mathrm{m}$, spanning two stories) featuring an external, second-story control room with an observation window overlooking the hall. To simulate a rear-centre, second-balcony view of the concert hall, a row of eight chairs was placed at the rear of the hall, facing a television approximately $6\,\mathrm{m}$ away that played the stimulus at a resolution of 3840×2160 at $60\,\mathrm{Hz}$.

Concert lighting conditions were replicated by closing the shutters on the slit windows and adjusting the three-section ceiling lights, which have five discrete intensity levels. The front section behind the television was set to 25% to simulate stage illumination, while the middle and rear sections over the chairs were set to 0%. This resulted in a measured light intensity of $0\,\mathrm{lm}$ at the centre of the row of chairs.

Six participants were recruited from researchers at our institution, including authors of this paper. Participants were invited to choose a chair. When viewed from the television's position, the chairs in the row were (from left to right) occupied: one to five, empty, seven, and empty. This simulates a row of audience members at symphonic concerts, where several may be grouped together and others separated by empty chairs. Participants were informed of the study's aims and instructed to behave as they would during a symphonic performance in a concert hall. This includes behaviours such as applause before and after the performance (cued by the concert audience's applause in the video) and sitting still during the performance.

3.2 Measures

QoM time series were generated from five data types collected using four device types. A data type is duplicated, and two QoM features were calculated from the accelerometers, thus resulting in seven QoM features in total. An overview of devices and data types is provided in Table 1, and example frames from the imaging data types are shown in Fig. 3.

To replicate the viewing angles and distances used in previous symphonic research concerts, all cameras were elevated by $9.6°$ relative to the chair row. The two Panasonic cameras were placed in the control room overlooking the hall, $17.4\,\mathrm{m}$ from the chairs. As glass is opaque to thermal infrared radiation, the thermal cameras could not be placed in the control room. Instead, the Box was mounted on a stand in front of the control room ($12.8\,\mathrm{m}$ from the chairs), and the Pocket was placed on top of the television ($6.14\,\mathrm{m}$ from the chairs). An infrared light source, positioned beside chair one and $5.9\,\mathrm{m}$ from chair four, was centred to illuminate chair four (Fig. 3; a full diagram of positioning can be found in the supplementary materials).

Prior to the participants' arrival, an activated AX3, mounted on a chest-worn elastic strap (typically used for ECG recording), was placed on each chair. Participants put on the strap after choosing a chair, and a researcher checked its fit.

All cameras record video data internally. Thermal data was recorded using a custom Real-Time Streaming Protocol (RTSP) Python logger. [2]

