


USC University of
Southern California

**Sharing Civil Engineering Data:
Technical, Cultural, and Practical Lessons
from 3 Open Databases**

**2026 NSF Workshop on Research Data
Management in Construction Engineering:
Challenges and Opportunities**

May 07-08, 2026 –Michigan Technological University

Lucio Soibelman PhD, NAC, Dist. ASCE
Fred Champion Estate Chair in Engineering Professor
Astani Department of Civil and Environmental Engineering
University of Southern California



**Astani Department of Civil &
Environmental Engineering**

1


USC University of
Southern California

Overview

Three datasets · common themes

In this talk

 <p>PLAID</p> <p>Plug-Load Appliance Identification Dataset (Electrical fingerprints @ 30 kHz)</p>	 <p>EMBED</p> <p>Energy Monitoring through Building Electricity Disaggregation (12 kHz, Los Angeles)</p>	 <p>STPLS3D</p> <p>3D point-cloud segmentation dataset (BMVC 2022 oral) (>1,000s of downloads)</p>
--	--	---

Soibelman | NSF Workshop | Michigan Tech | Overview



**Astani Department of Civil &
Environmental Engineering**

2



PLAID — Plug-Load Appliance Identification Dataset

Published in Nature Scientific Data · Zenodo/Figshare

What is PLAID?

- Fills a critical gap in detailed electrical characterization of domestic appliances.
- Electrical load disaggregation algorithms need appliance 'fingerprints' from whole-house measurements.
- Prior datasets reported only low-resolution metrics (e.g., average power).
- PLAID collects synchronized voltage & current waveforms at 30 kHz for 100+ appliances.
- Published via Nature Scientific Data (2020).

<https://www.nature.com/articles/s41597-020-0389-7>

30 kHz
Sampling rate

100+
Appliances

2010-18
Collection span

Zenodo / Figshare
Repository

Soibelman | NSF Workshop | Michigan Tech | PLAID



Astani Department of Civil & Environmental Engineering

3



PLAID — Experience & Lessons Learned

From Mario Berges

Advantages

✓ Peer-review process (Nature Sci. Data) improved documentation quality. · Reviewers acted as first real users — identified gaps in usability. · Mature repositories (Zenodo, Figshare) after 2016 made sharing easy. · Dataset became widely used for load-disaggregation benchmarking.

Disadvantages / Challenges

⚠ 2010 attempt was difficult — little infrastructure existed. · High data volume (30 kHz waveforms) required specialized storage. · Ongoing maintenance and user-support burden after publication.

Key Lessons

💡 Invest in documentation from day one — it doubles as the usage guide. · Submit to a data journal: the review process keeps you accountable. · Use established repositories; avoid self-hosting.

Soibelman | NSF Workshop | Michigan Tech | PLAID · Lessons



Astani Department of Civil & Environmental Engineering

4


USC University of
Southern California

EMBED — Energy Monitoring through Building Electricity Disaggregation

embed-dataset.org · ACM e-Energy 2018

What is EMBED?

- Designed for electricity disaggregation (NILM) research.
- Data from 3 apartment units in Los Angeles, CA.
- Aggregate power + appliance-level ground truth (smart plugs & light sensors).
- Fully labeled appliance events including sub-state transitions.
- Supports event-based and non-event-based (HMM) NILM methods.
- Benchmark for classification, clustering & disaggregation algorithms.

embed-dataset.org | <https://dl.acm.org/doi/10.1145/3208903.3208939>

Dataset Specifications

Parameter	Value
Sampling rate (V & I)	12 kHz
Power metrics	P, Q, harmonics (9th)
Plug-load sampling	~1–2 Hz
Buildings	3 apartments, Los Angeles
Duration	14–27 days each
Labeling	Full appliance events + sub-states

Soibelman | NSF Workshop | Michigan Tech | EMBED



Astani Department of Civil & Environmental Engineering

5


USC University of
Southern California

EMBED — Experience & Lessons Learned

From Farrokh Jazadeh

✓ Advantages

- Fully labeled events enable both supervised & unsupervised learning.
- Shared benchmark improves comparability across research groups.
- Eliminates costly per-lab data collection.
- Establishes a framework for expanding contextual variety.

✗ Disadvantages

- Limited contextual variance & relatively short duration.
- Possible bias from specific LA apartment configurations.
- High-resolution data requires large storage and specialized infrastructure.

