感知形态的声音生成：基于计算机视觉的有形交互式乐器设计研究

叶宇兴，谭亮 *

（广州美术学院，广东 广州 10586）

摘　要：本文提出以材料形态为核心的数字乐器设计方法，探索计算机视觉在有形交互中的作用，使材料物理特征直接参与声音生成与控制。研究构建基于视觉感知的“形态—声音”映射框架：相机实时捕捉纸等无源材料的铺展、褶皱与位移，经平滑与归一化处理后建立与声学维度的对应关系。框架强调“技术隐身、材料显形”的设计理念，使乐器设计由“识别物体”转向“感知形态”，并将纸的文化语义（白、留白、留痕）融入声音表达与意义生成。研究验证了基于材料形态的直觉映射与可演奏性，提出可迁移、可复用的跨模态设计方法，为数字乐器的教育、展演及艺术疗愈提供新的实践路径与理论支撑。

关键词：新型乐器（NIME）；有形交互；计算机视觉；第三波 HCI

中图分类号：J0-05

DOI：https://doi.org/10.71411/-2025-v1i3-898

Sound Generation through Perceived Form: A Tangible Interaction Instrument Design Based on Computer Vision

Yuxing YE,Liang TAN*

(Guangzhou Academy of Fine Arts, Guangzhou 10586, China)

Abstract: his study proposes a digital instrument design method centered on material morphology, exploring the role of computer vision in tangible interaction so that physical material properties directly participate in sound generation and control. It constructs a vision-based “morphology–sound” mapping framework: a camera captures the unfolding, folding, and displacement of passive materials such as paper, which are smoothed and normalized to establish correspondences with acoustic parameters. The framework embodies the principle of “invisible technology, visible

material,” shifting digital instrument design from object recognition to form perception, and integrates the cultural semantics of paper—whiteness, blankness, and trace—into sonic expression and meaning-making. The study demonstrates the intuitive mapping and playability of material morphology and proposes a transferable, cross-modal design approach offering new pathways for digital instrument education, performance, and art therapy.

Key word: NIME; Tangible Interaction; Computer Vision; Third-Wave HCI

近二十年来，新型乐器（New Interfaces for Musical Expression, NIME）研究已由“设备创新”逐渐转向对体验、意义与情境的综合建构。有形交互（tangible interaction）等理论框架成为重要的概念支撑，即通过将比特嵌入可触摸、可操作的物质载体，使身体与材料属性直接参与信息操控，从而重新确立物质性在交互体验中的核心地位。在更广义的 HCI 谱系中，“第三波”趋势转向关注情感体验、在地实践与意义生成，推动数字乐器设计由工程实现转向以体验与身体为核心的交互过程。数字乐器的设计愈发体现出与实践和审美语境相互生成的特征。

随着计算机视觉的成熟，NIME 的关注点已由控制器革新转向“映射”范式：输入与声音的映射不仅是技术中介，更影响可演奏性与意义表达。Reactable 以标记追踪验证了该路径的潜力。Wessel 指出，低时延与低方差带来的“亲密控制”（control intimacy）使演奏者在动作与声音之间建立连续的信任感。计算机视觉进一步使物体形态可被稳定捕捉，其面积、轮廓与纹理可转化为连续控制量，使日常材料成为可演奏媒介。然而，现有计算机视觉音乐系统仍主要聚焦于对象识别与位姿追踪，对材料形变、纹理与光影质感等连续物理属性的映射潜力缺乏系统探索。

据此，本文以材料感知形态为核心，提出一种基于计算机视觉的有形交互式声音生成方法：通过影像分析捕捉纸张等日常材料的铺展、弯曲与褶皱等形态变化，并将其转译为连续、可解释的声音控制信号，使材料的物理状态成为声音生成的主动维度。本文旨在探讨如何通过视觉感知将物质形态转化为声音生成的感知过程，拓展有形交互在数字乐器设计中的表现边界。研究将“技术隐身、材料显形”的命题具体化为三项关注：其一，物质形态与声音反馈之间是否存在可直觉感知的对应关系？其二，基于形态的操作能否在短时学习中形成具身体感与可控性的演奏体验？其三，纸的能指（形态、质地）如何在演奏中唤起所指（如留白、痕迹），从而体现“材料—声音”关系的非任意性。