[2] https://github.com/Hughav92/HIKVISION_RTSP_Thermal_Data

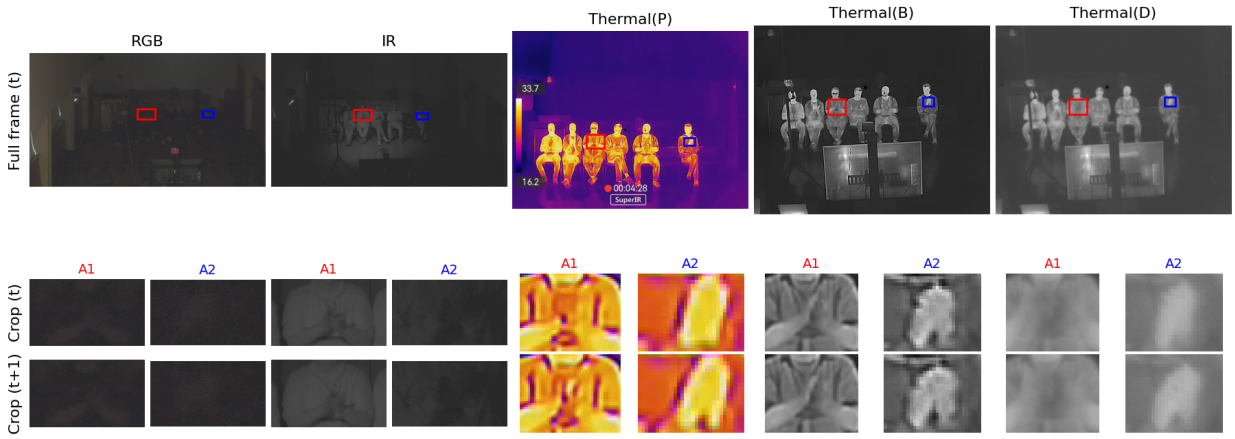


Figure 3. Applause detail across imaging types. Row 1 shows the full frame at 73.40 s, with highlighted areas (A1, A2) in Row 2 and the subsequent frame in Row 3. Thermal(B) and Thermal(P) show "smearing" in high-motion areas (A1) but less in A2, while DDE filtering (see Section 5) visibly enhances edge contrast in Thermal(B) and Thermal(P). Furthermore, low IR illumination limits the visibility of details. RGB and IR contrast is artificially enhanced for display.

Device	Data Type	QoM Signal Name	Spatial Resolution (Pixels)	Sample Rate (Hz)	GOP Size (Frames)	Codec	Bit Depth
Panasonic *HC-X2000*	RGB	RGB	3840×2160	59.97	30	*H.265*	8
Panasonic *HC-X2000*	IR	IR	3840×2160	59.97	30	*H.265*	8
HIKMICRO Pocket Thermography Camera *HM-TP42-3AQF/W-Pocket2*	Thermal Image	Thermal(P)	640×480	25*	50	*H.264*	8
HIKMICRO Thermographic Network Box Camera *HM-TD2067T-25/X*	Thermal Image Thermal Data	Thermal(B) Thermal(D)	640×512 640×512	50* 4*	250 *N/A*	*H.264* *N/A*	8 32
Axivity AX3 (worn on left-back ribs)	Tri-Axial Acceleration	|ACC| |Jerk|	*N/A* *N/A*	400* 10	*N/A* *N/A*	*N/A* *N/A*	10 10

* As reported in device descriptions/metadata. We experienced deviations from these reported figures. See Section 3.3.

Table 1. Overview of Measurements Collected.

3.3 Trimming and Alignment

Data types were trimmed and aligned using physical markers within the data. For imaging, a clapperboard marked the start and end of the simulation. The trim point was the exact frame following the clapper's closure. To align the acceleration data, a sequence of taps on the accelerometers was performed before and after the experiment, along with an audio recording (*Zoom H6* at 48 kHz) that captured the clapper's closure [16]. The audio recording served as a bridge to align the acceleration data with the visible claps in the imaging data types.

As trimming video without re-encoding is only precise to I-frames, and careless lossy re-encoding can introduce cascading distortion across GOPs, we developed a pipeline to obtain frame-precise trimming without distorting the derived QoM time series. [3] All frames were decoded into lossless *PNG* images, then losslessly re-encoded in *H.264* (with a static QP = 0), maintaining the original I-frame intervals.

Furthermore, probing the Presentation TimeStamps (PTS) of the Thermal(P) and Thermal(B) data reveals significant device clock jitter (available in the supplementary materials) accumulating to drift, despite specification and metadata reporting a fixed sampling period. To preserve timestamps when deriving QoM, the PTS of the re-encoded videos were shifted to match the original files. Although much more computationally expensive and storage-intensive than standard trimming (e.g., with *FFMPEG*), this lossless pipeline was considered necessary given the understudied relationship between re-encoding and QoM.

3.4 Quantity of Motion

An overview of QoM measures derived from the trimmed and aligned signals is shown in Table 1. For each imaging data type, this was a single measure, calculated as the sum of the frame-wise differences in pixel values (Eq. 1). This was taken for each participant (labelled by their chair position, AU1–AU7, from left to right) using a manually defined mask applied to the video frame to extract the area of interest. The QoM time series for each participant was normalised by the pixel size of their respective area of interest. From the acceleration data, two QoM time series were derived for each participant: the instantaneous acceleration magnitude (Eq. 3) and the instantaneous jerk magnitude (Eq. 2), both computed from a downsampled signal.

