Vision-based Acoustic Information Retrieval for Interactive Sound Rendering

Mattia Colombo *

DMT Lab, Birmingham City University

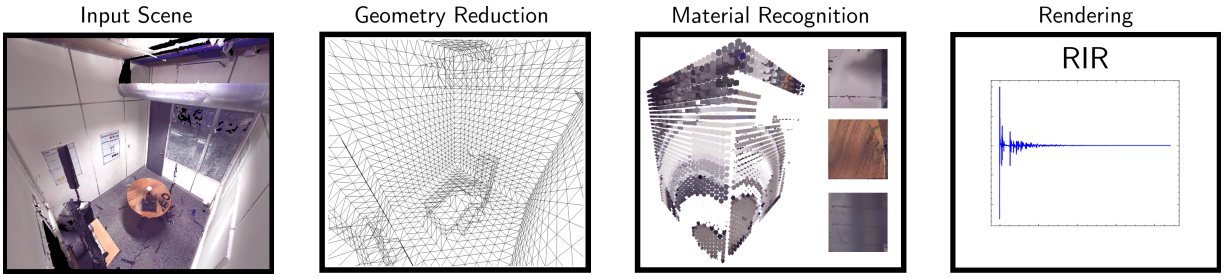

Figure 1: Envisioned pipeline for vision-based acoustic rendering. From left to right: virtual reconstruction of real space; its simplified geometry allows fast sound rendering; the material recognition stage analyses appearance of surfaces and tags acoustic materials to the reduced geometry; Finally the rendering stage produces Room Impulse Resposes (RIRs) that can be used for real time auralisations. RIRs are compact description of early and late reverberation expressed as power of sound reflections over time — they are a function of the acoustic space, position of source and listener and they can be used to propagate a dry audio source.

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.

Index Terms: Real-time simulations—acoustic rendering; mixed / augmented reality—scene understanding—computer vision

1 Introduction

The current state of interactive sound rendering allows for fast acoustic simulations, even on platforms with limited computational budgets, approximating the soundfield of any given environment, where a listener can experience realistic auditory interactions with virtual sound sources [9, 13]. Sound rendering can be considered a fundamental component to computer games technology, responsible for reproducing everyday sound emitted by objects or agents in a virtual scene and perceived by a listener. This poses the challenging task of reflecting basic acoustic principles to render such auditory interactions realistic. In the real world, sound propagates from a sound source to a listener, interacts with objects in the environment and with the environment itself arriving at the listener's ears [12]. Sound cues alone are sufficient to enable users in Virtual Environments (VEs) to pinpoint locations of sound-emitting entities in a scene by using auditory sound localisation, a natural ability associated with the human auditory system [14, 18].

As the acoustic principles that govern how sound propagates in space are difficult to reproduce in digital systems, many methods exist providing variable orders of approximations, depending on the application. Such approaches, emulate the wavefield of an environment, simulating how sound interacts with boundaries and scene objects. A subset of these can reproduce phenomena of sound, such as diffraction, reflection and refraction, which are determinant of realism as they emulate how waves bend around obstacles. Such phenomena make the simulated wavefield dependent on the accuracy of scene geometry and materials represented in a VE.

There is a large tree of techniques and methods to simulate sound propagation, reflecting acoustic properties to any given sound source in a VE, adapting to perceptual requirements and computational budgets available [7]. As a general rule, the more computational budget available, the more complex techniques can be employed, allowing realistic sound rendering. Finite-Difference Time Domain (FDTD) [8] or wave-based [17] methods, on this end of the spectrum, obtain high degrees of accuracy and realism, but often require precomputation stages or GPU implementations to produce acoustic simulations at interactive rates. On the other end of the spectrum there are fast geometrical acoustics methods, widely adopted in real-time application due to their low computational requirements and highly-parallelisable implementations [19], which reduce simulated sound waves to rays or beams, that are much simpler to compute. Finally, hybrid methods also exist to combine strengths of the main families.

Classic sound rendering has always been employed by acousticians and engineers to solve practical problems as it requires the work of experts to adjust parameters and define acoustic characteristics of a virtual scene. A constant here is the requirement of an accurate description of the environment, detailing geometry of architectural components and objects contained within with acoustic information such as acoustic energy absorption, reflection, or scat-

*ORCID: 0000-0002-4169-2045

tering — this is essential to model the behaviour of sound waves interacting with the environment.

