Acoustic Room Modelling using 360 Stereo Cameras

Hansung Kim, Luca Remaggi, Sam Fowler, Philip JB Jackson, *Member, IEEE* and Adrian Hilton, *Member, IEEE*

***Abstract*—In this paper we propose a pipeline for estimating acoustic 3D room structure with geometry and attribute prediction using spherical 360° cameras. Instead of setting microphone arrays with loudspeakers to measure acoustic parameters for specific rooms, a simple and practical single-shot capture of the scene using a stereo pair of 360 cameras can be used to simulate those acoustic parameters. We assume that the room and objects can be represented as cuboids aligned to the main axes of the room coordinate (Manhattan world). The scene is captured as a stereo pair using off-the-shelf consumer spherical 360 cameras. A cuboid-based 3D room geometry model is estimated by correspondence matching between captured images and semantic labelling using a convolutional neural network (SegNet). The estimated geometry is used to produce frequency-dependent acoustic predictions of the scene. This is, to our knowledge, the first attempt in the literature to use visual geometry estimation and object classification algorithms to predict acoustic properties. Results are compared to measurements through calculated reverberant spatial audio object parameters used for reverberation reproduction customized to the given loudspeaker set up.**

***Index Terms*—Indoor geometry reconstruction, Audio-visual processing, Room acoustic modelling, Geometrical acoustics**

I. Introduction

AUDIO-VISUAL data is the most familiar format of multimedia information acquired in our daily life. most cases, they are already paired as audio-video streams used in numerous fields such as media production and reproduction [1], games [2], and education [3]. Audio and image processing have been investigated as separate research areas, typically ignoring their synergy when they work together. Recently, some works have been proposed to exploit their multimodal information, for applications such as speaker tracking [4], speech recognition [5], and event detection [6]. In this paper, we apply computer vision techniques to support audio reproduction adapted to the acoustics of a specific location.

The motivation here stems from spatial audio reproduction, where knowledge of the acoustics of a space could allow for more accurate reproduction of a captured environment, or for reproduction room compensation techniques to be applied. In the acoustic design of spaces, either existing or at the planning stage, Room Impulse Responses (RIRs) can be used to predict aspects such as strong echoes or Reverberation Time (RT60) [7], which can improve the overall designed acoustic. Through application of spatial audio techniques these environments can also be reproduced/auralized, allowing the listener to experience a space without being there. Although RIRs can provide accurate information about the acoustics of a room at a specific location, they are inherently restricted to pre-existing spaces, and the number of required measurements for some applications can rapidly become impractical. Acoustic measurements are sometimes difficult to obtain, especially, considering our daily environments. For instance, recording setups may be too invasive to be deployed at a listener's home, and typical techniques, such as the swept-sine [8], may be too intrusive to be adopted by final users. Multiple set-up and recordings are required to get RIRs at multiple points in the room. If any RIRs in certain points are missing in the first recordings or major reflective objects in the room are moved, repeated setting and recordings in the room are inevitable.

Acoustic predictions from images offer an attractive alternative. In computer vision, estimating semantic room geometry is a classic problem with a wide range of applications. There have been many studies on room geometry estimation from small and simple visual sensors [9], [10]. With the progress of deep learning techniques, there has been good improvement in semantic 3D scene reconstruction to identify known objects in the scene together with the room geometry [11], [12]. One area of interest is the application of vision-based geometry estimation to compensation techniques for spatial audio reproduction in rooms [13]. If the RIR at the listening position for each loudspeaker is known, it is possible to adjust the loudspeaker signals to compensate for alterations in the frequency response, strong early reflections, or to some extent the level of reverberation [14]. This is particularly the case in the context of recent interest in object-based audio, where more control is passed to a renderer at the reproduction end [15]. For instance, recorded RIRs can be parameterized to generate Reverberant Spatial Audio Objects (RSAOs) [16]. However, by estimating the room geometry, predictions can be made when acoustic measurements are not available. This scenario also fits new research areas, such as mixed reality [17].

For simulation of an acoustic environment a robust method for obtaining room geometry is required. Recognition of 3D structure and material properties using Red-Green-Blue (RGB) [18], [19] or RGB+Depth (RGB-D) [20] images have been important problems. However, current approaches using normal perspective or RGB-D cameras have the following limitations for complete indoor semantic scene reconstruction. First, indoor scenes generally include textureless and non-Lambertian surfaces which result in errors in feature detection

H. Kim is with the School of Electronics and Computer Science, University of Southampton, Southampton, UK. e-mail: h.kim@soton.ac.uk

L. Remaggi is with Creative Labs UK. E-mail: luca_remaggi@cle.creative.com

S. Fowler, P.J.B. Jackson and A. Hilton are with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK. e-mail: a.hilton@surrey.ac.uk

and matching causing incorrect 3D geometry estimation. Second, normal or RGB-D cameras have limited field-of-views (FOV) capturing only a part of the whole environment. For a complete scene layout estimation, multiple input images from different views and a fusion technique are required.

Cuboid-based simplified room geometry modelling using a pair of spherical 360° cameras (hereafter referred to as 360 cameras) provides a potential solution for the above problems. For cuboid model reconstruction, room interiors are assumed to be composed of planar surfaces aligned to the main axes (Manhattan world) as introduced in [21]. Although not always the case, room layouts and larger objects in the room often fit with this assumption. It is well-known that human auditory perception is not sensitive enough to recognize differences of sound from small changes of geometrical details [22]. It is also known that visual information increases plausibility of rendered audio for an environment [23]. Complex representations of the scene require a high computational cost and run time, making acoustic simulation impractical in many cases. In contrast, approximated geometry allows the use of simple acoustic models to generate synthetic versions of the environment's acoustics in an efficient way. Therefore, simplified geometric models have been commonly used for room acoustics modeling [24].

In this paper, we propose a cuboid-based semantic room layout modelling pipeline for room acoustics estimation using a pair of off-the-shelf consumer 360 cameras. Two spherical 360 images captured by a pair of cameras are used to produce a complete scene model with semantic object information. The approach assumes that room interiors are composed of planar surfaces aligned to the main axes (Manhattan world) as proposed in [21]. Cuboid-based scene elements are detected and aligned to the main axes using stereo matching, and their object classes are predicted by a convolutional neural network trained for semantic segmentation (SegNet) [25]. This produces a complete scene model with a compact representation for acoustic predictions.

Preliminary versions of the methods presented in this paper have been previously published. In [26], indoor scene modelling with object and material attribute information using a multi-scale Convolutional Neural Network (CNN) has been proposed. In [27], a frequency-dependent acoustic prediction technique based on geometrical room modelling has been introduced. This is, to our knowledge, the first attempt in the literature to use visual geometry estimation and object classification algorithms to predict acoustic properties. However, cuboids are inferred from planar surface detection and materials are manually assigned in [27]. The main novelties of this paper distinguished from our preliminary works are:

- Cross-disciplinary integration of computer vision and audio processing to enable plausible acoustic simulation and adaptation of audio reproduction to the environment.
- Complete end-to-end system architecture proposed for acoustic room modelling using off-the-shelf consumer 360 cameras.
- Method proposed for acoustic room modelling using visual semantic segmentation and recognition (Section III-C and E)
- Method proposed for cuboid-based room layout and object geometry reconstruction using point-cloud occupancy (Section III-D)
- Objective evaluation of visually-estimated room acoustics conducted in terms of RSAO parameters.
- Public audio-visual data sets released with visual captures and ground-truth RIRs

The rest of this paper is organised as follows: Section II provides survey of recent works in relevant research fields and Section III overviews the proposed system and describes details of the proposed methods. Section IV introduces the system set up and data sets used for the evaluation. Experimental results and discussion are given in Section V, and Section VI makes conclusions of this paper.

II. Related Work

A. Approximated room geometry reconstruction

Indoor 3D scene reconstruction has been a long-standing area of research. Multi-view stereo and structure from motion methods using multiple photos or videos have been widely investigated [9], [28]. As low-cost RGB-D cameras have become readily available, various 3D reconstruction methods have been proposed using colour and range data. KinectFusion [29] made a great impact on real-time dense scene reconstruction with a RGB-D camera and has been extended for large scale scene modelling. Public RGB-D indoor datasets for the benchmark assessment have also been presented including ICL-NUIM [30], SUN3D [31], NYU [32], [33]. However, the limited FOV presents a challenging problem to ensure complete scene coverage for reconstruction as mentioned. In order to cover the occlusion and FOV problems, 3D scene completion was proposed by Song et al. [34]. From a given single RGB-D image, they build a semantically labelled 3D voxel structure including occluded and non-surface regions. This was extended to extrapolate 360 structure beyond the FOV [35]. However, the performance is strongly dependant on the training sets. Recently Dai et al. [36] proposed a self-supervised scene completion of RGB-D scans but this requires RGB-D video input.

Spherical imaging provides a solution to overcome this coverage problem. Schoenbein et al. [37] proposed a high-quality omnidirectional 3D reconstruction of Manhattan worlds from catadioptric stereo video cameras. However, these catadioptric omnidirectional cameras have a large number of systematic parameters including the camera and mirror calibration. In order to get high resolution spherical images with simple and accurate calibration and matching, Point Grey developed an omnidirectional multi-camera system, the Ladybug[1]. Spheron developed a line-scan camera, Spheron VR[2], with a fish-eye lens to capture the full environment as an accurate high resolution / high dynamic range latitude-longitude image. We used this Spheron VR for simplified scene modelling in our preliminary works [26], [38]. Li [39] has proposed a spherical

[1]Ladybug, https://www.flir.com/iis/machine-vision/spherical-vision-systems

[2]Spheron, https://www.spheron.com/products.html

image acquisition method using two video cameras with fish-eye lenses pointing in opposite directions. The biggest problem of the spherical stereo imaging from fish-eye lenses is large errors around epipoles and complex search along conic curves for stereo matching. This problem has been solved with accurate calibration and rectification. Various inexpensive off-the-shelf 360 cameras with two fish-eye lenses have recently become popular[3,4,5].

