Acoustic Information Retrieval for Interactive Sound Rendering in Virtual Environments

Mattia Colombo

A report submitted as part of the requirements
for the degree of Research to PhD in Computing

Birmingham City University

March 2024

Supervisors
Dr Carlo Harvey, Dr Maite Frutos-Pascual

Abstract

The planned thesis work involves adopting computer vision techniques in the process of decomposing complex scenes to recognise acoustic characteristics of space, determining physical and structural features of complex scenes. The experiments presented demonstrate applications of scene understanding techniques to game scenes and virtual reconstructions of real space to determine acoustic properties of scene geometry for automating realistic sound rendering, identifying the current state of automatic acoustic material recognition for virtual environments and proposing a novel evaluation framework to test objective and subjective accuracy against measurements from real environments. Proof-of-concept systems have been tested on state-of-the-art acoustic renderers to demonstrate their efficiency in offline procedures. Current directions are aimed at designing end-to-end pipelines for interactive, real-time applications, with the ambition of adopting computer vision to understand the acoustic space, even in contexts of dynamic geometry typical of Augmented Reality platforms, where the acoustic space is constantly updating based on the surrounding, real world.

Preface

The Acknowledgements section may be used to thank your supervisor, family, research funding bodies, or any other applicable individuals or institutions.

Declaration

I confirm that the work contained in this PhD project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged, and all verbatim extracts are distinguished by quotation marks.

Signed  Date: 1st March, 2024

Mattia Colombo

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

CNN	Convolutional Neural Network.
GA	Geometrical Acoustics.
VE	Virtual Environment.

Chapter 1

Methods for Acoustic Characteristics Retrieval from Complex Virtual Environments

The following Chapter introduces two systems to retrieve acoustic characteristics from space surrounding users in AR, providing fundamental building blocks for subsequent acoustic rendering techniques to simulate sound transmissions between an entity in AR space and the user. The acoustic characteristics extracted from the physical and virtual scene experienced by the user allow context-aware rendering, enabling the acoustic renderers discussed in Chapter **??** to approximate sound waves with geometrical primitives and compute propagation paths from sources to a listener, i.e. the user. Propagation paths, calculated by intersecting rays (or other primitives) with the environment, can respond to the physical properties of surfaces or scene entities, attributing characteristics to portions of the scene geometry. This allows propagation paths to model how acoustic energy propagating from an emitter in AR space is affected by diverse surfaces in space, enabling context-aware auditory interactions.

The following Chapter is structured around two main sections presenting novel workflows for acoustic information retrieval:

1. a **camera-based** system to understand complex scenes and project acoustic information based on visual renders,
2. and a **texture-based** extension that improves the above by abstracting away from camera render and analyses texture data from scene geometry.

The two methods predict acoustic characteristics of space from their visual representations to inform sound rendering and produce believable acoustic stimuli in interactive applications. The methods are tested on various complex scenes, ranging from authored virtual

scenes to reconstructions of physical space with LiDAR scanners, emulating input environments that are typically available to AR HMDs. The methods presented demonstrate applications of scene understanding techniques to virtual environments and digital reconstructions of real space to determine acoustic properties of scene geometry for automating realistic sound rendering, and they are evaluated on state-of-the-art acoustic rendering systems, measuring objective and subjective metrics relating to simulated soundfields.

1.1 Introduction

Modern approaches to audio rendering can be broadly categorised into geometrical acoustics (GA) methods (Savioja and Svensson, 2015) or finite or boundary element methods (FEM/BEM) (Hulusic et al., 2012). Finite elements methods, such as wave-based audio renderers, often require the positions of sound sources and listeners, as well as the scene geometry and associated materials, tagged with acoustic absorption coefficients for each material (Deines et al., 2006; Raghuvanshi and Snyder, 2014). This process is commonly performed manually, often at significant cost, due to the human-in-the-loop. This Chapter represents a first step towards creating an automatic process for the generation of such input data for pre-computed audio rendering pipelines in the absence of knowledge of geometry and material information of a complex scene. Specifically, we propose a proof-of-concept system for vision-based material information retrieval, which allows for tagging of an object's acoustic properties based on its image features, which are then mapped to frequency-dependent absorption coefficients. The system tags meshes in VE representing boundaries in sound propagation paths having a noticeable perceptual impact, facilitating the use of GA or FEM/BEM-based acoustic renderers on complex scenes. The contributions of this Chapter are:

- a novel methodology for acoustic material tagging using a camera-based system and computer vision techniques to infer from the scene;
- an alternative system that abstracts from the use of cameras and operates on reconstructed virtual geometry;
- objective and subjective methologies to evaluate the efficacy of the systems comparing against objective and subjective metrics;
- a proof-of-concept demonstration stemming from the development of the two systems

However, the accuracy of the acoustic simulation depends on material information assigned to the scene geometry. The scene geometry, tagged with frequency-dependent absorption and scattering information, determines how sound behaves in space and affects the resulting wavefield. In games development, the process of tagging materials with appropriate acoustic data often requires the work of experts, raising costs and resources needs for large scenes.

Advances in acoustic modelling propose automatic tagging of acoustic data to scene geometry using convolutional neural networks to tag acoustic materials from stereo photographs of real environments Li, Langlois and Zheng (2018). Alternatively, using a recent camera-based material tagging system tags geometry in VEs, applying scene understanding algorithms and filtering complex geometry based on its perceptual impact on the resulting acoustic model. At the core of these methods lies the problem of scene segmentation. In computer games, often, a set of meshes composes a scene, where each mesh represents an object in the scene. Acoustic materials are often assigned to all triangles composing a given mesh, allowing audio engineers to group scene geometry when assigning acoustic data. Hence, the resulting acoustic model's accuracy depends on the separation of the geometry, where ideal conditions would have each triangle mapped to its specific acoustic data. A naive approach would have the entire geometry mapping to a single acoustic material. Besides, the representation of materials in real and virtual environments adds further dimensions to the material tagging problem due to complex links between the visual representation of materials of an object and its perceptual effects on the soundscape of the environments in which it exists.

1.2 Camera-based Acoustic Material Tagging

1.2.1 System Overview

The camera-based acoustic material tagging pipeline is the first approach to attribute physical properties to portions of scene geometry representing a virtual environment. At a high-level overview, the pipeline adopts a Convolutional Neural Network to understand the scene features of the environment, expressed as pixel-wise semantic information predicted from camera renders obtained by a perspective camera within the virtual environment. The network generates a prediction map where pixel-wise semantic information maps to acoustic absorption information, e.g. a pixel belonging to a wooden table in the virtual scene may be attributed with "wood" semantics, which maps to absorption data related to the semantic material. Using camera transformation matrices, semantic information is projected onto the scene from camera space. The system has two phases: *training* and *inference*.

Training Phase The training phase used an environment with scene geometry tagged with ground truth acoustic materials, reproducing the workflow of audio engineers authoring virtual complex scenes by assigning acoustic properties to scene elements. The system is trained on these scenes by generating a set of camera renders with associated segmentation mapping to the correct pixel-wise category.

Inference Phase Once the network is fit on the ground truth set, it is deployed to a set of test scenes with no material tags, obtaining camera renders across uniformly spaced camera probes scattered around the walkable space of each scene. The scene is then tagged