Only recently, with the increase of processing power available in computers, it has gained popularity in computer games and immersive technology for entertainment and serious applications [27]. Augmented Reality (AR) technology can particularly benefit from this as the increase of processing allows sound rendering on mobile devices, enabling listeners to experience virtual sound sources propagating in reconstruction of real geometry, which is the main avenue that the planned thesis work aims to explore. Specifically, a pipeline for real time sound rendering Room Impulse Responses (RIRs) is envisioned, composed of components touching upon subfields intersecting with topics of the ISMAR community.

2 Related Work

The work presented by Kon and Koike [11] introduces a novel method to determine reverberation time of space, training a neural network on a dataset comprising images with associated IRs to map visual representations of en environment to acoustic parameters. This mapping represents a milestone in cross-modal rendering allowing the use of reverberation algorithms to tune acoustic high-level parameters, particularly useful in interactive applications where computational budgets do not allow acoustic rendering. Poirier-Quinot *et al.* presented a framework for sound rendering, which integrates into the Blender computer graphics toolset to generate acoustic simulations by manually tagging acoustic materials in geometry from complex scenes, generating on-the-fly auralisations [16].

Modern acoustic rendering leverages deep neural networks to automate the complex and multi-dimensioned mapping between high level attributes of an acoustic environment, and RIR generation. This is the case of the work presented by Tang *et al.* [26], whose approach captures audio recording and an estimate of the room geometry using commodity devices to predict acoustic characteristics of an environment via audio convolutional neural networks. Predicted acoustic characteristics can then be used to produce RIRs at interactive rates.

More recently, [25] Singh *et al.* introduced a novel method to estimate a RIR from a single image. Their system employs neural networks to estimate depth information from the input image, supplying it to an encoder and a generator, producing an RIR based on the learned mappings between RIRs and images combined with corresponding inferred depth data. The above mentioned works, inspired the author to shape the design of our pipeline by learning complex characteristics from real environments to estimate acoustic environments.

3 Contributions

Materials composing boundaries and scene objects are largely responsible for the perceptual difference between two soundscapes, and hence, sound renderers must consider them when computing sound propagation. This is because visual representations of a scene can describe physical characteristics associated with materials of boundaries and objects composing the scene. The initial milestone of the planned thesis work tackled the problem of material tagging in complex scenes [5]. It stems from Schissler *et al.'s* work [20], which pioneered acoustic material classification for sound propagation, leveraging virtual representations of real world scenes to determine frequency-dependent acoustic absorption coefficients.

3.1 Material Tagging

In acoustic rendering, a set of features is often associated with primitives of complex scenes, i.e. the smallest geometrical component composing complex scenes, and controls how the area inscribed in a given triangle interferes with a colliding incidental sound wave, absorbing or reflecting energy. Features like absorption or reflection coefficients, grouped by semantics, map to all triangles of a surface

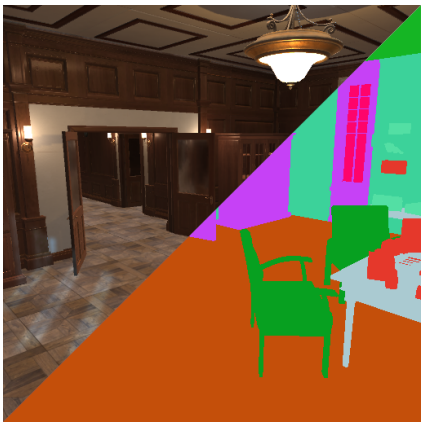

Figure 2: Render of an example complex scene in VE on the top left of the image. On the bottom right, semantic scene segmentation applied to the scene, mapping depicted objects to materials defined in Table 1. Predictions from the segmented image are then reprojected back into the scene, enabling acoustic renderers to automatically determine acoustic properties of surfaces.

determining acoustic features of space and objects within; e.g. by associating a floor with the "concrete" material, all triangles of the surface will have acoustic characteristic of a rigid material. Such mapping can be regarded as acoustic material tagging, and usually requires the work of engineers to tag a complex scene.

