Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360º Images

Hansung Kim, Luca Remaggi,
Philip J.B. Jackson and Adrian Hilton

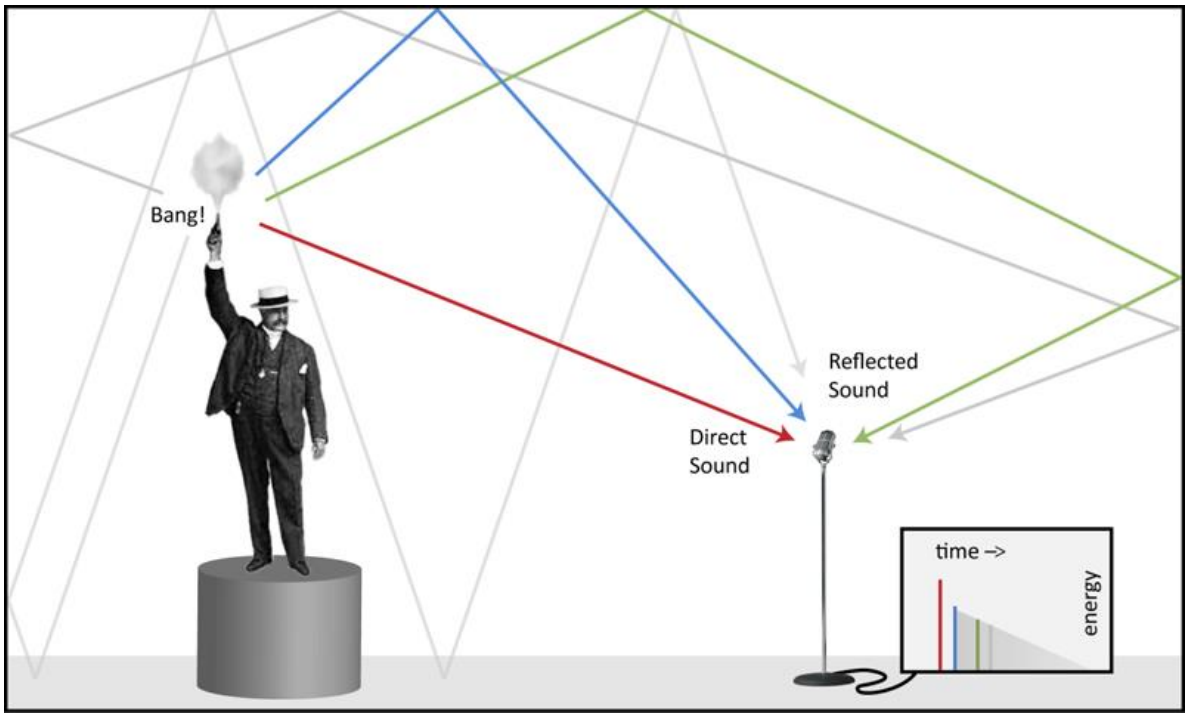
Centre for Vision, Speech and Signal Processing
University of Surrey, UK

Introduction

- Spatial Audio for Immersive Virtual and Augmented reality
  - Human perception relies on both audio and visual information
  - Spatio-temporal synchronisation of sound with visual information improves the sense of presence in VR/AR environments (Larsson 2010)

Introduction

- Best way to reproduce the acoustic of spaces
  - Measuring Room Impulse Response (RIR)
- Problem of RIR measurement for practical applications
  - Too invasive
  - RIR is valid only at a single point of measurement for a static scene



Introduction

- How to evaluate?
  - Coherency – audio-visual information
  - Plausibility (internal reference) – for VR applications
  - Authenticity (external reference) – for AR applications

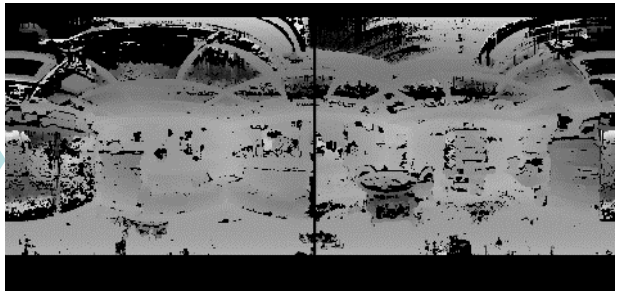
Overview

UNIVERSITY OF SURREY

- Goal
  - Simple and practical system to estimate room acoustic for plausible reproduction of spatial audio using 360º cameras
- Assumptions
  - Human audio perception is not sensitive enough to recognise differences of sound from the change of geometrical details (JUDD 1932)
- Contributions
  - Approximated room geometry estimation
  - Acoustic room modelling using visual semantic segmentation
  - Objective evaluation of estimated room acoustics
  - VR implementation