3.5 Signal Cleaning and Normalisation

Each measurement type was assessed for systemic noise artefacts and normalised to show comparable motion sig-

[3] https://github.com/Hughav92/Motion_Comparison_QoM_Imaging_Pipeline

nals across participants, as shown in Fig. 4. In the QoM generated from video recordings, I-frames produced regular or semi-regular peaks that obscured the motion signal. For IR and RGB, the I-frames are positioned at regular intervals (30 Frames), allowing their impact to be estimated from the median frame value sequence of these low-motion measurements and subtracted from the QoM time series. Thermal(B) and Thermal(P) both had I-frame peaks; however, frame loss and timing discrepancies made a median estimate impractical. The simplest method to compensate for these irregular I-frames was a moving median followed by a moving average, minimally viable at 5 sample points (5 Hz filtering) on Thermal(P) and 3 sample points (17 Hz filtering) on Thermal(B). Implementation details are available in the processing and analysis repository. [4]

For comparison to the slower measurements, the accelerometry measurements were filtered per axis (25 Hz low-pass zero-phase Butterworth) before calculating the magnitude for |ACC|. For |jerk| each axis of the acceleration data was filtered and downsampled to 10 Hz before differencing, as in Eq. 2, following prior practice [4]. Single dimensions of the tri-axial recordings were also decomposed using masked shifting with Empirical Mode Decomposition to extract a component resembling respiration [17] (see the supplementary material for details).

Normalisation between participants' measurements in a given QoM signal was performed using the range rather than the mean and parametric variance. Most signals were divided by the 99th or 99.5th percentile of the signal time series. This compensated for differences in motion intensity and visibility between participants, which is important for a validation exercise.

4. RESULTS

An overview of the QoM time series per participant and measurement type is shown in Figure 4. These measurements show an interval of applause, participants' sporadic movements during the music presentation, and a second round of applause, as visualised in the accompanying room audio spectrogram.

Before cleaning and normalisation, the imaging QoM clearly reflects the differences between measurements of emitted and reflected radiation. Notably, RGB provides little clear motion information, making it difficult to distinguish even applause from stillness. Furthermore, the encoded video formats (RGB, IR, Thermal(P), Thermal(B)) exhibit prominent I-frame spikes. While the thermal cameras' clock jitter and downsampling for plotting obscure many of the I-frame spikes in the Thermal(P) and Thermal(B) visualisations, the full time series found in the supplementary material confirms their presence throughout. These spikes are absent in Thermal(D), though a single temperature recalibration spike appears near 116 s.

After cleaning and normalisation, visual inspection reveals further facets. While RGB now shows the suggestion of applause, low illumination in the visible-light range resulted in faster, smaller movements due to increased noise. Conversely, IR, thermal, and jerk demonstrate substantially parallel descriptions of participants' motion. However, uneven IR illumination (Fig. 3) resulted in substantial differences in the QoM time series generated for participants on the edges of coverage: AU1 shows more noise during the music, and AU7 is barely present (last row). While the thermal measurements are robust to lighting issues, all three have additional artefacts: Thermal(B) and Thermal(D) have a distinct thermal sensor recalibration spike around 118 s, and Thermal(P) shows a substantial recording gap at 403 s during the second applause, when this camera momentarily paused. Such irregularities that span more than a frame risk missing information about the scene being recorded. The magnitude of acceleration did not substantially improve with normalisation because some of these sensor units appear to have become uncalibrated, resulting in gravity values that vary with sensor orientation.

4.1 Components of Audience Motion

Four analyses explored the quality of information for specific expected types of movement by a classical audience: Large Actions such as changes in seated posture, Applause, Stillness, and Respiration.

To compare these signals' effectiveness at capturing the overall profile of audience motion, the normalised QoM signals were downsampled to 10 Hz and cut to include the two intervals of applause. As we are evaluating the systems and not the behaviour, participant measurements were concatenated before calculating pairwise Pearson correlation coefficients. This concatenation directly compares what each signal captured of each participant at each moment, while the Pearson correlation assesses the shared information between these signals while retaining some proportionality. The pairwise results are shown in Fig. 5, including the relatively high correlations between |jerk|, Thermal(B) and (P), and IR.