Here the author proposed a study evaluating a pipeline to infer acoustic materials from game scenes, producing input for acoustic renderers, and comparing perceptual responses between manually assigned and automatically tagged acoustic materials. Table 1 shows example acoustic materials assigned to scene objects depicted in Figure 2 [12].

3.2 Superpixel Texture Tagging

The author reported on a system for determining acoustic characteristics of space from virtual reconstructions of real environments, adopting convolutional neural networks trained on classifying semantics of visual representations of surfaces [6]. Training and evaluation sample sets of material appearances and surfaces were assembled sampling from the OpenSurfaces dataset [1]. The pipeline introduces the decomposition of image textures, often describing material information of objects in VEs, into small image patches, see Figure 3. Schwartz and Nishino [23] demonstrate how small image patches contain enough information to determine their semantics, allowing the pipeline to associate them to acoustics materials, similar to the ones illustrated in Table 1.

4 Current Work

The avenue currently being explored aims at extending acoustic material classification to an end-to-end pipeline for real-time sound rendering. One of the obstacles ahead is handling large amounts of data from complex scenes: often the complexity of structures in VEs makes it difficult to simulate sound propagation maintaining low computational budgets at interactive rates, regardless of the acoustic rendering technique adopted. This is the case of modern AR technology allowing us to reconstruct space around the viewer with increasing resolution [3].

4.1 Geometry Reduction

The computational load associated with determining the behaviour of propagating sound in complex geometry grows with structural detail of scene objects. As demonstrated by Pelzer and Vorländer [15], it is possible to reduce input geometry of complex scenes to coarse

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

Table 1: Example acoustic materials expressed as mappings of semantic labels to α acoustic absorption over six frequency bands across the equivalent rectangular bandwidth scale ($125Hz$ to $4kHz$), defining surface reflection of incidental sound waves colliding with surfaces. They indicate two discrete levels of acoustic absorption, low α and high α to distinguish varying ρ specific mass among instances of the same materials (i.e. soft and rigid ceilings).

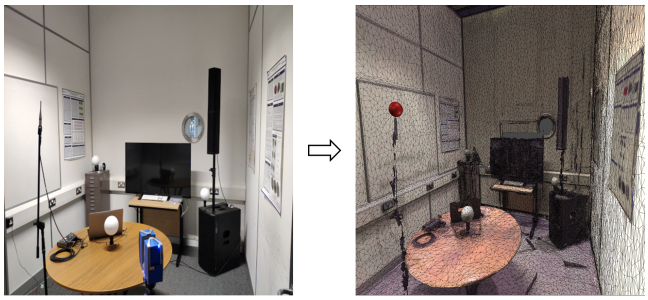

Figure 3: superpixel acoustic material classification: a real environment is scanned (left image), reconstructing its geometry in VE as textured meshes, representing scene objects, from point cloud data obtained from LiDAR scanners (right image). Image textures describing material information of each mesh are subdivided into small patches mapped to acoustic materials defined in Table 1.

levels of detail, maintaining the resulting acoustic simulation under a just-noticeable perceptual difference. Despite the fact that geometry reduction paradigms have been presented more than a decade ago by Siltanen [24], where techniques such as isosurface extraction, i.e. marching cubes [2] can simplify scene geometry, they have seen little improvement and experimentation. Especially considering recent material recognition paradigms, which can be essential to determine physical characteristics of the acoustic space.

Decomposing an input scene to a compact representation of its geometry and materials leads to significant efficiency and performance gains in sound rendering and allocation of more computational budget that can be used for acoustic material classification with convolutional neural networks. These would infer from visual representations of the scene, which map to primitives of the reduced geometry. Hence, the coarse decomposition of the input scene would always have up-to-date material information.

4.2 Prototype Acoustic Renderer

A sound renderer is being developed to test geometry reduction paradigms on virtual reconstructions of real environments, prototyping the pipeline illustrated in Figure 1 as an offline procedure: a real environment is reconstructed in VEs as a set of textured meshes. These are reduced to a simplified geometry with varying levels of resolution. The pipeline is being evaluated reproducing the testing methodology in the study described in Section 3.2.