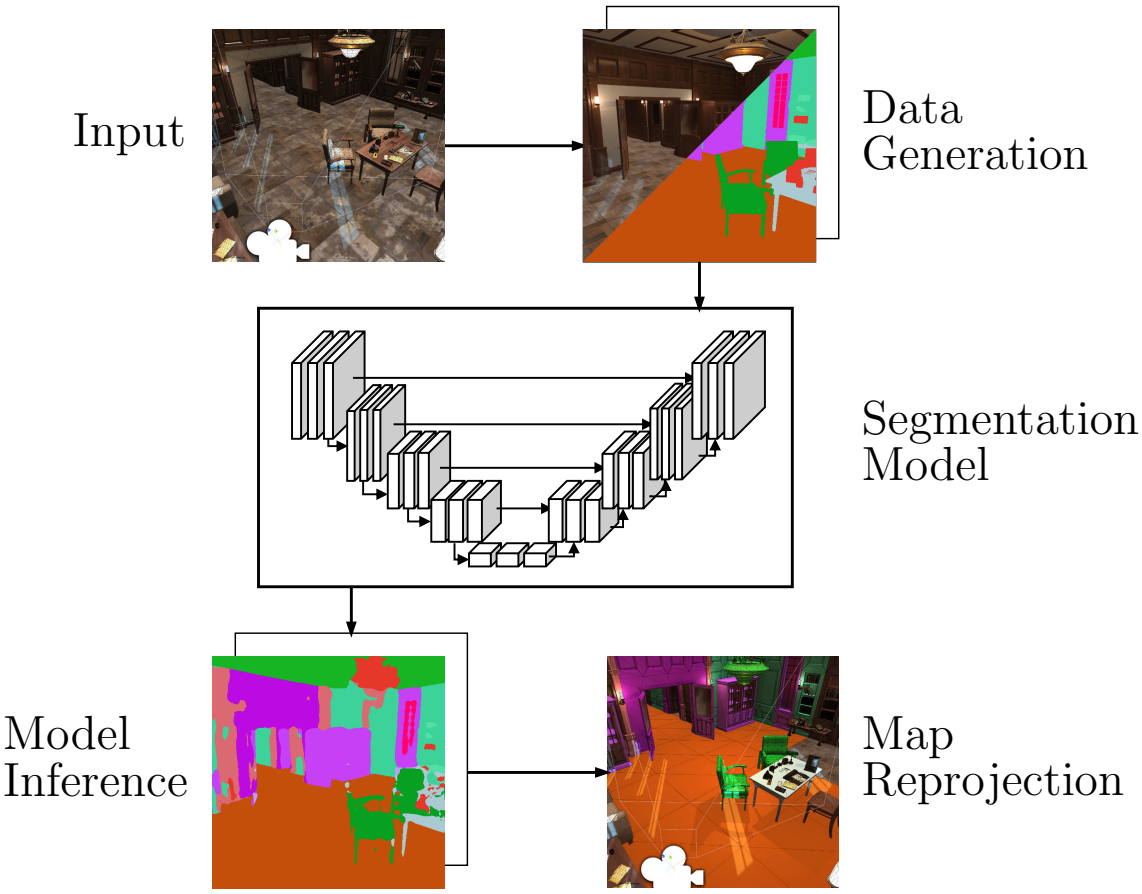


Figure 1.1: Overview of the proposed system: given a set of views captured in a VE, a convolutional neural network trained on samples from our scenes performs semantic segmentation. The predicted semantic maps are then reprojected onto objects in the virtual scene, associating predicted semantic classes with acoustic profiles that are attributed to the scene geometry. Tagged scene geometry provides input to acoustic renderers or physically-based audio engines for sound propagation or synthesis tasks.

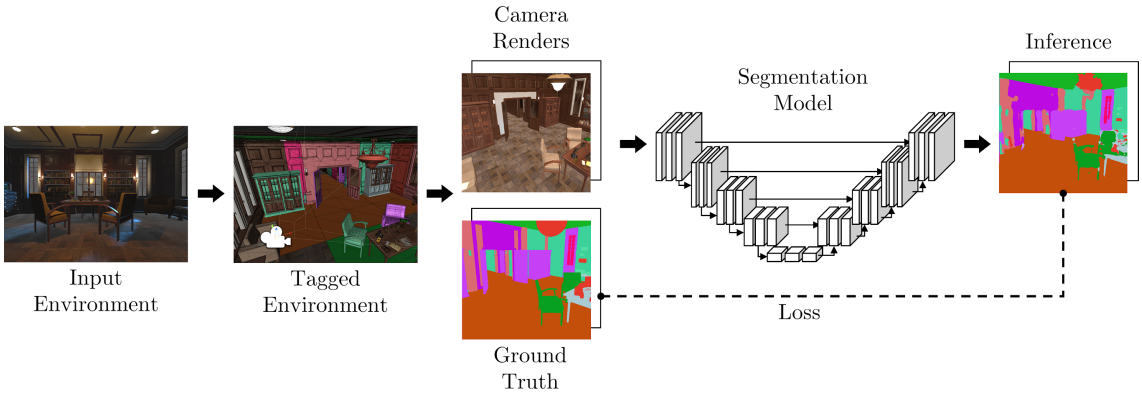


Figure 1.2: Training phase of the system pipeline: manual acoustic material tagging is performed on an input environment, generating pairs of camera renders and segmentation maps via ray casting, which are then used to train and evaluate the convolutional neural network.

by predicting segmentation maps and reprojecting acoustic material to virtual geometry.

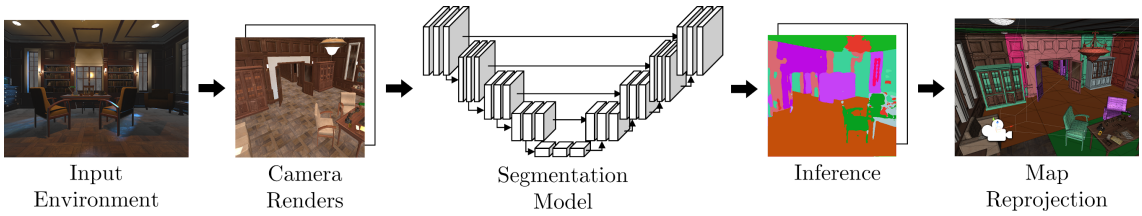


Figure 1.3: Inference phase of the system pipeline: camera renders are generated from an input environment, providing input to the convolutional neural network. With camera transformations, segmentation maps generated by the network are used to attribute acoustic properties via semantic class mapping to scene geometry.

1.2.2 Method

Based on advances in scene understanding and the current state of wave-based audio renderers, we design an architecture that enables the process of semantic mesh labelling in complex scenes and associating every category with a frequency-dependent acoustic absorption function. Methods based on perceptual metrics should consider only meshes that are relevant to the acoustic environment. Scene understanding methods and inference should be optimised depending on the scene geometry.

Input

We demonstrate the usage of the pipeline in two scenes: an open, urban environment, *City*, and an indoor wooden room, *Office*. City has $6.3M$ triangles and $8.6M$ vertices. Office has $3.3M$ triangles and $3.8M$ vertices. We define a set of classes using tables of measured acoustic absorption of construction materials, where materials are grouped in categories specifying a vector α of absorption coefficient values across an approximated equivalent rectangular bandwidth frequency scale ranging from $125Hz$ to $4kHz$. For every major material category that exists in our material database, we define two levels, representing the low and high bounds of mass density ρ in that category. Mass density is a physical property allowing for the acoustic properties of two objects made of the same material to be perceptually distinguishable Giordano and McAdams (2006). We define 23 material classes constituted by the two density levels for each of the 11 material categories and an additional class representing "air", see Table.

Data Generation

We implement the core material tagging system in Unity using a camera located across probe points of a complex scene. Segmentation masks associated with each view are generated by ray-casting through each point of C_n, the *near* camera clipping plane, to ∞. For this case, we exclude wavelength-based strides to maximise segmentation accuracy. The areas where rays intersect with C_f, the *far* camera clipping plane, are labelled as air; objects that are hit by a ray determine the pixel value of the mask, which points to

the corresponding material. The dataset consists of 3500 labelled images with 512×512 pixel resolution, split into 3000 training images and 500 validation images. In City and Office, rendered views are generated in different regions of the environments. The different regions delimit spaces for the collection of training and validation data. For each delimited region, sets of points are scattered to cover the walkable space. The camera position is interpolated across points in these sets and rotated between 0 and 2π along the azimuth and between 0 and π along the elevation.

Semantic Segmentation Model

A convolutional neural network is used to discriminate materials of objects represented in the camera-rendered views. This task is performed with pixel-level semantic segmentation using a ResNet-34-based UNet Ronneberger, Fischer and Brox (2015). The ResNet backbone offers a topology that is easy to train and has excellent generalisation performance. It also provides a compromise between accuracy and the number of parameters He et al. (2016). The model, pre-trained on the ImageNet dataset, is fine-tuned for 70 epochs minimising focal loss Lin et al. (2017). Table **??** shows the information on the networks trained, including the total number of parameters, F_1-score, intersection-over-union (IOU) and the number of epochs.

Table 1.1: CNNs used to produce segmentation maps with the camera-based system, detailing architecture type and performance metrics.

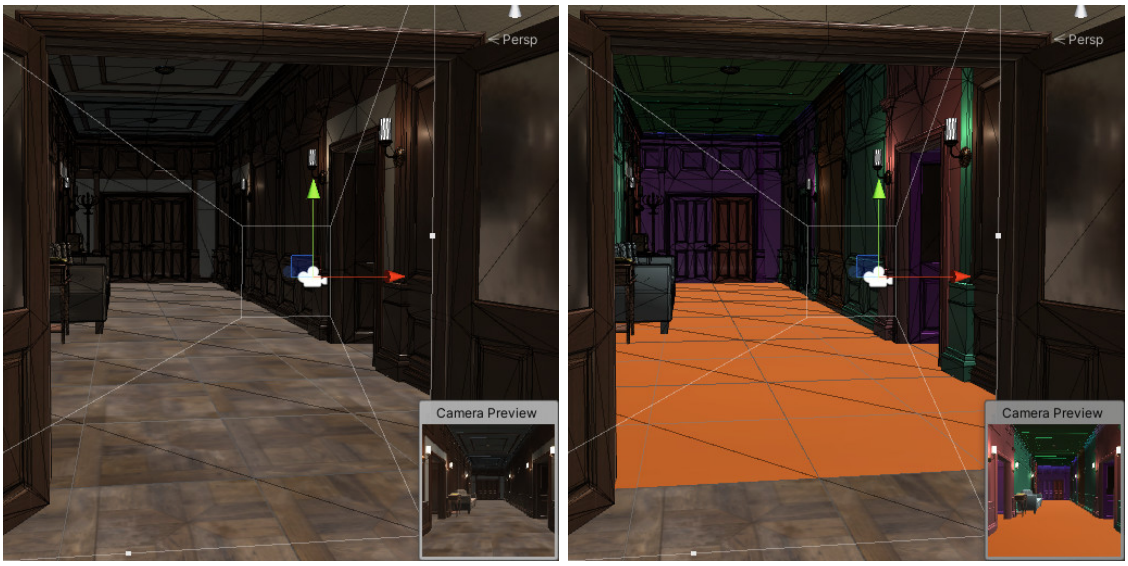
Architecture	**Backbone**	**Capacity**	**F_1-score**	**IOU**	**Epochs**	**Loss**
Unet	ResNet-34	$24.5M$	0.51	0.58	70	$2 * 10^{-3}$
Unet	ResNet-50	$32.6M$	0.54	0.47	70	$7 * 10^{-4}$

Model Inference

The output of the model is an $m \times n \times k$ matrix M, where m and n are the input image resolution and k is the number of classes. For each pixel, the k channels encode a probability distribution across the classes. Per-pixel classes are determined with the member having the highest presence probability, reducing M to an $m \times n$ matrix where entries encode the semantic class (see Table 1.2). In addition, counts of unique entries in M determine the number of pixels describing the associated material. With scaling based on the distance between a target object and C_n, this allows material exclusion below a threshold.