💡 Lessons Learned

- High-resolution collection demands significant storage planning upfront.
- Wireless ground-truth sensors (smart plugs) prone to disconnection — consider eGauge-type solutions.
- Detailed sub-state appliance labeling is valuable but extremely time-intensive; consider trade-offs.
- High-level or no labeling may be preferable if it enables longer collection periods.

Soibelman | NSF Workshop | Michigan Tech | EMBED · Lessons



Astani Department of Civil & Environmental Engineering

6


USC University of
Southern California

STPLS3D — Semantic Terrain Points Labeling

Synthetic 3D

BMVC 2022 (Oral) · stpls3d.com · Meida Chen

What is STPLS3D?

- Large-scale 3D point-cloud dataset for outdoor semantic & instance segmentation.
- Reconstructed via aerial photogrammetry across ~16 km² of varied landscapes.
- Combines real-world sites (USC campus, Orlando convention center, etc.) with synthetic data.
- Synthetic data generation pipeline using 3D scene engines + UAV simulators.
- Accepted as oral at BMVC 2022; dataset publicly available at stpls3d.com.
- Co-organized Urban3D challenges at ECCV 2022 and ICCV 2023.

www.stpls3d.com | <https://arxiv.org/abs/2203.09065>

Soibelman | NSF Workshop | Michigan Tech | STPLS3D

~ 16 km²

Area covered

1,000s

Downloads to date

2

International challenges
(ECCV 2022 · ICCV 2023)

Oral

BMVC 2022 acceptance



**Astani Department of Civil &
Environmental Engineering**

7


USC University of
Southern California

STPLS3D — Experience & Lessons Learned

From Meida Chen

 **Advantages**

- Significantly boosted research visibility — thousands of downloads.
- Community adoption: other teams benchmark and cite the paper.
- Organized two international challenges (ECCV 2022, ICCV 2023).
- Accelerated progress in 3D segmentation research.

 **Key Lesson**

Better documentation from the start would have reduced effort for both authors and users significantly — this is the single most impactful improvement possible.

 **Disadvantages**

- Continuous website maintenance required.
- Regular technical support & question-answering from diverse user base.
- Different downstream uses generate a wide variety of questions.
- Thousands of users amplify the support workload.

“People have different uses for the data, so they often have various questions about it.”

— Meida Chen



**Astani Department of Civil &
Environmental Engineering**

Soibelman | NSF Workshop | Michigan Tech | STPLS3D · Lessons

8


USC University of
Southern California

Comparing the Three Datasets

Similarities and differences at a glance

Attribute	PLAID	EMBED	STPLS3D
Domain	Electrical (appliances)	Electrical (buildings)	3D Computer Vision
Resolution	30 kHz V & I waveforms	12 kHz V & I + 1–2 Hz plugs	Point-cloud (aerial)
Labels	Appliance fingerprints	Full event + sub-state labels	Semantic & instance
Repository	Zenodo / Figshare Nature Sci. Data	embed-dataset.org ACM e-Energy	stpls3d.com BMVC 2022
Pub. period	2016–2018	2018	2022
Key lesson	Peer-review improves documentation	Storage & labeling trade-offs	Documentation from day one

Solbelman | NSF Workshop | Michigan Tech | Comparison



**Astani Department of Civil &
Environmental Engineering**

9


USC University of
Southern California

Universal Lessons for Open Engineering Datasets

Synthesized from all three experiences

01 **Documentation is not optional**

 Invest heavily before release. Documentation becomes the user manual — poor docs multiply support requests.

04 **Plan for ongoing maintenance**

 Budget time for updates, user support, and keeping links/downloads alive. Impact compounds but so does work.

02 **Use established repositories & data journals**

 Zenodo, Figshare, Nature Scientific Data. Peer review keeps you accountable and improves quality.

05 **Match resolution to reusability**

 Ultra-high-resolution data offers flexibility but creates storage, distribution, and labeling burdens. Define the trade-off deliberately.

03 **Public data amplifies research impact**

 Visibility, citations, benchmarking, and community challenges all grow with open access.

06 **Infrastructure matters**

 2010 was hard; 2018+ was easy. Leverage today's mature ecosystem of tools and journals.