5 Future Directions

Upon evaluation of a prototype acoustic renderer described in Section 4, several directions are identified as future work. This is where this programme of study would benefit the most from different point of view of members of the ISMAR community, to better identify subfields of AR that the thesis could target.

5.1 Fast Acoustic Rendering

The nature of AR technology combining virtual objects with reconstructions of real environments emphasise the potential of the method discussed in Section 4.2 in achieving sound rendering at interactive rates due to the decomposition of the input scene using geometry reduction. The key contribution which would result from designing such as a system is the use of computer vision techniques to re-build the acoustic material space in the environment in real time: the dynamic geometry of AR would provide a great use case here.

The acoustic material recognition would be based on networks classifying surface appearance via feature extraction. Figure 1 shows example image patches from a real conference room: these images provide input to a feature extractor, whose features are categorised to semantic labels, i.e. “wood” or “metal”, to determine acoustic materials, see Table 1. Finally, upon classification of an acoustic material, judged by the classification probability score, it is tagged to its corresponding portion of the acoustic volume.

The system would constantly build the acoustic space around the listener as they walk around a scene, inferring materials of primitives generated by the geometry reconstruction component, see Figure 1. The sound renderer computes reflection paths, as unseen parts of the scene are discovered and acoustic materials are assigned to it, constantly assembling RIRs that can be used to propagate sound emitted from virtual sources in the scene.

Despite the computational budget required to achieve such level of accuracy in the resulting acoustic model, it would simulate a crucial aspect of virtual objects in AR space: its propagating sound would respect the surrounding scene. For instance, the user would be able to experience the perceptual change in a virtual speaker talking in a small carpeted room as they walk out to an open environment.

The architecture would integrate material recognition at the low-level routines of the algorithm, rendering portions of space as the user discovers their surroundings and removing the need for waiting for a complete scene to be sent to a sound renderer. In addition, acoustic parameters would be cached based on emitter and listener positions across longitude and latitude points in space.

5.2 Subjective evaluation

One important aspect to be considered for future directions is the perceptual aspect of sound rendering. The listener is arguably the most important link in the chain of sound rendering, ultimately determining the efficacy of the overall pipeline: the evoked sense of realism and presence is the overarching goal of rendering. As mentioned in Section 4.1, one can drastically reduce structural detail in complex scenes whilst maintaining the same perceptual response from auralisations. However, considering novel material recognition paradigms, research is yet to investigate perceived presence and realism under varying frequency resolution of tagged acoustic materials, as well as varying amounts of surface sampled from an input environment.

Beside psycho-acoustic aspects, real time sound rendering must consider additional factors associated with the listener such as the anatomy of their head and torso, as in natural auditory perception they affect propagating sound emitted from a scene object arriving at the listener's ears [21]. These aspects are usually modelled with Head-Related Transfer Functions (HRTFs), which are usually specific to subjects' varying physiognomy, but approximations exists to cover a wide range of listeners [22].

5.3 Learning from real space

An important discussion point raising from Section 5.1 is the need for data in the process of acoustic material recognition. As shown in Section 3.1, materials are determinant in sound rendering as they define acoustic characteristics of geometry for sound propagation. Recognising materials is a task that has been providing challenge to the computer vision community due to the multi-dimensioned nature of surface appearance: intrinsic physical properties, lighting, surface shape and many other factor affect the resulting visual representation of materials [23]. With recent advances in the field however, we can overcome many of these challenges by understanding semantics associated with visual representations of surfaces.

In the evaluation discussed in Sections 3.2, the author was able to develop vision-based systems to automatically understand materials in a conference room with convolutional neural networks trained on surface appearances. However, applying the system to a comprehensive set of scenes would require the network to be trained on an extensive dataset to cover the diversity of materials in the real world.

Amongst most notable contributors to this field is Matterport, who provide technology and increasing amounts of open data on virtual reconstructions of real space [4]. Their data makes a great training environment for the proposed pipeline for learning to discriminate materials in virtual reconstructions that are typical of AR technology presented by Izadi *et al* [10].

