A Collaborative Sound Installation Using Projected Geometry and Spatial Interaction

Hani Alshamrani
University of Technology Sydney
Sydney, NSW, Australia
hani.alshamrani@student.uts.edu.au

Sam Ferguson
University of Technology Sydney
Sydney, NSW, Australia
samuel.ferguson@uts.edu.au

Andrew Johnston
University of Technology Sydney
Sydney, NSW, Australia
andrew.johnston@uts.edu.au

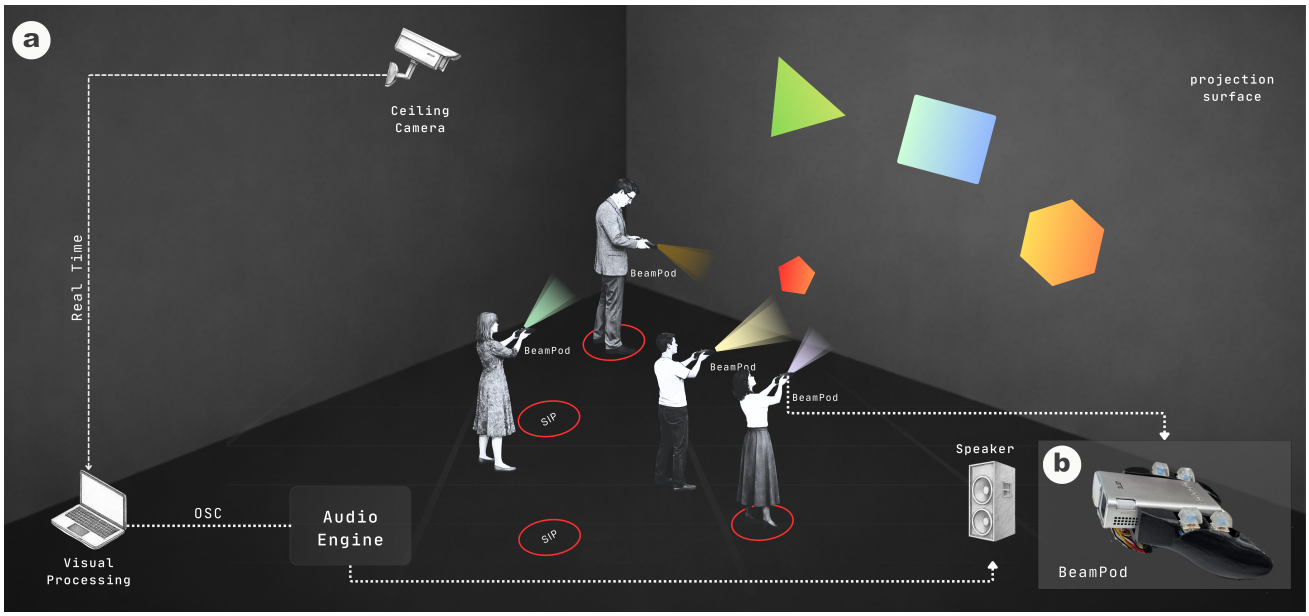

Figure 1: (a) System overview. Four performers carry BeamPods through a 4 m × 3 m interaction space containing five Spatial Interaction Points (SIPs). Each BeamPod projects a unique geometric shape onto the shared wall. A ceiling-mounted camera captures the projections, extracts geometric features, and sends control data via OSC to the audio engine. (b) A BeamPod: handheld projector with onboard Raspberry Pi.

Abstract

We present a collaborative audio visual installation where up to four performers create music by using portable handheld projection systems called BeamPods to project shapes onto a projection surface. A ceiling-mounted camera tracks the projections and extracts shape identity and geometric cues, including position, scale, and distortion, to drive real-time sound synthesis. The projected shape acts as a visible, portable musical voice, allowing performers to see their own contribution and coordinate with others in a shared visual field. The installation is structured around five Spatial Interaction Points (SIPs), discrete floor positions, that define discrete interaction states, while continuous geometric features support expressive control. When performers converge in the same region, the shared pitch mapping makes their contributions converge musically, supporting coordination through spatial negotiation. Designed for walk-up participation in public or workshop settings, the system supports collaborative music-making without prior musical training. This paper reports the design rationale, interaction model, and implementation of this system.

Keywords

collaborative music, spatial interaction, sound installation, projection interface, computer vision

1 Introduction

What if you could project a musical control interface onto a shared wall and shape sound together by moving through the room? Projection has often served as visual output in interactive music systems, typically alongside tangible or tagged inputs [13, 19]. Handheld projector interaction has precedent in HCI for co-located collaboration and situated visualization [8, 18]. We explore a less common configuration: performers carry handheld projectors used as an active input signal for collaborative music-making. The projected shape functions simultaneously as the visible control interface and the sensed control signal for sound (Figure 1).