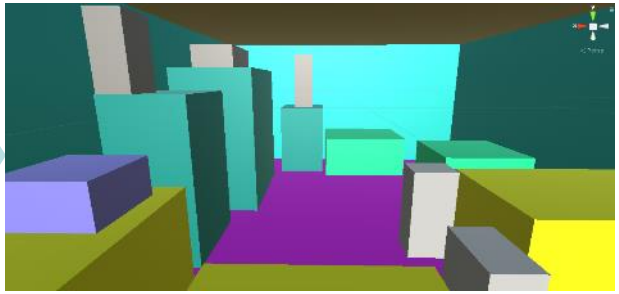
Overview

360 stereo image pair

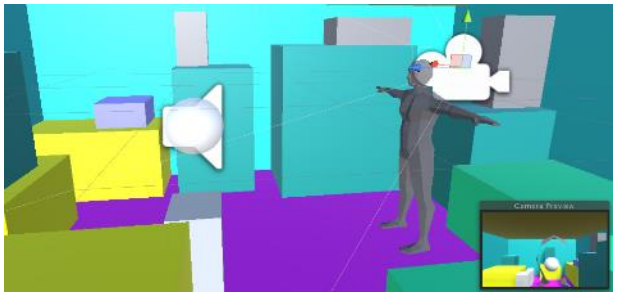


Depth estimation

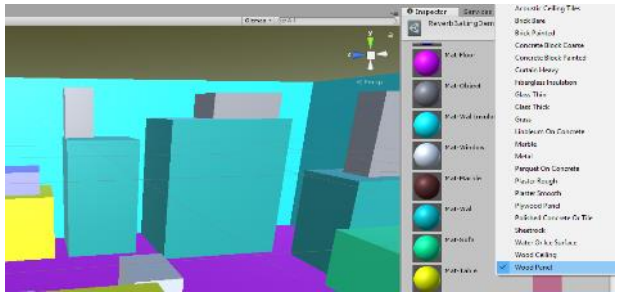
Object segmentation



3D room layout reconstruction



VR scene with spatial audio

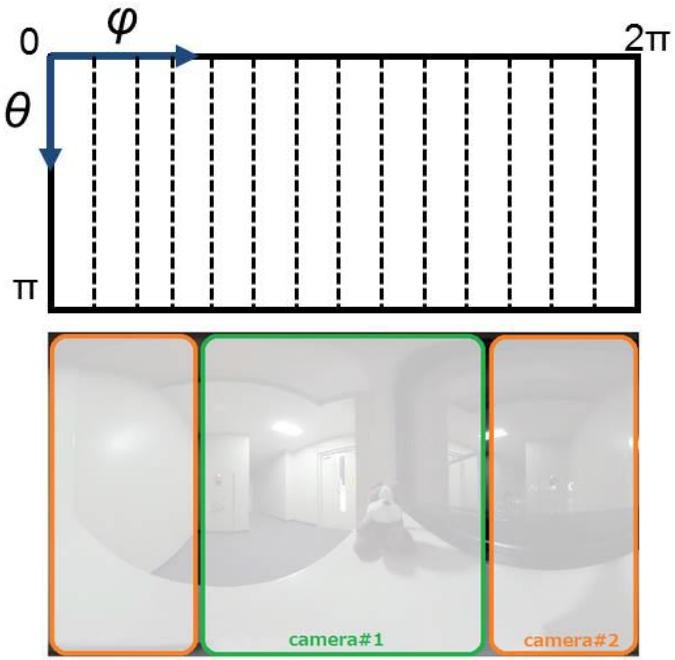


Acoustic material mapping

Capture system

- Vertical 360 Stereo Capture
  - Simple 1D matching for depth estimation
  - Real-world scale depth without calibration
  - Less occlusion between cameras
  - Higher accuracy for side regions