Dense depth map estimation pipeline using a narrow-baseline video clip captured by a 360 camera was proposed by Im et al. [40], but the reconstructed scenes have large incomplete regions due to the occlusion from the camera. Acoustic rendering requires an air-tight structure so that the sound does not escape from the space. One shortcoming of learning-based scene completion and recognition using 360 cameras is the lack of ground-truth 360 scene data. Recently, a few 3D 360 datasets have been released such as Stanford 2D-3D-Semantics dataset [41] and Matterport 3D [42], but the Stanford 2D-3D-Semantics comprises only 6 academic buildings and Matterport 3D covers only 90 private homes. Moreover, they are not directly captured from 360 cameras but rendered from 3D point clouds. Therefore, we propose to build an approximated complete geometry with cuboids and estimate object materials using 360 semantic segmentation by cubic projection and image decomposition to utilise existing RGBD datasets.

B. Semantic segmentation

Semantic segmentation aims to label every pixel in the image into a set of known classes. Zhu et al. [43] provide a good survey of semantic segmentation methods using RGB images. After the breakthrough results on ImageNet [44], the traditional pipeline of semantic object classification has been replaced by CNN [45]. Semantic segmentation CNN architectures continue to evolve, improving segmentation accuracy whilst shortening training times, reducing complexity and providing real-time performance. Modern CNNs for semantic segmentation often use an encoder-decoder architecture to exploit features learnt in object classification architectures [46] which internally encode input images into low resolution feature maps. After pre-training on large datasets [47] the produced feature maps provide strong representations of scene objects. Semantic segmentation networks [25], [48]–[50] implement decoder components that learn to map these low resolution feature maps to input resolution predictions for classification.

CNNs have been used for semantic object detection and segmentation in various ways [51]–[53]. Eigen and Fergus [19] proposed an hierarchical fully convolutional networks (FCN) architecture composed of three scales. The first scale is VGG-FCN [52], and its output is up-sampled, concatenated with a higher resolution version of the input images at the next scale. The same process occurs at the interface between the second and third scales. Noh et al. [48] build deconvolutional and unpooling layers on top of VGG 16 [46], utilising the pooling indices of encoded feature maps during upsampling. SegNet [25] follows a similar scheme but reduces the number of training parameters by discarding the fully connected layers of VGG 16. This proves to lower memory consumption and improve training and inference times whilst maintaining state-of-the-art performance.

[3]Insta360, https://www.insta360.com/

[4]Go Pro MAX, https://gopro.com/en/gb/shop/cameras/max/CHDHZ-201-master.html

[5]Ricoh Theta, https://theta360.com/en/

C. Acoustic room modelling

Accurate models of a RIR characterizing an acoustic environment provide acoustic attributes that allow to reproduce the acoustics of the space under investigation [7]. Classical state-of-the-art methods to approximate RIRs are the image source [54] and the ray tracing [55] methods. Although these methods can accurately model the early reflections, they are not able to correctly approximate the late reverberation, due to their point-based representation of the reflections. Furthermore, they do not consider phenomena like diffraction, thus a small surface at low frequency will be wrongly modeled with pure classical methods [56]. During the last decades new approaches have been proposed to better model RIRs, with a particular attention on approximate the low frequency modal propagation. With this purpose, popular approaches are the finite difference time domain (FDTD) [57], [58] and the digital waveguide (DWG) [59]–[61]. Other approaches focus on mainly modeling the late reverberation, by approximating its Gaussian statistics [62]. The power spectrum that is indicative of the size of the space and absorbing power of the materials is also typically considered using the Sabine's equation [63]. In this article, we employ an acoustic room model that joins the strengths of state-of-the-art approaches, by forming a standard hybrid approach [56]: the early reflections are modeled by using the image source method [54], the later reflections by using the ray tracing [55], and the late reverberation by following a statistical approach [63].

III. Proposed method

A. System overview

In this research, we propose a simple and efficient method to estimate acoustic RIRs from visual capture. The examples presented in the block diagram of Fig. 1 show the process for acoustic room modelling in a normal room environment. A full surrounding scene is captured by spherical 360 cameras at two different heights. Each image is mapped to a latitude-longitude (equirectangular) image and they are aligned to the room coordinate axes by cubic projection and line alignment. Then the process is split into two processes: 3D reconstruction and semantic object segmentation. Depth of the scene is estimated by stereo matching between two images. For semantic scene segmentation and object recognition, the spherical 360 image is projected onto a cube centred on the camera giving perspective images. Object regions are detected with SegNet from each projected perspective image, and the labels are back-projected to the original equirectangular format. Labelled cuboid structure is reconstructed from the depth information and semantic segmentation. This geometry

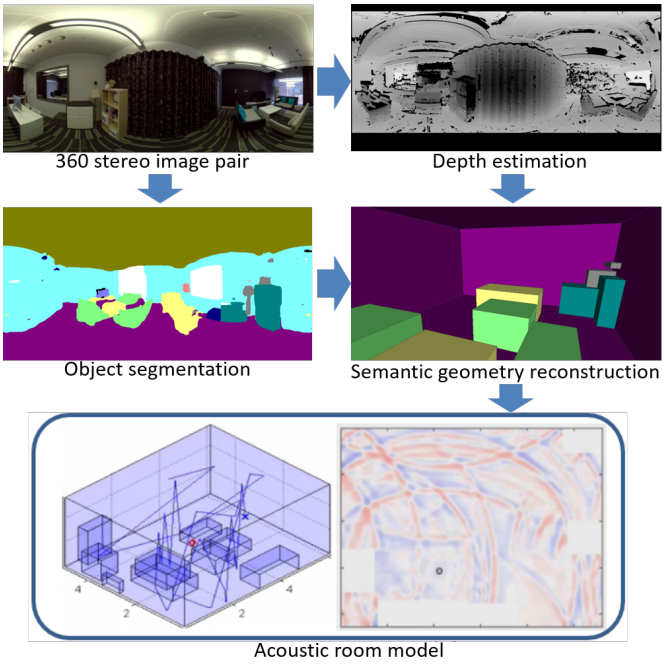


Fig. 1. Block diagram of the proposed system

and object information is used as an input to the acoustic room modelling pipeline. Frequency dependent acoustic simulations are broken down into three sections: early reflections derived from an image source model (ISM) (providing a deterministic early temporal response); later reflections and onset of the reverberant decay follow a ray tracing approach; and the late reverberant tail using Gaussian shaped and filtered white noise, with an envelope based on the decay of the preceding solution.

B. Visual capture system and Manhattan-world alignment

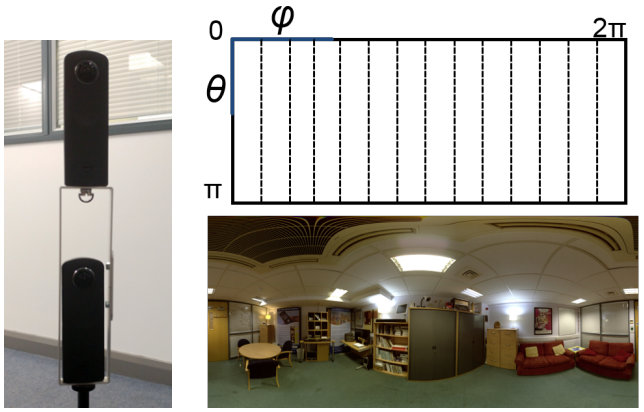
To recover 3D scene information, the scene is captured as a vertical stereo image pair. It can be captured by one camera at two different heights for a static scene or by a vertically aligned camera pair at the same time as shown in Fig. 2 (a). Two Theta S cameras by Ricoh are used in our experiments because it provides well-aligned seamless stitching with minimal distortion in mapping to the Spherical coordinates among the 360 cameras introduced in Section II-A. Photos acquired from two pre-calibrated fish-eye lenses are stitched to each other to generate an equirectangular projection image as illustrated in Fig. 2 (b).

A vertical stereo set up is used rather than typical horizontal stereo. There are several advantages in using vertical stereo for 360 captures:

1) Depth can be estimated by simple 1D stereo matching along the vertical longitude line, while horizontal stereo requires a complex search along conic curves [64]

2) The other paired camera is visible and occlude large portion of the scene in the horizontal stereo set up.

3) Depth errors resulting from incorrect stereo matching increase as the elevation angle to the baseline decreases as reported in [65]. This error diverges to infinity at the poles (blind spot). The vertical stereo system makes these blind spots on the ceiling and floor which are less important and can be easily concealed by neighbouring information, while the horizontal stereo system makes the blind spots on the side which may include important scene information.