The central insight is that the projected shape on the wall encodes the performer's spatial location and state. Because projection geometry changes predictably with the performer's location in the room—through shifts in scale, keystone distortion, and screen location—the system can infer which spatial interaction point the performer occupies without tracking their body directly.

NIME '26, June 23–26, 2026, London, UK

The installation features five Spatial Interaction Points (SIPs) marked in the space, each mapped to a pitch from a constrained set chosen for novice-friendly harmony, while four selectable shape types produce four distinct synthesizer voices.

A key design challenge is distinguishing intentional position changes from incidental movement. Our system uses an *intentional registration* strategy: a new SIP becomes active only when multiple conditions are met simultaneously, allowing free movement without abrupt sonic changes while requiring projection geometry to satisfy confidence and stability conditions before confirming a new SIP. This approach is designed for walk-up use without prior instruction.

This paper makes three contributions:

(1) A technique for inferring discrete performer spatial state from projection geometry, using the projected shape on a shared wall as the primary sensing signal rather than body tracking or wearable sensors.
(2) An intentional registration strategy that gates spatial state changes using prediction confidence, temporal consistency, and geometric stability.
(3) A walk-up musical interaction that combines discrete SIP pitch selection with continuous timbral modulation from projection geometry (position, area, tilt) during movement.

This paper presents the system design and technical implementation; formal user evaluation is planned as future work. The following sections situate this work in related research, then describe the interaction model, implementation, and discuss limitations and future directions.

2 Related Work

2.1 Collaborative Musical Interfaces

Collaborative musical interfaces are a long-standing topic within NIME, including instruments designed for group performance and novice participation [4, 5, 12]. Blaine and Fels [4] discussed design considerations for such systems, including interdependence between performers and the balance between accessibility and expressive range. They argue that collaborative instruments often face tensions or tradeoffs between being easy to initiate, but still remaining expressive over time.

Tabletop tangible interfaces are a common approach to multi-user musical interaction, where multiple people share a surface and manipulate physical tokens. The reacTable [13] uses objects on a shared tabletop as control tokens with coupled visual feedback, supporting co-located multi-user performance. Similarly, Audiopad [19] used tracked objects on a tabletop surface to control synthesis parameters, with projection providing visual feedback of the sonic state. Both systems provide a shared visual workspace, making performers' actions visible to each other during co-located play [24]. Beyond tabletops, collaborative installations have also explored how shared physical spaces can scaffold group music-making [17].

Recent work has examined the social dynamics that emerge within collaborative systems. Xambó et al. [24] studied peer learning in reacTable sessions, finding that the visibility of others' actions supported imitation and turn-taking. Tez and Bryan-Kinns [21] report that constraints can shape how performers coordinate and negotiate control in collaborative music-making. These studies suggest that collaboration in NIME systems involves not only technical coordination but also social awareness and mutual learning, supporting short-lived ensemble formations in public or workshop settings.

2.2 Spatial and Embodied Musical Interaction

Musical interaction research has increasingly treated the body and surrounding space as part of the instrument rather than a neutral container. Work on embodied music interaction argues that movement, posture, and spatial positioning can shape musical experience beyond discrete button-like gestures [10, 14]. Early systems demonstrated this potential at room scale. Rokeby's Very Nervous System, discussed in Transforming Mirrors [20], used camera-based tracking to translate whole-body movement into sound, establishing that rooms themselves could function as responsive instruments without requiring worn sensors.

Recent work on augmented reality musical instruments has explored mobility and spatial positioning as expressive dimensions. Wang et al. [23] identify mobility, space, and sound as key axes for AR musical experience, though their work relies on head-mounted displays. Spatial augmented reality (SAR) offers an alternative by projecting digital content directly onto physical environments. Arslan et al. [2] present a SAR interface for musical interaction using fixed projection and vision-based sensing to control actuated acoustic instruments.

Across these approaches, projection typically serves as visual output, displaying feedback or control interfaces, while sensing relies on tracking performers' bodies or gestures. The use of performer-carried projection as a primary input signal, rather than as output or fixed infrastructure, remains largely unexplored in prior work. In multi-user settings, this also raises questions about how spatial positioning might support social coordination among performers.