1 研究框架

1.1 有形交互与第三波 HCI

有形交互以物质媒介重构人与数字信息的关系：Ishii 系统化提出将比特嵌入日常物品与空间，使用户以操控实体的直觉直接操控数字信息，赋予交互以物理与感知维度；Hornecker 随后将焦点由界面操作扩展至用户体验，强调材料、空间与社会互动的协同。与此同时，第三波 HCI 将范式由效率转向体验，更关注情感、美学与文化；Bødker 指出其旨在跨越理性 / 情感与工作 / 生活的二元，使技术融入日常经验与意义生成。据此，数字乐器设计应被理解为一种体验性实践——不止输入—输出的工程实现，更是审美与

文化表达的媒介。

1.2 新型乐器与映射美学

新型乐器（NIME）以电子与人机交互技术重构“演奏—声音—表达”关系，强调身体与技术共演及感知体验。随 HCI 发展，研究由硬件创新转向体验设计，乐器从信号装置转变为激发身体与文化意义的审美媒介。“映射”指输入与输出的关系设计；“映射美学”关注这种关系的感知与意义，追求动作与声音间的情感逻辑与审美共鸣。Wessel 与 Wright 提出“亲密控制”，强调以低时延和连续手势实现表演者与系统的细腻互动；Hunt 与 Wanderley 提出多参数、层级映射以增强可玩性与表现力；进一步，Wanderley 将映射置于数字乐器控制与交互的核心，强调其对表达与感知的设计意义。

1.3 计算机视觉在有形交互中的应用

计算机视觉（CV）为有形交互开辟新路径：系统利用摄像头与算法实时提取位置、姿态、形状等特征并映射至声音生成，形成“感知—映射—反馈”闭环。ReacTable 的 reacTIVision 以下置摄像头识别标记，通过 TUIO/OSC 实现无接线控制；Stitch 捕捉编织动作，将针法与速度转为声音参数；MuGeVI 以手势识别生成多通道控制信号。CV 因此成为通用传感器，无需布线即可连接物质形态与声音控制。但视觉链路需在延迟与方差间权衡并抑制抖动；研究表明 10–20 ms 延迟已影响演奏时序与体验，低时延与低方差是可演奏性前提。同时，光照、遮挡等易致干扰，需通过平滑与归一化提升追踪稳定性。

1.4 媒介的意义生成与符号关系

媒介理论下，麦克卢汉的“媒介即讯息”揭示媒介形式自带意义生成力；Gell 的“物的代理性”与 Verbeek 的“技术媒介化”进一步表明，物与技术并非中性通道，而是参与经验与诠释的行动者。皮尔斯将符号划分为像似符号（icon）、指示符号（index）和象征符号（symbol），分别对应符号与其所指对象之间的相似关系、因果联系与社会约定三种意义指涉方式。由此，物质特性同时承载媒介信息与文化指向，在视觉与声音的交互中，其形态变化、痕迹与质感成为多层意义的触发点，从而构成有意义、可解释的材料—声音关系，而非随意的映射对应。

2 系统设计

2.1 设计目标与原则

本研究构建一套基于计算机视觉的有形交互数字乐器：相机解析材料形变、位置与纹理并映射至合成，使纸等日常材料的物理变化直接控声。原则：（1）材料直接性—遵循有形交互，保持纸的尺度、柔性与手感，外置摄像头远程感知，无需结构改造；（2）感知一致映射—映射决定可演奏性与可习得性，非接触视觉使操作直接被解释为输入，并据常见联想（更大 / 展开→更厚重或更低沉）实现直觉控制；（3）体验整体与意义建构—依第三波 HCI，将材料—映射—声音组成连续感知回路，“处理纸”即“演奏”，使纸成为具审美表达的声学界面而非中性通道。