(a) Camera set up (b) Equirectangular projection image

Fig. 2. Visual capture system

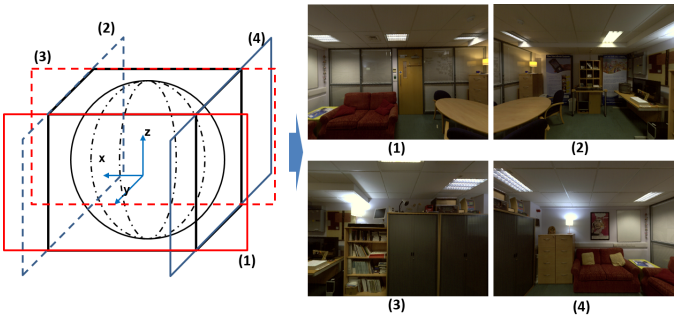


Fig. 3. Cubic projection and projected side images

Even though the baseline of the vertical stereo camera system is perpendicularly aligned to the ground, the spherical coordinate of each 360 camera can be misaligned either to each other or to the world (room) coordinate system. For image alignment to the room coordinate (Manhattan-world) system, cubic projection and Hough-line based optimisation as proposed in [26] are utilised. A Manhattan-world cubic projection image gives good advantages in piece-wise planar scene reconstruction as it generates central-point perspective projection images in which horizontal and vertical lines in the scene are aligned to horizontal and vertical directions in each projection image, respectively, and the lines aligned to the depth direction converges to the image centre as observed in Fig. 3. The optimal α (X-axis), β (Y-axis) and γ (Z-axis) rotation angles to align the image to the Manhattan-world are found by Eq. (1), where k represents the k-th face image in the cubic projection, H the lines detected by the Hough line detection, and C the cubic projection of the image I. The Hough lines are categorised into general Hough lines H, horizontal Hough lines H^h, and vertical Hough lines H^v, where

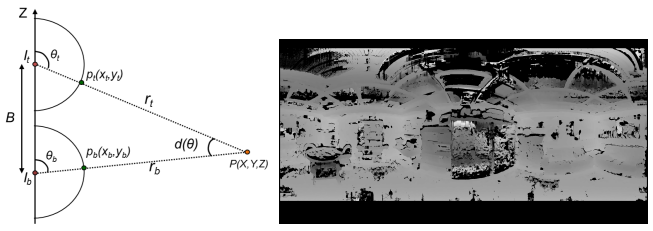
horizontal and vertical Hough lines represent those detected parallel and perpendicular to the horizon within 1° of angle tolerance. The optimal rotations are detected by maximising the following term. We use a greedy search with multi-scale sampling within the range of -45° $< (\alpha,\beta,\gamma) <$ 45° as the alignment is repeated every 90°.

$$(\alpha_{opt},\beta_{opt},\gamma_{opt}) = \underset{\alpha,\beta,\gamma}{\operatorname{argmax}} \sum_{k=1}^{6} \frac{|H_k^h(\alpha,\beta,\gamma) \cup H_k^v(\alpha,\beta,\gamma)|}{|H_k(\alpha,\beta,\gamma)|}$$
$$H_k(\alpha,\beta,\gamma) = H(C_k(R(\alpha,\beta,\gamma)I(x,y,z))) \tag{1}$$

C. Semantic scene segmentation and object labelling

Semantic segmentation of the scene is performed with SegNet [25] to provide pixel-wise object labels of the scene. SegNet is a deep fully convolutional neural network architecture for semantic segmentation, designed to be efficient during training and inference whilst maintaining state-of-the-art performance. The network employs an encoder-decoder architecture, applying the first 13 convolutional layers of VGG16 [46] to encode an input image into low resolution feature maps before upsampling (decoding) them into sparse feature maps. SegNet's novel decoder utilises max-pooling indices memorised during encoding to upsample, reducing the memory required during inference significantly. Per-pixel class probabilities are the final output of the system after a multi-class soft-max classifier is applied to the decoder's final output.

The SegNet implementation [25] provides a model trained on the SUN RGB-D indoor scenes dataset [66] to semantically segment structure and objects in images of indoor scenes (only the RGB colour channels are used in the architecture). To determine object labels, a captured spherical (equirectangular) image is projected onto planes on a unit cube in the Cartesian domain to provide six perspective images of the scene. Each plane is set to 4:3 aspect to match to the trained SUN RGB-D dataset format, and to compensate recognition error at the image boundaries. Due to cuboid alignment, two of the images are directed towards the ceiling and floor and are classified as such. The other images are individually inferred using the trained model to provide four semantically segmented images. Figure 3 shows the extended cubic projection and an example of projected side face images for the equirectangular image in Fig. 2 (b). All six output labelled images are back-projected to provide a fully labelled equirectangular image the same dimensions as the captured spherical image. Utilising this backprojection technique permits the use of segmentation models trained on standard indoor scene datasets without requiring currently unavailable large scale labelled spherical image datasets. Finally, the labelled image is refined by morphological opening process [67] to separate partially connect objects with the same label and smooth object boundaries. Small regions are eliminated to generate simplified scene structure. Each labelled region is indexed to be considered as independent object reconstruction in Section III-D.



(a) Spherical stereo geometry (b) Disparity (depth) map

Fig. 4. Depth estimation using a pair of spherical stereo images

D. 3D Geometry Reconstruction

3D geometry of the scene is reconstructed using dense stereo matching with spherical stereo geometry illustrated in Fig. 4 (a). If we assume the angles of the projection of the point P onto the spherical 360 image pair displaced along the Z-axis are θ_t and θ_b respectively, then the angle disparity d of point $p_t(x_t,y_t)$ can be calculated as $d(\theta) = \theta_t - \theta_b$. The distance of the scene point P from the top camera is calculated by triangulation as Eq. (2), where B is the baseline distance between the camera's center of projection.

$$r_t = B / \left(\frac{\sin\theta_t}{\tan(\theta_t + d)} - \cos\theta_t \right) \tag{2}$$

Stereo matching can be carried out along the epipolar lines which are vertical column lines in the equirectangular vertical stereo images. Recently deep learning based stereo matching algorithms have made a good progress in depth estimation from image pairs [68]–[70] as well as classical matching algorithms [71]. Deep learning-based methods require a large training dataset and long processing time for equirectangular image pairs. As only an approximated geometry is required in the proposed pipeline, we use a classic dense stereo matching method [65] incorporating a region-diving technique [72] which quickly produces reliable disparity fields for the complete scene by detecting occluded and ambiguous regions based on bi-directional consistency and the ordering constraint. Figure 4 (b) shows the disparity map estimated from Fig. 2 (b). Black regions indicate occlusion or unmatched areas. $0° \leq \theta < 5°$ and $165° < \theta \leq 180°$ regions have been cropped because depth from disparity near the epipole areas (blind spots) is unreliable.

All image points on the equirectangular image are projected to the 3D space using the spherical stereo geometry and form a 3D point cloud. This point cloud is segmented into object clusters based on the labels assigned in Section III-C. Kwon et al. [73] proposed cuboid fitting algorithm for 3D point clouds using least squares optimisation. Nguatem et al. [74] and Li et al. [75] used plane detection as primitives of cuboids for outdoor LiDAR scans. However, plane-based approaches do not work well for the proposed pipeline because: 1) indoor objects have more arbitrary shapes than outdoor structures; 2) reconstruction using stereo matching has more errors on surface reconstruction than LiDAR scans. We propose an occupancy based cuboid reconstruction method. Instead of detecting planes or major axes, cuboid primitives aligned to a Manhattan world are fitted to the point cloud clusters. The

volume of the cuboid is decided by the 3D point occupancy in the cluster. In order to eliminate outliers from depth estimation and segmentation errors in the cluster, we exclude 10% of the farthest points from the centre of cluster. Finally the volume of the reconstructed cuboids are refined by the physical stability [21]. For example, floating cuboids above the ground violate the law of gravity. Any cuboid which is not supported by another stable object is extended to the ground to retain the physical stability. Cuboids near the wall are also extended to the wall because objects commonly abut the wall and these objects increase the complexity of the scene and may cause resonance in the sound field rendering.

E. Acoustic room modelling

The acoustic RIR modelling was achieved using a geometrical acoustic approach [7]. Whilst this method is generally more accurate for medium to large scale spaces, the technique is suited to medium to high frequencies and provides a useful estimate of time and direction of arrival of predicted reflections [7]. The implementation of this model was introduced in [27], and allows the generation of synthetic RIRs assuming omnidirectional microphones.

For each source and microphone pair, the model was broken down into 3 sections for efficiency. The first early reflections were modelled using an image source method technique [76], which provides a more deterministic estimation of the early temporal response than stochastic methods. The later reflections were modelled stochastically using a ray tracing approach [77], with the scattering coefficient used to determine the probability of specular and non-specular reflections. The temporal threshold separating early and late reflections was calculated as the median of the second order reflection times of arrival (TOAs). This was done to model most of the first order early reflections by employing the image source method. On the other hand, the late reverberation was modelled as Gaussian white noise, with an exponential decay defined by the ray-traced solution. The response was calculated in octave bands from 63 Hz to 8.0 kHz, with a summation providing the wideband result. The reverberation onset time (also known as the mixing time) was calculated from the visually estimated room geometry. The estimated room volume V and total reflective surface area S were combined to calculate a model-based perceptual mixing time [15], [78]:

$$T_{\text{mix}} = 20 \cdot \frac{V}{S} + 12, \tag{3}$$

in milliseconds. An example of a simulated RIR is reported in Figure 5.

The acoustic properties of the materials are selected indirectly, based on output from visual classification algorithms, such as the one we propose in Section III-C. The classification labels each surface's object class, for which a corresponding material is defined. However, even the same material can have slightly different acoustic coefficients according to the density or surface condition. A list of acoustic absorption and scattering coefficients per material are given in [79]. Median values of the coefficients per material are taken for the acoustic absorption and scattering coefficients in our experiments.