Solbelman | NSF Workshop | Michigan Tech | Lessons



**Astani Department of Civil &
Environmental Engineering**

10

1. Data Licensing

Telling users what they can actually do with your data

Without a clear license, downstream users are in a legal gray zone — even if you intend the data to be freely reusable.

CC0 — Public Domain

- No restrictions whatsoever.
- Anyone can use, modify, sell.
- Maximum reach and reuse.
- Best for raw factual data.

CC-BY — Attribution

- Free to use and redistribute.
- Must credit the original authors.
- Most common for research data.
- Balances openness with credit.

CC-BY-NC — Non-Commercial

- Free for academic/research use.
- Commercial use prohibited.
- Protects against exploitation.
- May limit some reuse scenarios.

Recommendation:

Attach a license file to every dataset release. If in doubt, use CC-BY — it is the de facto standard for NSF-funded research data.

Only ~40% of published datasets include an explicit license — the majority leave users uncertain about legal reuse rights.

2. Privacy & Ethics

Protecting participants when sharing real-world building and appliance data

Key Ethical Obligations

IRB / Ethical Approval

Data collected in homes or workplaces requires Institutional Review Board approval before collection begins.

Informed Consent

Participants must be told what is collected, stored, and shared — and agree in writing.

Anonymization

Remove or hash identifiers: apartment numbers, occupant schedules, or device names that could re-identify residents.

Data Minimization

Collect only the resolution and duration you actually need. EMBED's 2-week window was a deliberate choice.

Relevant to EMBED & PLAID

- EMBED: collected in 3 private LA apartments — IRB approval required.
- PLAID: domestic appliance data may reveal occupancy patterns.
- NSF requirements: many programs mandate a Data Management Plan (DMP) addressing privacy.
- Publishing via Nature Scientific Data includes an ethics review as part of peer review.
- Consider synthetic or aggregated releases for sensitive high-resolution data.


USC University of
Southern California

3. Long-Term Sustainability & Link Rot

Ensuring data remains accessible years after publication

The problem: URLs break, hosting costs money, and institutional priorities change. Even well-cited datasets can become unreachable within 5-10 years of publication.

Assign a DOI

- Use Zenodo or Figshare to get a permanent Digital Object Identifier.
- DOIs resolve even if the underlying URL changes.
- Required by many journals and funding agencies.

Version and Archive Releases

- Tag each dataset version (v1.0, v2.0) with a separate DOI.
- Archive the exact files used in each published paper.
- Prevents reproducibility failures when datasets are updated.

Deposit with a Library / Repository

- Institutional repositories (e.g., USC Libraries) provide long-term stewardship.
- NSF-funded data can go to NSF's own repositories.
- Avoids dependency on personal websites (like stpls3d.com).

Write a Data Availability Statement

- Include in every paper: where the data is, its DOI, and access conditions.
- Journals increasingly require this for acceptance.
- Helps users find and cite data correctly.



**Astani Department of Civil &
Environmental Engineering**

13


USC University of
Southern California

4. Data Standards & Interoperability

Making datasets usable across research groups and tools

Each of our three datasets uses different formats, sampling rates, and metadata schemas. Without alignment to community standards, reuse requires custom preprocessing per dataset.

Dataset	Data Type	Format	Standards Status
PLAID	30 kHz V & I waveforms	MATLAB .mat files	No unified NILM standard
EMBED	12 kHz V & I + 1-2 Hz plug data	CSV + HDF5	Partially compatible with NILM-Eval
STPLS3D	3D point clouds	LAS/LAZ (aerial), custom synth	No single 3D benchmark format

Recommendations:
Adopt community formats early (e.g., NILM-Eval for energy data, LAS/LAZ for point clouds). Document your schema with a README or data dictionary. This dramatically reduces the preprocessing burden for new users.

NSF's FAIR principles (Findable, Accessible, Interoperable, Reusable) now guide data management expectations for funded projects.



**Astani Department of Civil &
Environmental Engineering**

14

5. Citation & Academic Credit for Data

Aligning incentives with open sharing

The Incentive Problem

Academic careers are built on paper citations, not data downloads. If sharing data does not yield career credit, researchers are disincentivized from doing it — even when it clearly benefits the field.

The Opportunity

STPLS3D received thousands of downloads and significant citations — demonstrating that a high-quality public dataset can generate as much (or more) academic impact than a traditional conference paper.

