Revolutionizing Immersive Technologies through Spatial Audio, Material Recognition, and Room Acoustic Modelling

Nilay Jain
California State University, Fullerton, California, USA
njain12@csu.fullerton.edu

INTRODUCTION

The advancements in virtual reality (VR) and augmented reality (AR) technologies have been remarkable, revolutionizing various industries and significantly impacting our increasingly digital world. Despite their distinctive features, the terms VR and AR are often used interchangeably, and now the focus has shifted towards mixed reality (MR), a blend of both technologies, exemplified by products such as Hololens.

In light of these technologies' growing interest and potential applications, it is crucial to investigate how they create an immersive experience for users. A detailed examination of the underlying mechanisms is required to provide a thorough understanding of this. This research explores critical components contributing to immersion in VR/AR environments: spatial audio reproduction, material recognition, and room acoustic modeling.

1 APPLICATION DOMAIN

Virtual Reality (VR) and Augmented Reality (AR) have revolutionized the digital world, providing immersive experiences and expanding into Mixed Reality (MR) applications. Notable examples include Snapchat Filters, Google ARCore, and Pokemon GO, offering users unique perspectives to interact with the world. The convergence of VR and AR has further extended the capabilities of these technologies.

In the realm of VR and AR, audio plays a crucial role in providing an immersive user experience. Therefore, in this research paper, we present a simple yet effective approach for detecting the approximate acoustic properties of a room by utilizing 360° cameras for VR/AR applications. The paper delves into the various techniques and approaches for achieving accurate and efficient acoustic modeling for VR/AR environments. Additionally, the paper underscores the importance of plausibility and authenticity in reproducing realistic acoustics for these applications.

The proposed method is evaluated by comparing the generated Room Impulse Responses (RIRs) with the measured RIRs in natural environments. This paper thus contributes to developing more realistic and immersive VR/AR environments by improving the accuracy and authenticity of acoustic modeling.

2 WORKS DESCRIBED IN THE ORIGINAL PAPER *(Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360◦ Images)*

This research paper proposes a novel method to estimate acoustic properties using a 360° camera setup in various environments, including Meeting Room (MR), Usability Lab (UL), Listening Room (LR), and Studio Hall (SH). The proposed method uses two 360° images to capture a panoramic view of the environment, which is then processed to detect the room's geometry and acoustic properties, such as minimal reverberation, uniform sound distribution, appropriate sound levels, low background noise, and minimal echoes.

In achieving this, the proposed method employs depth estimation from the panoramic images and semantic labeling using a Convolutional Neural Network (CNN) to generate a 3D geometric model of the environment. Synchronized spatial audio is reproduced based on the detected acoustic properties and geometric scene of the room. Room impulse responses (RIRs) are generated to validate the proposed method's accuracy, which measures

several acoustic properties such as reverberations and echoes.

The results of the proposed method are compared with the processed audio by rendering the generated scenes with the RIRs. The findings suggest that the proposed method is effective in accurately estimating the acoustic properties of the environment, which can help design spaces with optimal sound quality. The proposed method can be further extended to other environments and applications, such as virtual and augmented reality environments, where high-quality spatial audio is required.

Objectives and Methodologies

The present research proposes a comprehensive approach to estimate the acoustic properties of a given space and measure the synthesized audio. The proposed approach comprises several stages: visual capture and pre-processing, semantic segmentation, depth estimation, 3D modeling, and spatial audio rendering. Let us briefly discuss these methodologies' objectives in brief:

2.1 Visual Capture and Pre-Processing

The present study proposes a two-camera setup to capture the 3D information of a room scene. The setup employs a horizontal perspective to reduce stereo-matching errors caused by texture distortion and minimize camera stoppage. Two 360° cameras are arranged vertically to capture the environment's spatial information. The use of this camera setup enables the panoramic view of the environment to be processed to detect the geometry of the room and objects within it. Avoiding the horizontal setup in the proposed camera setup helps obtain accurate 3D environmental information. The camera setup's effectiveness in capturing the 3D information of a room scene has significant implications for various applications, such as virtual reality, augmented reality, and acoustical engineering. Overall, the proposed camera setup, shown in Fig.1, offers a comprehensive and practical approach for capturing the 3D information of a room scene.

Fig.1

2.2 Semantic Segmentation

The proposed technique divides the scenes into labeled regions with pre-defined classes. Object detection has been challenging due to the difficulty of determining material properties like roughness, density, and surface thickness. Therefore, the approach suggests using object recognition and mapping object categories. The SegNet model, trained on the SUN RGB-D indoor scenes dataset, is used for semantic segmentation and object labeling. The proposed method offers a practical approach to labeling and classifying objects in indoor scenes, thus improving the accuracy and precision of the 3D modeling process.