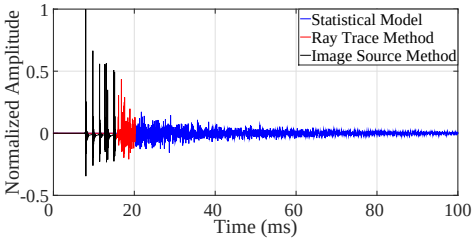


Fig. 5. RIR simulated by joint techniques.

IV. System set up and datasets

To support research into audio-visual data processing and demonstrate the performance of the proposed room acoustics estimation method, we made a dataset web page publicly available with our 360 image and corresponding ground-truth RIR data acquired in various indoor environments. Captured data, source code, and music and speech samples convolved with recorded (groundtruth) and estimated RIRs are all available at: http://cvssp.org/data/s3a/public/AV-Analysis2/.

A. Datasets

The proposed pipeline was evaluated over different spaces: Spatial audio listening room (hereafter referred to as LR, Fig. 6 (a)), Usability lab (hereafter UL, Fig. 1), Meeting room (hereafter MR, Fig. 2 (b)), and Studio hall (hereafter ST, Fig. 6 (b)). These four datasets were recorded with a bi-circular array of 48 omnidirectional microphones, to enable spatial analysis and RSAO parameterization as described in [16]. For comparison, our pipeline's acoustic modelling tool (described in Section III-E) synthesized RIRs to a virtual array of omnidirectional microphones at matching locations. The LR is an acoustically controlled listening environment with loudspeakers surrounding a central listening position, and reverberation time (RT60) of about 220 ms, averaged over the octave bands between 250 Hz and 8 kHz. The UL (RT60 about 280 ms) and MR (RT60 about 270ms) are by design more representative of typical domestic living room environments. ST is a large hall with a RT60 of about 910 ms.

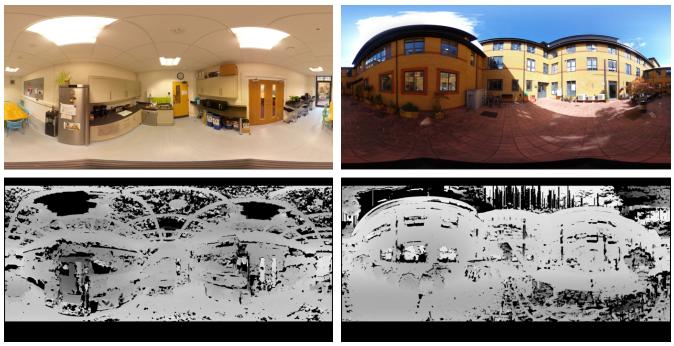
Two additional datasets, Kitchen (hereafter KT, Fig. 6 (c)) and Courtyard (hereafter CY, Fig. 6 (d)), were employed to test our pipeline across a broader variety of spaces. For the acoustic captures a First-order Ambisonics microphone was utilized to yield B-Format RIRs. The KT is a narrow and long room which has different acoustics, and the CY is an outdoor space surrounded by walls. As these signals require an alternative pipeline for the parameterization [80], these results are excluded from the acoustic evaluation that follows, for the sake of consistency.

B. Ground-truth acoustic measurements

For each test environment a series of RIR measurements was taken using the swept sine method [8]. These RIRs were utilized to then generate RSAOs [15], that were employed as ground-truth for our vision-based RSAO production.

RIR recording was performed using 48 microphones, evenly spaced around two concentric circles of radii 8.5 cm and

(a) Listening room (LR) (b) Studio (ST)



(c) Kitchen (KT) (d) Courtyard (CY)

Fig. 6. Four 360 capture datasets (For each dataset, Top: Upper camera image; Bottom: Estimated disparity). Black regions in the estimated disparity maps are unknown regions due to ambiguous matching or stereo occlusion.

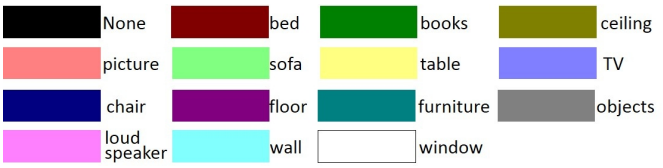
(a) Microphone array (b) Recording with loudspeakers

Fig. 7. Ground-truth acoustic measurement for UL

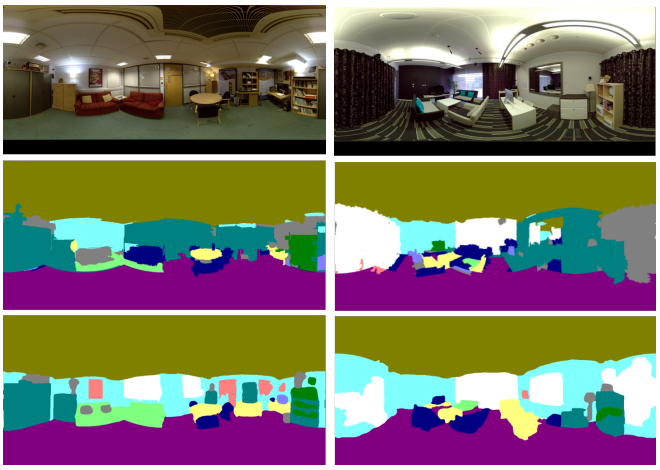
10.6 cm, respectively, to form a custom array [81] as shown in Fig. 7 (a). Furthermore, at the center of the circular array, a soundfield microphone was placed. This recorded additional RIRs that were used to avoid the up-down ambiguity produced by the planar circular array. Both living room style environments had loudspeaker setups based on an ITU 5.0 surround sound setup [82], whereas the LR included a high channel count setup, formed by 32 loudspeakers [83]. Figure 7 (b) shows a snapshot of ground-truth recording for UL. In ST, four loudspeakers were deployed for the measurements, at a height of 1.50 m. Three of them were 2 m away from the microphone array and azimuth of 0 and ± 45 degrees, whereas the fourth one was at 0 degrees azimuth and 3 m distance.

V. Experimental Evaluation

The test scenes have been captured by two Theta S cameras which have their own built-in fisheye cameras internally



(a) Object colour index



(b) Meeting room (MR) (c) Usability lab (UL)

Fig. 8. Object Labelling results (in (b) and (c), Top: Original image, Middle: Result by Eigen-based method [19], Bottom: Result by SegNet-based method)

calibrated. The baseline distance between two cameras has been set as 11cm and the image resolution is 3000×1500. RIRs generated from the room geometry are parameterised following the RSAO concept [15]. These parameters were compared with the ground-truth extracted from the recorded RIRs, to evaluate the room geometry estimation accuracy.

The 3D reconstruction and recognition process ran on a PC with an Intel Core i7 3.40 GHz CPU and 32G RAM. It took less than 5 mins for the whole geometry reconstruction process including pre-processing, depth estimation and cuboid reconstruction for any data set. The semantic segmentation took around 3 mins on an NVIDIA Tesla M2090 GPU with 5GB memory run in parallel. In a real environment, the whole process from camera setting to the final model output can be done within half an hour. Considering the real RIR measurements in Section IV-B takes about a half day per room including system set-up, recording and tidy-up, the proposed approach is much simpler and faster than current audio-based approaches.

A. Object Labelling

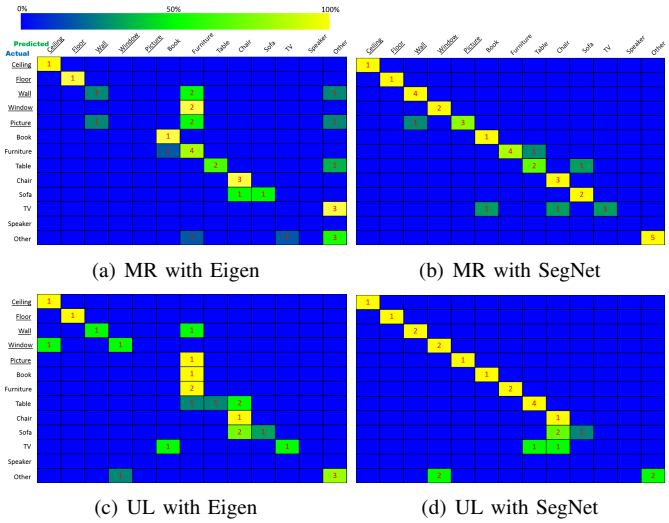
For the evaluation of semantic scene and object labelling, we compared the results of SegNet for spherical imaging in Section III-C (SegNet) with the results of Eigen and Fergus [19] (Eigen) trained for NYUDepth v2 dataset. The objects were labelled with the 15 classes indexed in Fig. 8 (a). We mapped the object labels to material properties as shown in Table I, and assigned frequency-dependent absorption coefficients in the material list given in [79].

Fig. 8 (b) and (c) show the original images and estimated object class labels for the MR and UL. We can clearly observe

TABLE I
MATERIAL MATCHING TO OBJECT LABELS

Object	Material	Object	Material
None	Transparent	Chair	Wood panel
Bed	Heavy fabric	Floor	Heavy fabric
Books	Paper	Furniture	Wood panel
Ceiling	Wood panel	Object	Plastic
Picture	Wood panel	Loud Speaker	Plastic
Sofa	Heavy fabric	Wall	Smooth plaster
Table	Wood panel	Window	Thick glass
TV	Metal		

that the proposed SegNet-based method produced more accurate and meaningful segmentation than Eigen. Most objects have been correctly classified including windows and mirrors. However some small objects were missing due to the post-processing of the SegNet. Figure 9 shows a 13×13 confusion matrix ("None" and "Bed" were not considered) for object recognition results. Underline items are objects consisting the room boundaries and other items are objects in the room. The Blue-Yellow colour map represents the ratio of recognised items. We manually generated the ground-truth segmentation label map, and we considered the estimated semantic labels are correct for a certain ground-truth region if over 70% of pixels are correctly labeled in the region. For the MR scene, some Wall, Window and Pictures were recognised as Furniture with Eigen but most of them were correctly labeled with the proposed SegNet-based method. In the MR scene, 29 of 34 objects (85.3%) were correctly recognised with the proposed method while only 17 objects (50%) were correctly recognised with Eigen. The UL scene was more challenging because the lighting condition was bad and the scene was captured not at the centre but at the side of the room. Some false labels are observed between Furniture, Table, Chair and Sofa with Eigen. Some sofas were mislabelled as Chair with SegNet because only the side or back of those sofas were visible in the scene. In manual ground-truth generation, curtains were annotated as Objects because they are not listed in the index. However, they were predicted as Window with both methods. In the UL scene, 18 of 24 objects (75%) were correctly recognised with the proposed method while only 13 objects (54.2%) were correctly recognised with Eigen.