How to Maximize Citation Credit

Publish a Data Paper

Submit to Nature Scientific Data, Data in Brief, ASCE Computing Data Paper, or Scientific Data. Peer review validates the dataset as a formally citable artifact.

Assign a DOI + Citation File

A DOI makes the dataset citable in reference lists. Add a CITATION.cff or BibTeX entry so users cite correctly with one click.

Track Dataset Citations

Use Google Scholar, Semantic Scholar, or DataCite to monitor citations. Present this data in tenure and promotion dossiers.

Organize Competitions

Hosting a challenge (like Urban3D at ECCV/ICCV) generates dozens of citing papers in a single event — STPLS3D is proof.

6. Dataset Versioning

Managing updates without breaking reproducibility

PLAID evolved across three major releases (2010 → 2016 → 2018). Without a clear versioning strategy, users cannot reliably reproduce results from papers that cite earlier versions.

2010 — PLAID v0 (Sample)

- Small appliance sample released.
- No formal repository — file sharing only.
- No DOI, no documentation standard.
- Users had no way to cite it formally.

2016 — PLAID v1.0

- First versioned release on Zenodo.
- Added voltage waveforms for 55 appliances.
- Accompanied by a technical report.
- Enabled early NILM benchmarking work.

2018 — PLAID v2.0 (Final)

- Published in Nature Scientific Data.
- 100+ appliances, full documentation.
- DOI assigned; peer-reviewed data paper.
- Became the de facto community benchmark.

Best Practices for Versioning:

Tag each release with semantic version numbers (v1.0, v2.0). Assign a separate DOI per version. Keep a changelog noting what changed and why. Never silently overwrite a published release — papers citing that version must remain reproducible. Notify users via the repository's mailing list or GitHub.


USC University of
Southern California

7. The True Cost of Data Collection

Making the case for shared datasets more concrete

"Eliminates the need for each group to collect their own data" sounds compelling — but how compelling? Putting rough numbers on the cost makes the ROI argument far more persuasive for an NSF audience.

PLAID Collection Costs	EMBED Collection Costs	STPLS3D Collection Costs
<ul style="list-style-type: none"> Custom 30 kHz synchronized V&I acquisition hardware. Lab technician time: instrument setup, calibration, QA. 100+ appliances sourced, tested, and documented. Estimated total: \$50,000-\$100,000+ for a comparable new effort. 	<ul style="list-style-type: none"> 3 instrumented LA apartments over 2-4 weeks each. Wireless smart plugs, light sensors, central data loggers. Ongoing sensor monitoring to prevent dropouts. Researcher time for sub-state appliance labeling. Estimated total: \$30,000-\$60,000 per comparable campaign. 	<ul style="list-style-type: none"> Aerial photogrammetry flights over ~16 km² of terrain. UAV operations, permits, and pilot costs. Manual semantic + instance labeling of point clouds. Synthetic scene generation pipeline development. Estimated total: \$80,000-\$150,000+ for equivalent coverage.

Key Takeaway:
Sharing a single high-quality dataset can save the community hundreds of thousands of dollars in redundant collection effort. This is a compelling ROI argument for NSF and other funders — and a strong motivation to invest in proper maintenance over time.



Astani Department of Civil & Environmental Engineering

17


USC University of
Southern California

Questions & Discussion

PLAID
nature.com/articles/s41597-020-0389-7

EMBED
embed-dataset.org

STPLS3D
stpls3d.com

Lucio Soibelman · NSF Workshop · Michigan Tech

Special thanks to the PhD students who contributed their data and experiences:



Astani Department of Civil & Environmental Engineering

18



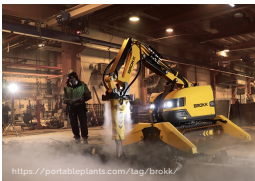
Towards a Sensory-Enhanced Teleoperation Interface for Demolition Robots Through End-User Engagement

1

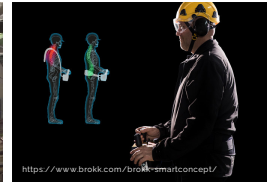
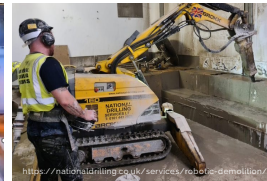
1