Map Reprojection

Using the segmented images, meshes are labelled by raycasting through C_n divided in strides. Based on the distance of every Mesh Renderer Unity object inside the camera frustum, the stride size is determined by the lowest structural dimension of each mesh, scaled according to its distance to C_n. This allows consideration of filtering objects by wavelength, $\lambda = 0.7m$, from the reprojection process. This is because some objects are too



(a) Input Camera Render (b) Re-projected Acoustic Materials

Figure 1.4: Map Reprojection performed on an input camera render: based on a predicted segmentation map, acoustic materials determined by pixel-wise semantic information is re-projected onto scene objects captured by the camera.

small to have a significant impact on the human perception of the soundscape Pelzer and Vorländer (2010). Through this level of detail (LOD) graduation, we reduce the analysed scene geometry, excluding structures having smaller perceptual impacts on the resulting acoustic model. Among the factors affecting the performance and accuracy of acoustical simulation methods is the polygon count of the acoustic VE, dependent on the complexity of a scene and the presence of detail and small objects. In acoustic environments, smaller structures on surfaces tend to induce scattering of incidental high-frequency waves reflected, and they are neglected by lower frequencies whose wavelength is greater than the structure dimensions. As a consequence, the amplitude of lower frequencies is more likely to be affected by first-order room modes, given by walls and large boundaries, affecting the frequency response of the sound field resulting in a more noticeable perceptual effect. As opposed to frequencies higher than the Schroeder Frequency, which tend to scatter chaotically Kuttruff (2016); Blauert (1997). A pilot study of this perceptual optimisation demonstrates up to 123% of performance gains in offline and real-time acoustic modelling implementations. Small structures on surfaces can therefore be excluded from modelling processes. Results of this process can be seen in Fig. **??** where smaller objects than this λ of 0.7m do not receive a material tag. Considering meshes inside the camera frustum, marching cubes algorithms should enable the geometry reduction considering the size of cubes dependent on the wavelength of the lowest frequency Pelzer and Vorländer (2010).

CASE WHERE MULTIPLE PIXELS PROJECT ONTO THE SAME OBJECT

Table 1.2: Common acoustic absorption coefficients with ranges (low-high) of α absorption characteristics across the frequency bands for those material types. It should be noted that these are regressions and averages of generally adopted materials and existing measurement tables; realistic or surveying acoustic simulations should adopt absorption measurements of real materials.

Material	**Low** α	**High** α
Glass and glazing		
Masonry walls		
Stud-work & lightweight walls		
Wood & wood panelling		
Floors		
Panels & doors		
Other		
Wall treatments & construction		
Ceilings		
Mineral wool & foams		
Audience & seating		

1.2.3 Acoustic Materials

Preliminary Evaluation

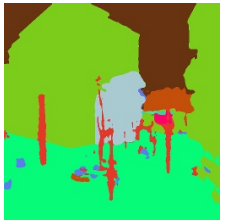
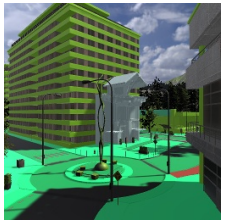
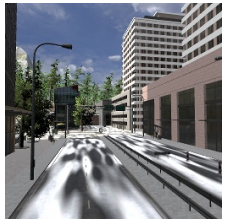
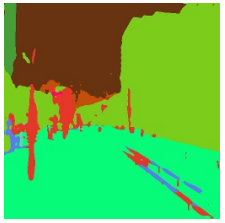
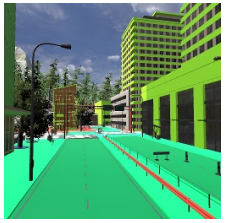
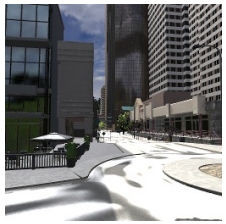
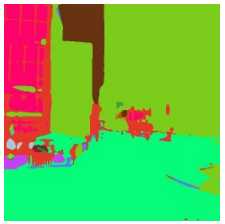
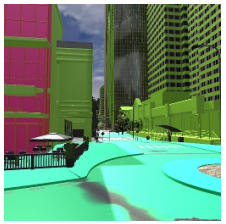
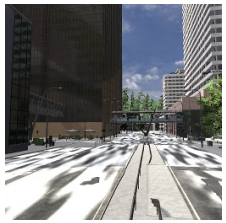
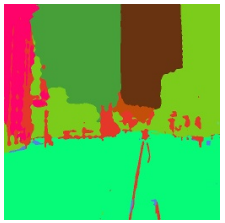
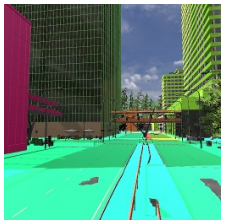
An acoustic renderer is used to test the validity of this method by producing auralisations of Office. City is excluded because of its computationally-expensive scene complexity. We employ a state-of-the-art acoustic renderer Raghuvanshi and Snyder (2014), integrated into Unity, to generate a model of the acoustic environment in which all meshes having a potential impact on the VE are included. The renderer determines per-mesh absorption information based on the texture meta-data as per Default behaviour. A sound source and listener are placed at human height in the scene; the listener captures a $30s$ chirp signal sweeping logarithmically from $20Hz$ to $20kHz$ emitted by the sound source to measure an IR. Maintaining the same settings and positions of source and listener, we repeat the procedure supplying meshes and absorption information inferred by our system, Tagged. We compare the two IRs generated by the former (Default) and the latter acoustic model (Tagged) through comparisons of their deconvolved frequency responses, see Fig.

1.2.4 Results

The model inference takes an average of $400ms$ and the re-projection process takes an average of $96ms$. These figures are quoted per camera probe that is used to generate acoustic labels for surfaces in the scene. Images to be inferred are of a fixed size from the scene frame buffer, 512×512 pixels. The time taken for inference is largely invariant to typical scene complexities such as shape, polygon count, materials etc. The Office scene requires 12 probes to completely tag the environment, requiring ∼6s to complete the tagging process. The City scene shown, has extra complexity and requires use of solutions to the Art Gallery problem to deduce the minimum number of probes to cover the space and tag all objects. As shown in Fig. **??**, acoustic properties can be associated with geometry in the scene, and can be tagged from camera probes. These materials are used in an acoustic rendering process, either directly in game audio engines or external offline acoustic renderers. This can result in more realistic aural spatialisation, using IRs to encode early and late reflections. An example of this offline rendering is shown in Fig. **??**. Currently our approach works by providing inference for camera views within the scene. These camera views are manually placed and would need to be placed in many positions in order to tag materials accurately for the entire scene. This process still requires a human-in-the-loop and needs to be addressed to ensure the goal of having this system as an end-to-end autonomous vision based material tagging system. To extrapolate materials tagged to the entire scene, solutions to the Art Gallery problem would optimise the number of predictions required Devadoss and O'Rourke (2011); Bajuelos et al. (2008). Considering the polygons encapsulating the walkable space W of a scene, minimum vertex guard algorithms suggest that $\lfloor n/3 \rfloor$, where n indicates the total vertices of W, is the least number of positions from where the entire scene can be seen. Based on the depth of the camera, additional intermediate positions $\mathbf{p}$ might be needed to accurately represent

objects, this also depends on the number of pixels per object allowing the the neural network to infer materials from the set of camera views that facilitate the whole scene to be visible. For each camera probe position, rotation steps are needed to ensure that all points of W are inside the camera frustum. For an omni-directional camera probe, these rotation steps $\mathbf{r}_\theta$ for azimuthal steps and $\mathbf{r}_\phi$ for elevation steps should cover the space in 2π azimuth and π elevation. The resulting complexity of the material tagging process for the scene would then be determined by $O(\lfloor n/3 \rfloor + \mathbf{p} + \mathbf{r}_\theta + \mathbf{r}_\phi)$.