2. 系统设计

系统部署于单一台面：固定俯视摄像头覆盖交互区，中央放置普通纸为主介质，左右为独立声道扬声器（图 1）。演奏者在区内对纸进行平移、折叠、揉搓、摊平等操作；摄像头实时采集其位置与形态，立体声即时回放，使形态变化与空间化听感在同一物理范围内被感知。无需屏幕或指示装置，材料操作与声音反馈构成连续感知回路，“处理纸”自然被体验“演奏”。

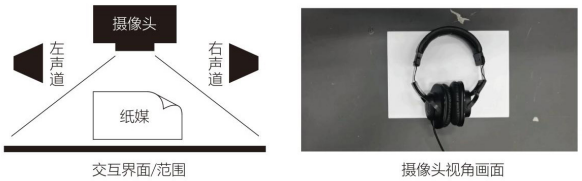

图 1 左：基于俯视布局的交互示意；右：摄像头实际视角画面。

视觉感知模块通过固定摄像头逐帧采集交互区域影像，并基于 YOLOv8 对纸区域定位。系统结合轮廓与纹理分析提取四类连续特征参数（图 2）：X、Y 表示纸张在俯视平面的质心位置，反映平移状态；S 表示有效轮廓面积，用于刻画铺展、折叠或撕裂程度；F 表示褶皱度，用以量化纸面由平整到高纹理的变化。参数 {X, Y, S, F} 逐帧输出，用于实时刻画纸张的位置、尺度与表面能量。系统不改造材料本体，所有控制信号由外部视觉推断，以保留纸媒的物理特性与触感。特征数据经 OSC 输入 TouchDesigner，在软实时环境中完成平滑与去噪，抑制光照、遮挡与抖动干扰，并在归一化与重标定后映射至 Bitwig PolyGrid（图 3）的合成参数，实现视觉特征到声音实时响应。