Findings – Challenges in Traditional Demolition

Site and Task Conditions

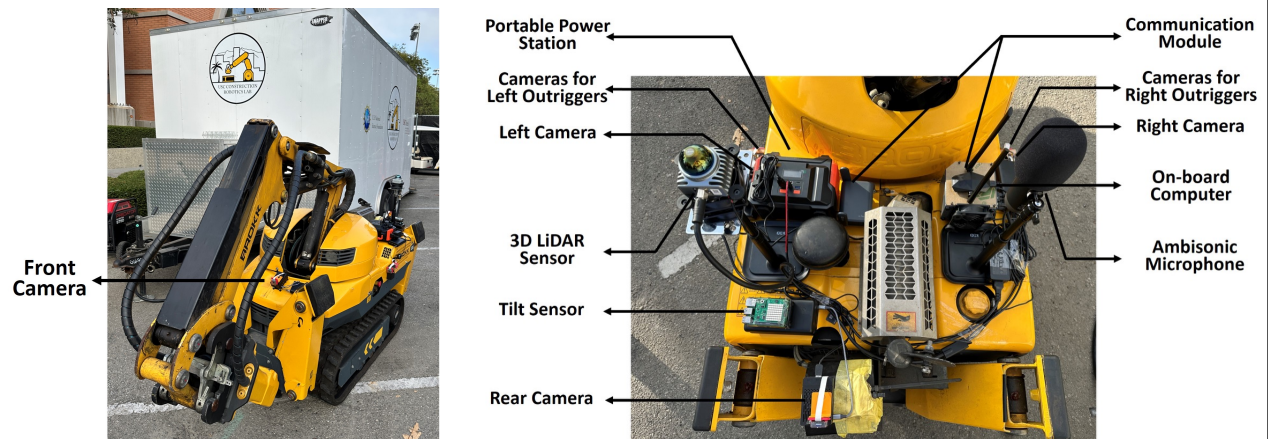


Safety and Well-Being



2

Robot



3

Teleoperation Interface



4

Data

• Interview Data

- Stakeholder Advisory Panel (SAP)
- Demolition San Antonio Tradeshow

• Physiological Data

- Heart Rate (HR) and Heart Rate Variability (HRV) [Polar H10]
- Electrodermal Activity (EDA) [Empatica EmbracePlus]
- Eye-Gaze [Kexxu Eye-Tracker]

• Ergonomics

- Posture [Rapid Upper Limb Assessment – RULA]

5

Data

• Subjective Assessments

- Workload [NASA Task Load Index – NASA TLX]
- Usability [System Usability Scale – SUS]
- Presence [Presence Questionnaire – PQ]
- Anxiety [State Trait Anxiety Inventory – STAI]
- Trust [Trust Survey]
- Self-Efficacy [Self-Efficacy Survey]
- Fatigue [Level of Fatigue Survey]
- Pain [Musculoskeletal Discomfort Survey]

6

Data

• Behavioral, Performance, and Safety

- Situational Awareness [Situation Awareness Global Assessment Technique – SAGAT]
- Time to Complete
- Driving Performance
- Operation Performance
- Number of Collisions
- Collision Avoidance Behaviors
- Worker Avoidance Behaviors

7

Interview Data – Stakeholder Advisory Panel (SAP)

“I've seen people get hurt; I've seen people get killed on a job”

“I started seeing the floor cracking and falling underneath me ... It was pretty scary”

“Even outside, guys hit power lines when they're not paying attention”

“I've lost half my hearing breaking limestone rock when I was 15 years old”

“I'm having problems getting young guys to apply for jobs”

“He had a hearing aid ... they were afraid that he wasn't going to hear a machine backing up. So, we had to fire him”

8

8

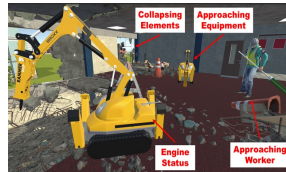
Findings – Teleoperation Interface Design Suggestions



