3D Room Geometry Reconstruction Using Audio-Visual Sensors

Hansung Kim[1], Luca Remaggi[1], Philip JB Jackson[1], Filippo Fazi[2] and Adrian Hilton[1]
[1]CVSSP, University of Surrey, Guildford, GU2 7XH, UK
[2]University of Southampton, Southampton, SO17 1BJ, UK
h.kim@surrey.ac.uk, l.remaggi@surrey.ac.uk,p.jackson@surrey.ac.uk,
Filippo.Fazi@soton.ac.uk, a.hilton@surrey.ac.uk

Abstract

In this paper we propose a cuboid-based air-tight indoor room geometry estimation method using combination of audio-visual sensors. Existing vision-based 3D reconstruction methods are not applicable for scenes with transparent or reflective objects such as windows and mirrors. In this work we fuse multi-modal sensory information to overcome the limitations of purely visual reconstruction for reconstruction of complex scenes including transparent and mirror surfaces. A full scene is captured by 360° cameras and acoustic room impulse responses (RIRs) recorded by a loudspeaker and compact microphone array. Depth information of the scene is recovered by stereo matching from the captured images and estimation of major acoustic reflector locations from the sound. The coordinate systems for audio-visual sensors are aligned into a unified reference frame and plane elements are reconstructed from audio-visual data. Finally cuboid proxies are fitted to the planes to generate a complete room model. Experimental results show that the proposed system generates complete representations of the room structures regardless of transparent windows, featureless walls and shiny surfaces.

1. Introduction

3D indoor scene reconstruction has been an important research topic for practical applications as various sensors became available in our daily lives. Computer vision techniques using visual sensors such as cameras and RGBD sensors have played an important role in geometry reconstruction and various approaches have been proposed in the 3D vision community. Recovering geometric information from a single photograph relies on learnt cues such as silhouettes, shading and texture [26]. Recent progress in deep learning has accelerated this field [38, 33], but this approach still works in very limited environments and relies on large corpuses of training data for similar scenes. Stereo or multi-view reconstruction from multiple images is a widely used approach for general scene reconstruction [29, 31]. However, visual reconstruction does not work for featureless regions such as white wall or reflective surface where a unique matching pair cannot be defined. In real world indoor environments, transparent and uniform surfaces are common resulting in poor performance of visual reconstruction. Recently, Kinect-like RGBD sensors provides good depth information for an indoor scene with featureless regions [5, 6]. However, active depth sensors are limited and also fail for transparent and mirror surfaces which are easily observed in common indoor scenes.

In contrast to vision, detection and localisation of geometrical surfaces (reflectors) is a new area of research in audio processing. The methods that can be found in the literature were mainly proposed during the last decade. It is possible to spot an evolution in the state-of-the-art, starting from older methods which approached the problem from a 2D point of view [8, 2], to arrive to more advanced methods, where walls, ceiling and floor positions were estimated as planes in the 3D space [35, 9, 28]. Usually, the acoustic signal that is employed to perform the reflector localisation is the so called room impulse response (RIR). A RIR characterises the acoustic of an environment with respect to source and receiver positions [21]. It can be described as superimposition of Dirac deltas, in the time domain, representing the direct sound and reflections that arrive at the microphone. In [8, 2, 35, 9], the feature extracted from multiple RIRs, to localise the reflectors, was the times of arrival (TOAs) of the reflections. However, in [28], it was also demonstrated that a combination of TOAs and directions of arrival (DOAs) can improve the results. Previous work also proposed the combination of TOA with additional features, for instance, the time differences of arrival (TDOAs) among microphones [34].

This audio-based approach can provide a solution for geometry reconstruction of transparent or highly specular mirror surfaces. Audio and vision-based approaches are complementary to each other for scene reconstruction. Vision-based approaches are strong at recovering dense geometry

for cluttered scenes with visual and geometrical features but weak at recovering featureless, transparent or mirror surface regions. In contrast, audio-based approaches can localise acoustic reflectors such as glass regardless of their visual characteristics though they are weak at reconstructing complicated scenes or detailed geometry.

A few researchers have tried to combine audio and visual sensors to reconstruct 3D scenes, especially in the field of underwater scene reconstruction. Murino and Fusiello used sonar and camera to model underwater scenes [24]. Lagudi et al. proposed alignment method for the integration of stereo cameras and an acoustic camera for underwater 3D data capture [22]. Recently Ye at al. attached a ultrasonic sensor to Kinect to reconstruct glasses in normal indoor scenes [39]. Hussain et al. proposed a room layout estimation method using a single photo and acoustic echoes [16]. However, these approaches have the following limitations. First, sonar or ultrasonic sensors are not common in our daily lives. Special equipment is required to use these approaches. Second, normal cameras and ultrasound sensors have limited field-of-views (FOV) capturing only a part of the whole scene. For a complete scene estimation, multiple inputs and fusion technique are required.

In this paper, we propose a cuboid-based complete (airtight) room geometry estimation method using an off-the-shelf 360° camera, normal speaker and microphones. The approach assumes that room interiors are composed of piece-wise planar surfaces aligned to the main axes (Manhattan world). Section 2 introduces related previous works and Section 3 presents the proposed method. Experimental results and discussion are given in Section 4, and Section 5 makes conclusions of this work.