2.3 Vision- and Light-Based Musical Interfaces

Computer vision has enabled a wide range of approaches to musical interaction, from tracking bodily movement to recognising objects and shapes. Machine learning has enabled flexible mapping of vision-derived features to musical control [9], and subsequent systems have applied depth sensors [11] and full-body tracking [22] for musical interaction. Shape recognition offers particular affordances for musical control. Levin and Lieberman [15] explored hand silhouette contours as input for audiovisual performance in The Manual Input Sessions, mapping geometric properties such as area, compactness, and position to sound parameters. Their work demonstrated that shape features extracted through computer vision could drive both discrete note triggering and continuous timbral modulation.

Light-based interaction has also been explored for musical applications. Ma et al. [16] developed Pharosphones, a system enabling audience participation through mobile phone flashlights tracked by computer vision. They frame the visible nature of light as a means of rendering individual actions perceptible to others. Recent work has extended vision-based musical control to diverse input materials, including knitting patterns [6] and the audification of camera imagery capturing light, shadow, and material patterns [7]. In many vision-based musical systems, projection is primarily used as an output modality, providing visual feedback of sonic state, as in reacTable and Audiopad. The use of projection itself as a primary input signal for musical interaction has received limited attention in prior NIME work. Prior systems often either constrain performers to fixed surfaces, sense bodies invisibly, or use projection primarily for feedback. This leaves

a gap in systems that treat performer-carried projection as a primary input signal for spatial music-making in shared space, which the following sections address.

3 System Design

3.1 Interaction Model

The installation is deployed in a 4 m × 3 m interaction area containing five Spatial Interaction Points (SIPs) marked on the floor (Figure 2a). One to four performers each carry a BeamPod, a handheld projector that casts a geometric shape onto a shared wall (Figure 1). Performers interact by walking through the space, pointing the BeamPod at the shared wall, and adjusting aim and orientation, which reshapes the projection. A ceiling-mounted camera observes the wall and detects each projected polygon. Rather than tracking performers' bodies or requiring worn sensors, the system infers each performer's discrete spatial state from the geometry of their projection: as a performer moves through the room or adjusts the BeamPod's aim, changes in position, scale, and keystone distortion of the projected shape provide cues used to estimate the performer's current SIP.

Interaction is built on three layers:

- **Identity layer (who):** Each performer is associated with a distinct geometric shape (triangle, square, pentagon, hexagon), chosen because polygons are legible at distance, easy to distinguish, and yield stable geometric descriptors under the perspective distortion introduced by off-axis projection. Four shapes match the maximum group size, each aligning a visible token with a distinct sonic voice. Identity is encoded solely through shape type, a constraint discussed in Section 5.4.
- **Spatial state layer (where):** The space is structured around five SIPs, each mapped directly to a single pitch from a constrained set chosen to reduce dissonance. Standing on a SIP selects its associated pitch; moving away holds the last registered pitch until a new SIP is confirmed. The rationale for discrete rather than continuous spatial states is discussed in Section 3.2.
- **Expressive layer (how):** While SIPs provide discrete pitch selection, continuous geometric features extracted from the projection support expressive variation. Walking closer to the wall decreases projection area; changing the BeamPod's orientation alters the observed geometry. These continuous cues are mapped to synthesis parameters that modulate timbre and intensity, providing expressive control without requiring screen-based UI or parameter menus. The specific mappings are described in Section 3.3.

Each detection emits an event containing performer identity (`shape_id`), discrete SIP state (`zone`), and continuous geometric descriptors (`x`, `y`, `area`, `tilt`) that shape the resulting sound through discrete note selection and continuous parameter control. The projected shape is simultaneously the input signal that the system senses and the visible representation of a performer's musical presence. Because musical activity is shared on the wall, performers can coordinate by observing how their shapes move relative to others.

3.2 Spatial Design

We use five SIPs to structure exploration and keep the system learnable for walk-up novices [17]. With too few SIPs, the musical space becomes repetitive; with too many, performers must remember a complex layout and the sensing pipeline must reliably separate closely spaced SIP states from projection geometry alone. Five SIPs strike a balance between variety and learnability. Zones 1–5 correspond one-to-one with SIP 1–SIP 5. We intentionally use discrete SIPs rather than continuous zones. Continuous location control can be expressive, but it can make interaction boundaries harder to interpret in walk-up contexts, where performers may be unsure when a musical change will occur [3]. Discrete SIPs create clear, repeatable musical states—a performer can learn "this spot produces this pitch" and return to it deliberately. Discrete states also simplify the sensing task, since the system distinguishes five spatial categories from projection geometry rather than inferring fine-grained continuous location.