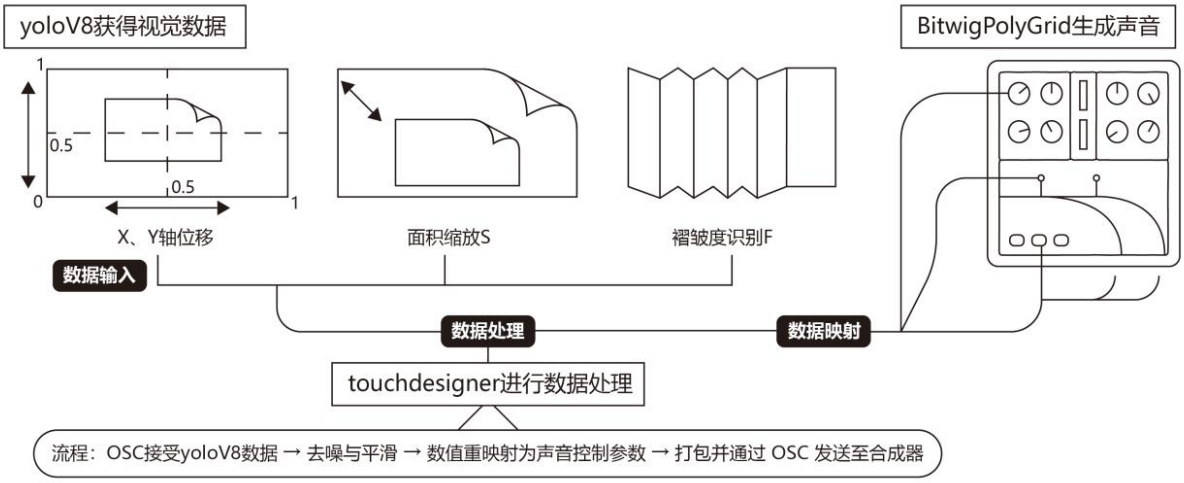

图 2 系统数据处理与映射流程。

图 3 Bitwig PolyGrid 声音合成模块系统。

为实现从视觉形态到声学表征的可解释控制，本文将归一化后的四个视觉特征 {S, F, X, Y} 分别对应至“基频与能量框架、纹理与谐波复杂度、立体声像分布、时域深度线索”四个声学维度（见表 1）。映射遵循“直觉一致—连续可控—低耦合”的原则：

表 1 视觉特征与声学维度映射关系。

视觉特征	声学维度	主要映射目标	感知效果（方向示意）
面积 S	基频/能量	基频相关节点、主音量	面积增大时，声音趋于厚实，基频分布范围扩大；面积减小时，声音更紧致，能量下降。
褶皱度 F	纹理/谐波复杂度	滤波截止、噪声混合、失真驱动	褶皱度增加时，音色变得粗粝，高频和非谐波成分增强。
水平位移 X	立体声像	声像（Pan/Balance）	向左或向右移动纸张时，声音的声像在左右声道之间平滑偏移。
垂直位移 Y	时域“深度”	延时 time/feedback（轻耦合滤波）	向上或远处移动纸张时，声音的延时和回响增强，听感上呈现更远的空间感。

系将视觉参数 {X, Y, S, F} 映射至声音合成的四个维度：① 面积 S 控制基频与能量，接入 Pitch Quantize → Octaver 与 Wavetable/Sine/Phase-1 的 Pitch，联动幅度包络（AD）level，形成“体量—响度”对应；纸面摊展时基频上移、能量增强；折叠或撕裂、音高收敛、声能减弱。② 褶皱度 F 调制 SVF 与 Low-pass MG 的截止频率，并驱动噪声与失真强度；揉搓或折叠增强时，高频与非谐波分量上升，音色更粗粝，映射曲线经 Curves 非线性整定以保持颗粒稳定。③ 水平位移 X 控制 Mixer 的 Pan/Balance，实现声像在左右声道间的动态漂移。④ 垂直位移 Y 映射至 Delay 的 Time 与 Feedback，并耦合 Sallen-Key 滤波截止，形成“距离—扩散/衰减”线索。纸张上移或远离时，延时拉长、反馈增强、频率略降，产生更远空间感（系统以 Delay + 滤波替代混响）。系统层面，YOLOv8 负责目标检测，TouchDesigner 进行参数滤波与重标定，PolyGrid 实现合成调制，三者构成从视觉形态到听觉表征的实时跨模态管线。纸的位移、尺度变化与质感由被动感知对象转化为声学生成的主动驱动量，形成“材料—动作—声音”的演奏闭环。

3 用户测试与设计反思

3.1 用户测试

为检验前文提出的三项关注，测试共邀请 24 名来自美术学院的学生参与，其中 8 人具音乐学习或演奏基础（如钢琴、打击乐或声乐），5 人无艺术创作实践经验（主要从事理论方向研究），其余为视设计方向。

本研究采用六个离散动作（图 4）作为操作引导；结果按条件（P1–P4）而非单个动作报告。在 TouchDesigner 中以 Filter CHOP 设四档平滑，每个动作演示 5 分钟。记录平均延时与抖动度两项客观指标，并收集掌控感（1–7）用以综合评估“亲密控制”与“直觉映射”的感知表现；所有指标在参与者与动作层面取平均后进行统计。

从表 2 可见，延时越低，响应越直接，但抖动度较高；平滑度越高，反馈更稳定，却伴随明显的声音滞后。该关系直接影响参与者的“亲密控制感”与“直觉映射度”，两者皆源于流畅且可预测的控制体验。由于系统以持续发声的合成器为核心，低延时虽能带来较强的惯性控制，