As audience stillness is also of musical interest, a similar analysis was performed to assess whether these signals agree on when audience members are particularly still during music presentation (see supplementary materials for further description and plots). Each set of signal values was cut to the interval of music presentation and smoothed and downsampled to 1 Hz in imitation of the analysis of the stilling response. These time series were clipped to their lowest third before being concatenated for correlation. Given the heavy skew of the time series, the shared timing of relatively low motion values was assessed using Spearman correlations of rank, and pairwise values are reported in the bottom plot of Fig. 5. These values indicate that the measures did not consistently agree on the timing of low-motion periods.

The applause measurements were analysed using the highest available frequency per signal, as shown in Table 1, while true clapping rates were assessed by watching the thermal video. Fig. 6 shows the measurements by a participant clapping at 115 BPM before slowing. Rather than a simple oscillation, clapping motion produces an uneven double cycle of hands moving slowly apart and quickly together, and the power spectra of most QoM measurements

[4] https://github.com/finn42/motion-comparison-signal-prep

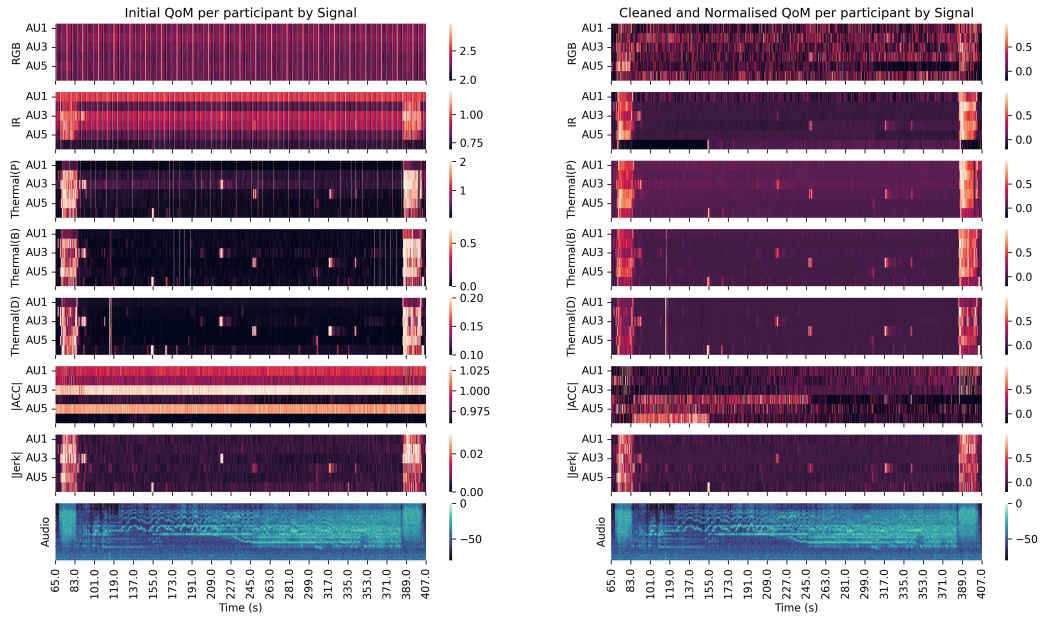


Figure 4. All Quantity of Motion time series for the six participants, as initially evaluated (left) and after cleaning and normalisation (right). Last plot in each column shows a constant Q spectrogram of the room audio from the Zoom H6 recorder, including applause in the stimulus video and by participants between 65.5 s and 82.9 s and 388.9 s and 406.9 s, and the music performed in between.