2.3 Depth Estimation

Depth information of the object in the scene can be estimated using the concept of spherical stereo geometry, as shown in Fig.2. In a 360° vertical camera setup, real-scale depth estimation from stereo images can be achieved by matching 1D vertical lines, as opposed to the conventional method of reconstructing depth from perspective stereo images, which necessitates complex camera calibration processes. This approach simplifies the

depth estimation process and eliminates the need for internal and external camera calibrations, thus improving the overall efficiency and accuracy of the 3D modeling process.

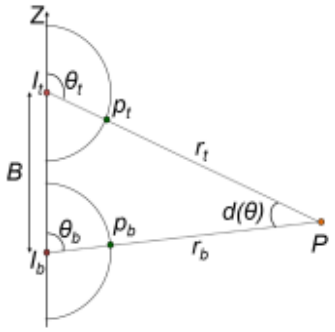

Fig.2

2.4 3D Modeling and Spatial Audio Rendering

After obtaining depth estimates from stereo images, the 2D vertices of captured images are plotted in 3D space. Object labels are then used to group these points into clusters, forming a point cloud. The point cloud is utilized to reconstruct block structures, generating an approximate geometric representation of the scene. This methodology provides an efficacious approach to generating 3D models of indoor environments, improving accuracy and precision in the modeling process.

The Google *'Resonance Audio' package* simulates spatial audio, providing access to 22 different object types and their corresponding acoustic properties. While determining the acoustic attributes of objects using only visual data is challenging, the object labels are mapped to the materials available in the resonance audio library. This methodology improves the accuracy and realism of spatial audio reproduction, providing a practical approach to generating immersive audio experiences. The table below shows the material matching the object:

Object	**Material**		**Object**	**Material**
Ceiling	Wood panel		Furniture	Heavy curtain
Book	Sheetlock		Chair	Wood panel
Floor	Parquet		Object	Metal
Window	Thick Glass		Wall	Smooth Plaster
Sofa	Heavy curtain		Table	Wood panel
TV	Metal		Unknown	Transparent

Table 1

Key Finding and Contributions

This research paper presents an innovative system for estimating room acoustics to create spatial audio for VR/AR applications. The system uses 360-degree images to estimate room geometry and acoustic properties. A simplified 3D model of the scene is generated through depth estimation from captured images and semantic labeling using a convolutional neural network. Frequency-dependent acoustic predictions of the scene characterize the acoustic properties of the environment. Spatially synchronized audio is produced based on the scene's estimated geometric and acoustic properties. The reconstructed scenes are rendered with synthesized spatial audio as VR/AR content. The estimated room geometry and simulated spatial audio results are evaluated against actual measurements and audio calculated from ground-truth Room Impulse Responses recorded in the rooms. The proposed method offers a robust recognition of room geometry and object materials to simulate an acoustic environment in VR/AR platforms. This system provides a promising alternative to direct Room Impulse Response measurements, which can be invasive in practical VR/AR applications.

3 WORKS DESCRIBED IN PAPER 1 *(Material Recognition for Immersive Interactions in Virtual/Augmented Reality)*

In order to deliver a truly immersive experience through spatially synchronized audio, visualization of real-world scenes, and haptic feedback, it is essential to have a deep understanding of the materials comprising the surfaces in the rendered environment. This research paper aims to focus specifically on identifying materials in real-world images captured by a 360° camera setup, as these materials significantly impact the optical and acoustic properties of the rendered objects. By delving into the techniques and methodologies used in the original paper, the aim is to provide a comprehensive analysis of material identification for creating genuinely realistic and immersive virtual environments.

Objectives and Methodologies

In this paper, we present a novel approach that employs the DPT architecture for object tracking. This architecture adaptively determines the importance of various patch resolutions. We also propose several techniques to introduce the components of this architecture. Our evaluation demonstrates the effectiveness of our approach on the LMD and OpenSurfaces datasets:

3.1 Cross-Resolution Feature Extractor

A pioneering approach to object identification with the integration of the DPT architecture with CAM-SegNet. By shifting focus from the entire image to smaller regions or patches, our method enhances computational efficiency. Leveraging a multi-layer perceptron (MLP) within the DPT component further augments learning and comprehension of these patches. Notably, we introduce the dynamic backward attention transformer (DBAT), replacing conventional convolutions with self-attention mechanisms. This integration empowers the program to detect objects adeptly, capitalizing on the self-attention concept for refined feature acquisition within the patches.