(a) MR with Eigen (b) MR with SegNet

(c) UL with Eigen (d) UL with SegNet

Fig. 9. Comparison of confusion matrices with Eigen-based method and SegNet-based method for Meeting Room (MR) and Usability Lab (UL)

B. Geometry Reconstruction

The 3D geometry of each test scene was reconstructed with the spherical stereo geometry in Eq. (2) from the estimated disparity map in Fig. 6. The region-diving technique [72] efficiently produced accurate depth maps for reliable regions by masking out ambiguous areas, such as texture-less regions. The erroneous regions most visible in the equi-rectangular depth maps in Fig. 6 are ceilings and floors near the poles in the spherical coordinate system, where the areas are stretched and exaggerated when they are converted into longitude-latitude images. Those ceiling and floor regions near the poles are not very important in scene understanding (both semantically and geometrically) as stated in Section III-B.

Figure 10 shows the reconstructed cuboid-based models with colour-coded object labels. The ceilings and floors were coloured with the wall label in order to represent the room layout as one cuboid. For efficient geometry representation, Picture and Window labels were merged to Wall, and Book labels are merged into Furniture in the final scene reconstruction. Table II shows the manually measured ground-truth room layouts and comparative evaluation of estimated errors in room diagonal against the ground-truth (Err) and number of reconstructed objects(# Obj). The layout estimation errors by the proposed method are smaller than those by our previous plane-based method (Kim16 [26]). Kim16 shows equivalent estimation for the LR dataset, which has four large, clean walls (2 % of room diagonal error for both methods) but poor estimation for ST (14.8 % diagonal error) and KT (23.6 % diagonal error) because it failed to detect one side wall due to a featureless wall and transparent door. In MR, it falsely recognized a side face of another object in the middle of the room as the boundary. Otherwise, the estimated room dimensions are very close to the ground-truth (0.2 % error). The errors of the wall positions in the ST scene (0.55m in width and 0.32m in length) are not large considering the room size, but the height of the room was estimated incorrectly (0.8 m error) due to the uneven ceiling with rails and air-conditioning ducts. The depth estimation accuracy for spherical stereo is inversely-proportional to distance and decreases with elevation angle as shown in [65]. The Err for CY is not based on volume error but area error (width $\times$ length), as the CY scene does not have a ceiling. It is difficult to quantitatively evaluate the detection and reconstruction performance for individual objects in the rooms because ground-truth models of the objects are unavailable, but the proposed method could retrieve more objects in the scenes compared with Kim16. The cuboid primitives represent the approximate structure of the rooms well though some small or thin objects are missing.

TABLE II
EVALUATION OF ROOM LAYOUT AND OBJECT RECONSTRUCTION

Data	Ground-truth	Kim16 [26]			Proposed		
	Dim (m^3)	Dim (m^3)	Err (Diag, %)	# Obj	Dim (m^3)	Err (Diag, %)	# Obj
MR	5.61×4.28×2.33	6.15×4.71×2.88	11.21	9	5.52×4.35×2.36	0.23	16
UL	5.57×5.20×2.91	6.12×4.96×2.90	2.91	9	5.92×4.95×2.95	1.28	11
LR	5.64×5.05×2.90	5.85×5.11×2.92	2.36	3	5.77×5.17×2.98	2.39	3
ST	17.08×14.55×6.50	5.96×17.70×6.85	14.84	3	16.53×14.87×5.70	1.74	4
KT	6.64×3.46×2.67	4.12×3.55×2.71	23.56	11	6.95×3.41×2.70	3.14	15
CY	19.00×10.10×–	18.25×9.72×–	(3.91)	12	18.51×9.61×–	(3.08)	13

C. Reverberant Spatial Audio Object (RSAO) Parameters

The RSAO representation of RIRs [15] is exploited to evaluate the room geometry reconstruction algorithm, in terms of acoustic property estimation. In fact, RSAOs represent the scene acoustics in a way that was proved to be coherent with the human perception of spatial sound [15]. The RIRs estimated through the pipeline proposed in Section III were thus parameterized. These parameters were also calculated, for comparison, from the acoustically measured RIRs. The RSAO representation of the room's acoustic response [15] is exploited to evaluate the room geometry reconstruction algorithm, in terms of estimated acoustic properties. The RSAO represents the scene acoustics in a way that has been shown to be coherent with human perception of spatial sound [15]. The RIRs generated through our pipeline in Section III were thus parameterized. For reference, RSAO parameters were also calculated from the RIRs recorded at the array of omnidirectional microphones, i.e., for datasets MR, UL, LR and ST.

Different RSAO parameters describe different parts of the RIR [16]: the RIR direct sound and the early reflections are described by parameters defined as their TOAs and Directions Of Arrival (DOAs); the reverberation parameters are defined as octave band late energy decays. To calculate the TOAs from the recorded RIRs, the clustered-dynamic programming projected phase-slope algorithm (C-DYPSA) [81] was employed. It is based on the DYPSA algorithm [86], which was used to locate peaks on each of the 48 microphone RIRs. A clustering technique was then employed to eliminate outliers, considering every k-th reflection, over the 48 DYPSA outputs. The mean of the inlier TOAs corresponds to the TOA parameter $\overline{\tau}_{k,l}$, with l being the loudspeaker index. Both azimuth and elevation DOAs were calculated by applying a delay-and-sum beamformer (DSB) to the RIRs [87]. To avoid the up-down ambiguity given by the planar microphone array, a Soundfield microphone was placed at the center of it, to record B-Format signals. From the B-Format Z-channel, the up-down information was thus recovered. To apply the DSB, the RIRs were first segmented, by applying a Hamming window (heuristically obtained length of 2.5 ms for UL, LR and ST, 0.8 ms for MR), for each RIR, centered at $\overline{\tau}_{k,l}$. The simulated RIRs were generated by placing a single virtual microphone at the room position where the center of the microphone array was placed for the recordings. In this case, TOAs and DOAs were calculated directly from the image source positions.

Finally regarding the late reverberation part of the RIR, the RT60 was calculated for each octave band between 250 Hz and 8 kHz, by analyzing the first 20 dB of decaying late energy after the mixing time [15]. This approach was used for both recorded and simulated RIRs.

D. Acoustic Results and Discussion

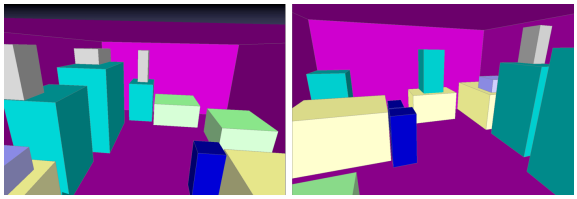
For comparison, the RSAO parameters were calculated for both simulated and recorded RIRs. This comparison was made by employing two groups of objective evaluation metrics. The first one (i.e. composed of TOA and DOA) investigates the early reflection estimation accuracy. The second one evaluates the estimated late reverberation, by calculating RT60 errors.

1) Evaluation Metrics: Coherent with the evaluation we did for our preliminary work [27], the TOA parameter errors $\varepsilon_{k,l}^{\mathrm{TOA}}$ are calculated as the absolute value of the difference between the TOA obtained from the simulated and the recorded RIRs, considering the direct sound ($k=0$) and each k-th early reflection, separately. The evaluated error E_k^{TOA} is then calculated as the median over the L available loudspeakers and averaged over all the reflections to obtain $\overline{E}^{\mathrm{TOA}}$. Similarly, the DOA errors $\varepsilon_{k,l}^{\mathrm{DOA}}$ are obtained as the absolute value of the difference between the DOA calculated from the simulated and recorded RIRs. This is done for both azimuth and elevation, separately. As for E_k^{TOA}, also the evaluated error E_k^{DOA} is obtained as the median over the L loudspeakers, and the provided results $\overline{E}^{\mathrm{DOA}}$ is the mean over all the early reflections. Finally, the RT60s for both simulated and recorded RIRs are estimated, and averaged over the octave bands between 250 Hz and 8 kHz. The median of the averaged RT60s is then calculated over the L loudspeakers, to obtain $\overline{\mathrm{RT60}}$.