2. Related Work

2.1. Visual geometry scene estimation

Simplified scene modelling has been a long-standing area of research since the FACADE system introduced an approach for modelling architecture from a number of photographs [7]. Sinha et al. used feature matching and Structure-from-Motion methods with line and vanishing point detection for interactive 3D architectural modelling from photo collections [32]. To achieve fully automatic reconstruction, various methods have been proposed over the last decade such as piece-wise planar depth map fusion [15], axis aligned depth map integration [13], cuboid fitting [25], inverse constructive solid geometry [37], etc.

As inexpensive off-the-shelf 360° cameras become popular in our daily lives[1][2], various 3D reconstruction methods for 360° images have been proposed [30, 4]. Kim and Hilton used an industrial spherical camera for simplified scene modelling [18] and extension with object recognition [17]. Spherical stereo images are captured and converted into cubic projection image with façade alignment, then cuboid-based scene structure is reconstructed using plane detection. However, it does not work well for walls with uniform appearance, windows and mirrors which are common in indoor scenes because it was originally designed for a large scale outdoor scene where plenty of image features exist. The vision-based room geometry reconstruction part in this work has been motivated from these works but modified for fast indoor scenes reconstruction with a pair of low-cost consumer 360° cameras.

2.2. Major reflector localisation using audio sensors

In audio, the concept of acoustic scene analysis usually refers to major reflector localisation, such as walls, ceiling and floor. This is due to the challenging issue of identifying, within an acoustic signal, reflections related to the sound bouncing off small objects, thus having a low signal-to-noise ratio (SNR). Reflector localisation methods are categorised into two groups [28]: "image-source reversion", that exploits TOA information to revert the image source method [1] and determine the reflector position; "direct localization", that directly localises the reflector, without estimating any other room acoustic element first.

The method in [35] uses the image-source reversion approach to localise reflectors in 3D, by maximising the probability of a point in the space to be the image source position. However, a large number of putative points needed to be investigated. The main contribution of [9] was an algorithm to label the reflections from a distributed microphone array, where the reflector order would otherwise be ambiguous if compared among different microphone recordings. Nevertheless, the algorithm used for the image source localisation failed when applied to microphone arrays that are compact in space. Remaggi et al. proposed three image source reversion methods in [28]: the image source direction and ranging (ISDAR), and two variants of it. They exploited a combination of both TOA and DOA to determine the image source position. However, only the first reflection in time was analysed. In this paper, an evolution of ISDAR is proposed, that is able to localise multiple reflections.

One of the first attempts to employ, instead, the direct localisation approach was proposed in [20]. The authors mapped reflections from a linear microphone array to the related reflecting objects. This method provided an accurate analysis of the scene, nonetheless, the microphone array was assumed to be exactly parallel to the reflector. In [28], Remaggi et al. also proposed a direct localisation method, that exploited quadratic surfaces constructed by considering the reflection TOAs. It was demonstrated to provide high performance in localising reflectors. However, the ap-

[1] Samsung Gear 360, http://www.samsung.com/global/galaxy/gear-360/

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

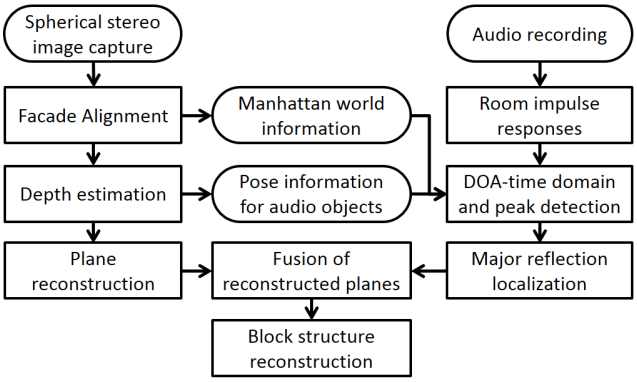


Figure 1. Block diagram of the proposed system

proach is limited by the computational complexity

3. Proposed Method

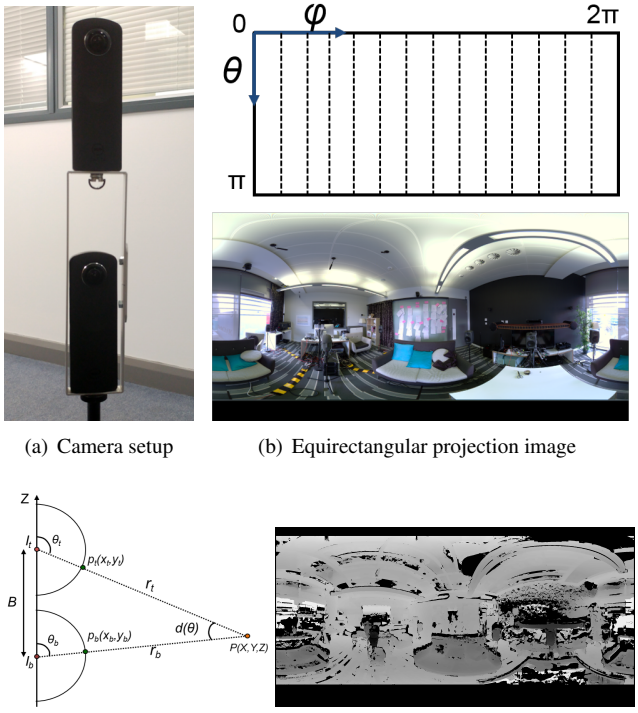
3.1. System overview

Figure 1 shows a pipeline for building a complete room model with cuboid estimated from 360° cameras and audio sensors. A full surrounding scene is captured as a pair of vertical stereo images by two 360° cameras. The captured spherical images are mapped to equirectangular images and aligned to the room coordinate axes (Manhattan world). Depth information of the scene is retrieved by stereo matching and axis-aligned planar regions are detected. In parallel, acoustic RIRs are recorded by employing a compact microphone array. A super-directive array beamformer (SDA) analyses the acoustic energy arriving at the microphone array in time, from every direction, in both azimuth and elevation. Images representing the energy in the angle-time domain are then generated, and a 2D peak-picking algorithm is employed to detect the reflections. 2D reflector planes are built on the detected reflections and refined with 2D planes from the vision sensor. Finally cuboid proxies are fitted to the planes to generate a complete room model.

