An Audio-Visual Method for Room Boundary Estimation and Material Recognition

Luca Remaggi
CVSSP, University of Surrey
Guildford, UK
l.remaggi@surrey.ac.uk

Hansung Kim
CVSSP, University of Surrey
Guildford, UK
h.kim@surrey.ac.uk

Philip J. B. Jackson
CVSSP, University of Surrey
Guildford, UK
p.jackson@surrey.ac.uk

Adrian Hilton
CVSSP, University of Surrey
Guildford, UK
a.hilton@surrey.ac.uk

ABSTRACT

In applications such as virtual and augmented reality, a plausible and coherent audio-visual reproduction can be achieved by deeply understanding the reference scene acoustics. This requires knowledge of the scene geometry and related materials. In this paper, we present an audio-visual approach for acoustic scene understanding. We propose a novel material recognition algorithm, that exploits information carried by acoustic signals. The acoustic absorption coefficients are selected as features. The training dataset was constructed by combining information available in the literature, and additional labeled data that we recorded in a small room having short reverberation time (RT60). Classic machine learning methods are used to validate the model, by employing data recorded in five rooms, having different sizes and RT60s. The estimated materials are utilized to label room boundaries, reconstructed by a vision-based method. Results show 89 % and 80 % agreement between the estimated and reference room volumes and materials, respectively.

CCS CONCEPTS

• **Computing methodologies → Object detection**; **Object recognition**; *Instance-based learning*; *Feature selection*;

KEYWORDS

Audio-Visual, Material Recognition, Room Boundary Estimation, KNN, Acoustic Absorption Coefficient

ACM Reference format:
Luca Remaggi, Hansung Kim, Philip J. B. Jackson, and Adrian Hilton. 2018. An Audio-Visual Method for Room Boundary Estimation and Material Recognition. In *Proceedings of 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Seoul, Republic of Korea, October 26, 2018 (AVSU'18),* 7 pages.
https://doi.org/10.1145/3264869.3264876

1 INTRODUCTION

Humans rely on understanding the audio-visual characteristics of the environments to interact with the world [53]. In most cases, inputs to the human sensing system are paired audio and video signals [50]. Therefore, machines should be provided with capability of analyzing audio-visual scenes. This would allow them to emulate the human perception of the world, that is the benchmark for many of the current artificial intelligence technologies. This is important for different areas of application. For instance, it can be exploited by robots to autonomously carry out tasks [24]. Virtual reality (VR) and augmented reality (AR) are currently two major topics in the audio-visual research [5, 7, 36]. Several VR software development kits (SDKs) are freely available to be used by researchers. Classically, they mainly focused on the visual experience [17, 52]. However, recently, some toolkits have been extended with the inclusion of spatial audio tools [19, 39]. Models that describe both geometry and materials composing the environments are thus of great interest for VR and AR, since they allow to approximate real room acoustics [21, 26, 43]. Studies demonstrated that high-quality sound reproductions improve the perceived similarity to reference environments [8, 44].

3D room boundary estimation has been an important research topic for practical applications. Computer vision techniques using visual sensors have played a major role in geometry reconstruction. Recovering geometric information from a single perspective photograph or 360 image relies on geometrical cues such as lines and texture [51, 56, 57]. 3D reconstruction from stereo or multiple images is widely used for general scene reconstruction [45, 46]. Kinect-like Red Green Blue Depth (RGBD) sensors also provide good depth information for an indoor scene estimation [9, 10]. However, vision sensors fail for transparent, reflective or featureless uniform surfaces which often arise in common indoor scenes. Several methods were proposed in the audio signal processing community to localize the room boundaries [1, 14, 41], and detect small objects near the listening position [42]. These are of great interest, especially in those conditions where vision sensors fail. A few studies combined audio and visual sensors to reconstruct 3D geometry such as sonar + camera [37], Kinect + ultrasonic sensor [58], acoustic echoes + single photo [23], and acoustic echoes + 360 photo [27] .