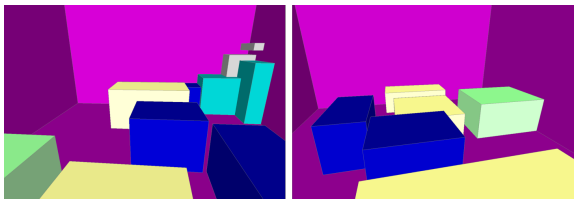
To understand how perceptually similar the estimated RIR is from the related recorded one, we define the just noticeable differences (JNDs) for the evaluation metrics. For $\overline{E}^{\mathrm{TOA}}$, this is considered to be 2.2 ms, corresponding to 75 cm [84]. For the azimuth $\overline{E}^{\mathrm{DOA}}$ the JND is set to 15°, i.e. an average angle calculated for audio-visual spatial coherence [85]. Regarding the elevation $\overline{E}^{\mathrm{DOA}}$, the JND is set to 35° [84]. Finally, the JND for the $\overline{\mathrm{RT60}}$ was chosen to be the 20 % [88].

2) Early Reflection Results: The evaluation of the estimated early reflections is performed by employing the TOA and DOA errors $\overline{E}^{\mathrm{TOA}}$ and $\overline{E}^{\mathrm{DOA}}$, both in terms of azimuth and elevation, as described in Section V-D1. In Fig. 11, the results are reported, by comparing the method that we propose in this article with our previous approach that we presented in [27]. To our knowledge, there is no other method yet that generates RIRs from visual captures.

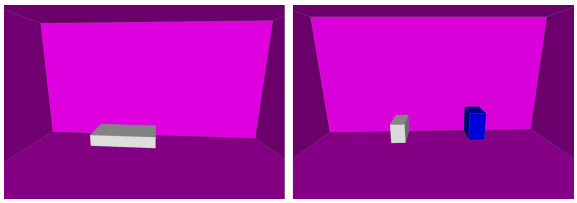
In general, both the proposed method and our previous approach generate errors that are below the related JNDs.



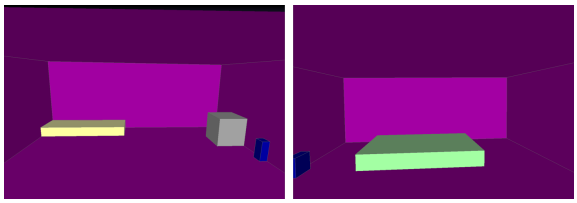
(a) Meeting Room (MR)



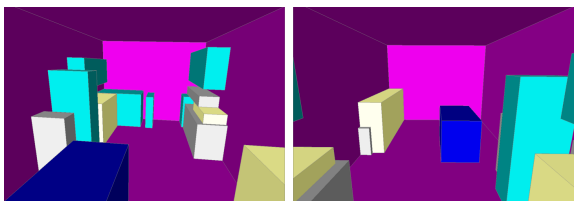
(b) Usability Lab (UL)



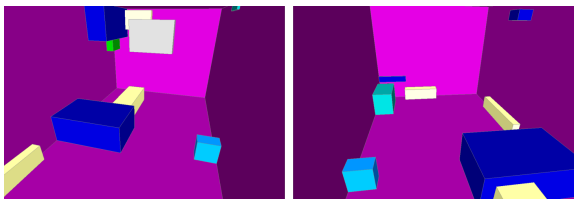
(c) Listening Room (LR)



(d) Studio hall (ST)



(e) Kitchen (KT)



(f) Courtyard (CY)

Fig. 10. Semantic room geometry estimation results (Left: Looking front, Right: Looking back, the color indicates object class, as per Fig. 8 (a).)

This means that humans cannot perceive, in terms of early reflection TOAs and DOAs, any difference with respect to a RIR recorded at the same rendering position. Regarding UL, although the novel approach produces a slightly higher error in terms of rendered reflection TOAs, though still within the JND limit, it greatly improves the elevation DOA, with the error median that is equal to zero. Finally, for the ST dataset, as reported in Table II, our previous approach in [27] failed

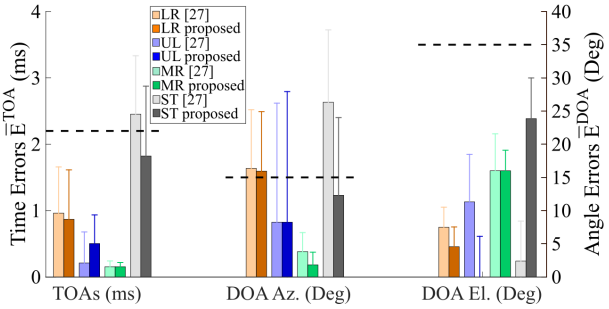


Fig. 11. Median and standard errors for the estimated RIR reflections' TOAs, azimuth (Az.) DOAs and elevation (El.) DOAs. The LR's dataset results are reported in orange, in blue the UL's, in green the MR's, and in grey the ST's. The lighter color bars refer to our old method [27], whereas the darker bars to the method proposed here. The dashed lines show the JNDs [84], [85].

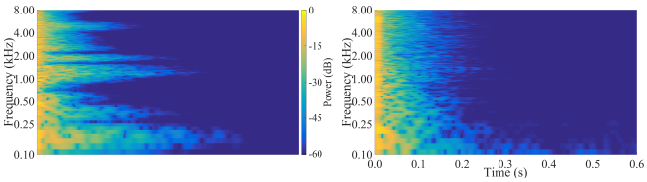


Fig. 12. Spectrograms of a estimated (top) and recorded (bottom) RIR.

to estimate the room boundary. The generated space is much narrower than the groundtruth, which results in wrong TOA and DOA Az. The new proposed method, instead, succeed to estimate the ST geometry. To understand the improvement given by the new method estimation, we also calculate the errors for the previous method estimated geometry. There, errors were 0.25 ms and 11° greater than the JNDs for TOAs and azimuth DOAs, respectively. Instead, the new proposed approach produces errors that are now lower than the respective JNDs.

The only error that is 1° over the JND limit, also for the new proposed method, is the azimuth DOA error related to the LR dataset. This is due to the fact that LR was an empty listening environment, with thirty loudspeakers clumped on a metal structure around the perimeter. The microphone array captured reflections produced by the sound encountering these loudspeakers during its propagation. However, a choice was made, and to improve the accuracy in localizing large planes, loudspeakers were not modeled with the proposed method. In fact, they were small and too close to the room boundaries. Therefore, they were not considered during the estimated RIR generation.

3) Late Reverberation Results: The late reverberation is analyzed by observing the reverberation time error $\overline{\text{RT60}}$, obtained by comparing a recorded RIR to the estimated one, as it was described in Section V-D1. Furthermore, as general indication of the quality of estimation that can be reached, the spectrogram of an estimated RIR, together with its related recorded version, reported in Fig. 12. There, it can be observed that, a part from some artifacts, the overall decay is similar at almost all the frequencies.

In Fig. 13, results related to all the datasets are reported. There, it is evident the error decrease (of about 71 % for LR, 81 % for UL, 44 % for ST, and 37 % for MR) given by the new

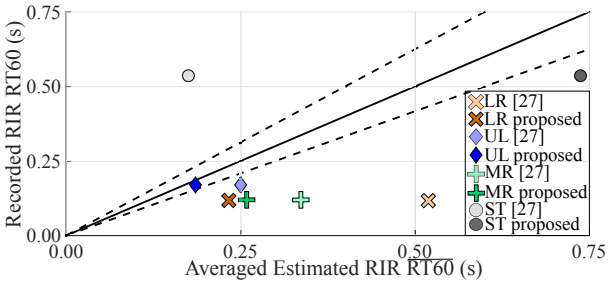


Fig. 13. Comparison between the RT60s obtained from recorded RIRs and the ones estimated from 360 images. The dashed lines show the JND [88]. Circles refer to the ST dataset, diamonds to UL, crosses to LR, and pluses to MR. The color legend is defined as in Fig. 11.

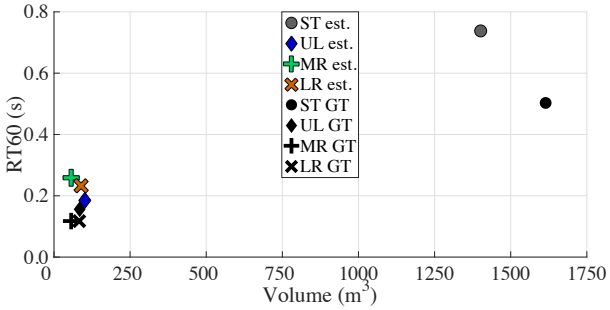


Fig. 14. Relationship between room volumes and estimated $\overline{\text{RT60}}$s. In black, the groundtruths are reported.

proposed method with respect to our previous one [27]. This improvement is given by the fact that the RT60 depends on the room volume [89]. The new method takes into account this, by modeling the furniture by predicting its overall volume. Instead, in [27], we estimated the furniture by analyzing the related cuboid spatial boundaries. In UL the estimated RT60 is now inside the JND band. This means that, for UL, humans would not notice the difference from a recorded RIR, in terms of RT60. Also considering LR, MR and ST, the estimated RT60s are now much closer to the JND band (i.e. about 100 ms away from the JND limits), when compared to our previous method's results [27]. The estimated RT60 errors are mainly given by the material recognition algorithm that, being based only on vision, cannot evaluate the acoustical properties of the modeled geometry with high accuracy (see Fig. 9). This is also the reason why, as described in Section III-E, the acoustic reflector materials were selected among those types that could more realistically be identified by visual classification algorithms. However, it is important to note that this is, to our knowledge, the first attempt in the literature to use a visual object classification algorithm to predict acoustic properties.