3.2. Visual capture and depth estimation

The scene is captured with two Ricoh Theta S cameras on a bracket as shown in Fig. 2 (a) to recover 3D information of the whole surrounding scene at a time instance. The Theta S camera automatically stitches photos acquired from two pre-calibrated fish-eye lenses to generate an equirectangular projection image as illustrated in Fig. 2 (b). We use a vertical stereo setup rather than horizontal stereo because: (1) Stereo matching for depth estimation can be simplified to a 1D search in the equirectangular images; (2) the paired camera occludes less important areas (ceiling or floor); (3) epipoles where accurate stereo matching is impossible appear on the ceiling and floor.

To align the coordinates of camera and audio systems



(a) Camera setup (b) Equirectangular projection image

(c) Spherical stereo geometry (d) Disparity map

Figure 2. Spherical stereo camera system

to the room reference frame (Manhattan world coordinate system), the Façade alignment algorithm [17] using Hough line detection is applied to the captured images. Depth information is recovered with spherical stereo geometry as illustrated in Fig. 2 (c) and Eq. (1). Stereo matching can be carried out for the aligned equirectangular image pairs to find corresponding pairs of image points (p_t) and (p_b).

$$r_t = B / \left(\frac{\sin \theta_t}{\tan(\theta_t + d)} - \cos \theta_t \right) \tag{1}$$

Any stereo matching algorithm can be utilised. We use a block matching method incorporating a region-diving technique which produces reliable disparity fields by detecting occlusion regions and ambiguous regions based on bi-directional matching and the ordering constraint [19]. Figure 2 (d) 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. Serious depth errors are observed at windows, mirror, white table and rear wall behind TV in the scene. Speakers and microphone array are also observed. The speaker and microphone locations are calculated from the depth map and delivered to the audio processing pipeline with the room coordinate system.

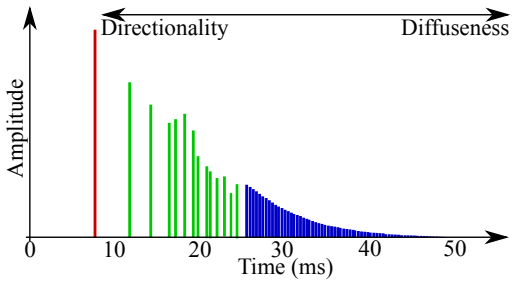


Figure 3. Schematic representation of a RIR, highlighting its three component: the direct sound (red), the early reflections (green) and the late reverberation (blue). Figure modified from [36].

Figure 4. Microphone array (left) and loudspeaker (right).

3.3. Audio capture system and pre-processing

A RIR is an acoustic signal, carrying information about the environment in which it is recorded. As shown in Fig. 3, it is composed of three elements [21]: the direct sound, revealing the position of the sound source; the early reflections, conveying a sense of the environmental geometry; and the late diffuse reverberation, indicating the size of the environment [36]. Localisation of major acoustic reflectors, such as walls, ceiling and floor, can be achieved by analysing the early reflection part [28].

A compact microphone array (shown in Fig. 4 left), that is composed of 48 microphones lying on two concentric circles having radii 85 mm and 106 mm, respectively, is employed to record RIRs, together with a loudspeaker (Fig. 4 right). This kind of microphone configuration is chosen to have a high resolution in estimating the azimuth DOA of the sound. The swept-sine method [11] for a large frequency range is used for sound recording since it is known to be robust against background noise.

To analyse the room acoustics, the DOA of the acoustic energy over time can be visualised by applying beamforming algorithms to the recorded multi-channel RIRs. A visualisation similar to [27] is achieved by steering a SDA beamformer. In this paper, the SDA proposed in [3], to observe the azimuth angle only, is improved by allowing the observation of both azimuth and elevation direction, with a resolution of one degree. Before being processed by the SDA, the RIRs are high-pass filtered at 1 kHz, to avoid the poor directivity factor of the SDA for the low frequencies.

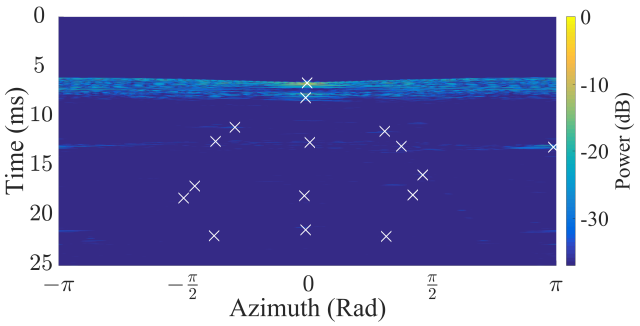


Figure 5. Example of energy analysis in the azimuth DOA-time domain (background figure). The white crosses are the reflection DOAs and TOAs detected by the angle-constrained peak-picking algorithm.