Ricoh Theta

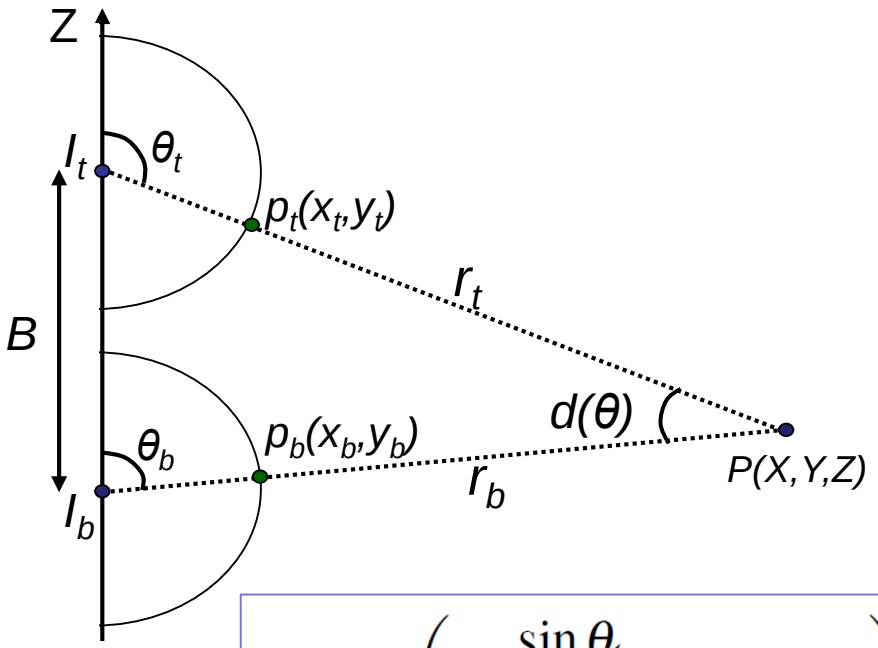


Equi-rectangular image

Captured vertical stereo images

Geometry Reconstruction

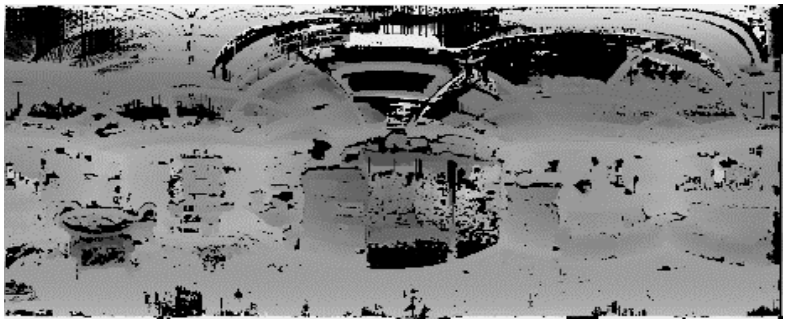
- Spherical stereo geometry
  - Feature-based dense block matching method*



$$r_t = B / \left(\frac{\sin\theta_t}{\tan(\theta_t + d)} - \cos\theta_t \right)$$

$$d(\theta) = \theta_t - \theta_b$$

Org image

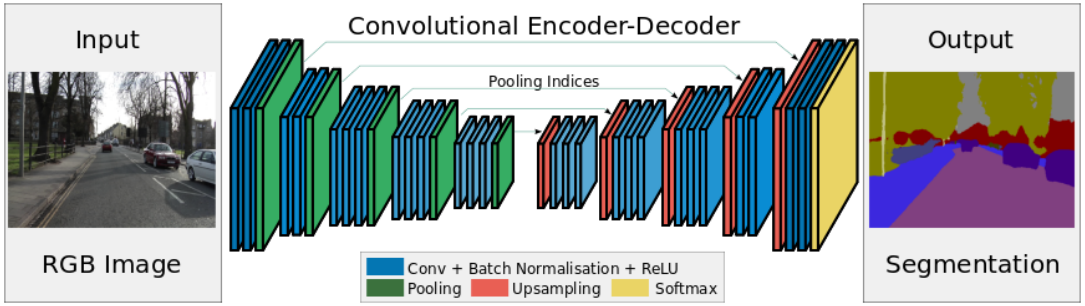


Sparse depth with occlusion

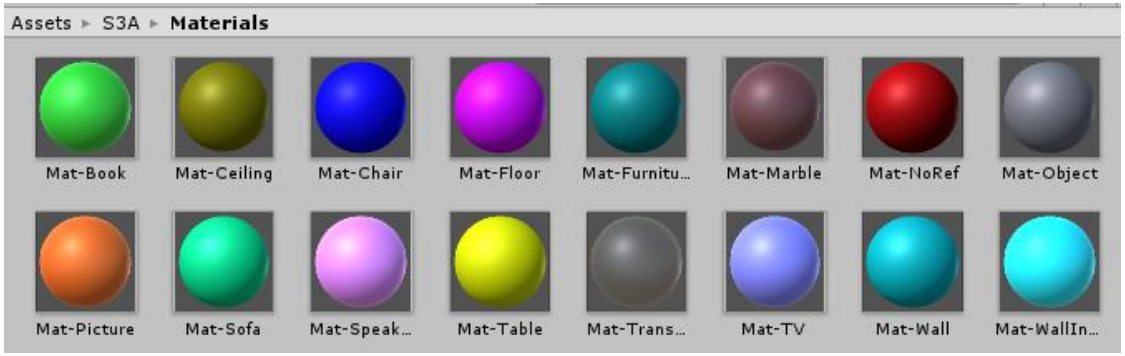
H. Kim and A. Hilton, "3D scene reconstruction from multiple spherical stereo pairs", *International Journal of Computer Vision*, 2013

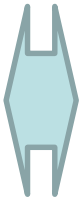
Object and material recognition

- Semantic object segmentation and mapping to acoustic materials
  - SegNet* for semantic image segmentation