A key challenge is that natural movement produces transient changes in the projected shape. When performers walk, turn, or re-aim the BeamPod, the detected geometry shifts even if they do not intend to switch SIP, which can lead to unstable behaviour if the system updates state immediately. To address this, the system uses an *intentional registration* strategy: SIP changes are treated as a registered state rather than a continuously updated estimate, and a new SIP becomes active only when the system produces the same SIP estimate across multiple consecutive frames. The detailed logic is described in Section 4, but the design goal is that performers should be able to move freely without accidental switching, so spatial transitions feel deliberate and performer-controlled.

Table 1: Mapping from detected projection features to musical parameters.

Feature	Mapped parameter	Perceptual effect
`shape_id`	Voice assignment	Performer identity
`zone`	Pitch	Spatial state selection
`x`	Stereo pan	Left–right position
`y`	Filter cutoff	Brightness
`area`	Amplitude	Loudness
`tilt`	Detune / wobble	Timbral instability
`swipe`	One-shot accent	Shape-specific hit

3.3 Sound Design and Mapping

Sound is designed around four distinct voices, one per shape. Each voice is a separate synthesizer layer with a complementary musical role (e.g., melodic lead, harmonic pad, textural sparkle, rhythmic bass), designed to reduce masking and help performers identify individual contributions within the ensemble. Each SIP maps directly to a single pitch from a D minor pentatonic set (D, F, G, A, C), chosen to reduce the likelihood of dissonance during exploratory play in novice-oriented performance contexts [4]. With five SIPs and five pitches, the mapping is one-to-one: each spatial state selects exactly one note.

The detection system transmits control data to the audio engine via Open Sound Control (OSC) as `/shape/data [shape_id, zone, x, y, area, tilt, swipe]`. Discrete fields determine voice assignment and pitch selection, while continuous geometric descriptors modulate timbre. Table 1 summarises the mapping used in the current prototype. Parameters are normalised and smoothed (100–150 ms): $x \rightarrow$ pan, $y \rightarrow$ LPF cutoff, area$\rightarrow$ amplitude, tilt$\rightarrow$ detune depth; swipe triggers a short accent.

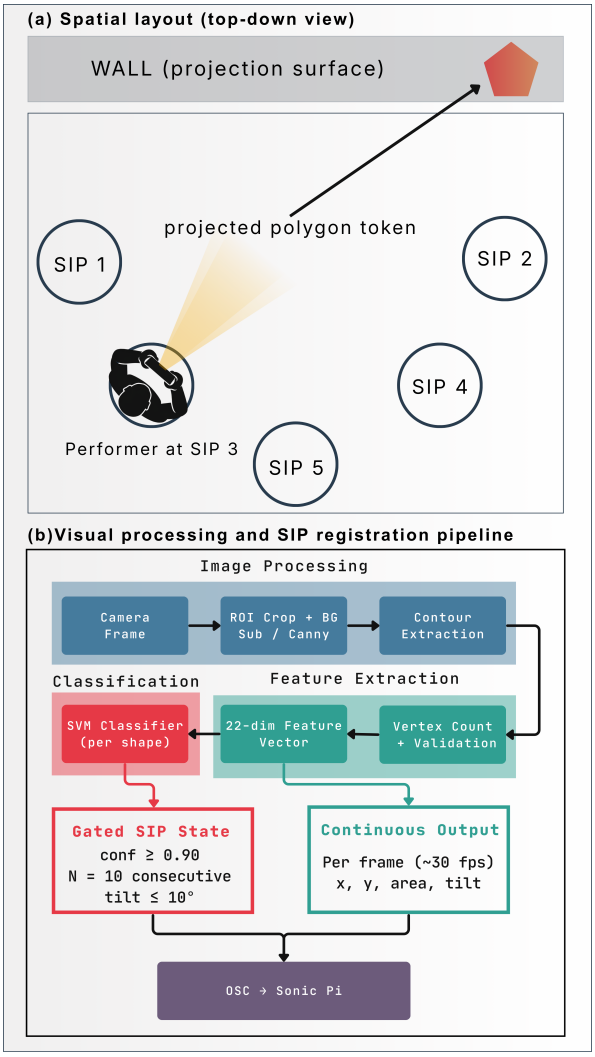

Figure 2: (a) Top-down view of the spatial layout showing five SIPs and a performer projecting onto the shared wall. (b) Visual processing and SIP registration pipeline, showing the separation between gated discrete SIP state and continuous expressive output.