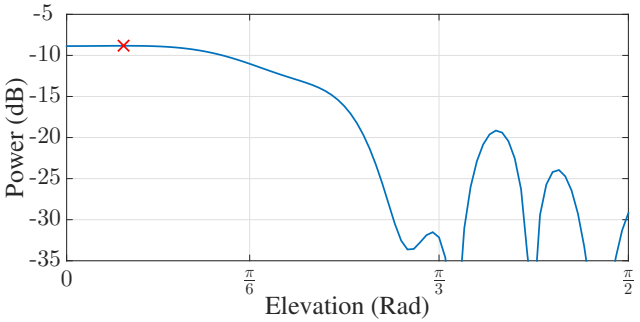


Figure 6. Example of the beamformed signal power for different elevations, regarding the reflection at 180 degrees azimuth DOA of Fig. 5. The red cross is the detected reflection elevation DOA.

The beamformed RIRs, for every azimuth DOA and 0° of elevation, are visualised by positioning them adjacently one to each others. The obtained image is depicted in Fig. 5.

3.4. Reflector localisation from audio recording

Considering the azimuth DOA-time domain image shown in Fig. 5, any peak-picking algorithm can be employed to detect the positions of the energy peaks. Here, we employed a simple method based on adaptive thresholding [14]. Detecting the energy peaks means to find the azimuth DOAs and TOAs of the reflections. To avoid a computationally expensive search, the peak-picking algorithm is constrained within ranges of the image: the first 25 ms of the beamformed RIRs; $\pm 5°$ around the loudspeaker and opposite directions; a dataset-dependent angular range having size of $45°$, both on the left and on the right of the frontal direction. The time constraint can be made since, in a typical living room-like environment, 25 ms can be considered as the limit between RIR early reflections and reverberation [23]. The first angular constraint is made by knowing the loudspeaker and microphone array positions, estimated from the vision. The second angular constraint, instead, depends on the position of loudspeaker and microphone array with respect to the room geometry. This information is

retrieved from preliminary results given by vision, and by assuming a shoebox-like environment.

For every estimated reflection azimuth DOA, the RIRs are then beamformed by steering the SDA towards every elevation DOA between 0° and 90°. Observing the energy at the estimated TOAs, the elevation DOAs are estimated by calculating the energy maximum position, as illustrated in Fig. 6. Knowing the sound speed in air, the TOAs are then converted into distances, and, together with azimuth and elevation DOAs, they define image sources in the 3D space, similarly to what was done in the ISDAR algorithm [28]. An image source is the mirrored position of the sound source with respect to the reflector [1], hence, the reflection point is finally localised as the midpoint between the image source and the centre of the microphone array (instead, in the loudspeaker-image bisection algorithm (LIB) [28], the midpoint between the image source and the main source was considered to localise the planar reflector).

The power carried by each reflection is also extracted, as the peaks of amplitude in the beamformed signals, and then normalised. By setting a threshold at 0.5, every reflection having normalised amplitude lower than the threshold is discarded. In this way, weak reflections that do not arrive directly from the major reflectors are discarded.

3.5. 3D geometry reconstruction

For 3D geometry reconstruction, the block world reconstruction method in [18] is modified to accommodate audio and visual data. Piecewise planar elements are reconstructed from the estimated depth information from visual capture and estimated reflector locations from audio capture. The input spherical image is segmented into superpixels by the graph-based segmentation method [12], and optimised planes with fitted bounding boxes are reconstructed by the total least squares (orthogonal regression) fitting algorithm [10]. Unreliable or not-aligned ($\sigma^2_{n_i} > 0.2Rad$) planes are eliminated, then close planes are merged into one plane to simplify the scene. One problem on detecting reflection points using a microphone is that it is not possible to have an accurate estimation of the reflected surface size. Therefore, the reflection point from the audio sensor is projected to the segmented superpixel image from the vision sensor, corresponding segments are then checked that they can be reconstructed as an aligned plane with more generous threshold ($\sigma^2_{n_i} < 0.3Rad$). If it passes the alignment test, a new plane is assigned to the the superpixel region. If it fails, an arbitrary plane with a size of 50 cm × 50 cm is generated at the reflection point. The newly generated planes by the reflection points are merged and refined with planes from the vision sensor with the same refinement algorithm. Final 3D geometry of the room is reconstructed by fitting cuboids into the plane elements. In order to get an air-tight model of the room, the farthest planes in each direction are considered as the boundary of the room (walls, ceiling and floor) and their surface normals are set to the inside of the room. All other planes are used for cuboid structure generation by the outward extrusion process from the camera capture position and occupancy of point cloud [18], and the surface normals are set outward of the cuboid.

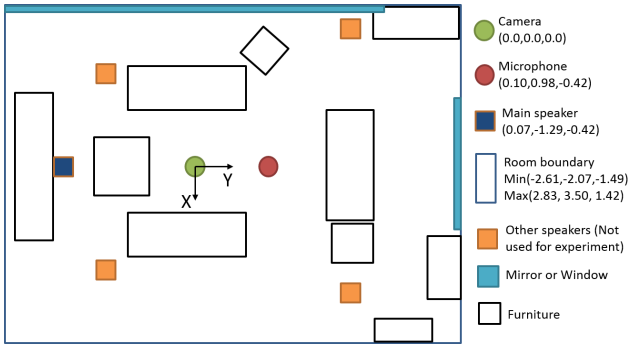


Figure 7. Ground-truth structure and setup of UR2 in Fig. 2 (b). Positions are manually measured by a laser measure.