6 Summary

In conclusion, it is presented the current state of the work towards the PhD programme in sound rendering for interactive application using vision-based material recognition. In light of the contributions reported on acoustic material classification methodologies evaluated on virtual reconstructions of real space, potential future directions are identified, which the author extols as points of contact with the ISMAR community. With this doctoral consortium application the author wishes to contribute to the community by raising interesting discussions and benefit from constructive criticism and feedback from experienced researchers, as well as networking with other consortium participants.

References

[1] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on graphics (TOG)*, 32(4):1–17, 2013.

[2] P. Bourke. Polygonising a scalar field, 1994.

[3] M. Boussaha, B. Vallet, and P. Rives. Large scale textured mesh reconstruction from mobile mapping images and lidar scans. In *ISPRS 2018-International Society for Photogrammetry and Remote Sensing*, pp. 49–56, 2018.

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[5] M. Colombo, A. Dolhasz, and C. Harvey. A computer vision inspired automatic acoustic material tagging system for virtual environments. In *2020 IEEE Conference on Games (CoG)*, pp. 736–739. IEEE, 2020.

[6] M. Colombo, A. Dolhasz, and C. Harvey. A texture superpixel approach to semantic material classification for acoustic geometry tagging. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

[7] E. Doukakis, K. Debattista, C. Harvey, T. Bashford-Rogers, and A. Chalmers. Audiovisual resource allocation for bimodal virtual environments. *Computer Graphics Forum*, 37(1):172–183, 2018. doi: 10.1111/cgf.13258

[8] B. Hamilton and S. Bilbao. Fdtd methods for 3-d room acoustics simulation with high-order accuracy in space and time. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2112–2124, 2017.

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

[10] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 559–568, 2011.

[11] H. Kon and H. Koike. Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. *Journal of the Audio Engineering Society*, may 2018.

[12] H. Kuttruff. *Room acoustics*. Crc Press, 2016.

[13] E. Lakka, A. G. Malamos, K. G. Pavlakis, and J. A. Ware. Spatial sound rendering–a survey. *IJIMAI*, 5(3):33–45, 2018.

[14] T. Lokki and M. Grohn. Navigation with auditory cues in a virtual environment. *IEEE MultiMedia*, 12(2):80–86, 2005.

[15] S. Pelzer and M. Vorländer. Frequency-and time-dependent geometry for real-time auralizations. In *Proceedings of 20th International Congress on Acoustics, ICA*, pp. 1–7, 2010.

[16] D. Poirier-Quinot, M. Noisternig, and B. F. Katz. Evertims: Open source framework for real-time auralization in vr. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, AM '17. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3123514.3123559

[17] N. Raghuvanshi and J. Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014.

[18] J. L. Rubio-Tamayo, M. Gertrudix Barrio, and F. García García. Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal Technologies and Interaction*, 1(4):21, 2017.

[19] L. Savioja and U. P. Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015. doi: 10.1121/1.4926438

[20] C. Schissler, C. Loftin, and D. Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE transactions on visualization and computer graphics*, 24(3):1246–1259, 2017.

[21] C. Schissler, A. Nicholls, and R. Mehra. Efficient hrtf-based spatial audio for area and volumetric sources. *IEEE transactions on visualization and computer graphics*, 22(4):1356–1366, 2016.

[22] D. Schönstein and B. F. Katz. Hrtf selection for binaural synthesis from a database using morphological parameters. In *International Congress on Acoustics (ICA)*, 2010.

[23] G. Schwartz and K. Nishino. Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1981–1995, 2019.

[24] S. Siltanen, T. Lokki, L. Savioja, and C. Lynge Christensen. Geometry reduction in room acoustics modeling. *Acta Acustica united with Acustica*, 94(3):410–418, 2008.

[25] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori. Image2reverb: Cross-modal reverb impulse response synthesis. *arXiv preprint arXiv:2103.14201*, 2021.

[26] Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha. Scene-aware audio rendering via deep acoustic analysis. *IEEE transactions on visualization and computer graphics*, 26(5):1991–2001, 2020.

[27] Z. Zhang, N. Raghuvanshi, J. Snyder, and S. Marschner. Ambient sound propagation. *ACM Trans. Graph.*, 37(6), Dec. 2018. doi: 10.1145/3272127.3275100