使参与者感到“纸牵着声音走”，但此时动作与声音的因果同步被削弱，纸媒不再被感知为声源本身，而成为外在的控制界面，从而导致材料与声音的一体感下降，直觉映射度减弱。为进一步理解参与者对系统的感知机制，研究在实验后进行了半结构式访谈。访谈以三项研究关注（RQ1–RQ3）为核心构架（见表 3）。

图 4 纸媒交互示意。

表 2 研究结果。

研究关注（RQ）	访谈提纲问题	综合主题摘要（24 人）
RQ1 直觉映射是否成立	操作时能否自然理解声音变化？声音反馈是否与你的动作想象一致？	大多数参与者在无提示下即可理解形态与声音的对应关系，常以“纸的展开更亮”“揉搓更噪”为例说明映射直觉；部分人提到初期需短暂适应，但整体认为“动作—声音”关系清晰自然。
RQ2 可演奏性与身体体验	你是否能在短时间内进入“可演奏”的流动状态？操作中是否感到节奏或身体连贯性？	超过三分之二参与者表示在约 2 -3 分钟内即可建立控制感并进入演奏节奏；部分人指出过度延时导致“被拉着走”的滞感；多数反馈认为系统在低延时条件下具有明显的身体可玩性与流动体验。
RQ3 材料语义与声音意义	纸的形态或质地是否影响你对声音的理解？你会把纸的特性与声音联系起来吗？	多数参与者将纸视为“有记忆的材料”，声音被理解为“留白”“痕迹”“时间的折叠”等象征；少数人提到材料特征强化了声音的“物性”，使“材料—声音”关系具象且富含文化意涵。

注：平均延时（记录 Filter 输入帧时刻 T_1 与输出帧时刻 T_2，计算 $\Delta t = T_2 - T_1$，并以 Analyze CHOP 取平均）与抖动度（Filter 输出在进入目标窗 ±5% 后 1 s 内的标准差，归一化至 0–1）。

表 3 研究关注、引导问题与参与应摘要。

研究关注（RQ）	访谈提纲问题	综合主题摘要（24 人）
RQ1 直觉映射是否成立	操作时能否自然理解声音变化？声音反馈是否与你的动作想象一致？	大多数参与者在无提示下即可理解形态与声音的对应关系，常以“纸的展开更亮”“揉搓更噪”为例说明映射直觉；部分人提到初期需短暂适应，但整体认为“动作—声音”关系清晰自然。
RQ2 可演奏性与身体体验	你是否能在短时间内进入“可演奏”的流动状态？操作中是否感到节奏或身体连贯性？	超过三分之二参与者表示在约 2–3 分钟内即可建立控制感并进入演奏节奏；部分人指出过度延时导致“被拉着走”的滞感；多数反馈认为系统在低延时条件下具有明显的身体可玩性与流动体验。
RQ3 材料语义与声音意义	纸的形态或质地是否影响你对声音的理解？你会把纸的特性与声音联系起来吗？	多数参与者将纸视为“有记忆的材料”，声音被理解为“留白”“痕迹”“时间的折叠”等象征；少数人提到材料特征强化了声音的“物性”，使“材料—声音”关系具象且富含文化意涵。

参与者普遍能够在无提示情境下理解纸的形态与声音之间的对应关系，并在短时演奏中建立控制感与投入度；多数反馈显示系统映射直觉清晰、具备可演奏性。同时，纸的物质与文化特征进入了声音解释过程，表现为书写、留白、记忆与时间等多重联想，验证了材料—声音映射的非任意性及其在意义生成层面的潜力。