INPUT – CONTROLLERS



FEEDBACK - VISION



FEEDBACK - SOUND



FEEDBACK - HAPTICS



PERSONALIZATION



SAFETY FEATURES



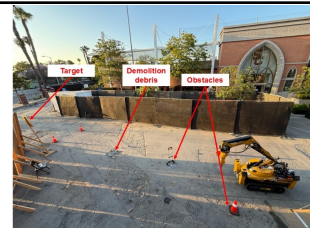
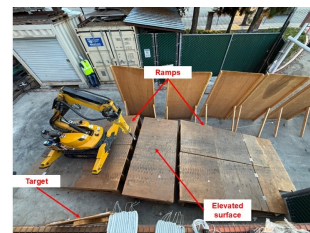
LEVELS OF AUTOMATION



COORDINATION AND COLLABORATION

9

Obstacle Course



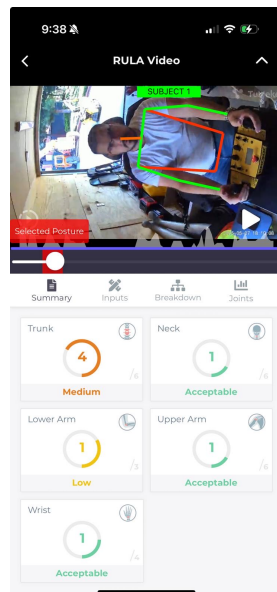
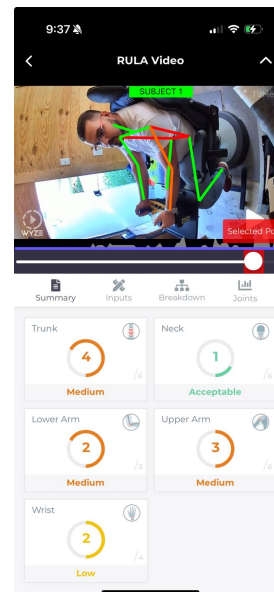
10

Behavioral, Performance, and Safety



11

Posture



12

Subjective Assessments

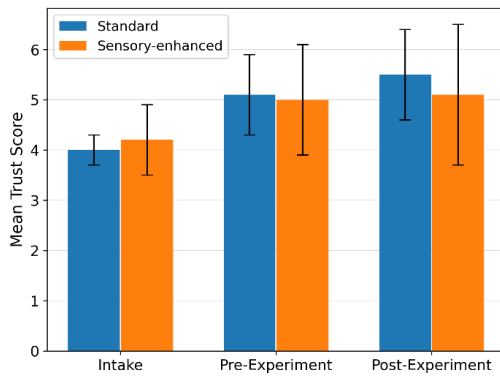


Figure 7. Mean trust scores by interface type and measurement time. Error bars represent ± 1 standard deviation.

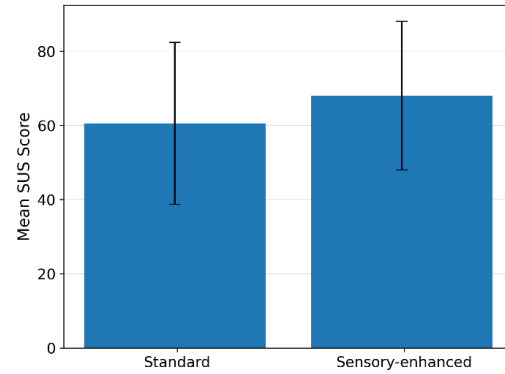
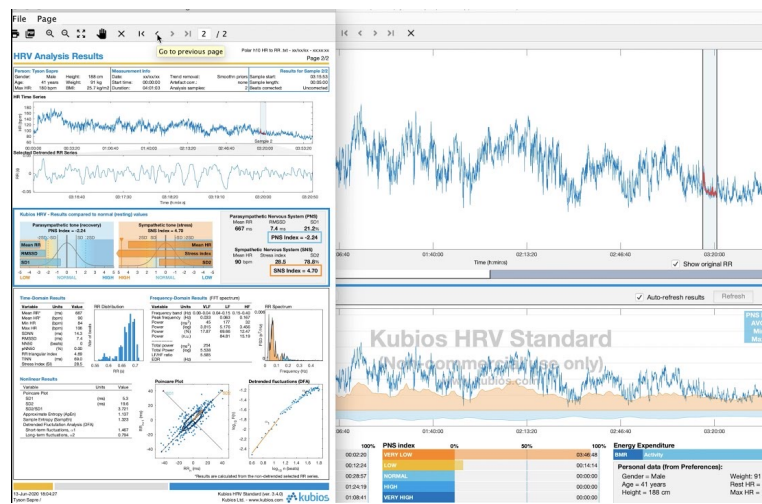


Figure 8. Mean SUS scores by interface type. Error bars represent ± 1 standard deviation.