  - Mapping materials and acoustic attributes**





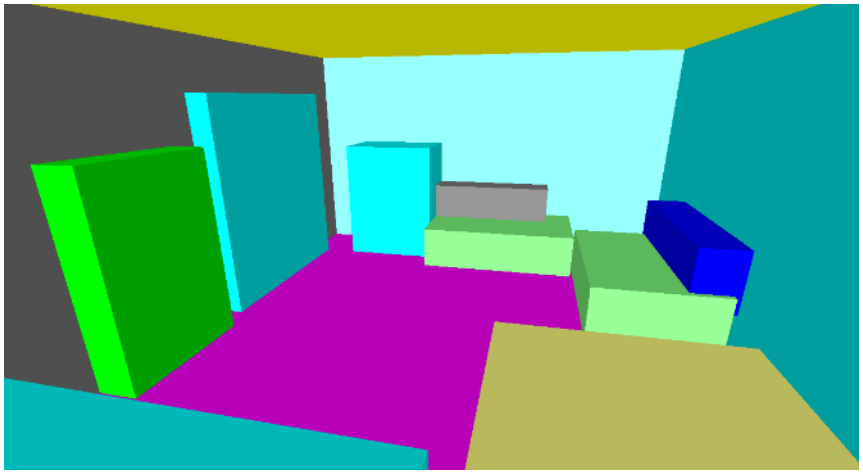
Transparent
Acoustic Ceiling Tiles
Brick Bare
Brick Painted
Concrete Block Coarse
Concrete Block Painted
Curtain Heavy
Fiberglass Insulation
Glass Thin
Glass Thick
Grass
Linoleum On Concrete
Marble
Metal
Parquet On Concrete
Plaster Rough
Plaster Smooth
Plywood Panel
Polished Concrete Or Tile
Sheetrock
Water Or Ice Surface
Wood Ceiling
Wood Panel

* V. Badrinarayanan, A. Kendall and R. Cipolla "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *IEEE Trans. PAMI*, 2017.
** T. Cox and P. D'Antonio, Acoustic absorbers and diffusers, third edition: theory, design and application. CRC Press, 2016.

3D room modelling

- Final 3D room geometry reconstruction
  - Cuboid fitting and labelling
    - Fitting with point cloud occupancy

Reconstructed Room geometry with texture



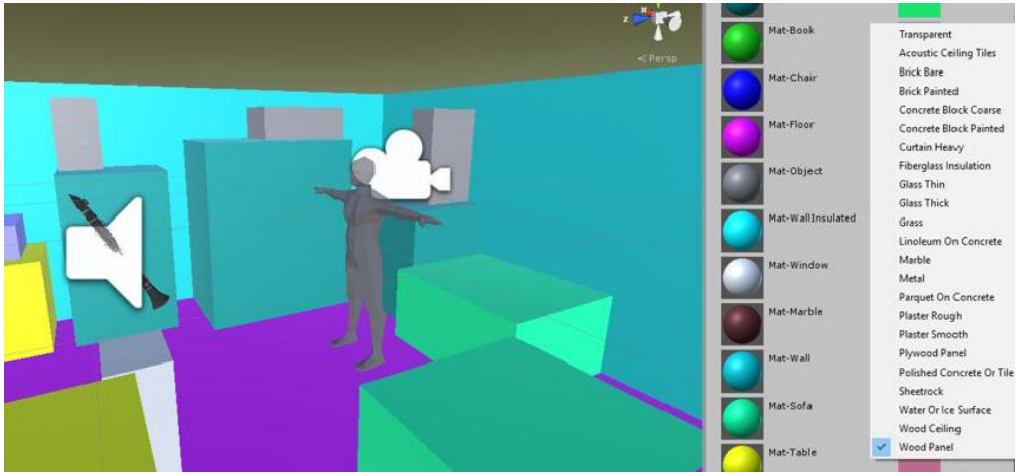
Room geometry with object labels

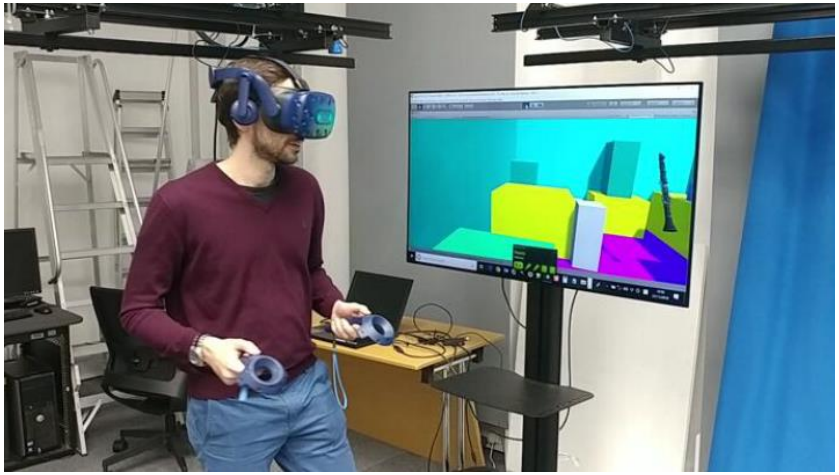
bed books ceiling chair floor furniture objects

picture sofa table TV unknown wall window

VR Scene with Spatial Audio

- Metadata format
  - OBJ for geometry
  - JSON for scene and acoustics information
- VR Platform
  - Unity with Google Resonance Audio package
  - Alternative option: Unreal / Steam Audio