Table 1.3: cog preliminary

Render	Input	Segmented	Tagged	Render	Input	Segmented
		Office			City	
A						
B						
C						
D						
legend				legend		

1.2.5 Discussion

Acoustic modelling and audio rendering methods can benefit from research and development of computer vision methods. The current status of this work does not eliminate the human-in-the-loop; however, it can generalise and operate on large sets of complex scenes. As a result, artistic and creative workflows for level design can benefit from automated material tagging system that is agnostic of scene complexity and allow for easy integration of wave-based acoustic renderers. The next steps planned for this work include the development of a generalised system to perform material tagging in complex scenes. This will consider optimisation methods to allow the inference of entire scenes automatically with the minimal set of camera probes to consistently tag every acoustically congruent object that is contributory to the VE.

One crucial advantage of a camera-based acoustic material recognition system is the potential to tailor the recognition of materials based on their appearance to a defined ecosystem of material appearances expressed by a set of virtual environments. Although this aspect contrasts the goal of CNN of providing generalisable models, it addresses the problem of the large variance found in visual features of materials by constraining the network within a range of material appearances that interest VE designers. Resembling workflows common in Neural Radiance Fields (Mildenhall et al., 2020), game designers or VE artists would need to provide exemplary scenes tagged with a customisable set of acoustic and semantic materials to the camera-based system, enabling acoustic material tagging of unseen or novel scenes sharing the same nature. With the rising availability of computational resources allocated to rendering VEs, performing a forward pass with a ResNet50 CNN (He et al., 2016) is becoming feasible at runtime, unlocking the potential for online sound propagation on dynamic scenes.

Case for artistic material tags

1.3 Texture-based Acoustic Material Tagging

Here, we propose a novel architecture for tagging acoustic material in virtual environments, which improves upon recent work by abstracting away from camera-based systems and tests vision-based material recognition methods in real environments. By working on virtual reconstructions of complex scenes, the approach is agnostic of the technology and architecture of the target platform. Despite the significant progress made in sound propagation over the last decades, there are still many limitations in simulations for indoor and outdoor environments due to the complexity of the factors that describe a wavefield Liu and Manocha (2020). This improved system expands from the camera-based methods by contributing towards:

- a more efficient application of acoustic rendering to virtual environments;
- a novel architecture for recognising materials from textured meshes in complex scenes,

reducing the need for manual tagging of acoustic materials and eliminating the needs for camera-based workflows;

- an objective evaluation of the architecture conducted on a virtual reconstruction of a real conference room.

1.3.1 Method

We present a method for processing scene geometry generating acoustic models by predicting materials of objects composing complex scenes. A breakdown of the system shown in Figure **??** illustrates how visual representations of the environment map to acoustic data used by sound renderers to model sound propagation in a scene. The unwrapped texture of each mesh in the scene geometry provides representations of their materials, which provide input for a convolutional neural network. The network, based on features extracted from textures, recognises different material labels in textures and maps them to acoustic data from a database, expressed as acoustic materials.

Material Recognition

According to Schwartz and Nishino (2019), small image patches contain enough information to distinguish materials and hence, we decompose input image textures into small image patches.

Training

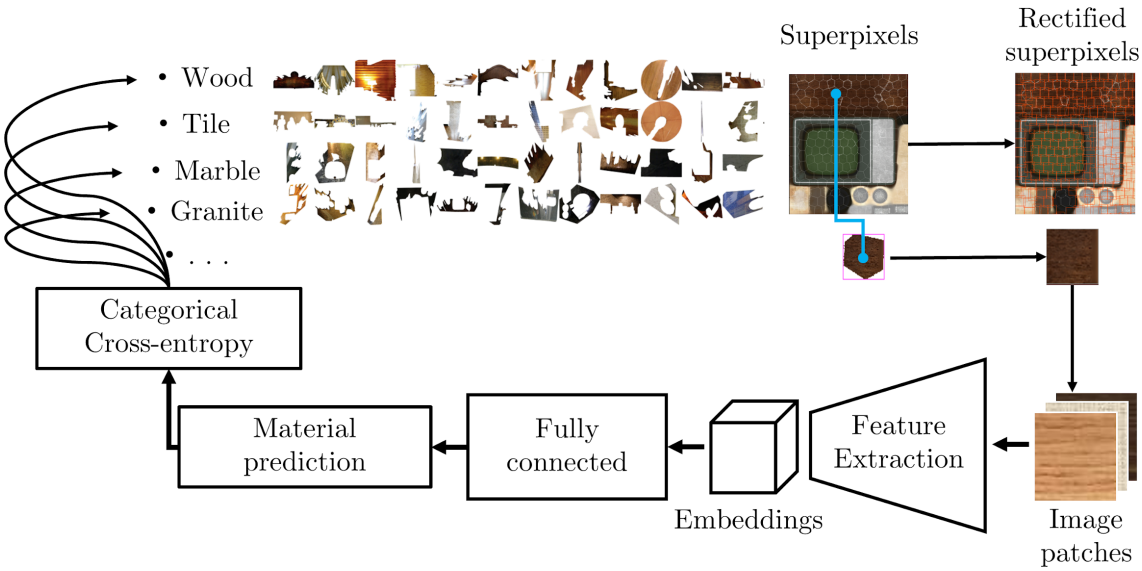


Figure 1.5

We determine the visual material space by applying transfer learning to the OpenSurfaces dataset Bell et al. (2013), which comprises 36 classes of surface photographs. The SLIC algorithm Achanta et al. (2012) segments input surface photographs into a set of superpixel labels, which determine regions correlated with boundaries of objects. From these resulting

superpixel labels, we then generate rectified image patches encapsulating their contours through edge detection Ding and Goshtasby (2001). The rectified image patches are fed through a ResNet50 He et al. (2016), used as a feature extractor for a classification network using a standard fully connected layer to predict class labels based on embeddings of 32×32 pixel resolution input patches. We train the network on 13677819 input patches, composing a train set of about 9.1M images and an evaluation set of about 4.5M, adopting the standard Adam optimiser **?** to update the weights initialised from the ImageNet dataset Deng et al. (2009). The model usually converges in 45 epochs with a training and validation accuracy of about 0.94 and 0.83 respectively.

Inference

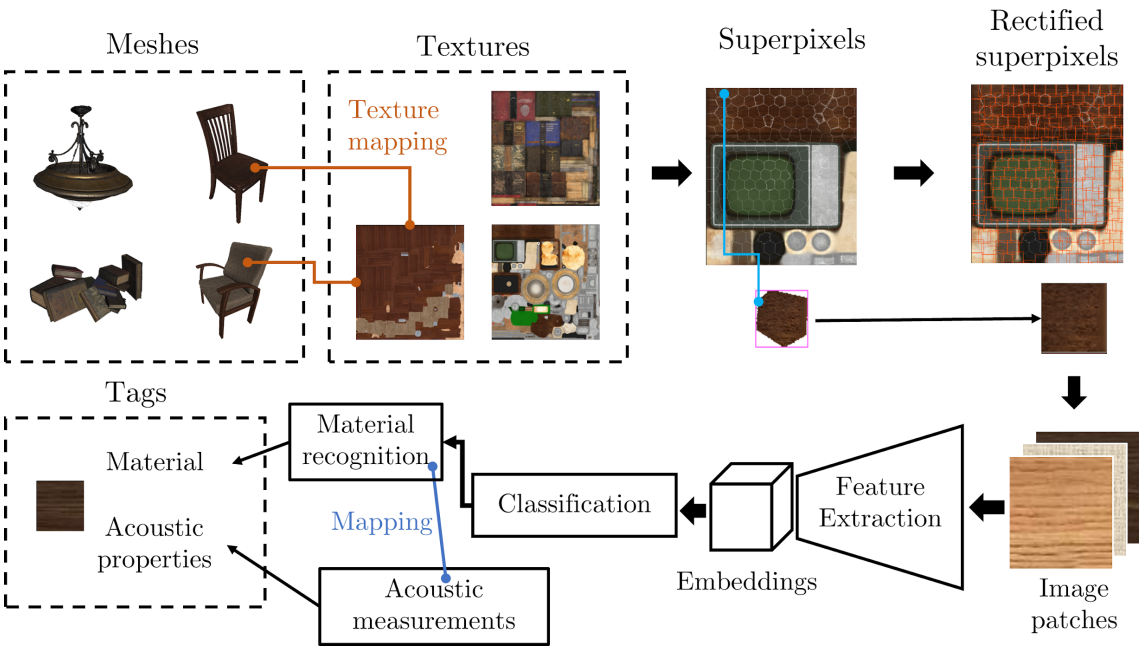
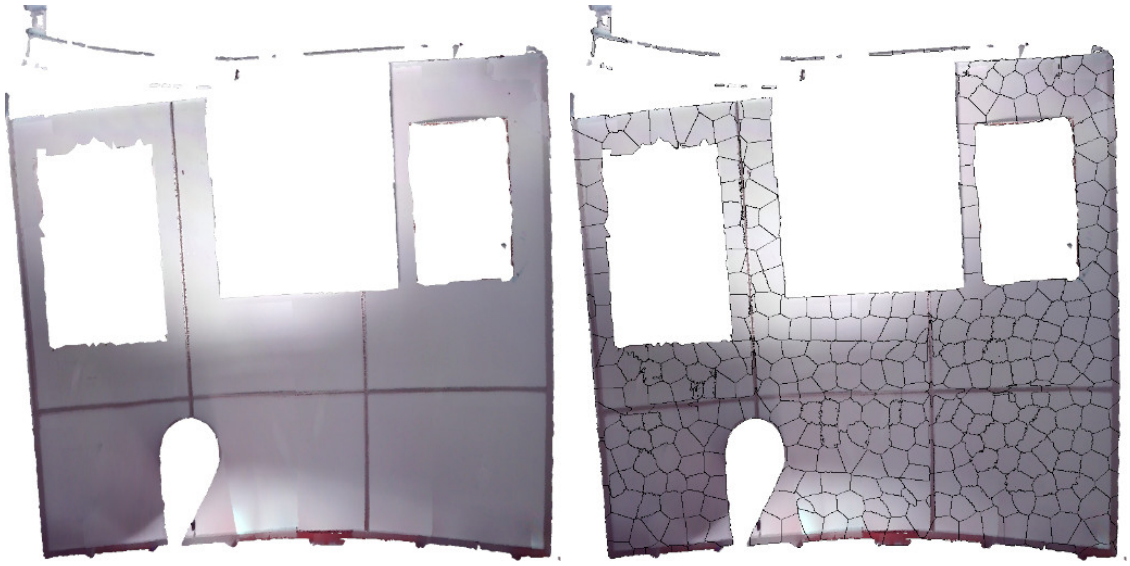