Nevertheless, an authentic, indistinguishable reproduction is often not required and, for many applications, such as virtual reality (VR), the creation of a plausible scene is sufficient [90] (i.e. authenticity is described by the JNDs). Lindau *et al*. described plausibility of a virtual environment as "a simulation in agreement with the listener's expectation towards a corresponding real event" [91]. Several studies about reproduction of virtual room acoustics are nowadays targeting plausibility rather than authenticity [15], [92], [93]. Since the late reverberation provides the listener with the impression of the room size [62], here, similar to [15], we evaluate the plausibility by observing the coherence between estimated RIR RT60s and related room sizes. Fig. 14 shows the estimated RT60s plausibility, since RIRs in the three rooms (i.e. UL, MR and LR) having similar (small) sizes are generated with similar (short) RT60s. On the other hand, RIRs related to ST, that is a large room, are estimated by having longer RT60s.

4) Ablation Experiment Result: Finally, we provide a better insight of the effect that the geometry estimation and material recognition accuracy have on the room acoustic estimation through ablation studies on the MR and UL datasets. We compare the results of the proposed pipeline (we will refer to it as "Full") to the results with errors manually introduced in individual components in the pipeline. To analyse the importance of objects within the scene for sound rendering, an empty room model which represents only room boundaries without any object in the scene is tested (referred as "Empty"). To analyse the impact of the scene scale accuracy, we include results with the models which have been 1.2 times scaled up ("E-Scale"). To analyse the influence of the detected object shape and locations, the model with major objects shrunk by 30% in scale and location in the given coordinate system while keeping the original room boundary ("E-Obj") is also tested. Finally, the estimated room model has been manually modified by assigning wrong materials (Plastic for MR and Glass for UL) to the three largest objects in the scene ("E-Mat") to analyse the error produced by wrong material estimation.

The results are reported in Table III. For MR, The result clearly shows the importance of objects in the scene as the "Empty" model generates large errors, both in terms of early reflections (TOAs and DOAs) and reverberation. The scaling error does not affect the early reflections for MR, since the early reflections are produced by furniture that has been kept at the similar distances from the listener, but it has a large impact on the reverberation. The main effect on the reverberation, however, is given by the use of wrong materials.

For UL, the "Empty" model only affects the late reverberation because the early reflections come from the floor, wall and ceiling in the tested source-microphone location. This is confirmed by the results related to the wrong-scale model ("E-Scale") because there is an increase of the early reflection errors as the room boundaries are moved away from the listener position. Errors in object size and location ("E-Obj") also influence the early reflections as objects block some early reflection on the room boundary. Finally, as expected, wrong materials generate the largest error in terms of reverberation.

The results of this ablation study provide a quantitative context in terms of the present evaluation metrics that is coherent with the findings established in the room acoustics literature [79], [94]. The geometrical factors are important for the estimation of early reflection, but sound rendering is relatively insensitive to small errors in the shape and size of objects as found in previous research [22], [24]. Errors in material attribute have a significant impact on the reverberation RT60 error [95].

TABLE III
EVALUATION OF THE EFFECT OF GEOMETRY ESTIMATION AND MATERIAL RECOGNITION ACCURACY ON THE ROOM ACOUSTIC PROPERTIES.

	Meeting Room					Usability Lab				
	Empty	E-Scale	E-Obj	E-Mat	Full	Empty	E-Scale	E-Obj	E-Mat	Full
TOA Error (ms)	1.7	0.1	0.2	0.1	0.1	0.1	0.2	1.2	0.1	0.1
DOA Az. Error (Deg)	142	7	7	7	7	8	157	49	8	8
DOA El. Error (Deg)	10	10	10	10	10	4	21	23	4	4
RT60 Error (ms)	146	125	117	139	89	92	90	82	97	56

VI. CONCLUSIONS

In this work, we have proposed a practical audio-visual approach to room acoustic estimation for audio rendering in novel scenes. A practical single-shot 360 stereo imaging is proposed to estimate the approximate large scale room and object geometry as a cuboid approximation together with estimates of surface material properties. This avoids the requirement for measurement of room acoustic RIRs at multiple locations with the room. The estimated model of room geometry and material properties is demonstrated to allow plausible audio augmentation of the scene by adaptation and rendering of audio sources.

Room geometry is estimated through vertical stereo systems using commercial off-the-shelf speherical 360 cameras. The captured images are aligned to the principal axes of the room coordinate, and semantic objects in the scene are detected by a convolutional encoder-decoder network. The final semantic cuboid-based room structure with object labels is reconstructed from the point cloud with object attribute. The acoustic RIR is simulated using a geometrical acoustic approach. Experiments were conducted by comparing the RSAO parameters of the simulated RIRs with the ones extracted from recordings. Results show that the proposed method outperforms our previous one, presented in [27], by confining the simulated early reflections' errors within their JNDs, and reducing the late reverberation error of 46 % (averaged over the four tested datasets). It was also found that the RT60s are now estimated coherently with the room size. Through the ablation study, it was observed that the geometry estimation is important for early reflection while material attribute has more impact on reverberation. This allows perceptually plausible acoustic reproductions, with the listener being able to correctly associate a listened sound with the respective room size.

Future extension of this research will include robust material identification in the room geometry modelling to replace the current surface-to-material mapping. This work provides a step change to acoustic room model reconstruction using audio-visual data. This, in the future, could be applied to several application areas, such as music broadcast, games, VR, and augmented reality. Furthermore, the proposed method may be used for research applications, such as source separation and localization.

ACKNOWLEDGEMENTS

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1), the BBC as part of the BBC Audio Research Partnership, and Audio-Visual Media Research Platform (EP/P022529/1). Details about the data underlying this work are available from: http://dx.doi.org/10.15126/surreydata.00812228.

REFERENCES

[1] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. J. Hughes, D. Menzies, M. F. S. Gálvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An audio-visual system for object-based audio: From recording to listening," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, Aug 2018.

[2] S. Wang and S. Dey, "Adaptive mobile cloud computing to enable rich mobile multimedia applications," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 870–883, June 2013.

[3] I. Cheng, A. Basu, and R. Goebel, "Interactive multimedia for adaptive online education," *IEEE MultiMedia*, vol. 16, no. 1, pp. 16–25, Jan 2009.

[4] Q. Liu, W. Wang, T. de Campos, P. J. B. Jackson, and A. Hilton, "Multiple speaker tracking in spatial audio via phd filtering and depth-audio fusion," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1767–1780, July 2018.

[5] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2304–2308.

[6] N. Takahashi, M. Gygli, and L. V. Gool, "Aenet: Learning deep audio features for video analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 513–524, March 2018.

[7] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[8] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*, Feb 2000.

[9] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Reconstructing building interiors from images," in *Proc. ICCV*, 2009.

[10] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proc. CVPR*, 2015.

[11] A. Wang, J. Lu, J. Cai, G. Wang, and T.-J. Cham, "Unsupervised joint feature learning and encoding for rgb-d scene labeling," *IEEE Trans. Image Processing*, vol. 24, pp. 4459–4473, 2015.

[12] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *Proc. CVPR*, 2013, pp. 1352–1359.

[13] D. Menzies and F. M. Fazi, "A perceptual approach to object-based room correction," in *Proc. of AES Convention*, 2016.

[14] M. A. Poletti, T. Betlehem, and T. D. Abhayapala, "Higher-order loudspeakers and active compensation for improved 2d sound field reproduction in rooms," *J. Audio Eng. Soc.*, vol. 63, no. 1/2, pp. 31–45, 2015.

[15] P. Coleman, A. Franck, P. J. B. Jackson, R. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77, 2017.

[16] L. Remaggi, P. J. B. Jackson, and P. Coleman, "Estimation of room reflection parameters for a reverberant spatial audio object," in *Audio Engineering Society Convention 138*, 2015.

[17] Y. Ohta and H. Tamura, *Mixed reality: mergin real and virtual worlds*. Springer Publishing Company, Incorporated, 2014.

[18] Q. Hao, R. Cai, L. Zhang, Y. Pang, F. Wu, Y. Rui, and Z. Li, "Efficient 2d-to-3d correspondence filtering for scalable 3d object recognition," in *Proc. CVPR*, 2013.

[19] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. ICCV*, 2015.

[20] S. Gupta, P. ArbelÃ¡ez, R. Girshick, and J. Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *International Journal of Computer Vision*, pp. 1–17, 2014.

[21] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *Proc. of ECCV*, 2010.

[22] D. B. JUDD, "Chromaticity sensibility to stimulus differences," *Journal of the Optical Society of America*, vol. 22, no. 2, pp. 72–72, Feb 1932.

[23] W. Bailey and B. M. Fazenda, "The effect of visual cues and binaural rendering method on plausibility in virtual environments," in *Proc. of the 144th AES Convention*, Milan, Italy, 2018.

[24] S. Siltanen, T. Lokki, L. Savioja, and C. Lynge Christensen, "Geometry reduction in room acoustics modeling," *Acta Acustica united with Acustica*, vol. 94, no. 3, pp. 410–418, 2008.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[26] H. Kim, T. Campos, and A. Hilton, "Room layout estimation with object and material attributes information using a spherical camera," in *Proc. 3DV*, 2016.

[27] H. Kim, R. Hough, L. Remaggi, P. Jackson, , A. Hilton, T. Cox, and B. Shirley, "Acoustic room modelling using a spherical camera for reverberant spatial audio objects," in *Proc. 142nd AES*, 2017.

[28] S. N. Sinha, D. Steedly, R. Szeliski, M. Agrawala, and M. Pollefeys, "Interactive 3d architectural modeling from unordered photo collections," in *Proc. of SIGGRAPH ASIA*, 2008.

[29] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Proc. of ISMAR*, 2011.

[30] A. Handa, T. Whelan, J. McDonald, and A. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. of ICRA*, 2014.

[31] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using sfm and object labels," in *Proc. of ICCV*, 2013, pp. 1625–1632.