Table 1. Ground-truth and setup for the datasets ((x,y,z) in m)

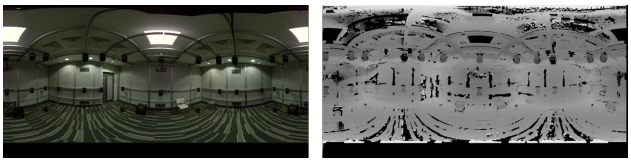
Data	Room size (m)	Mic. pos.	Loudspeaker pos.
UR2	(5.44,5.57,2.91)	(0.1,0.98,-0.42)	(0.07,-1.29,-0.42)
LR	(5.60,5.00,2.90)	(-0.05,-0.01,-0.62)	(-0.05,-1.95,-0.62)
MR2	(4.28,5.57,2.32)	(2.09,1.11,-0.46)	(2.21,-0.75,-0.46)
S1	(14.55,17.08,6.50)	(-2.65,0.25,-0.41)	(-2.65,2.25,-0.41)

4. Experiments

4.1. System set up and datasets

We used a pair of Ricoh Theta S cameras which have built-in calibrated fisheye cameras as shown in Fig. 2 (a). We captured the scene with an inter-camera baseline distance of 11 - 27 cm according to the room size and image resolution of 3000 × 1500. We tested the proposed pipeline on four different indoor rooms. The Usability room 2 (UR2) in Fig. 2 (b) is similar to a normal living room environment with furniture, TV and 5.1 channel speakers. One whole side of the room is glass and there is a big mirror on the front wall. Figure 7 shows manually measured ground-truth room structure and camera/microphone setup for UR2. The origin of world coordinates was set as the location of the top camera. There are six loudspeakers in the room but we used only one frontal loudspeaker for audio recording.

Figure 8 shows other three scenes with the top image of vertical stereo pairs and estimated disparity maps. The Listening room (LR) is a simple room with a few small objects and many loudspeakers, the Meeting room2 (MR2) is a more cluttered scene with various objects, and the Studio1 (S1) is a large hall. The images were captured just before the audio setup to provided clean room capture for visual reconstruction. Table 1 gives the ground-truth room dimensions and positions of the audio sensors.

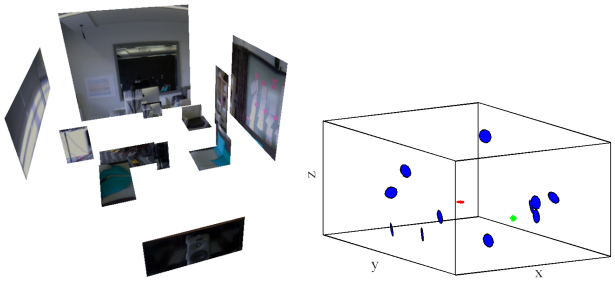


(a) Listening Room (LR)

(b) Meeting Room2 (MR2)

(c) Studio1 (S1)

Figure 8. Other datasets (left: Image, right: Estimated disparity)



(a) Reconstructed planes (b) Localised reflectors

Figure 9. Initial elements detected from audio-visual sensors

4.2. Room geometry reconstruction results

We assume that there is neither a reference nor ground-truth model for the reconstruction process. The positions of the microphone and loudspeaker should be calculated for reflector estimation from the audio recording and registered to the reference camera coordinate system. Estimated position of the microphone and loudspeaker of UR2 from the depth map in Fig. 2 (d) are (0.05, 1.05, -0.45) and (0, -1.12, -0.43), with errors of (0.05, 0.07, 0.03) and (0.07, 0.17, 0.01), respectively. The estimation of the microphone position is relatively accurate but the loudspeaker position has 17cm error in the Y direction which is caused by the error in depth estimation. The reflector localisation would be affected by this position error.

Figure 9 shows reconstructed planes and reflector positions for UR2 estimated from visual and audio sensors. 14 planes have been reconstructed from the 360° camera capture in Fig. 9 (a). As expected from the depth map in Fig. 2 (d), the ceiling saturated by lights, the side wall with glass and the featureless rear wall behind the TV could not be reconstructed from the vision sensor. The front wall with a big mirror could be reconstructed owing to the surrounding regions. 11 reflection points have been detected from the audio sensors as shown in Fig. 9 (b) (Green and red dots indicate the loudspeaker and microphone, respectively.) and Table 2. Comparing with the ground-truth model, locations of Points 1,2,7,8 and 9 correspond to the walls, and Point 11 to the ceiling. Points 3 and 4 are from the left sofa and Points 5 and 6 from the right sofa. One of two reflections in each sofa is the 2nd order reflection because the sofas have concave joints with the seat and back. Point 10 is from the shiny table between the loudspeaker and microphone. We can observe that the ceiling, walls and table which could not be detected by the vision processing were detected by the audio sensors. As we assume the ground-truth is unknown, the reflection points are converted to planes by projecting to the corresponding superpixels or arbitrary planes, then refined with planes from the vision sensor as proposed in Section 3.5. Figure 10 shows snapshots of the original plane elements from the vision sensor (left) and the final 3D cuboid model reconstructed from the planes with audio-visual inputs (right). We can see that all 6 boundaries have been reconstructed and the missing table was recovered from the combination of audio-visual sensors. Table 3 shows the errors of estimated object surfaces against the ground-truth (GT) measurements. Large objects such as walls and sofas were detected from both audio (A) and visual (V) sensors. Most objects are with $<$ 10 cm errors to the GT but a few objects have relatively large ($>$ 20 cm) errors. It is difficult to say which sensor is more accurate or superior in accuracy, but it is clear that audio and visual sensors are complementary in reconstruction.