3.2 设计反思

系统以“技术隐身、材料显形”为设计目标，在用户测试中展现出多项积极特征。多数参与者在无提示下即可理解形态与声音的对应关系，表明映射逻辑具直觉性与一致性；系统具备良好的上手性与沉浸感，部分反馈形容为“像有生命的演奏体验”。纸作为文化符号，其形态变化激发了“留白”“痕迹”“不可逆”等象征性联想，显示材料在声音生成中具有独特表达潜力。然而，系统仍存在局限：视觉识别易受光照与环境干扰影响，纸张稳定性不足以支撑高重复性操作；材料—声音的语义关联在跨文化语境下亦缺乏统一解释。总体而言，该框架在无硬件嵌入的条件下实现了材料形态到声音的即时转译，具备低门槛与高可塑性，适用于即兴演出、装置艺术及教育体验等语境。后续将从提升鲁棒性、引入语义提示与扩展可控维度等方向优化，以深化材料—声音之间的表现与文化指涉。

4 结论与未来工作

本文提出一种基于计算机视觉的材料感知型数字乐器系统，尝试以纸张的物质形态为声音生成的核心输入，回应“技术隐身、材料显形”的设计命题。通过小规模用户测试与质性分析，初步验证了形态与声音之间的直觉映射、系统的演奏可控性，以及材料语义在声音理解中的非任意性。研究显示，纸的视觉与文化特性可有效激发声音的象征性解读，拓展了数字乐器中材料—声音关系的表现力。未来工作将聚焦三个方向：一是提升系统对复杂环境下形态变化的识别鲁棒性；二是丰富声音层次与演奏机制，增强持续使用中的表达深度；三是进一步引入语义引导与叙事结构，构建具象征张力与沉浸体验的材料—声音交互范式。

参考文献：

[1] ISHII H., ULLMER B. Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms [C]// Proceedings of CHI ’97: Conference on Human Factors in Computing Systems. New York: ACM, 1997: 234–241.

[2] BØDKER S. When Second Wave HCI Meets Third Wave Challenges [C]// Proceedings of NordiCHI 2006. New York: ACM, 2006: 1–8.
[3] HUNT A., WANDERLEY M. M., PARADIS M. The Importance of Parameter Mapping in Electronic Instrument Design [C]// Proceedings of the 2002 Conference on New Interfaces for Musical Expression (NIME). Dublin: NIME, 2002: 88–93.
[4] WESSEL D., WRIGHT M. Problems and Prospects for Intimate Musical Control of Computers [J]. Computer Music Journal, 2002, 26(3): 11–22.
[5] JORDÀ S., KALTENBRUNNER M., GEIGER G., BENCINA R. The reacTable* [C]// Proceedings of the International Computer Music Conference (ICMC). Barcelona: ICMC, 2005: 579–582.
[6] HORNECKER E., BUUR J. Getting a Grip on Tangible Interaction: A Framework on Physical Space and Social Interaction [C]// CHI 2006: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2006: 437–446.
[7] TORRE G., ANDERSEN K., BALDÉ F. The Hands: The Making of a Digital Musical Instrument [J]. Computer Music Journal, 2016, 40(2): 22–34.
[8] BOSEN K., OVERHOLT D. Stitch: A Knitting-powered Musical Interface Using Computer Vision [C]// Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2024). Utrecht: Zenodo, 2024: 390–394.
[9] YANG Y., WANG Z., LI Z. MuGeVI: A Multi-Functional Gesture-Controlled Virtual Instrument [C]// Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2023). Mexico City: Zenodo, 2023: 536–541.
[10] JACK R. H., MEHRABI A., STOCKMAN T., et al. Action–Sound Latency and the Perceived Quality of Digital Musical Instruments: Comparing Professional Percussionists and Amateur Musicians [J]. Music Perception, 2018, 36(1): 109–128.
[11] CHEN F., WANG X., ZHAO Y., et al. Visual Object Tracking: A Survey [J]. Computer Vision and Image Understanding, 2022, 222: 103508.

作者简介：

叶宇兴，男，硕士，研究方向为人工智能与交互艺术设计。

通信作者：

谭亮，男，教授，博士生导师，研究方向为数字艺术与交互设计。