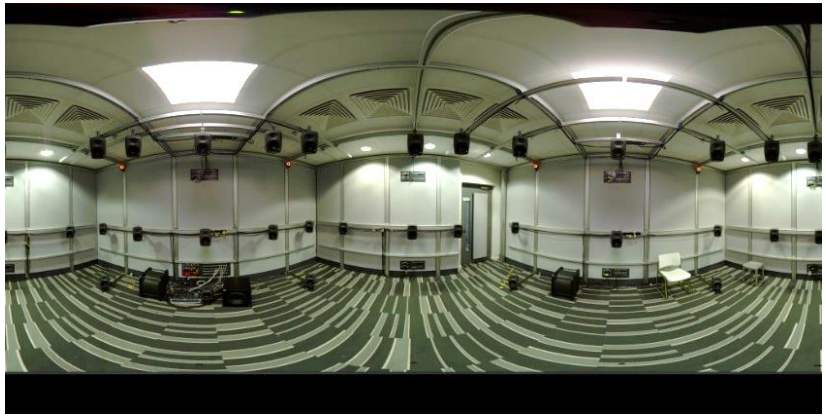


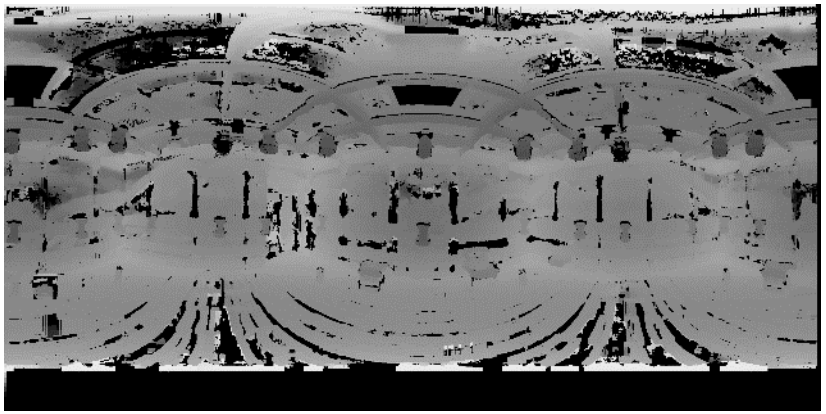


Experiments

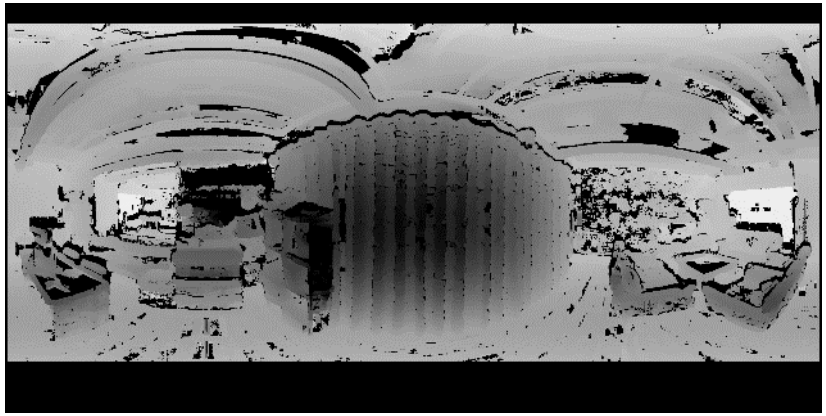
UNIVERSITY OF SURREY

- Datasets and estimated depth maps





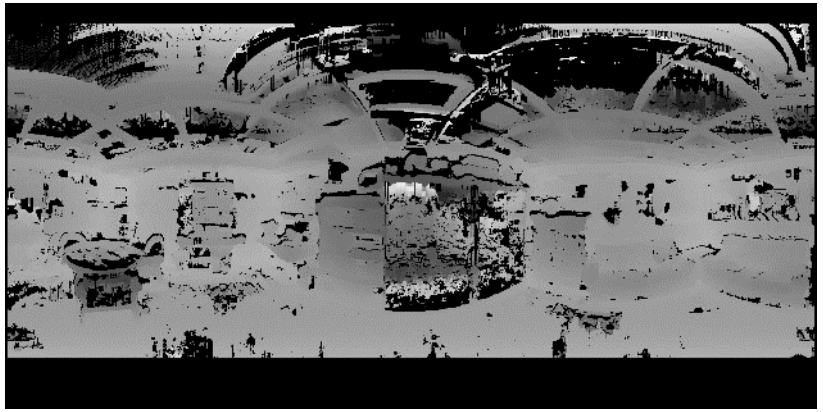
Listening room (LR)