During the last decade, several vision-based approaches were proposed to tackle the problem of identifying surface materials. In [31], Bayesian generative models were proposed to exploit features such

AVSU'18, October 26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5977-1/18/10…$15.00
https://doi.org/10.1145/3264869.3264876

as color and micro-texture. Kernel descriptors were then utilized in [22], together with a Nearest Neighbor (NN)-based approach. Later, accuracy was improved by employing convolutional neural networks [4]. These were also employed in [25], where multi-class classification was proposed to identify both the object type and attributes. However, performance provided by all the methods above was not of high-quality, particularly, when cross-dataset scenarios where analyzed. As it happens for humans, vision-based features are not strong enough to accurately classify surface materials by themselves. Recently, approaches have been proposed trying to exploit acoustic features. In [32], a robotic finger was built to emulate the action of "knocking" on objects. Although the experiments provided good results, this approach is limited by the availability of dynamic robots. Later, in [34], it was proposed to analyze the acoustic characteristics of the reflective surfaces at specific sound frequencies. However, it was suggested to use different frequencies depending on the tested material, thus making it not generalizable.

In this paper, we propose an audio-visual method to localize the room boundaries and classify them in terms of their material. A block diagram representing the proposed method is depicted in Figure 1. Three are the main novelties:

- A method to extract acoustic reflector absorption coefficients from recorded spatial room impulse responses (RIR);
- A method for acoustic reflector material classification;
- A combination of the proposed material recognition method with a vision-based off-the-shelf room boundary estimator, to associate room geometry and materials.

The rest of the paper is organized as follows: Section 2 discusses the theoretical concepts behind the proposed approach; Section 3 describes the vision-based method proposed to reconstruct the room boundaries; Section 4 presents the novel methods to classify the material of acoustic reflectors; in Section 5 experiments and results are discussed; Section 6 draws the overall conclusion.

2 THEORETICAL BACKGROUND

The material recognition method that we propose is based on determining the material acoustic absorption coefficients, given RIRs. In this section, we describe the theory that is behind our approach.

2.1 Room Impulse Response

A sound propagating within a reverberant environment is observed, at the listening position, as filtered version of the produced sound [29]. With n being the discrete time domain variable, and $s_j(n)$ the sound produced at the j-th source, we can write the signal recorded by the i-th microphone as:

$$s'_i(n) = r_{i,j}(n) * s_j(n) + a_{i,j}(n), \tag{1}$$

where "$*$" denotes convolution, and $a_{i,j}(n)$ is the additive Gaussian noise. $r_{i,j}(n)$, by acting as a filter, shapes the received sound, and it is widely known as RIR [29].

A RIR is an acoustic signal, carrying information about the environment in which it is recorded. It is generallt considered as being composed of three elements [29]: the direct sound; the early reflections; and the late diffuse reverberation. From this classical decomposition, the RIR from source j to sensor i can be defined as superimposition of bursts, delayed by $n_{k,i,j}$ samples, with k enumerating the reflections:

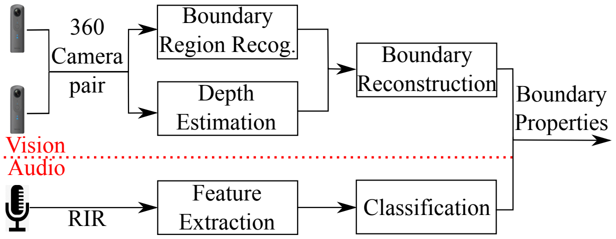

Figure 1: The proposed audio-visual method to reconstruct the room boundaries and determining the related materials.

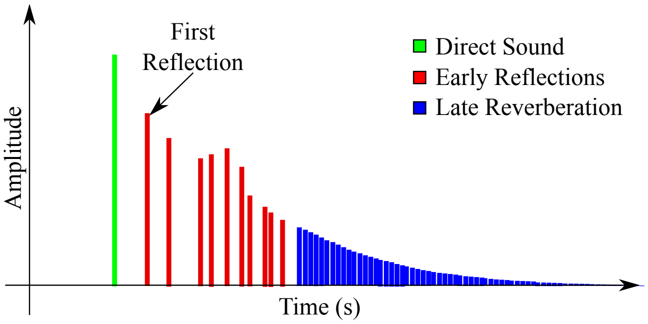

Figure 2: A graphical representation of a RIR, with the three components highlighted by different colors.

$$r_{i,j}(n) = \sum_{k=0}^{T_m} h_{k,i,j}(n - n_{k,i,j}) + l(n), \tag{2}$$

where $h_{0,i,j}(n)$ represents the direct sound, $h_{k,i,j}(n)$ the discrete early reflections, and $l(n)$ is the late reverberation modeled as exponentially decaying Gaussian noise; T_m is the k-th peak corresponding to the last reflection before the mixing time. A graphical representation of a RIR is depicted in Figure 2, where the red bursts (i.e. the early reflections $h_{k,i,j}(n)$) are different acoustic paths between source j and microphone i, within the recording environment.