Figure 1.6

Given the set of textured meshes in a scene, we unwrap textures as images to predict materials represented. The trained ResNet50 extracts features from input image textures in complex scenes, whose embeddings enable the classifier to predict class labels associated with each input superpixels. The most frequent prediction maps to an acoustic measurement database, defining the output acoustic material. On average, the classifier takes $11.2s$ to determine the acoustic material for a given texture, see Figure **??**, divided in $3.8s$ for generating rectified patches and $7.3s$ to extract features and compute the mapping.

Acoustic Material Mapping

Material labels inferred from textures are associated with acoustic measurements of absorption coefficients. For every label, a one-to-many mapping groups measurements of the given material. Following the methodology in Kim et al. (2020), we use median frequency-dependent values to determine acoustic absorption, defining acoustic materials. A single



(a) Input Texture (b) SLIC Superpixels

Figure 1.7: SLIC Superpixel computation on real texture from a scanned space.

acoustic material maps to each given mesh, associating a vector of acoustic absorption coefficients to its triangles, determining the overall acoustic mapping accuracy to depend upon the mesh separation of the scene geometry. In the example texture shown in Figure **??**, "Tile" defines the acoustic material, as per predictions shown in Figure **??**.

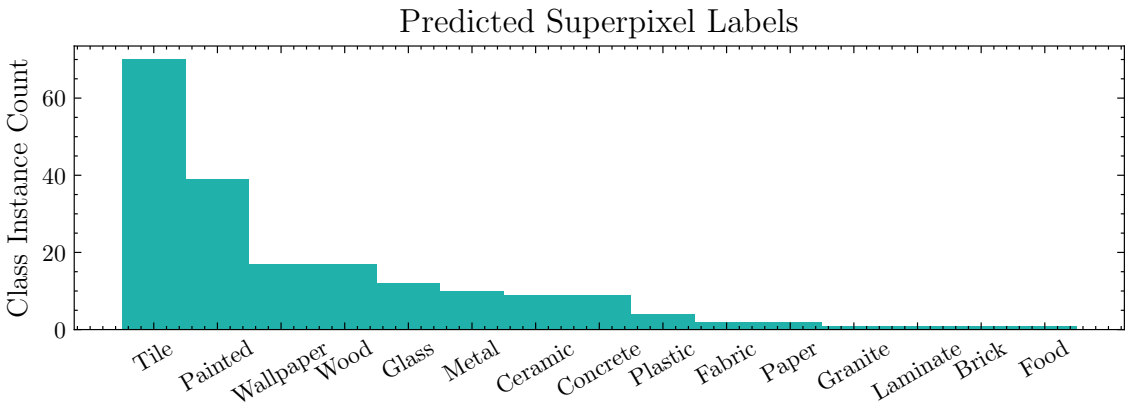


Figure 1.8: Predicted Class labels from input image patches computed from the texture shown in Figure 1.7.

1.3.2 Preliminary Evaluation

We compare acoustic models estimated with state-of-the-art wave-based audio renderers, testing whether the proposed system for the automatic tagging of scene geometry has a significant impact on the resulting acoustic model. The experimental evaluation uses a real environment as a benchmark for testing how predicted models express acoustic parameters of measurable space.

(a) Input Camera Render (b) Re-projected Acoustic Materials

Figure 1.9: Map Reprojection performed on an input camera render: based on a predicted segmentation map, acoustic materials determined by pixel-wise semantic information is re-projected onto scene objects captured by the camera.

Scene Geometry Reconstruction

Wavefields are simulated from a real conference room that is $3.5346m$ deep, $2.8367m$ wide and $3.5149m$ tall. Theoretically, the dimensions determine a Schroeder frequency of $261Hz$. The room is reconstructed using the Unity game engine with $2.9M$ of triangles and $6.6M$ vertices. We reconstruct the geometry of a conference room to deploy the proposed system in a real environment and conduct the subsequent subjective experimental evaluation. We adopt a LiDAR scanner, FARO Focus^{3D} X300, to capture several point clouds scans of the room across 8 positions: 4 position points for each corner of the room at about $1m$ height and 4 additional positions at about $0.2m$ height to capture furniture and materials from different angles and enhance the spatial resolution. We then segment the reconstructed mesh, obtained from the scanner, to ensure that every scene object is represented by a separate textured mesh.

RIR Measurements

The experimental evaluation compares simulated wavefields generated from the reconstructed environment to a sample measurement of the real counterpart's wavefield. We compare wavefields using Room Impulse Responses (RIRs) to describe acoustic properties dependent on geometry and materials surrounding a sound transmission between a source and a listener Stan, Embrechts and Archambeau (2002). We capture the environments' acoustic characteristics emitting and recording an exponentially swept sine, ranging from $20Hz$ to $20kHz$, from a listening to a receiving point that is consistent between the real and reconstructed environments, see Figure **??**. Applying inverse filtering, we recover the

RIRs from the captured signal Holters, Corbach and Zölzer (2009). Given the single listener position, all RIRs are mono. In the conference room, a public address system, the dB Technologies ES 1002, emitted the exponentially swept sine converted from a laptop using an Audient ID14 DAC and ADC, which captured the signal back through an omnidirectional measurement microphone, the Earthwork M30. With Steam Audio **?**, we simulate wavefields of the reconstructed environment. This allows synthesis of wavefields based on acoustic geometry, considering absorption coefficients expressed across three frequency bands: low, medium and high. All simulations share the same resolution of 65536 and 16384 direct and secondary rays, respectively, with 256 bounces off solid geometry. With the same setup, we synthesise three wavefields with different acoustic materials: the *generic* with a single acoustic material for the entire scene geometry; the *tagged*, with acoustic materials assigned through manual material tagging; and *ours*, using the proposed automatic tagging.