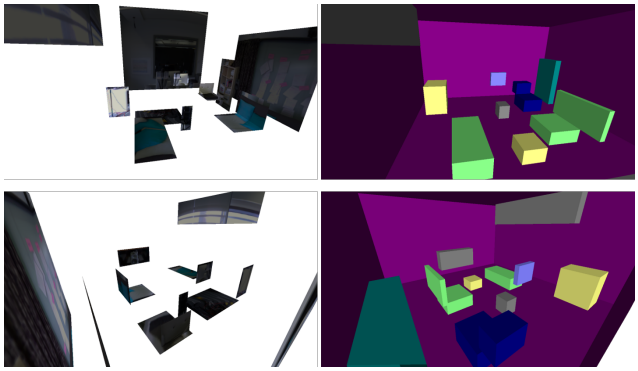


Figure 10. Block-based reconstruction results for UR2 (left: plane elements from the vision sensor, right: final model)

Figure 11 (a) shows the results of LR. 19 planes were detected from the 360° camera including loudspeakers on the ceiling. However, the ceiling itself could not be reconstructed due to the loudspeakers and their frames. The room size estimated from 4 walls is 4.9 m $\times$ 5.71 m, which have

Table 2. Detected reflector positions for UR2

	X plane							Y plane		Z plane	
Point	1	2	3	4	5	6	7	8	9	10	11
X	-2.57	-2.88	-1.92	-1.54	1.61	2.04	2.99	0.02	0.21	0.12	0.06
Y	-0.70	0.29	0.16	-0.13	-0.20	0.12	0.52	3.20	-2.06	-0.17	-0.11
Z	0.10	-0.74	-0.92	-0.45	-0.45	-0.79	-0.47	-0.03	-0.12	-1.20	1.46

Table 3. Evaluation of detected object plane locations in UR2 (Unit in $meter$. "GT" is Ground-truth, "A" is from audio and "V" from visual sensors.)

Dir.	Object	Source	GT	Estimated	Error
X	Right wall	A+V	2.83	2.81	0.02
	Left wall	A+V	-2.61	-2.57	0.04
	Bookshelf	V	2.31	2.38	0.07
	Right sofa	A+V	1.79	1.53	0.26
	Left sofa	A	-1.45	-1.54	0.09
Y	Front wall	A+V	3.50	3.48	0.02
	Rear wall	A	-2.07	-2.06	0.01
	Monitor	V	2.05	1.87	0.18
	Left table	V	1.82	1.96	0.14
	Front sofa	V	2.45	2.38	0.07
	TV	V	-1.75	-1.65	0.10
Z	Floor	V	-1.49	-1.53	0.04
	Ceiling	A	1.42	1.46	0.04
	Front sofa	V	-1.01	-1.11	0.10
	Right sofa	A+V	-1.01	-1.19	0.18
	Left sofa	V	-1.01	-1.11	0.10
	Centre table	A	-0.95	-1.20	0.25

around 60cm error in both direction. On the other hand, 6 reflection points were detected from the audio sensors. The azimuth-time domain of the beamformed audio signal is reported in Fig. 12 (a). The ceiling that was missed by vision sensors was detected from the audio sensors. However, two false reflection points were detected: one in front of the rear wall, probably corresponding one to a higher order reflection, the other one to a wrong estimation of one of the loudspeaker positions. Due to these false reflections, phantom volumes were generated in the final cuboid model. Object (1) in the rendered scene is a phantom object from the false reflector, Object (2) is the plane detected from the vision sensor which should be a wall but modelled as an object cuboid due to the false reflection detected behind this wall. Strong multiple order reflections in a shoebox-like environment disturbed detection of 1st-order reflections.

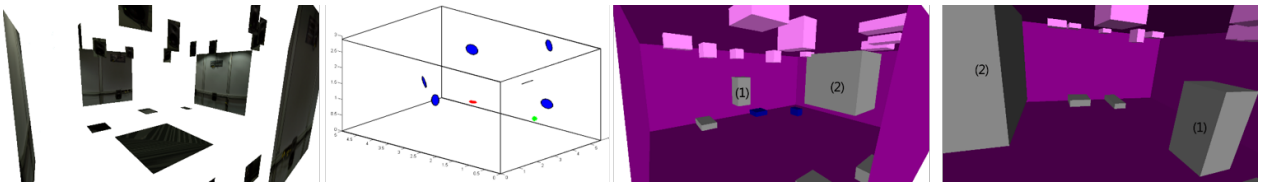
In MR2 results illustrated in Fig. 11 (b), 19 planes were detected from the camera. Most of the objects and walls were detected but one side wall with windows (object (2)) was misplaces and the cabinets (object(1)) were not reconstructed due to stereo matching errors. Instead, seven reflection points were detected from the audio sensors: three corresponding to the walls, one each to the ceiling, chair, cabinet and loudspeaker. The azimuth-time domain of the beamformed audio signal is reported in Fig. 12 (b). One of the two missing cabinets (object (1)) in the scene was reconstructed from audio. the reflection point on the window (object (3)) could retrieve the right position of the side wall and the wrong window plane detected from the camera remained as a phantom volume (object (2)). Object (4) looks like a phantom volume but it is a position of different loudspeaker used for the audio recording. The estimated room size was $4.34\,\mathrm{m}\times5.01\,\mathrm{m}\times2.28\,\mathrm{m}$, which was a bit shorter to the Y direction than the ground-truth because the original wall could not be detected due to many cluttered objects near the wall.