3.2 Dynamic Backward Attention Module

The technique incorporates a backward attention module that generates attention masks for every pixel in the image. These attention masks play a crucial role in combining the features obtained from different patch resolutions. By leveraging these masks, our approach effectively integrates cross-resolution information, enabling more comprehensive and accurate feature representation at the pixel level.

3.3 Feature Merging Module

Feature merging module is a crucial part that helps DPT architecture to learn more complementary features compared to its backbone encoder. The attention module is used to detect vital information from the combined features of different patch resolutions.

Key Findings and Contributions

Our proposed Dynamic Patch Training (DPT) architecture significantly improves material segmentation in real-world images for realistic virtual environments. DPT dynamically determines patch resolution dependency, effectively handling variations in material appearance. The Cross-Resolution Feature Extractor merges adjacent features, enhancing learning. The Dynamic Backward Attention Module predicts per-pixel attention masks for comprehensive feature representation. The Feature Merging Module integrates relevant information, improving feature learning. Experimental evaluations demonstrate superior material segmentation accuracy and real-time performance of DPT. Our research advances virtual reality technologies, particularly in immersive audio synthesis, enhancing user experiences.

4 WORKS DESCRIBED IN PAPER 2 *(Room Acoustic Properties Estimation from a Single 360◦ Photo)*

Estimating room impulse responses (RIRs) for real spaces in room acoustic modeling is a resource-intensive process. However, a pioneering computer vision approach utilizes a single 360-degree photo to approximate acoustic material properties. This method reconstructs 3D geometry through monocular depth estimation and semantic scene completion, estimating material properties using transformer-based dense material segmentation. Virtual simulation in Unity with the Steam spatial audio plug-in validates the approach, enabling accurate RIR estimation and material property assessment. The comparative evaluation confirms its transformative and cost-efficient nature, seamlessly integrating computer vision and acoustic modeling for real space applications.

Objectives and Methodologies

The proposed approach incorporates several key techniques to estimate room impulse responses (RIRs) and material properties in real spaces:

4.1 Monocular Depth Estimation:

The proposed approach utilizes a modified U-Net shape encoder-decoder model to estimate depth from a single 360-degree image. By focusing on supervised learning and enhancing stability, the model achieves higher accuracy in predicting realistic scenes. The encoder adopts ResNet50 as the backbone, while the decoder consists of convolution and bilinear upsampling layers. The depth estimation is trained using a combination of Structural Similarity (SSIM) loss and dense depth loss, providing a robust depth map representation.

4.2 Materials Recognition::

Inspired by transformer architectures, the materials recognition module employs a windowed self-attention strategy to extract features from different patch sizes within a single network. The network dynamically determines the patch size based on the input image, enhancing flexibility. Attention masks are predicted to aggregate features, followed by feature pyramid network utilization for shape recovery and pixel-wise material label prediction. This modification significantly improves pixel accuracy in material estimation compared to the preliminary work.

4.3 Semantic 3D Scene Completion:

Building upon previous research, the semantic 3D scene completion module reconstructs a voxel structure by projecting points from estimated depth maps into 3D space. The process includes partitioning the 3D coordinates into multiple view parts and applying semantic scene completion using EdgeNet360. The final output is a complete 3D model with semantic labels, replaced by material labels inferred from the materials recognition module.

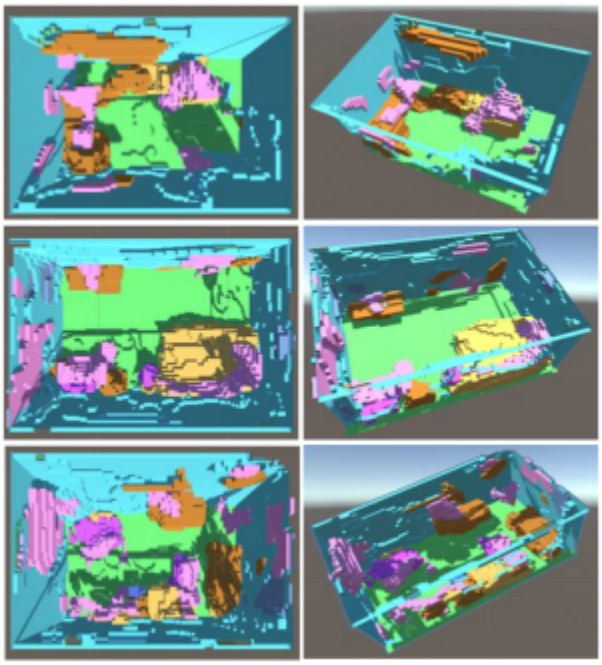

Fig. 3

4.4 Sound Rendering and Room Acoustics Evaluation in a Virtual Space:

The 3D semantic scene is imported into Unity with the Steam Audio plug-in for room acoustics simulation and sound rendering in a virtual space. Binaural sound simulation, incorporating head-related transfer functions (HRTFs), enables the analysis of binaural room impulse responses (BRIRs). Objective metrics such as Early Decay Time (EDT) and Reverberation Time (RT60) are utilized to assess room acoustic properties. EDT measures energy decay from early reflections, while RT60 characterizes reverberation, with average values calculated across specific frequency bands.