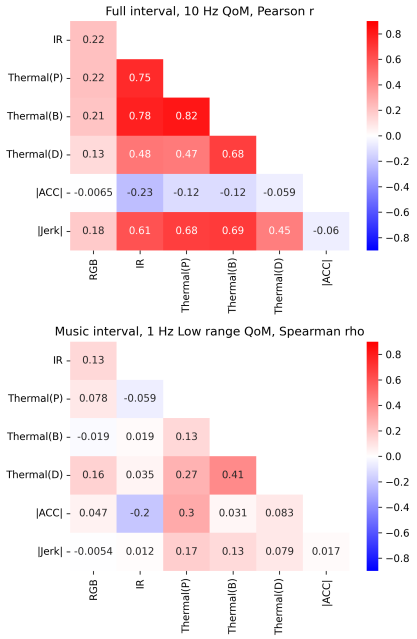


Figure 5. Correlations between QoM features over time and participants. Above are pairwise Pearson correlations between signals over the full measurement interval (including applause), sampled at 10 Hz. Below are Spearman correlations for clipped QoM features focusing on low motion during music presentation, sampled at 1 Hz.

showed the highest spectral energy at twice the applause frequency. Clap impacts are most obvious in the IR signal, with clear peaks around hand collision, while Thermal(B) has more energy at twice the clapping rate. Clapping frequency energy in Thermal(P) is much weaker because the

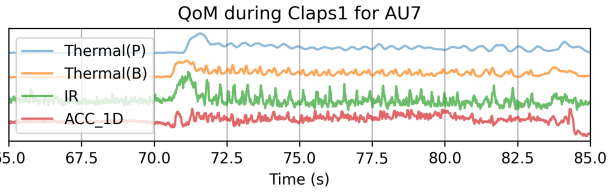


Figure 6. QoM signal values for one participant during clapping: accelerometer (single dimension), IR, Thermal(B) and Thermal(P). Clapping rate was around 115 BPM or 1.9 Hz before tapering off.

rate is close to the cutoff of the filtering effects on cleaning video I-frame peaks, whereas |ACC| displays claps irregularly. Consequently, detailed capture of applause actions may require different tools.

As expected, the torso-mounted accelerometers captured respiratory sequences from five of the six participants during the music. The quality was easily disrupted and not robust to applause; however, the shape of oscillations during quiet sitting showed many recognisable characteristics such as roughly 1 s inspiration phases and attentive respiration rates. None of the video-based measurements captured a comparable component during music presentation.

5. DISCUSSION

Each strategy for measuring audience motion has specific limitations and advantages for ecological concert conditions (Table 2). Results indicate that thermography, IR, and body-worn accelerometry are viable methods for generating similar motion information in the concert hall. However, each system highlights different behaviours.

These measures also depend upon the concert hall envi-

(a) Sensitivity to Audience Behaviours

Method	Large Actions	Applause	Stillness	Respiration
RGB	low light results in a low signal-to-noise ratio	barely separable from stillness in low light	*measures did not agree, highly dependent upon sensor configuration*	not available
IR	clear where illuminated	sharpest clap impacts		not available
Thermal (image)	robust, suited to posture change	frequency present, but smearing and filtering blur fine timing		not available
Thermal (data)	robust	too slow for clap timing		not available
Accelerometry	higher-resolution timing, |jerk| comparable to imaging data types, |ACC| not comparable to imaging	hand motion not directly measured, |ACC| irregular, other positions perform better		available depending upon placement and participant behaviour

(b) Practical Contingencies

Method	Illumination	Spatial Considerations	Temporal Considerations
RGB	needs active lighting, not usually found at measured concert levels	higher-resolution, but unusable at these light levels	$\sim$60 Hz
IR	needs *even* coverage, edge participants poorly described	higher-resolution, degrades where coverage is uneven	$\sim$60 Hz
Thermal (image)	independent (emitted radiation)	lower-resolution, distant bodies merged with DDE	$\sim$25 Hz to 50 Hz
Thermal (data)	independent (emitted radiation)	lower-resolution	$\sim$4 Hz to 8 Hz, regulatory ceiling (sub-9 Hz)
Accelerometry	independent	local to torso	up to 400 Hz