4 Implementation

4.1 System Overview

The implementation consists of a vision-based sensing pipeline and an audio engine connected via Open Sound Control (OSC). Interaction input is provided by up to four standalone BeamPods that project polygons onto the shared wall. Each BeamPod is a handheld projector driven by an onboard Raspberry Pi, which renders a fixed polygon assigned to that performer. On-device controls for shape and colour selection are used during setup only and are not transmitted to the sensing or audio system.

A ceiling-mounted wide-angle camera captures the shared wall. Video frames are processed using Python and OpenCV on a dedicated computer, which extracts geometric features and estimates SIP state. Control data is sent via OSC to Sonic Pi[1], which synthesises sound in real time. The room is kept dim to improve contour stability.

[1]Sonic Pi: https://sonic-pi.net/ [1].

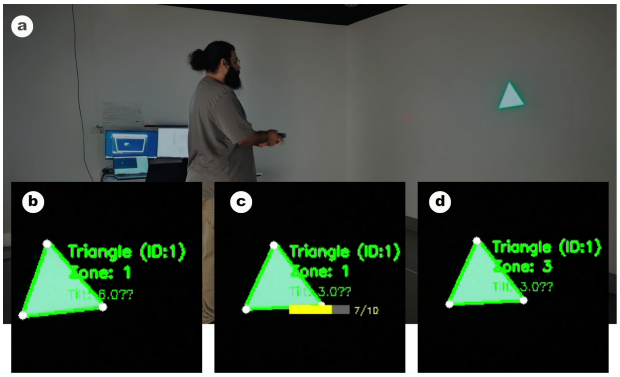

Figure 3: SIP transition process. (a) Performer projecting a triangle onto the shared wall. (b–d) Detection view showing the registration pipeline: (b) triangle registered at Zone 1, (c) confirmation accumulating as the performer stands at SIP 3 (7 of 10 consecutive frames required, subject to tilt and confidence gating), (d) transition confirmed to Zone 3.

4.2 Visual Processing and SIP Registration

The sensing pipeline runs in a real-time loop (Figure 2b). Each frame is converted to grayscale and restricted to a user-defined wall region (ROI). Shape segmentation uses two methods: when a background frame is available, we apply background subtraction followed by Gaussian blur, Otsu thresholding, and morphological opening/closing to obtain a clean binary mask. If no background is captured, the system falls back to Canny edge detection with morphological cleanup. Contours are extracted from the resulting mask and approximated as polygons using `approxPolyDP`.

Shape identity is determined by vertex count (3–6 vertices), with additional validation to reject spurious detections. We filter contours by area, extent, and internal angles, confirming a shape only after M=3 consecutive frames and rejecting duplicates within a fixed radius of existing detections.

For each accepted polygon we extract a 22-dimensional geometric feature vector capturing position, scale, bounding geometry, edge statistics, and asymmetry. We also compute a vision-based tilt proxy from keystone distortion: `tilt` is defined as the normalised deviation of a reference edge from horizontal. The reference edge is taken from the bottom two vertices for triangles, squares, and pentagons, and from the two vertices just below the top peak for hexagons. Shapes are flagged as tilted when `tilt` exceeds 10°.

SIP estimation is formulated as a supervised classification task. An SVM classifier with an RBF kernel maps the 22-dimensional feature vector to one of five SIP labels and outputs a confidence score. To reduce variability across shape types, we collect training data per shape type and train the SIP classifier on the combined dataset. SIP updates are treated as a registered state: a new SIP becomes active only when three conditions are met—the projection is not tilted ($\texttt{tilt} \leq 10°$), the classifier confidence exceeds 0.90, and the same SIP is predicted for N consecutive frames (we use N=10), tuned through iterative rehearsal. Figure 3 illustrates this process, showing a triangle transitioning from Zone 1 to Zone 3 as the confirmation counter accumulates across consecutive frames. Continuous descriptors (`x`, `y`, `area`, `tilt`) are transmitted at the camera frame rate (≈30 fps in our setup) regardless of SIP updates, preserving expressive modulation during movement.

Training data was collected by having a single operator stand at each SIP and project each shape type for approximately 30 seconds, yielding roughly 192 samples per SIP per shape (3,835 samples total across four shapes and five SIPs). All samples were collected under the same lighting and camera conditions used during performance. Five-fold cross-validation on this data yields classification accuracies of 99.2% (triangle), 98.9% (square), 98.6% (pentagon), and 99.7% (hexagon). Misclassifications occur primarily between adjacent SIPs, consistent with closer spatial proximity producing more similar projection geometry.