Usability Lab (UL)

Experiments

- Datasets and estimated depth maps



Meeting room (MR)

Studio (ST)

Experiments

- Computational load
  - Geometry reconstruction
    - Processor: Intel Core i7 3.40 GHz CPU with 32G RAM
    - Processing time: around 5 mins
  - Semantic segmentation
    - Processor: NVIDIA Tesla M2090 GPU with 5GB RAM
    - Processing time: around 3 mins

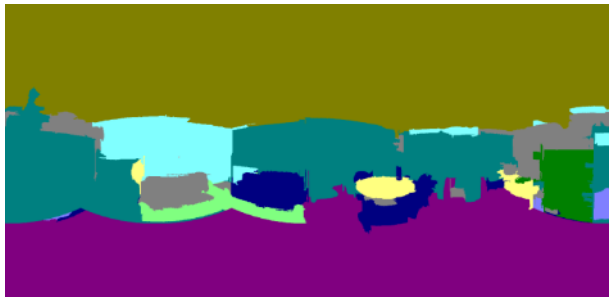
- Evaluation of room layout reconstruction

Ground-truth (m^3)	Estimated (m^3)	Error (%)
$5.61 \times 4.28 \times 2.33$	$5.52 \times 4.35 \times 2.36$	1.3
$5.57 \times 5.20 \times 2.91$	$5.92 \times 4.95 \times 2.95$	27.0
$5.64 \times 5.05 \times 2.90$	$5.77 \times 5.17 \times 2.98$	7.6
$17.08 \times 14.55 \times 6.50$	$16.53 \times 14.87 \times 5.70$	13.2

Experiments

- Object recognition and segmentation results






Eigen (ICCV 2015)

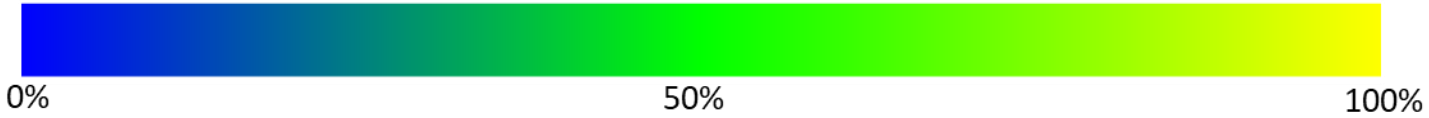
SegNet

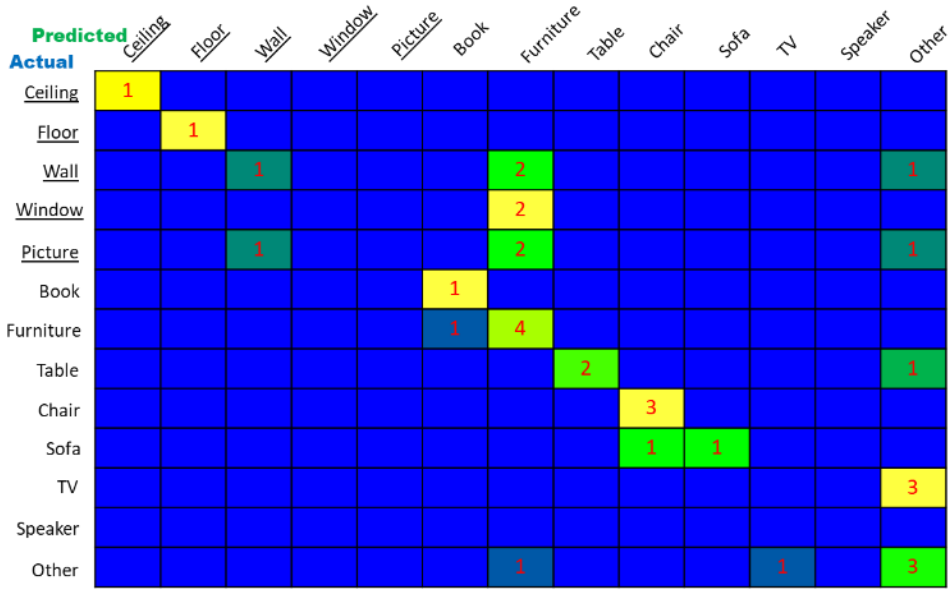
bed books ceiling chair floor furniture objects