Table 2. Method profiles for measuring audience body motion in this study. Profiles are contingent, not prescriptive. What each sensor makes visible depends on the device, environment, and behaviour, and the rows record trade-offs rather than verdicts. For stillness, the measures did not agree, so no per-method entry is given.

ronment. Thermography's measurement of emitted radiation is more reliable than illumination-dependent IR. However, thermal cameras rarely offer comparable spatial resolution to commercially available RGB and IR devices and miss details of action timing retained by more intrusive accelerometers. QoM calculated from closer crops of video recordings to regions of interest, such as the hands, may make these signals clearer but cannot compensate for the devices' sampling limits.

Another challenge of measuring classical audience motion is that there is very little movement to capture [4]. These signals did not strongly agree on when participants' motion quantity was particularly low, suggesting that the camera-based measurements can not reliably capture the extremes of stillness evaluated with other systems [1].

This study was also our first opportunity to characterise thermography for motion tracking, which revealed both advantages and challenges. When thermal data is mapped into image formats, the bit depth of pixel values is drastically reduced. To compensate, it is common to perform image processing onboard the device to preserve and/or enhance adjacent contrasts. Digital Detail Enhancement (DDE) boosts high-spatial-frequency, non-noise components while smoothing low-spatial-frequency components (see, e.g., [18]), thereby exaggerating morphological edges in areas of contrast. Fig. 3 shows this effect on the visibility of hand motion in the visualised frames (corresponding DCT spatial-frequency spectra are available in the supplementary material). This form of processing has implications for motion tracking of larger groups of people and over larger distances. In this simulation, although participants were evenly spaced and relatively large in the thermal camera frame, DDE increased QoM values when calculated using Eq. 1 due to larger pixel-value differences at exaggerated edges. However, thermal recordings we have obtained from symphony concert halls show individual audience members in distant seats becoming "merged" within a single contiguous border, resulting in the suppression of high-spatial-frequency detail, such as the separation between multiple individuals' hands and heads.

Unlike our short simulation, prolonged classical concerts involve significant changes in body temperature due to larger numbers of people, heat transfer to seating and clothing, and thermal occlusion by hair, which can alter the thermal scene. While frame-differencing QoM is robust to these shifts, fixed-temperature mapping ranges must be chosen carefully to prevent objects of interest from drifting out of bounds. While thermal "smearing" may affect the investigation of motion at the microscale, at the timescales examined in this study ($\sim$6 min), it does not appear to alter the apparent relative similarity of the QoM contours.

When working with thermal data, the speed of the measured motion is a key consideration. Our HIKMICRO box cameras were unable to obtain thermal data sample rates above 8 Hz. This is likely a hardware limitation, which aligns with the regulatory framework of the Wassenaar Arrangement, which classifies microbolometer-based focal plane array sensors with variable focal lengths and sample rates above 9 Hz as dual-use military/civilian technology [19, pp.111,115]. Devices exceeding this rate are subject to export and access control within the 42 signatory countries. While sub-9 Hz rates suffice for generating QoM time series of slower, macroscale behaviours (e.g. posture changes), they may not be suitable for behaviours such as rapid applause. Consequently, regulatory and geographical constraints on access to equipment may directly influence the range of audience behaviours that can be analysed using thermal data.

6. CONCLUSION

This study assessed several methods for measuring audience body motion during symphonic performance. Thermal imaging is a viable alternative to illumination-dependent IR and intrusive body-worn accelerometers for tracking macroscopic behaviours, though its specific limitations make method selection highly dependent on the environment and behaviour of interest. Understanding the specific devices and signals produced is likewise important, particularly where signals and noise magnitudes are similar (e.g. unevenly illuminated IR).

Future work from a music technology perspective should prioritise comparative studies in ecological concert settings that generate simultaneous, multi-sensor datasets. More broadly, it should examine how device signal characteristics—for instance, the understudied relationship between video encoding and QoM—shape the resulting measurements.