5 Discussion

5.1 Projection Geometry as Spatial Sensing

This work demonstrates that keystone-distorted projection geometry can serve as a sensing primitive for discrete room-scale spatial state, without body tracking or instrumented floors. This approach requires only a standard camera observing a shared wall. The performer's spatial state is encoded in the projected shape itself: as the performer moves through the room, changes in position, scale, and keystone distortion provide discriminative cues for classifying which SIP the performer occupies. Because each SIP induces a distinct combination of centroid position, scale, and keystone asymmetry, the resulting feature distributions are separable enough for supervised classification.

Using contour geometry as a musical control source has precedent: Levin and Lieberman show that contour-derived features support both continuous modulation and discrete event detection [15]. Our use of projection keystone distortion extends this lineage by analysing projected polygons whose distortion encodes performer position.

A natural question is why geometric features beyond centroid position are needed. Centroid position captures where the shape falls on the wall, but it does not capture the projector's pose relative to the wall. In our pipeline, projected area provides a proxy for distance-to-wall, while the tilt proxy captures off-axis orientation via keystone distortion. Together with edge and asymmetry descriptors, these cues help separate SIP states that can overlap in centroid space. Jordà et al. similarly highlight that computer-vision tracking can detect token orientation rather than treating tokens as points [13].

5.2 Comparison with Related Approaches

This system shares design goals with several prior collaborative musical interfaces but differs in its combination of sensing modality, performer mobility, and the role of the visual token.

Tabletop systems such as reacTable [13] track tangible objects on shared surfaces and use projection to present visual workspaces, enabling performers to see and respond to each other's actions. These systems support rich multi-user interaction but restrict performer mobility to the table boundary. Prior work on shared musical play shows that when individual contributions are difficult to perceive, performers rely more heavily on visual feedback [5], motivating our use of a legible per-performer projected shape as each player's visible musical presence (Figure 4).

Pharosphones [16] extends interaction to room scale using mobile phone flashlights tracked by a camera. Our system shares this principle of visible light as input, but the projected polygon carries richer geometric information (position, scale, distortion, vertex structure) supporting both identity classification and spatial state inference from a single visual signal.

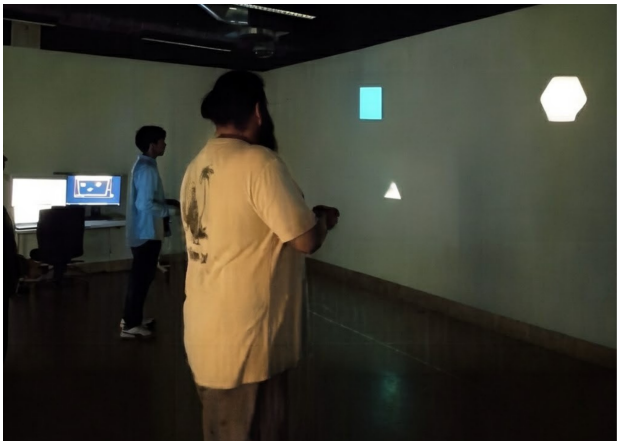

Figure 4: Three performers simultaneously projecting distinct geometric shapes (square, triangle, hexagon) onto the shared wall during a collaborative session. A monitoring station (left) displays the detection view in real time.

Vibrating Shapes [2] uses depth cameras and projectors, detecting hand-body intersection with 3D volumes to drive sound. Our system instead uses a single standard camera observing 2D projection geometry, deriving spatial state from keystone distortion rather than depth-based intersection. This requires less specialised hardware but sacrifices fine-grained volumetric interaction. Murray-Browne et al. note that invisible external influences can confuse performers and delineate player-controlled from externally controlled parameters [17]; our system similarly separates performer-driven continuous expression from system-gated discrete transitions, streaming continuous descriptors while gating SIP updates.

Coordination is supported by the shared SIP-to-pitch mapping: performers who occupy the same SIP produce the same pitch material, encouraging spatial negotiation.

5.3 Intentional Registration as a Design Strategy

A central challenge in spatial musical interfaces is balancing responsiveness with stability. Continuous vision-based tracking can produce unintended state changes when performers move between interaction regions. This is well documented: Han and Gold report that per-frame velocity thresholding produced repeated false triggers requiring debounce logic [11]. Trail et al. add a delay-based activation time so a performer must pause before a control engages, but note that a purely time-based approach can cause accidental activation when playing on a single target long enough to engage it [22].