picture sofa table TV unknown wall window

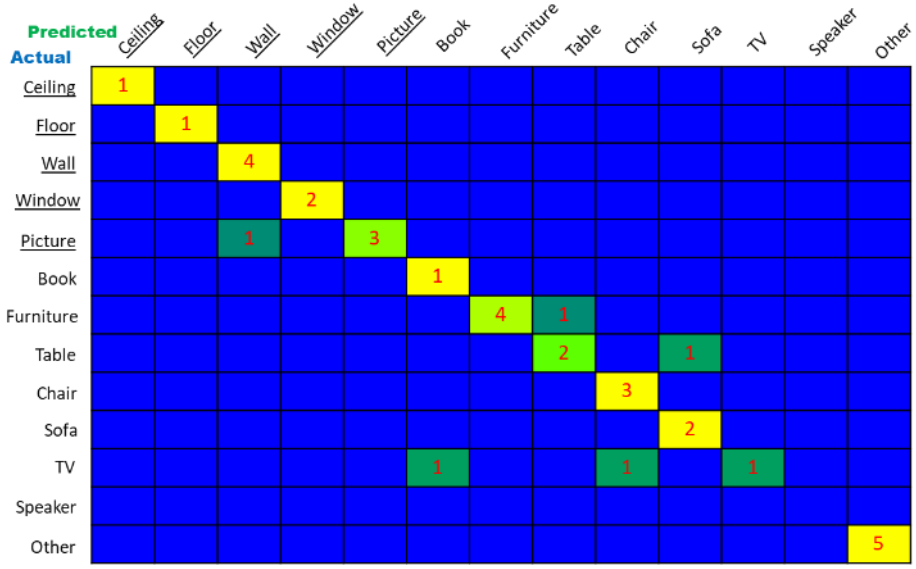
Experiments

- Object recognition result (heat map) for MR





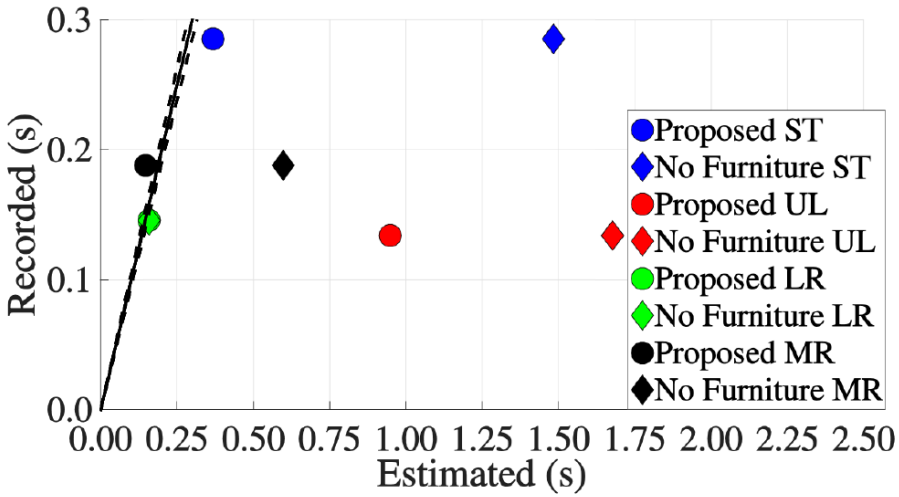
Eigen (ICCV 2015)



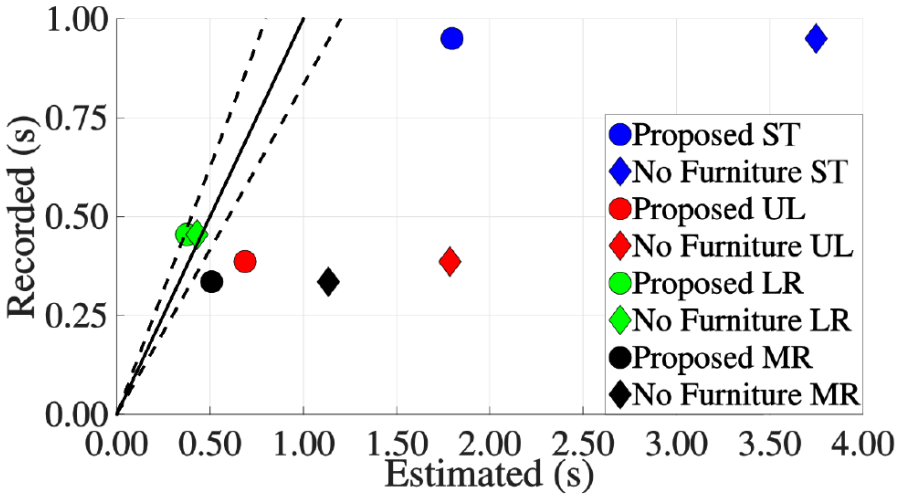
SegNet

Experiments

- Room Impulse Response (RIR)
  - Ground-truth RIRs vs. Estimated RIRs
  - Evaluation
    - Early Decay Time (EDT) – early reflections
    - RT60 – late reverberation
  - Just-Noticeable Difference (JND) level
    - 5% for the EDT (Vorlander 1995) and 20% for the RT60 (Meng 2006)