[32] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *Proc. of ICCV Workshop*, 2011.

[33] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. of ECCV*, 2012.

[34] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic Scene Completion from a Single Depth Image," in *Proc. CVPR*, 2017.

[35] S. Song, A. Zeng, A. X. Chang, M. Savva, S. Savarese, and T. Funkhouser, "Im2pano3d: Extrapolating 360 structure and semantics beyond the field of view," in *Proc. CVPR*, 2019.

[36] A. Dai, C. Diller, and M. Niessner, "Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans," in *Proc. CVPR*, June 2020.

[37] M. Schoenbein and A. Geiger, "Omnidirectional 3d reconstruction in augmented manhattan worlds," in *Proc. of IROS*, 2014, pp. 716 – 723.

[38] H. Kim and A. Hilton, "Block world reconstruction from spherical stereo image pairs," *Computer Vision and Image Understanding*, vol. 139, pp. 104–121, 2015.

[39] S. Li, "Real-time spherical stereo," in *Proc. ICPR*, 2006, pp. 1046–1049.

[40] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. Kweon, "All-around depth from small motion with a spherical panoramic camera," in *Proc. ECCV*, vol. 9907, 10 2016.

[41] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3D semantic parsing of large-scale indoor spaces," June 2016, pp. 1534–1543.

[42] A. X. Chang, A. Dai, T. A. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," *CoRR*, vol. abs/1709.06158, 2017. [Online]. Available: http://arxiv.org/abs/1709.06158

[43] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 12–27, 2016.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[45] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[47] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[48] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. of ICCV*, 2015, pp. 1520–1528.

[49] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of CVPR*, 2015, pp. 3431–3440.

[50] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Advances in Neural Information Processing Systems*, 2015, pp. 1495–1503.

[51] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *Proc. ICCV*, 2015.

[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015.

[53] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation." *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[54] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 708–730, 1979.

[55] A. Krokstad, S. Strom, and S. Sørsdal, "Calculating the acoustical room response by the use of a ray tracing technique," *J. of Sound and Vibration*, vol. 8, no. 1, pp. 118–125, 1968.

[56] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.

[57] K. Kowalczyk and M. van Walstijn, "Room acoustics simulation using 3D compact explicit FDTD schemes," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 34–46, 2011.

[58] S. Bilbao, "Modeling of complex geometries and boundary conditions in finite difference/finite volume time domain room acoustics simulation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1524–1533, 2013.

[59] L. Savioja and V. Välimäki, "Interpolated rectangular 3D digital waveguide mesh algorithms with frequency warping," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 11, no. 6, pp. 783–790, 2003.

[60] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley, "Acoustic modeling using the digital waveguide mesh - recent activities in articulatory vocal tract modeling, room impulse response synthesis, and reverberation simulation," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 55–66, 2007.

[61] E. De Sena, H. Hacihabiboğlu, Z. Cvetković, and J. O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1478–1492, 2015.

[62] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[63] L. L. Beranek, "Analysis of sabine and eyring equations and their application to concert hall audience and chair absorption." *J. of the Acoustical Society of America*, vol. 120, no. 3, pp. 1399–1410, 2006.

[64] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski, "Low-cost 360 stereo photography and video capture," *ACM Trans. Graphics*, vol. 36, no. 4, pp. 148:1–148:12, 2017.

[65] H. Kim and A. Hilton, "3d scene reconstruction from multiple spherical stereo pairs," *International Journal of Computer Vision*, vol. 104, no. 1, pp. 94–116, 2013.

[66] S. Song, S. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proc. of CVPR*, 2015.

[67] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson, 2017.

[68] J. Liu, H. Li, R. Wu, Q. Zhao, Y. Guo, and L. Chen, "A survey on deep learning methods for scene flow estimation," *Pattern Recognition*, vol. 106, p. 107378, 2020.

[69] X. Song, X. Zhao, L. Fang, H. Hu, and Y. Yu, "Edgestereo: An effective multi-task learning network for stereo matching and edge detection," *International Journal of Computer Vision*, pp. 910–930, 2020.

[70] C. Won, J. Ryu, and J. Lim, "Omnimvs: End-to-end learning for omnidirectional stereo matching," in *Proc. ICCV*, October 2019.

[71] X. Huang, C. Yuan, and J. Zhang, "Graph cuts stereo matching based on patch-match and ground control points constraint," in *Advances in Multimedia Information Processing – PCM 2015*, Y.-S. Ho, J. Sang, Y. M. Ro, J. Kim, and F. Wu, Eds., 2015, pp. 14–23.

[72] H. Kim and K. Sohn, "3d reconstruction from stereo images for interactions between real and virtual objects," *Signal Processing: Image Communication*, vol. 20, no. 1, pp. 61–75, 2005.
[73] S.-W. Kwon, F. Bosche, C. Kim, C. Haas, and K. Liapi, "Fitting range data to primitives for rapid local 3d modeling using sparse range point clouds," *Automation in Construction*, vol. 13, no. 1, pp. 67–81, 2004.
[74] W. Nguatem, M. Drauschke, and H. Mayer, "Finding cuboid-based building models in point clouds," in *Proc. of ISPRS*, 2012, pp. 149–154.
[75] M. Li, L. Nan, and S. Liu, "Fitting boxes to manhattan scenes using linear integer programming," *International Journal of Digital Earth*, vol. 9, pp. 806–817, 2016.
[76] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.
[77] M. Vorländer, "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *The Journal of the Acoustical Society of America*, vol. 86, no. 1, pp. 172–178, 1989.
[78] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," *J. Audio Engineering Society*, vol. 60, no. 11, pp. 887–898, 2012.
[79] T. Cox and P. D'Antonio, *Acoustic absorbers and diffusers, third edition: theory, design and application*. CRC Press, 2016.
[80] P. Coleman, A. Franck, P. J. B. Jackson, and D. Menzies, "Object-based reverberation encoding from first-order ambisonic RIRs," in *Proc. of the 142nd AES Convention*, Berlin, Germany, 2017.
[81] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: novel image source reversion and direct localization methods," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 2, pp. 296–309, 2017.
[82] Rec. ITU-R BS.775-3, "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunication Union, Geneva, Switzerland, Recommendation, 2012.
[83] C. Pike and M. Romanov, "An impulse response dataset for dynamic data-based auralisation of advanced sound systems," in *Audio Engineering Society Convention 142*, Berlin, Germany, 2017.
[84] J. C. Middlebrookes and D. M. Green, "Source localization by human listeners," *Annual Review of Psycology*, vol. 42, no. 1, pp. 135–159, 1991.
[85] H. Stenzel, P. J. B. Jackson, and J. Francombe, "Modeling horizontal audio-visual coherence with the psychometric function," in *Proc. of AES Convention*, 2017.
[86] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.
[87] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.
[88] Z. Meng, F. Zhao, and M. He, "The just noticeable difference of noise length and reverberation perception," in *Proc. of the ISCIT*, 2006.
[89] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," *Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 982–993, 2008.
[90] C. Pörschmann, P. Stade, and J. Arend, "Binauralization of omnidirectional room impulse responses - algorithm and technical evaluation," in *Proc. of DAFx, YEAR = 2017,*.
[91] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, vol. 98, no. 5, pp. 804–810, 2012.
[92] C. Pike, F. Melchior, and T. Tew, "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room," in *Proc. of AES Conference*, 2014.
[93] A. Neidhardt, A. I. Tommy, and A. D. Pereppadan, "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified brir sets," in *Proc.of AES Convention*, 2018.
[94] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 708–730, 2015.
[95] J. Cucharero, H. Tuomas, and T. Lokki, "Influence of sound-absorbing material placement on room acoustical parameters," *Acoustics*, vol. 1, no. 3, pp. 644–660, 2019.

Hansung Kim received his Ph.D. degree from Yonsei University, South Korea, in 2005. He was a researcher with Advanced Telecommunications Research Institute International, Japan, from 2005 to 2008. He was employed as a senior research fellow at the Centre for Vision Speech and Signal Processing, University of Surrey from 2005 to 2020. He is currently an Assistant Professor at the University of Southampton, UK. His research interests include 3D computer vision, audio-visual data processing, and virtual reality.

Luca Remaggi is Audio Research Engineer at Creative Labs, UK, working on spatial audio technologies for binaural products. Between 2017 and 2019, he was Research Fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, UK, where he also pursued his PhD, in 2017. His research interest was to investigate the multipath sound propagation combining acoustic and visual data, for applications in spatial audio and source separation.

Sam Fowler received his Ph.D. degree from the University of Surrey in 2020. His research at the Centre for Vision, Speech and Signal Processing investigated the analysis of human activity towards understanding the structure and affordance of an indoor scene. His research interests include 3D reconstruction, image segmentation, and human activity analysis.

Philip Jackson received his B.A. in Engineering from Cambridge University and Ph.D. from the University of Southampton, UK. He joined the Centre for Vision, Speech and Signal Processing (CVSSP, University of Surrey, UK) in 2002, where he is Reader in Machine Audition, enabling him to pursue broad interests from speech articulation and audio-visual perception to sound zones and object-based spatial audio [Scholar: bit.ly/2oTRw1C].

Adrian Hilton received the B.S. (Hons.) and D.Phil. degrees from the University of Sussex in 1988 and 1992, respectively. He is currently a Professor of computer vision and graphics and the Director of CVSSP, University of Surrey. He leads the Visual Media Research (VLab), which is conducting research in video analysis, computer vision and graphics for next-generation communication and entertainment applications. His research interests include robust computer vision to model and understand real world scenes.