Figure 11 (c) shows the results of S1. 9 planes and 7 points were detected from the 360° cameras and audio sensors, respectively. One wall was missing from the vision sensor. 3 reflection points correspond to the floor plane detected from the camera, 2 points were from each wall, 1 point from the chair and 1 point is unknown. However, the audio sensor could not detect the front and real walls because they were to far, thus, the related signals were recorded, at the microphone position, with low SNRs. The azimuth-time domain of the beamformed audio signal is reported in Fig. 12 (c). One phantom object was generated in the *X* direction. The estimated room size was $15.76\,\mathrm{m}\times16.09\,\mathrm{m}\times5.34\,\mathrm{m}$ which is close to the ground-truth ($< 1.3\,\mathrm{m}$ error to each direction) considering the volume of the scene.

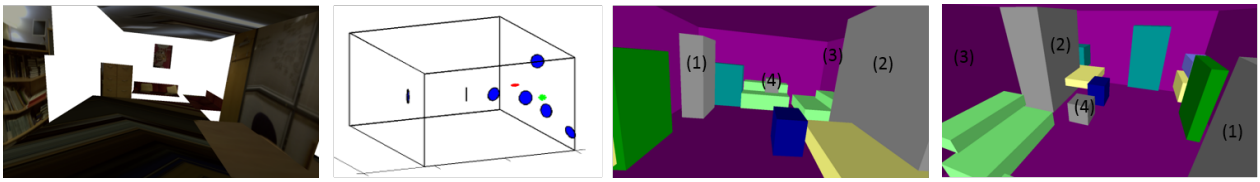
5. Conclusions

We have developed a solution combining acoustic and visual sensors for the challenging task of room geometry reconstruction with transparent and mirror surfaces. The experiments show that the two modalities act in a complementary way for surface reconstruction. For the objects which are not correctly detected through existing vision-based methods, the acoustical information provides reliable localisation. On the other hands, objects having high acoustic absorption coefficients, thus that are not detected by the audio analysis, are reconstructed through the vision sensors.

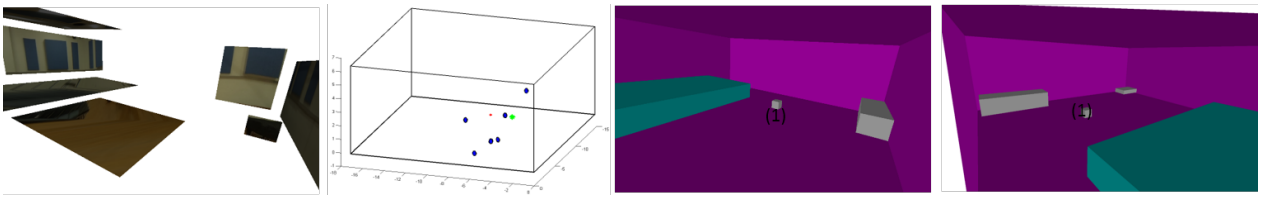
This work is still in progress rather than a definitive evaluation. Future work may continue within the scene analysis research area, by studying new techniques for object and material recognition, in order to aid acoustic reflection analysis. The peak-picking algorithm used for reflector detection is simple, hence, it introduces false reflectors into the analysis. Therefore, more sophisticated algorithm to detect maxima given 2D images is required. Finally, optimal way



(a) LR

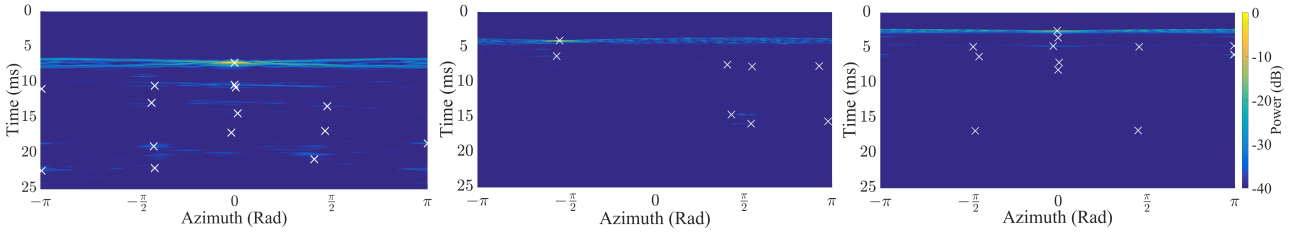


(b) MR2



(c) S1

Figure 11. Reconstruction results for other datasets (1st col.: Reconstructed planes from vision sensor, 2nd col.: Localised reflectors from audio sensor, 3rd and 4th col.: Snapshots of final block models



(a) LR (b) MR2 (c) S1

Figure 12. Energy analysis in the azimuth DOA-time domain for three analysed datasets (background figures). The white crosses are the reflection DOAs and TOAs detected by the angle-constrained peak-picking algorithm.

to fuse planes from vision sensors and reflectors from audio sensors based on their reliability will be developed so that more accurate object localisation can be achieved.