Acknowledgments

This work was partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 262762. We thank the participants for their time and Kayla Burnim for technical support.

7. REFERENCES

[1] F. Upham, S. Høffding, and F. E. Rosas, "The Stilling Response: From Musical Silence to Audience Stillness," *Music & Science*, vol. 7, p. 20592043241233422, Sep. 2024.

[2] R. Haswell-Martin, F. Upham, S. Høffding, and N. Nielsen, "Embodied, Exploratory Listening in the Concert Hall," *Behavioral Sciences*, vol. 15, no. 5, p. 710, 2025.

[3] D. Swarbrick and J. K. Vuoskoski, "Collectively classical: Connectedness, awe, feeling moved, and motion at a live and livestreamed concert," *Music & Science*, vol. 6, p. 20592043231207595, 2023.

[4] F. Upham and F. E. Rosas, "When and How the Audience Moves," *Music & Science*, vol. 8, p. 20592043251398280, Sep. 2025.

[5] S. Høffding, N. C. Hansen, and A. R. Jensenius, "Music Research "in the Wild" – Introducing the MusicLab Copenhagen Special Collection," *Music & Science*, vol. 7, p. 20592043241294161, Sep. 2024.

[6] M. Leman, P.-J. Maes, L. Nijs, and E. Van Dyck, "What Is Embodied Music Cognition?" in *Springer Handbook of Systematic Musicology*, R. Bader, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 747–760.

[7] C. Small, *Musicking: The Meanings of Performing and Listening*, ser. Music/culture. Hanover: University Press of New England, 1998.

[8] M. C. Dobson and S. E. Pitts, "Classical Cult or Learning Community? Exploring New Audience Members' Social and Musical Responses to First-time Concert Attendance," *Ethnomusicology Forum*, vol. 20, no. 3, pp. 353–383, Dec. 2011.

[9] A. R. Jensenius, "Methods for studying music-related body motion," in *Springer Handbook of Systematic Musicology*, R. Bader, Ed. Springer Berlin Heidelberg, 2018, pp. 805–818.

[10] L. Theodorou, P. G. T. Healey, and F. Smeraldi, "Engaging with contemporary dance: What can body movements tell us about audience responses?" *Frontiers in Psychology*, vol. 10, 2019.

[11] G. J. Tattersall, R. M. Danner, J. A. Chaves, and D. L. Levesque, "Activity analysis of thermal imaging videos using a difference imaging approach," *Journal of Thermal Biology*, vol. 91, p. 102611, Jul. 2020.

[12] M. Vollmer and K.-P. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*, second edition ed. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2018.

[13] A. Camurri, R. Trocca, and G. Volpe, "Interactive systems design: a KANSEI-based approach," in *Proceedings of the international conference on new interfaces for musical expression*, Dublin, Ireland, 2002, pp. 30–37.

[14] I. E. G. Richardson, *The H.264 advanced video compression standard*, 2nd ed. Hoboken, NJ: Wiley, 2010.

[15] I. E. Richardson, *Coding Video: A Practical Guide to HEVC and Beyond*, 1st ed. Wiley, Sep. 2024.

[16] F. Upham and K. Burnim, "Accelerometer-based synchronisation protocol for music-aligned measurements of distributed datalogging wearables," in *Proceedings of the 22nd. Sound and Music Computing Conference*, ser. SMC2025, vol. 22. Graz, Austria: Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, 2025, pp. 415–420.

[17] R. Deering and J. F. Kaiser, "The use of a masking signal to improve empirical mode decomposition," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4. IEEE, 2005, pp. iv–485.

[18] F. Garcia, C. Schockaert, and B. Mirbach, "Real-time visualization of low contrast targets from high-dynamic range infrared images based on temporal digital detail enhancement filter," *Journal of Electronic Imaging*, vol. 24, no. 6, p. 061103, Sep. 2015.

[19] Wassenaar Arrangement Secretariat, "Public documents volume II: List of dual-use goods and technologies and munitions list (2025 corr.)," Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies, WA-LIST (25) 1 Corr., Jan. 2026.