Perceptual test

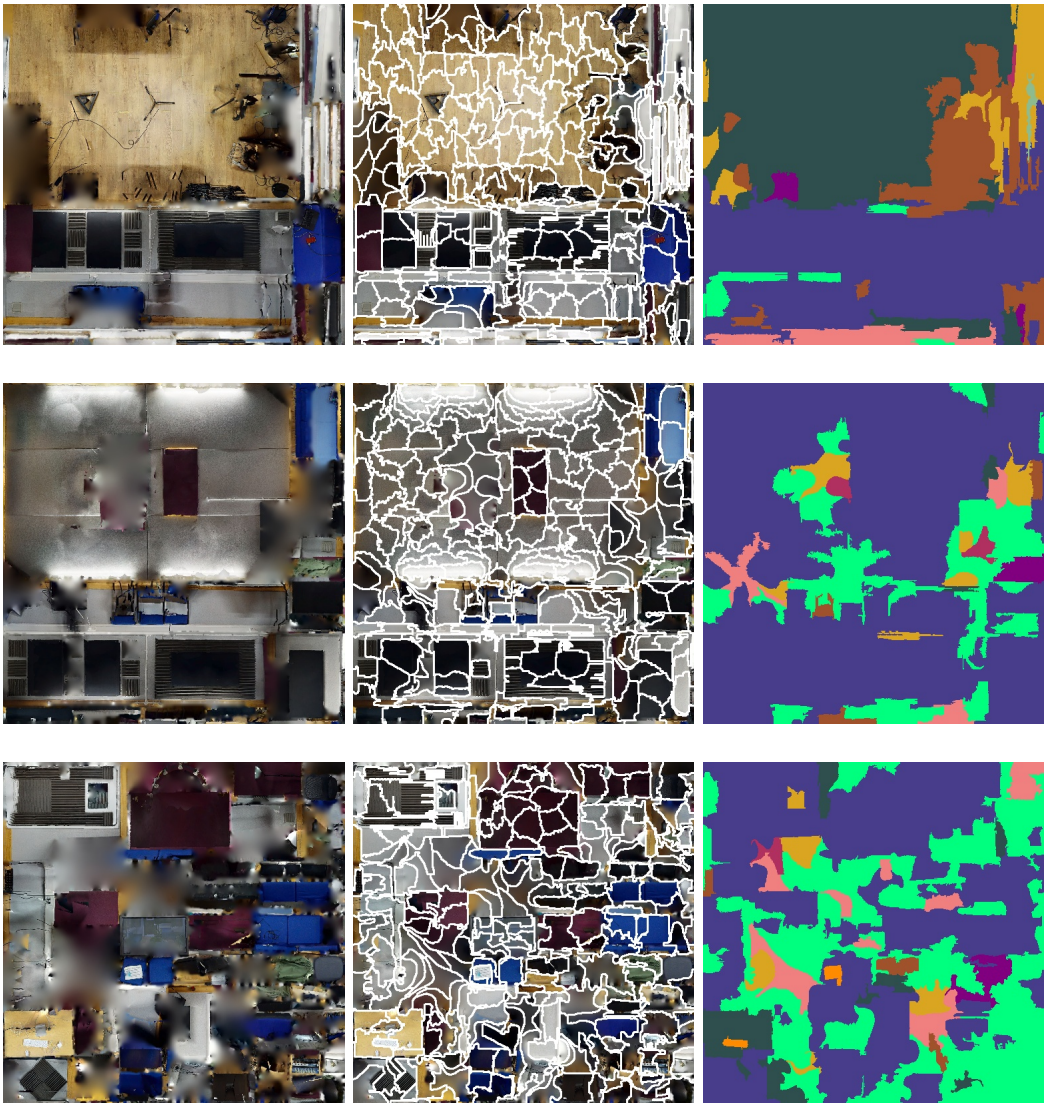
We conduct a perceptual comparison between simulated wavefields at the same positions of source and listener, see Figure **??**, using an automated perceptual metric learned on subjects' responses Manocha et al. (2020). The metric consists of a 14-layer deep neural network with filters trained on features extracted from paired input audio samples; it expresses a distance $D(x_{ref}, x_{per})$ between two input signals, where x_{ref} is a reference signal, and x_{per} is a perturbated signal. The function D considers factors including reverb and the ratio between direct and reverberated signal. We test whether the metric expresses a closer perceptual distance between the measured ground truth and the synthesised wavefields, by convolving the RIRs to samples from the evaluation subset of a database for acoustic scene classification, which comprises 15 different acoustic environments, generating a total of 18 minutes of audio over 1620 N samples Mesaros, Heittola and Virtanen (2016). The learned metric determines the distance between the ground truth convolutions and each of the relative simulated RIRs: for each audio sample k in N, we determine perceptual distances $D(x_{ref,k}, x_{per_i,k} \; \forall \; i \in \{generic, tagged, ours\})$, where ref and per_i are the measured and simulated RIRs.

1.3.3 Evaluation Results

figures.

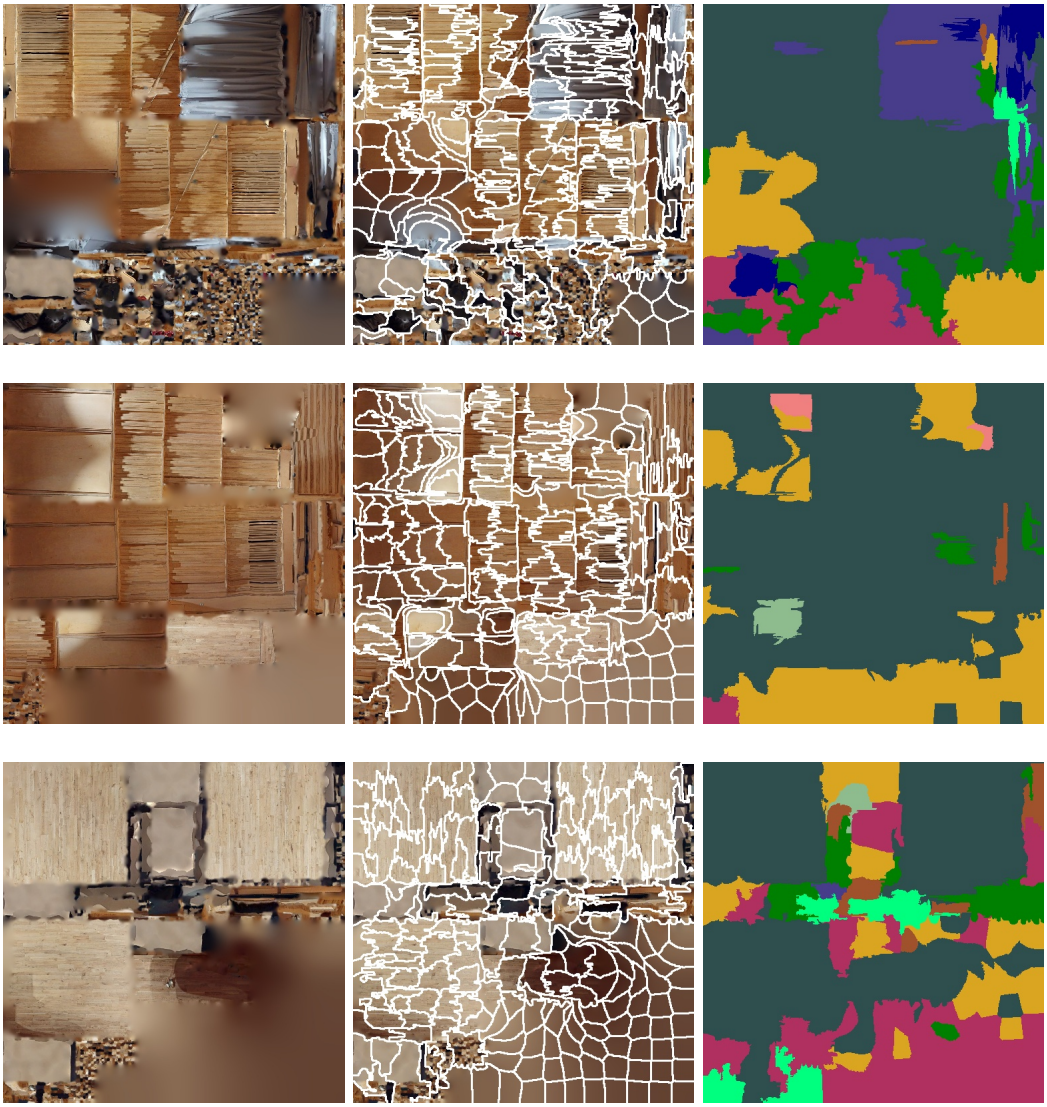
1.4 Proof of Concept Demonstration

A proof of concept implementation is designed to work on real-world scenes, gathering insights into the performance of the proposed material tagging prototypes in attributing acoustic properties to segments of the scene geometry of a given environment expressed as a set of triangulated meshes. The design focuses on scenes scanned with space reconstruction