Acknowledgements

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

References

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoustical Society of America*, 4(65):943–950, 1979.

[2] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro. Inference of room geometry from acoustic impulse responses. *IEEE Transactions on Audio, Speech and Language Processing*, 20(10):2683–2695, 2012.

[3] M. R. Bai and C.-C. Chen. Application of convex optimization to acoustical array signal processing. *J. of Sound and Vibration*, 332(25):6596–6616, 2013.

[4] L. Barazzetti, M. Previtali, and F. Roncoroni. 3d modelling with the samsung gear 360. pages 85–90, 2017.

[5] K. Chen, Y.-K. Lai, and S.-M. Hu. 3D indoor scene modeling from RGB-D data: a survey. *Computational Visual Media*, pages 267–278, 2015.

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

[7] P. Debevec, C. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and

image-based approach. In *Proceedings of SIGGRAPH*, pages 11–20, 1996.

[8] I. Dokmanić, Y. M. Lu, and M. Vetterli. Can one hear the shape of a room: the 2-D polygonal case. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[9] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli. Acoustic echoes reveal room shape. *Proc. of the National Academy of Science of the United States of America (PNAS)*, 110(30):12186–12191, 2013.

[10] D. Eberly. Least Squares Fitting of Data. `http://www.geometrictools.com/Documentation/LeastSquaresFitting.pdf`, 2016. [Online; accessed 13-July-2017].

[11] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. of the 108th Audio Engineering Society Convention (AES)*, Paris, France, 2000.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2), 2004.

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

[14] R. C. Gonzalez and R. E. Woods. *Digital image processing - Third edition*. Pearson, Prentice Hall, 2008.

[15] C. Hane, C. Zach, B. Zeisl, and M. Pollefeys. A patch prior for dense 3d reconstruction in man-made environments. In *Proceedings of 3DIMPVT*, pages 563–570, 2012.

[16] M. W. Hussain, J. Civera, and L. Montano. Grounding acoustic echoes in single view geometry estimation. In *Proc. AAAI*, pages 2760–2766, 2014.

[17] H. Kim, T. de Campos, and A. Hilton. Room layout estimation with object and material attributes information using a spherical camera. In *Proc. 3DV*, pages 519–527, 2016.

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

[19] H. Kim and K. Sohn. 3d reconstruction from stereo images for interactions between real and virtual objects. *Signal Processing: Image Communication*, 20(1):61–75, 2005.

[20] M. Kuster, D. de Vries, E. M. Hulsebos, and A. Gisolf. Acoustic imaging in enclosed spaces: analysis of room geometry modifications on the impulse response. *J. Acoustical Society of America*, 116(4):2126–2137, 2004.

[21] H. Kuttruff. *Room acoustics - Fifth edition*. Spon press, 2009.

[22] A. Lagudi, G. Bianco, M. Muzzupappa, and F. Bruno. An alignment method for the integration of underwater 3d data captured by a stereovision system and an acoustic camera. *Sensors*, 16(4), 2016.

[23] 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*, 60(11):887–898, 2012.

[24] V. Murino and A. Fusiello. Augmented scene modeling and visualization by optical and acoustic sensor integration. *IEEE Transactions on Visualization and Computer Graphics*, 10:625–636, 2004.

[25] W. Nguatem, M. Drauschke, and H. Mayer. Finding cuboid-based building models in point clouds. In *Proceedings of ISPRS*, pages 149–154, 2012.

[26] M. R. Oswald, E. Toeppe, C. Nieuwenhuis, and D. Cremers. A survey on geometry recovery from a single image with focus on curved object reconstruction. In *Proc. Conference on Innovations for Shape Analysis: Models and Algorithms*, 2011.

[27] L. Remaggi, P. J. B. Jackson, P. Coleman, and J. Francombe. Visualization of compact microphone array room impulse responses. In *Proc. of the 139th Audio Engineering Society Convention (AES)*, New York, USA, 2015.

[28] 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*, 25(2):296–309, 2017.

[29] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002.

[30] M. Schoenbein and A. Geiger. Omnidirectional 3d reconstruction in augmented manhattan worlds. In *Proc. IROS*, pages 716 – 723, 2014.

[31] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, pages 519–528, 2006.

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

[33] H. Su, H. Fan, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proc. CVPR*, 2017.

[34] S. Tervo, J. Pätynen, and T. Lokki. Acoustic reflection localization from room impulse responses. *ACTA Acustica united with Acustica*, 98:418–440, 2012.

[35] S. Tervo and T. Tossavainen. 3D room geometry estimation from measured impulse responses. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012.

[36] V. Välimäki, J. A. Parker, L. Savioja, J. O. Smith, and J. S. Abel. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(5):1421–1448, 2012.

[37] J. Xiao and Y. Furukawa. Reconstructing the worlds museums. *International Journal of Computer Vision*, 110(3):243–258, 2014.

[38] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Proc. NIPS*, 2016.

[39] M. Ye, Y. Zhang, R. Yang, and D. Manocha. 3d reconstruction in the presence of glasses by acoustic and stereo fusion. In *Proc. CVPR*, pages 4885–4893, 2015.