13

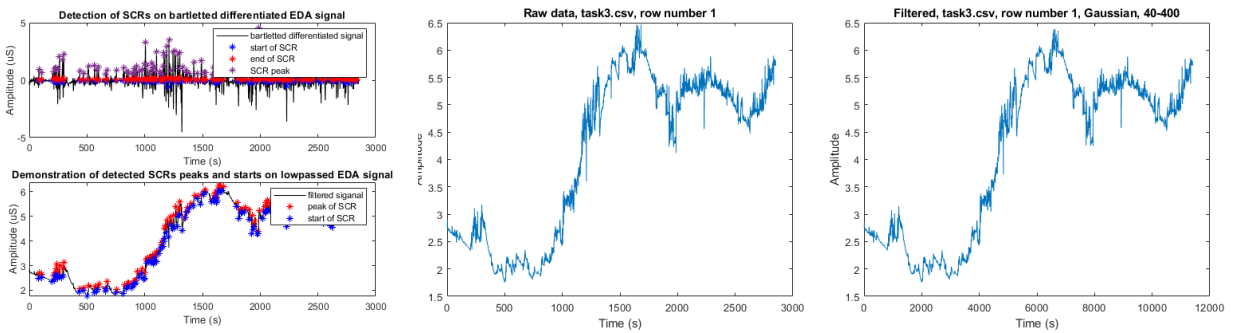
Heart Rate and Heart Rate Variability



Polar H10 Heart Rate Sensor

14

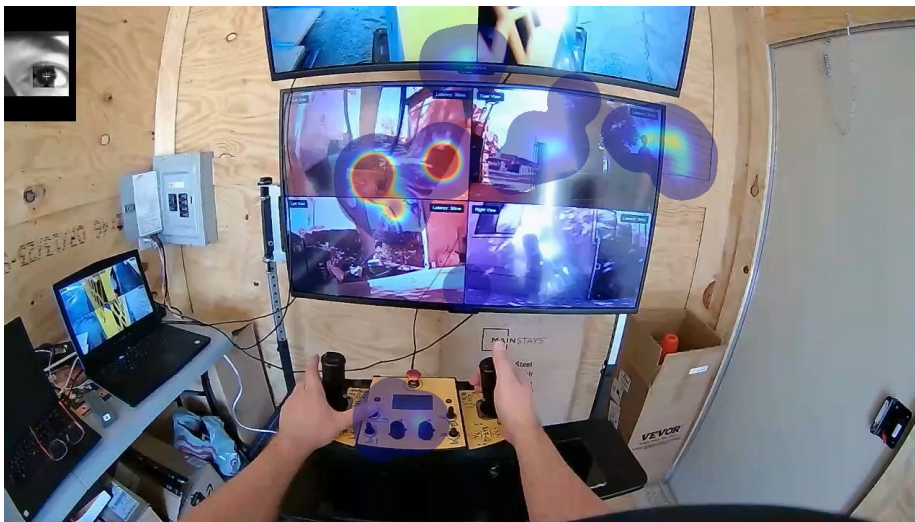
Heart Rate and Heart Rate Variability



Empatica Embrace Plus

15

Eye-gaze



Kexxu Eye-Tracker

16

Eye-Gaze



17

Acknowledgements



U.S. National
Science Foundation



USC University of
Southern California

USC Viterbi

*Sonny Astani Department of Civil
and Environmental Engineering*

BROKK®



NATIONAL
DEMOLITION
ASSOCIATION



18

18

Acknowledgments



- **NSF Grant # 2222572**
 - FW-HTF-R: Demolishing Barriers to Democratize Future Construction Operations by Providing Multi-Sensory Capabilities for Effective Remote Work

19

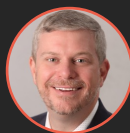
Collaboration



Dr. Burcin Becerik-Gerber
Professor and Chair, Civil Engineering, USC



Dr. Lucio Soibelman
Professor, Civil Engineering, USC



Dr. Shawn C. Roll
Professor, Occupational Therapy, USC



Dr. Gale M. Lucas
Research Associate Professor, Psychology, Computer Science, Civil Engineering, USC

20

20