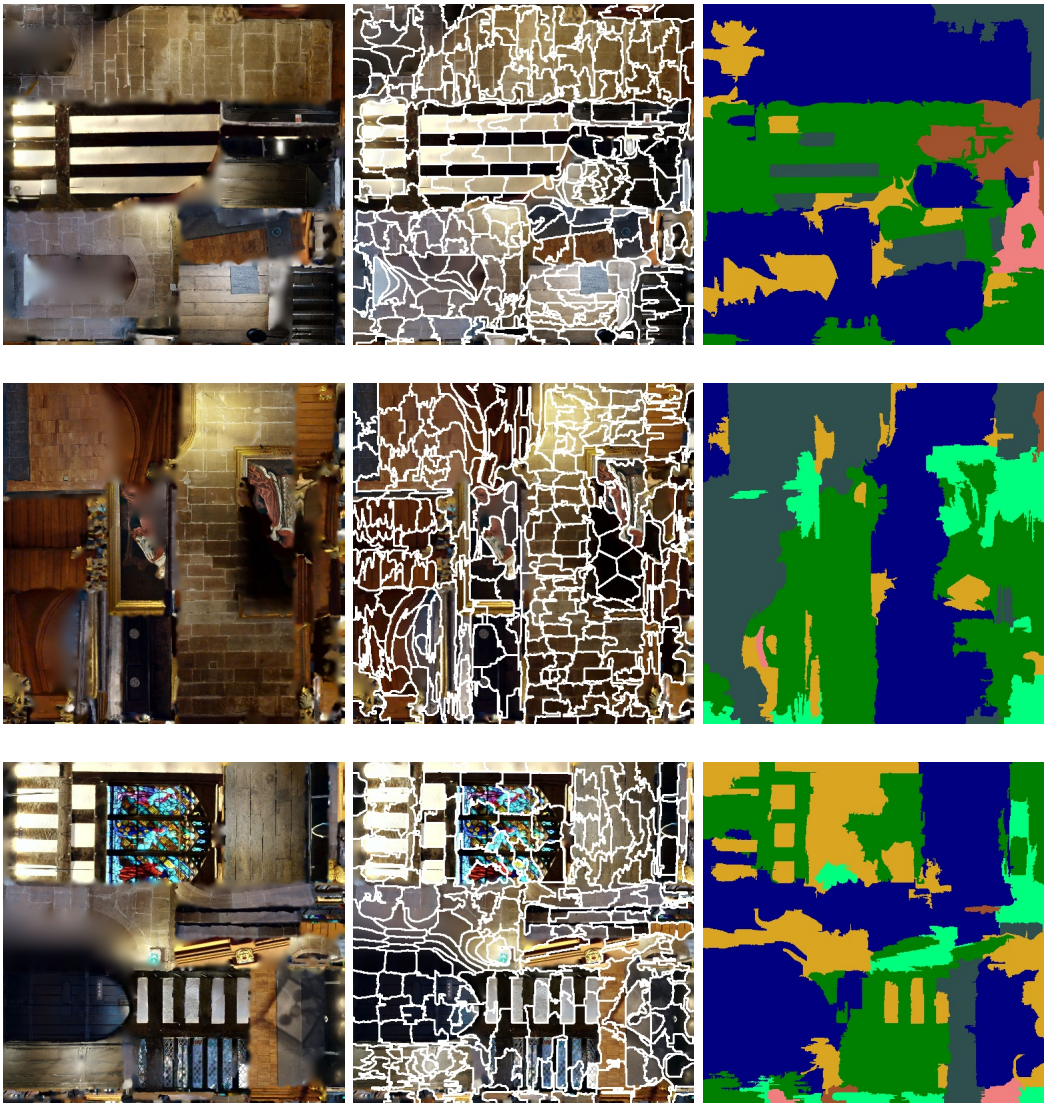
(a) Input Texture (b) Superpixel Segmentation (c) Material Tagging

Figure 1.10: Material Tagging performed on textures from the Mastering Suite scene.



(a) Input Texture (b) Superpixel Segmentation (c) Material Tagging

Figure 1.11: Material Tagging performed on textures from the Recital Hall scene.



(a) Input Texture (b) Superpixel Segmentation (c) Material Tagging

Figure 1.12: Material Tagging performed on textures from the St Mary's Guild Hall scene.

Scene	Dimensions	Volume	Triangles	T_{60}	Segments
Mastering Suite	ph	ph	ph	0.00	00
Recital Hall	ph	ph	ph	0.00	00
St Mary's Guildhall	ph	ph	ph	0.00	00

Table 1.4: Summary of the scenes used for the testing procedure of the acoustic material tagging prototypes.

techniques applied to real environments, emulating space reconstruction paradigms often adopted by AR HMDs.

The design prototype uses standard space-sensing cameras to produce triangulated meshes from a set of real environments. These virtual reconstructions are then processed, segmenting the scene of each environment by generating portions of scene geometry into a set of semantically meaningful sub-meshes expressing elements of the scene. Submeshes are tagged with acoustic material by assigning each scene element, referred to by its sub-mesh, a semantic acoustic material drawn from a generated list of materials defined by each of the two prototypes.

The two prototype systems are deployed on the virtual reconstructions of the real spaces, inferring acoustic materials across all scene elements. The evaluation measures the accuracy in assigning the correct semantic information to each scene element, comparing predictions to ground truth acoustic material tags.

1.4.1 Test Scenes

The test adopts three real spaces: a small, a medium, and a large environment, see Table **??**. These environments have diverse ecosystems of surfaces and materials that characterise and determine their respective soundscape and acoustic fingerprints.

1.4.2 Mesh Segmentation

Current space reconstruction algorithms that generate a triangulated mesh representing the physical space scanned have limited knowledge of the semantics of the reconstruction. Until the definition of a scene entity is given and a solution to distinguish physical scene entities in the reconstructed space, it is impossible to separate the triangulated mesh into semantically meaningful segmented.

Such limitation is beyond the scope of this thesis work. It is a problem that can be addressed with complex mappings between defining a scene entity belonging to a physical environment and its reconstruction as a triangulated mesh. A family of techniques exists to address the problem of segmenting 3D scenes expressed as triangulated meshes or point clouds. However, the complex nature of three-dimensional space reconstructions makes the problem an open challenge. The work by Chen, Golovinskiy and Funkhouser (2009) is a significant stepping stone towards the problem of defining a physical entity represented

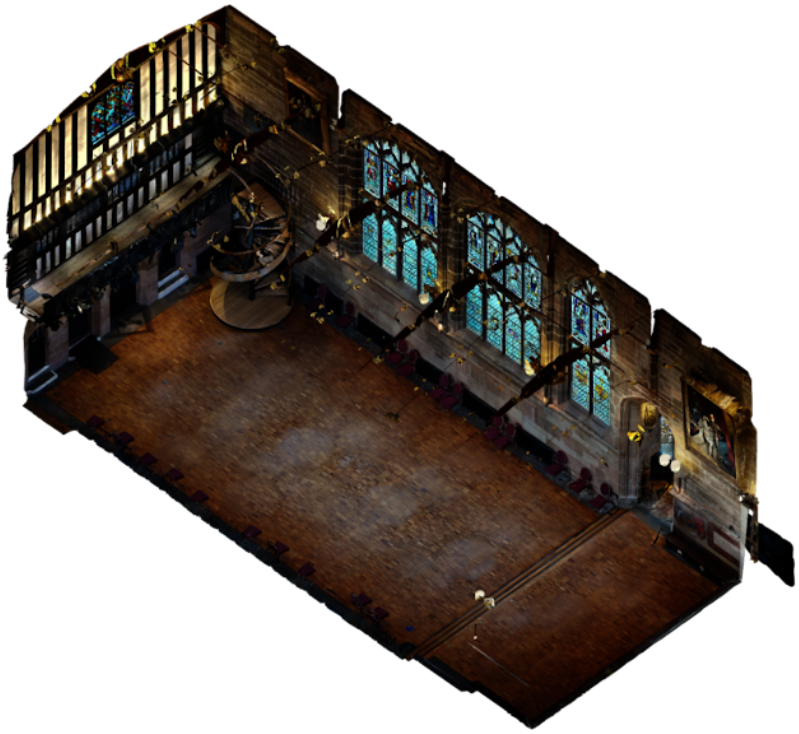


Figure 1.13: St Mary's Guildhall: a Medieval-style church in Coventry, West Midlands, England. The large environment has a unique soundscape, characterised by a recorded T_{60} reverberation time of over a second.

Figure 1.14: Recital Hall: a wooden space serving as a stage for musical performances. The soundscape is characterised by a recorded T_{60} reverberation time within a second.

Figure 1.15: Mastering Suite: a small audio production studio with acoustic treatments to minimise the effect of the soundscape on the sound reproduction system within.

(a) Input Camera Render (b) Re-projected Acoustic Materials

Figure 1.16: Map Reprojection performed on an input camera render: based on a predicted segmentation map, acoustic materials determined by pixel-wise semantic information is re-projected onto scene objects captured by the camera.

in a complex mesh or point cloud, as the authors gather annotations from humans to investigate the consistency in segmenting entities represented in reconstructed data.

Such experiments have shown a consensus in the segmentation process and provided a benchmark as an evaluation tool and starting point for automated methods. There are techniques available that address the problem, such as Rusu and Cousins (2011)'s work providing a set of algorithms in Point Cloud Library, as well as new methods leveraging geometrical CNNs to segment 3D meshes and point clouds, such as Feng et al. (2020)'s work. Despite these techniques being available, there are still large margins of errors when deployed on in-the-wild reconstructions that require fine-tuning and human authoring.

Hence, the mesh segmentation stage of the material recognition pipeline is performed manually, separating the reconstructed data into semantically meaningful segments with a minimum structural size of $70cm$, following Pelzer and Vorländer (2010)'s perceptual threshold for GA pipelines. Using Blender as a 3D data editor[1], the reconstruction obtained from the Matterport camera is separated into individual meshes, see Figure 1.17.

1.4.3 Manual Acoustic Material Tagging

The segmented meshes are then manually tagged, assigning acoustic absorption coefficients to the triangles of each sub mesh.

UV Mapping of scene geometry to texture segments.

1.5 Discussion

Both systems provide fundamental building blocks for acoustic material tagging for AAR pipelines.

Problem of mesh segmentation with semantics projection. Future work.