2.2 Acoustic Absorption Coefficient

In general, when a sound, during its propagation, encounters obstacles, i.e. acoustic reflectors, several physical phenomena may occur, such as scattering, refraction, diffraction, evanescent waves, etc [29]. However, here, we assume the specular component of the reflection to be dominant, as drawn in Figure 3.

The sound reflection factor, R, gives the ratio of the specularly reflected and incident sound pressure, i.e. p_r and p_i respectively, as $R = p_r/p_i$ [13]. By following the complex nature of the sound propagation, R incorporates both magnitude and phase information about the reflection. The absorption coefficient is defined as ratio of the absorbed and incident energy, and it can be formulated as [13]:

$$\alpha = 1 - |R|^2, \tag{3}$$

where $|R|$ is the absolute value operator. Since α is frequency dependent, we define α_b as the absorption coefficient referring to the b-th frequency octave band. In this paper, we consider the six octave

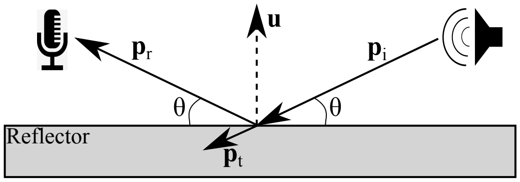

Figure 3: A schematic drawing of the physical behavior of sound impinging on an acoustic reflector. Loudspeaker and microphone are shown as icons, p_i is the incident sound wave, p_r the reflected one, p_t is the transmitted/absorbed sound, u represents the normal vector to the reflective surface, and θ is the angle of incidence.

bands between 125 Hz and 4 kHz. These are the frequencies that are mostly analyzed in the literature in terms of reflector material absorption coefficients [13, 55].

3 ROOM BOUNDARY RECONSTRUCTION

The room boundaries are estimated by using off-the-shelf 360° cameras. Two spherical panorama images captured at two different heights are used to produce a room layout by stereo matching. Ceiling, floor and wall regions are detected using convolutional neural network trained for semantic segmentation (SegNet) [2]. For those regions, shoe-box like room boundaries aligned to the main axes are reconstructed by depth estimation.

3.1 Boundary Region Recognition

To recover 3D room boundary information, the scene is captured as a vertical stereo image pair by Theta S cameras by Ricoh[1], which provides well-aligned seamless stitching with minimal distortion in mapping to the Spherical coordinates. Images acquired from two fish-eye lenses are stitched into an equirectangular projection image as shown in Figure 4 (a). To extract room boundary regions, a semantic segmentation of the scene is performed with SegNet [2], a deep fully convolutional neural network architecture employing an encoder-decoder architecture with the first 13 convolutional layers of VGG16 [47]. 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 as shown in Figure 4 (b). The captured spherical panorama image is projected onto a unit cube by the cubic projection [25] to provide six perspective images of the scene, because the SegNet works with a model trained on the SUN RGB-D indoor scenes dataset [49].

3.2 Depth Estimation

3D geometry of the room boundary is reconstructed using dense stereo matching with spherical stereo geometry [25]. Depth of the scene is calculated by triangulation from the estimated disparity and baseline distance between two camera positions. Disparity estimation can be carried out along the column lines in the vertical stereo image pair. Figure 4 (c) shows the disparity map estimated from Figure 4 (a). All depth points in the regions labelled as ceiling, floor and wall in Figure 4 (b) are projected to the 3D space and form a 3D point cloud. The volume of the cuboid is decided by the 3D point occupancy in the cluster using least squares optimisation [30]. In order to eliminate outliers from depth estimation in the point cloud, 10% of the farthest points from the mean of each plane cluster are excluded in the boundary estimation.