Our intentional registration strategy separates continuous expressive modulation from discrete spatial state updates. A new SIP is confirmed only when the projection is not tilted (below 10°, as extreme keystone deformation pushes features outside the training distribution), confidence exceeds 0.90, and the same SIP is predicted for 10 consecutive frames (≈330 ms at 30 fps) (Figure 2b). Continuous parameters (position, area, tilt) remain responsive at ≈33 ms per update, so performers retain timbral expressiveness during transient movement. Tez and Bryan-Kinns argue that constraints should be ecologically valid, reporting that abrupt interventions caused frustration [21]; our gating follows

this principle by filtering transient detections so SIP transitions occur only when stable and deliberate.

This tradeoff prioritises predictability over immediacy. Rokeby argues that predictability serves as the primary proof of interactivity, and that filtering improves reliability at the cost of expressive richness [20]. Our design gates discrete state updates (where trust matters most) while leaving continuous parameters responsive (where richness matters most), an important consideration in walk-up contexts [4].

5.4 Limitations, Failure Modes, and Scalability

Lighting dependency. The system requires a dimly lit room; ambient light degrades contour detection. This is a shared constraint across vision-based musical interfaces [11, 22]. These constraints are acceptable for fixed installations but reduce portability to ad-hoc spaces.

Calibration. The SVM classifier must be retrained if camera position or room geometry changes, requiring labelled samples at each SIP for each shape type (a short per-setup collection session). Training data was collected from a single operator; generalisation across different body types and holding styles remains untested.

Identity constraint. Identity is encoded through shape type, limiting the system to four simultaneous performers.

Failure modes. Projector bloom and motion blur distort contour boundaries; partial occlusion truncates polygons; extreme tilt compresses geometry outside trained distributions; overlapping projections merge contours; and non-white or textured walls reduce contrast. Several of these are partially mitigated by validation and confirmation filters (extent thresholding, angle checks, multi-frame confirmation), but they remain visible in edge cases.

Overlapping projections. When two performers project onto the same region of the wall, the merged contour can produce one of two behaviours during informal system testing. If the merged contour is classified as one of the trained shape types, both original shapes stop playing and a new voice begins; however, because the merged geometry lies outside the training distribution, SIP classification becomes unreliable, yielding unpredictable pitch behaviour. If the merged contour cannot be resolved into any trained shape, both performers' voices drop out entirely. This behaviour is a consequence of the single-camera, contour-based sensing approach: the system treats the wall as a shared visual field with no notion of separate performer regions. While this is a limitation for predictable musical control, it also makes spatial proximity audible: performers can hear when their projections collide or merge. Whether this audibility can function as an interactional resource in shared spatial interfaces is a question for future user studies.

Scalability. Five SIPs were chosen as a balance between musical variety and separability of states from projection geometry alone. Cross-validation on the collected dataset (Section 4) yields per-shape classification accuracies between 98.6% and 99.7%, with misclassifications occurring primarily between adjacent SIPs, consistent with closer spatial proximity producing more similar projection geometry. We expect that adding SIPs without increasing geometric separation would reduce the effective margin between classes.

Latency. Continuous control updates at ≈30 fps; gated SIP changes appear after ≈0.33 s. End-to-end audio latency additionally depends on OSC transmission and Sonic Pi scheduling and was not formally measured.

5.5 Future Directions

Future work will focus on formal evaluation with participants to study learnability, spatial coordination, and intentional registration effectiveness. We plan to investigate whether the approach generalises to other room geometries through auto-calibration. Additional identity mechanisms such as colour modulation or temporal patterning could support larger performer groups, while accounting for projector colour calibration and camera colour constancy under varying lighting. The spatial model could also be extended from discrete SIPs to graded proximity-based interaction, where distance between performers continuously modulates shared sound parameters [12]. Mutual awareness in shared spatial musical interfaces, particularly how performers perceive and respond to each other's projected shapes, is a related direction we plan to investigate in future user studies.

6 Conclusion

This paper presented a collaborative sound installation that uses handheld projection geometry as a sensing primitive for SIP-based musical interaction. By extracting geometric features from projected polygons on a shared wall, the system infers discrete performer spatial state without body tracking or wearables. An intentional registration strategy gates SIP transitions using confidence, temporal consistency, and geometric stability, prioritising predictable behaviour in walk-up contexts. The system demonstrates that projection can function simultaneously as a shared visual interface and as the primary sensing signal.

Ethics Statement

This paper reports system design and technical evaluation only and does not involve new studies with human participants or identifiable personal data; therefore, ethics approval was not required.

References

[1] Sam Aaron. [n. d.]. Sonic Pi. https://sonic-pi.net/. Accessed: 2026-02-08.