(a) EDT



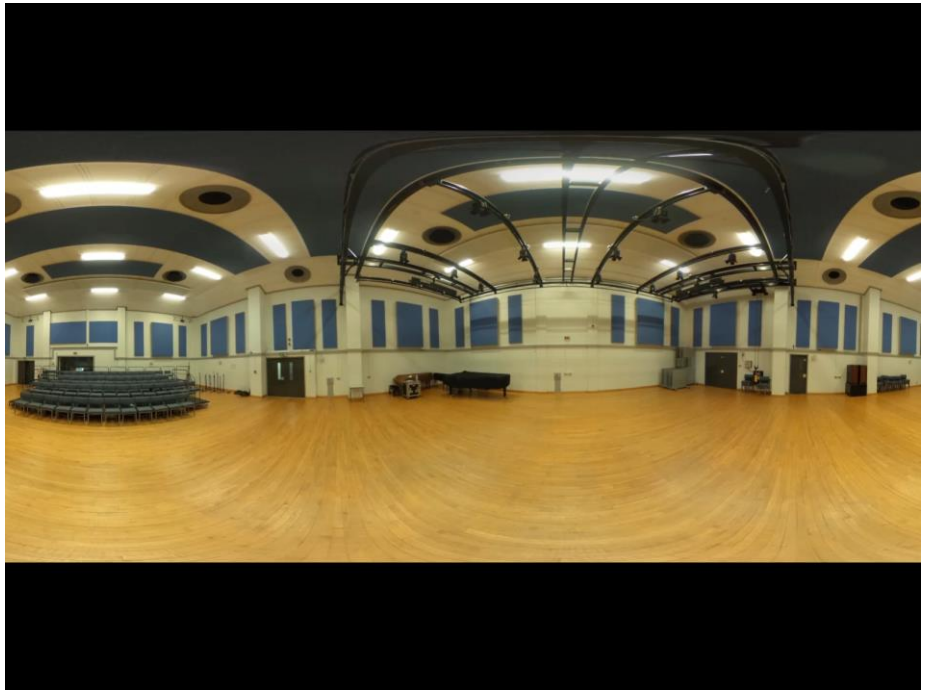
(b) RT60

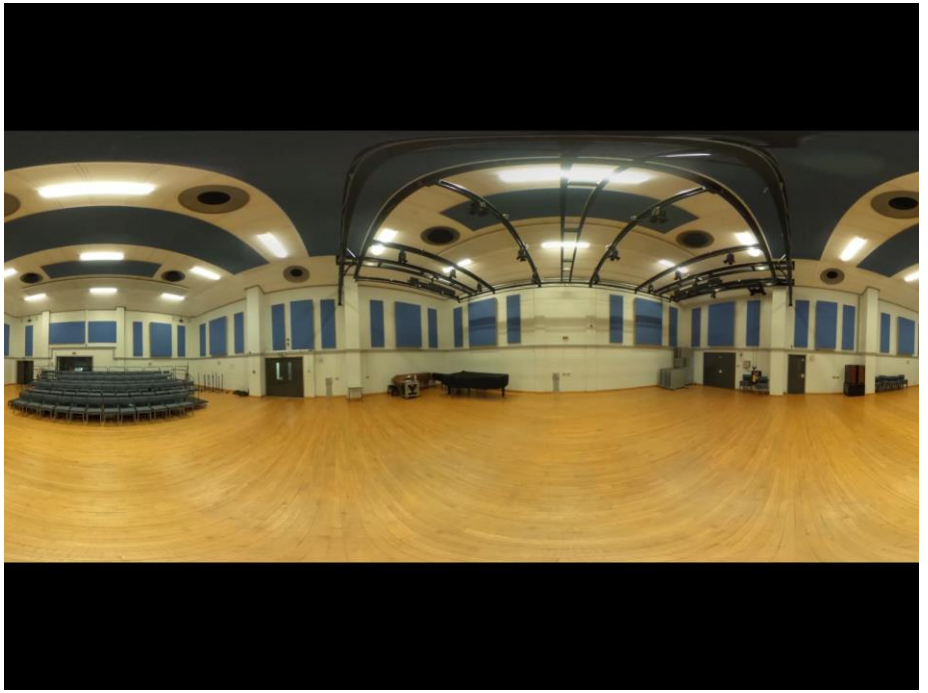
Experiments

UNIVERSITY OF SURREY

- Sound Rendering

  -Ground-truth RIR vs. Estimated RIR





VR Scene Rendering

UNIVERSITY OF SURREY

- VR Demo on Unity with Google Resonance Audio
  - Interactive real-time spatial audio rendering
  - Comparison of Open space vs. Room only vs. Room with objects

VR Scene Rendering

UNIVERSITY OF SURREY

- VR Demo on HTC VIVE Pro headset
  - Audio in this video has been recorded using an external speaker

Conclusion

UNIVERSITY OF SURREY

- Summary
  - Vision-based 3D structure and acoustic property estimation system
  - Reproduction of plausible spatial audio in VR/AR environment
  - VR implementation

- Future work
  - Robust material detection
  - Subjective evaluation of plausibility in VR reproductions
  - Perception with/without visual cue

UNIVERSITY OF SURREY

Thank you very much!

Hansung Kim

h.kim@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey, UK