Key Findings and Contributions

This groundbreaking research presents a novel approach for estimating room impulse responses (RIRs) and material properties in real spaces. The proposed computer vision-based method demonstrates the feasibility of utilizing a single 360-degree photo to approximate acoustic characteristics. By leveraging monocular depth estimation and semantic scene completion, an accurate 3D geometry model is reconstructed. The transformer-based dense material segmentation technique enables precise estimation of material properties for semantic objects within the scene. The virtual simulation on the Unity platform with the Steam spatial audio plug-in facilitates the estimation of acoustic properties and their evaluation against the real environment. This research contributes a cost-effective and efficient solution that integrates computer vision and acoustic modeling, revolutionizing the estimation of RIRs and material properties in real spaces, with broad implications for room acoustic modeling and design.

5 COMPARISON: ORIGINAL PAPER VS. NEW PAPERS

The research presented in the Original Paper, Paper 1, and Paper 2 collectively represents significant advancements in the field of estimating acoustic properties and material recognition for immersive interactions in virtual and augmented reality environments. These works build upon the foundation laid by the original paper, extending its concepts and methodologies to achieve more accurate and efficient results.

The 'Original Paper' proposes a novel method for estimating acoustic properties using a sophisticated 360° camera setup. It focuses on capturing panoramic views of diverse environments and processing the images to detect room geometry and acoustic characteristics. The paper introduces depth estimation and semantic labeling techniques, coupled with spatial audio rendering, to validate the accuracy of the proposed method. The findings effectively demonstrate the feasibility of estimating acoustic properties for optimizing sound quality in various spaces.

'Paper 1' takes inspiration from the 'Original Paper' and expands its scope by specifically addressing material recognition for immersive interactions in virtual and augmented reality. The primary objective is to identify materials in real-world images captured by a 360° camera setup, as these materials significantly influence the optical and acoustic properties of rendered objects. The paper introduces the Dynamic Patch Training (DPT) architecture, a novel approach that improves material segmentation accuracy. The paper enhances feature learning and representation by incorporating a cross-resolution feature extractor, a dynamic backward attention module, and a feature merging module. The evaluation demonstrates the effectiveness of the proposed approach on different datasets.

'Paper 2' complements the 'Original Paper' by tackling the challenge of estimating room impulse responses (RIRs) and material properties using a single 360° photo. Leveraging computer vision techniques, the paper reconstructs 3D geometry and estimates the material properties of natural spaces. It introduces monocular depth estimation, materials recognition with a transformer-based approach, and semantic 3D scene completion. The virtual simulation in Unity with the Steam spatial audio plug-in enables the evaluation of room acoustic properties. The proposed approach revolutionizes RIR estimation and material property assessment in natural spaces, providing a cost-efficient and efficient solution.

6 CONCLUSION

In summary, 'Paper 1' and 'Paper 2' build upon the work described in the 'Original Paper' by expanding the applications and techniques for estimating material properties, enhancing feature learning and representation, and introducing innovative approaches to estimating room impulse responses. These advancements enable more accurate and immersive virtual and augmented reality experiences, pushing the boundaries of acoustic modeling and design.

7 REFERENCES

[1] Hansung Kim; Luca Remaggi; Philip J.B. Jackson; Adrian Hilton (2019). Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360◦ Images. Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images | IEEE Conference Publication | IEEE Xplore

[2] Mona Alawadh; Yihong Wu; Yuwen Heng; Luca Remaggi; Mahesan Niranjan; Hansung Kim (2022). Room Acoustic Properties Estimation from a Single 360° Photo. Room Acoustic Properties Estimation from a Single 360° Photo | IEEE Conference Publication | IEEE Xplore

[3] Yuwen Heng; Srinandan Dasmahapatra; Hansung Kim (2023). Material Recognition for Immersive Interactions in Virtual/Augmented Reality. Material Recognition for Immersive Interactions in Virtual/Augmented Reality | IEEE Conference Publication | IEEE Xplore