[2] Cagan Arslan, Florent Berthaut, Anthony Beuchey, Paul Cambourian, and Arthur Paté. 2022. Vibrating shapes: Design and evolution of a spatial augmented reality interface for actuated instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Article 34, 25 pages.

[3] Ben Bengler and Nick Bryan-Kinns. 2013. Designing collaborative musical experiences for broad audiences. In *Proceedings of the 9th ACM Conference on Creativity & Cognition*. 234–242.

[4] Tina Blaine and Sidney Fels. 2003. Collaborative Musical Experiences for Novices. *Journal of New Music Research* 32, 4 (2003), 411–428. https://doi.org/10.1076/jnmr.32.4.411.18850

[5] Tina Blaine and Tim Perkis. 2000. The Jam-O-Drum Interactive Music System: A Study in Interaction Design. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. 165–173.

[6] Kate Bosen and Dan Overholt. 2024. Stitch: a Knitting-powered Musical Interface using Computer Vision. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 390–394.

[7] Jasmine Butt, Benedict Gaster, Nathan Renney, and Maisie Palmer. 2025. Entangling with Light and Shadow: layers of interaction with the pattern organ. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 46–55.

[8] Xiang Cao, Clifton Forlines, and Ravin Balakrishnan. 2007. Multi-User Interaction using Handheld Projectors. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*. 43–52.

[9] Rebecca Fiebrink, Dan Trueman, and Perry R. Cook. 2009. A Meta-Instrument for Interactive, On-the-Fly Machine Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 280–285.

[10] Rolf Inge Godøy and Marc Leman (Eds.). 2010. *Musical Gestures: Sound, Movement, and Meaning*. Routledge.

[11] Jihyun Han and Nicolas Gold. 2014. Lessons Learned in Exploring the Leap Motion Sensor for Gesture-Based Instrument Design. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 371–374.

[12] Sergi Jordà. 2005. Multi-user Instruments: Models, Examples and Promises. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 23–26.

[13] Sergi Jordà, Günter Geiger, Marcos Alonso, and Martin Kaltenbrunner. 2007. The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of the 1st international conference on Tangible and embedded interaction*. 139–146.

[14] Marc Leman. 2007. *Embodied Music Cognition and Mediation Technology*. MIT Press.

[15] Golan Levin and Zachary Lieberman. 2005. Sounds from Shapes: Audiovisual Performance with Hand Silhouette Contours in The Manual Input Sessions. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 115–120.

[16] Zhengyang Ma, Iurii Kuzmin, Duan Ruilei, and Raul Masu. 2024. Pharosphones: interactive audience participation using light. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 480–485.

[17] Tim Murray-Browne, Dom Aversano, Susanna Garcia, Wallace Hobbes, Daniel Lopez, Tadeo Sendon, Panagiotis Tigas, Kacper Ziemianin, and Duncan Chapman. 2014. The Cave of Sounds: An interactive installation exploring how we create music together. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 307–310.

[18] Biswaksen Patnaik, Huaishu Peng, and Niklas Elmqvist. 2024. VisTorch: Interacting with Situated Visualizations using Handheld Projectors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Article 208, 13 pages.

[19] James Patten, Ben Recht, and Hiroshi Ishii. 2002. Audiopad: a tag-based interface for musical performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 1–6.

[20] David Rokeby. 1995. Transforming Mirrors: Subjectivity and Control in Interactive Media. In *Critical Issues in Electronic Media*, Simon Penny (Ed.). State University of New York Press, Albany, NY, USA, 133–158.

[21] Hazar Emre Tez and Nick Bryan-Kinns. 2017. Exploring the Effect of Interface Constraints on Live Collaborative Music Improvisation. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 342–347.

[22] Shawn Trail, Michael Dean, Gabrielle Odowichuk, Tiago Fernandes Tavares, Peter Driessen, W Andrew Schloss, and George Tzanetakis. 2012. Non-invasive sensing and gesture control for pitched percussion hyper-instruments using the Kinect. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[23] Yichen Wang, Mingze Xi, Matt Adcock, and Charles Patrick Martin. 2023. Mobility, space and sound activate expressive musical experience in augmented reality. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. International Conference on New Interfaces for Musical Expression, 128–133.

[24] Anna Xambó, Eva Hornecker, Paul Marshall, Sergi Jordà, Chris Dobbyn, and Robin Laney. 2013. Let's jam the reactable: Peer learning during musical improvisation with a tabletop tangible interface. *ACM Transactions on Computer-Human Interaction* 20, 6, Article 36 (2013), 34 pages.