1.6 Conclusions

1.6.1 Contributions

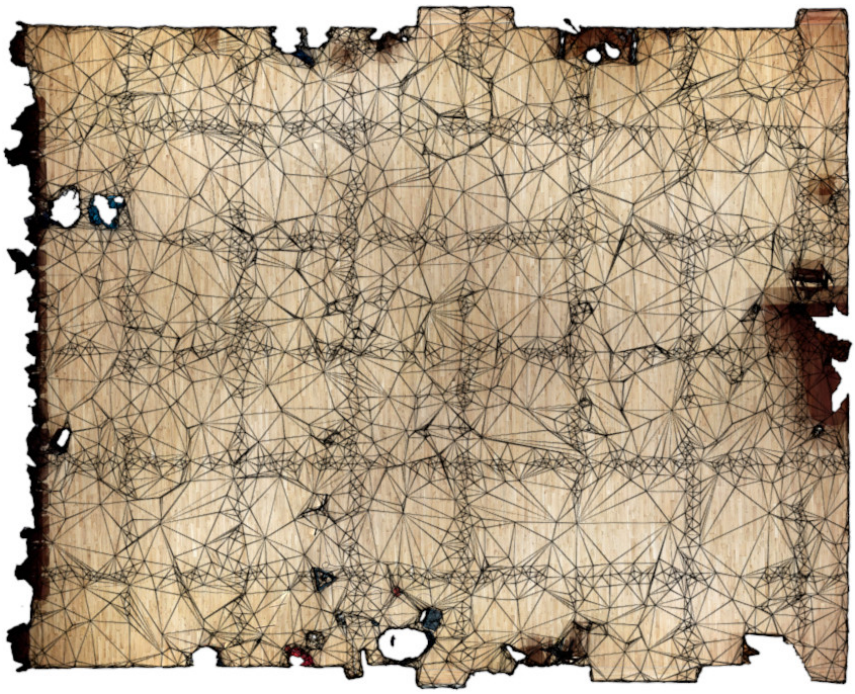
Camera-based system for acoustic material tagging.

1.6.2 Future Research Directions

Acoustic modelling and audio rendering methods can benefit from research and development of computer vision methods. The current status of this work does not eliminate the human-in-the-loop; however, it can generalise and operate on large sets of complex scenes. As a result, artistic and creative workflows for level design can benefit from automated material tagging system that is agnostic of scene complexity and allow for easy integration

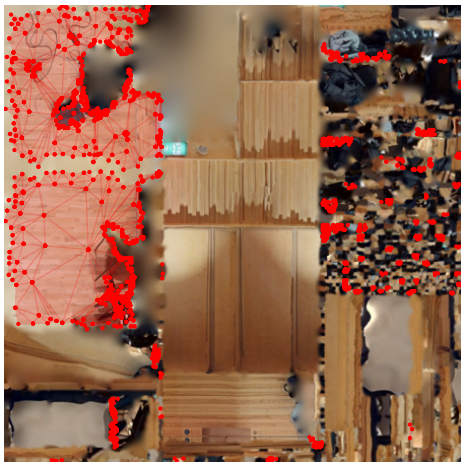
[1] https://www.blender.org/

Figure 1.17: Mesh segmentation process performed manually on a reconstructed scene.



(a) Scene Geometry: Floor

(b) Texture Mapping 1



(c) Texture Mapping 2

Figure 1.18: UV mapping of one scene geometry segment to many texture segments.

of wave-based acoustic renderers. The next steps planned for this work include the development of a generalised system to perform material tagging in complex scenes. This will consider optimisation methods to allow the inference of entire scenes automatically with the minimal set of camera probes to consistently tag every acoustically congruent object that is contributory to the VE. We addressed the problem of mapping acoustic absorption data to scene geometry in virtual environments for acoustic simulations. Methods and frameworks for material recognition have become efficient enough in recognising materials in the wild with varying factors of illumination, shape and surface characteristics. Despite their limited resolution in computer games applications, sound propagation systems benefit from acoustic material tagging. With the proposed system, we aim to integrate material recognition to wave-based methods to determine materials' acoustic properties. The next steps of this work aims to extend the system to broader scenes with larger sets of materials as well as improving the material recognition by performing multi-scale analysis of superpixels and adopting clustering paradigms to overcome the limitations of finite material space definitions. With the proposed system, we aim to employ these methods to eliminate the human-in-the-loop for the process of labelling materials for acoustic renderers or assist in artistic and creative processes for level design.

Bibliography

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. *Ieee transactions on pattern analysis and machine intelligence*, 34(11), pp.2274–2282. Available from: https://doi.org/10.1109/TPAMI.2012.120.

Bajuelos, A.L., Canales, S., Hernández, G. and Martins, A.M., 2008. Optimizing the minimum vertex guard set on simple polygons via a genetic algorithm. *Wseas transactions in information science and applications*, 5(11), pp.1584–1596.

Bell, S., Upchurch, P., Snavely, N. and Bala, K., 2013. Opensurfaces: A richly annotated catalog of surface appearance. *Acm transactions on graphics (tog)*, 32(4), pp.1–17.

Blauert, J., 1997. *Spatial hearing: the psychophysics of human sound localization*. MIT press.

Chen, X., Golovinskiy, A. and Funkhouser, T., 2009. A benchmark for 3d mesh segmentation. *Acm transactions on graphics (tog)*, 28(3), pp.1–12.

Deines, E., Bertram, M., Mohring, J., Jegorovs, J., Michel, F., Hagen, H. and Nielson, G.M., 2006. Comparative visualization for wave-based and geometric acoustics. *Ieee transactions on visualization and computer graphics*, 12(5), pp.1173–1180.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 ieee conference on computer vision and pattern recognition*. Ieee, pp.248–255.

Devadoss, S.L. and O'Rourke, J., 2011. *Discrete and computational geometry*. Princeton University Press.

Ding, L. and Goshtasby, A., 2001. On the canny edge detector. *Pattern recognition*, 34(3), pp.721–725.

Feng, M., Zhang, L., Lin, X., Gilani, S.Z. and Mian, A., 2020. Point attention network for semantic segmentation of 3d point clouds. *Pattern recognition*, 107, p.107446. Available from: https://doi.org/ https://doi.org/10.1016/j.patcog.2020.107446.

Giordano, B.L. and McAdams, S., 2006. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The journal of the acoustical society of america*, 119(2), pp.1171–1181.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.770–778.

Holters, M., Corbach, T. and Zölzer, U., 2009. Impulse response measurement techniques and their applicability in the real world. *Proceedings of the 12th international conference on digital audio effects (dafx-09)*. Italy: DAFX, pp.1–5.

Hulusic, V., Harvey, C., Debattista, K., Tsingos, N., Walker, S., Howard, D. and Chalmers, A., 2012. Acoustic rendering and auditory–visual cross-modal perception and interaction. *Computer graphics forum*. Wiley Online Library, vol. 31, pp.102–131.

Kim, H., Remaggi, L., Fowler, S., Jackson, P. and Hilton, A., 2020. Acoustic room modelling using 360 stereo cameras. *Ieee transactions on multimedia*, 1, p.1.

Kuttruff, H., 2016. *Room acoustics*. Crc Press.

Li, D., Langlois, T.R. and Zheng, C., 2018. Scene-aware audio for 360 videos. *Acm transactions on graphics (tog)*, 37(4), pp.1–12.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the ieee international conference on computer vision*. pp.2980–2988.

Liu, S. and Manocha, D., 2020. Sound synthesis, propagation, and rendering: A survey. *arxiv preprint arxiv:2011.05538*, 1(1), p.1.

Manocha, P., Finkelstein, A., Jin, Z., Bryan, N.J., Zhang, R. and Mysore, G.J., 2020. A differentiable perceptual audio metric learned from just noticeable differences. *arxiv preprint arxiv:2001.04460*, 1(1), p.1.

Mesaros, A., Heittola, T. and Virtanen, T., 2016. Tut database for acoustic scene classification and sound event detection. *2016 24th european signal processing conference (eusipco)*. IEEE, Europe: IEEE, vol. 1, pp.1128–1132.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. *Eccv*.

Pelzer, S. and Vorländer, M., 2010. Frequency-and time-dependent geometry for real-time auralizations. *Proceedings of 20th international congress on acoustics, ica*. pp.1–7.

Raghuvanshi, N. and Snyder, J., 2014. Parametric wave field coding for precomputed sound propagation. *Acm transactions on graphics (tog)*, 33(4), pp.1–11.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention*. Springer, pp.234–241.

Rusu, R.B. and Cousins, S., 2011. 3d is here: Point cloud library (pcl). *2011 ieee international conference on robotics and automation*. pp.1–4. Available from: https://doi.org/10.1109/ICRA.2011.5980567.

Savioja, L. and Svensson, U.P., 2015. Overview of geometrical room acoustic modeling techniques. *The journal of the acoustical society of america*, 138(2), pp.708–730.

Schwartz, G. and Nishino, K., 2019. Recognizing material properties from images. *Ieee transactions on pattern analysis and machine intelligence*, 1(1), p.1.

Stan, G.B., Embrechts, J.J. and Archambeau, D., 2002. Comparison of different impulse response measurement techniques. *Journal of the audio engineering society*, 50(4), pp.249–262.