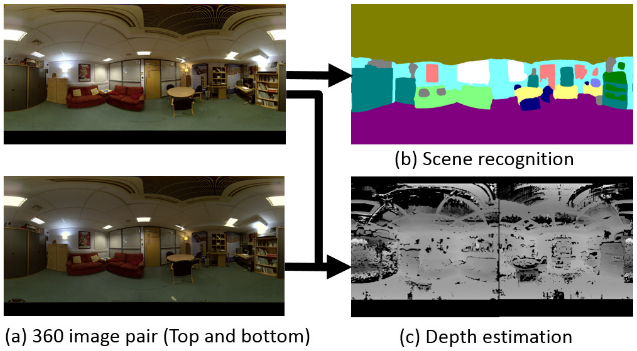

Figure 4: Room boundary estimation system.

4 MATERIAL RECOGNITION

Once the room boundaries are determined, it is important to estimate also their acoustic properties, to aid targeted applications, such as VR and AR. In this paper, we analyze large planar reflectors. Therefore, the acoustic properties depend mainly on the type of material composing them. We formulate this problem as classification, where the classes represent different materials.

The acoustic method for material classification utilizes the acoustic reflector absorption coefficients as features, evaluated in octave bands [29]. In the literature, absorption coefficients are typically measured within controlled environments, by employing experimental tools, such as reflective tubes [13]. However, here, since we use recorded RIRs, we do not aim to accurately estimate the absorption coefficients of the tested material directly. Instead, we aim to estimate them accurately enough to match with the ideal absorption coefficient values of the sample material measured in controlled environments [13, 55]. To do so, we take into account loudspeaker and microphone directivity, to compensate for the signal lost [35]. Furthermore, the angle of incidence is modeled [13].

4.1 Feature Extraction

The feature that we propose to use is the frequency-dependent absorption coefficient α_b. As already stated above, in Section 2.2, we observe α_b in the six octave bands between 125 Hz and 4 kHz. Therefore, the space under investigation is six dimensional.

To calculate the $\alpha_{b,k}$s, the k-th reflection $h_{k,i,j}(n)$ has to be segmented from $r_{i,j}(n)$. As it was also done in [41], these segments were obtained by applying a Hamming window $w(n)$, of length W, centered at the reflection time of arrival $n_{k,i,j}$:

$$h_{k,i,j}(n) = r_{i,j}(n) \cdot w(n - n_{k,i,j}). \tag{4}$$

The same process is applied to the direct sound, to obtain $h_{0,i,j}(n)$.

Both the direct sound (i.e. $k = 0$) and the k-th reflection segments are transformed into the frequency domain as $H_{k,i,j}(\omega) = \mathscr{F}\{h_{k,i,j}(n)\}$, where $\mathscr{F}\{\cdot\}$ represents the discrete time Fourier

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

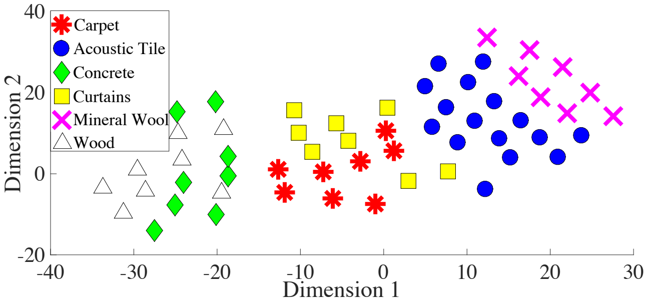

Figure 5: 2D visualization of the six-dimensional absorption coefficient data obtained from recorded RIR samples. The dimensionality reduction was performed by using the t-distributed Stochastic Neighbor Embedding.

transform operator. The complex frequency-dependent reflection factor of the k-th reflector can be then calculated as:

$$R(\omega) = \frac{H_{k,i,j}(\omega)}{H_{0,i,j}(\omega)} \cdot K^{\text{air}}(\omega) \cdot K^{\text{dir}}(\omega). \tag{5}$$

$K^{\text{air}}(\omega) \in \mathbb{R}$ and $K^{\text{dir}}(\omega) \in \mathbb{R}$ are frequency-dependent coefficients that compensate our model of $R(\omega)$ to take into account of the air absorption and loudspeaker directivity, respectively. $K^{\text{air}}(\omega)$ is obtained from the frequency-dependent sound absorption in air[2] reported in [33]. $K^{\text{dir}}(\omega)$ can be obtained by looking at the directivity characteristics on the employed loudspeaker datasheet.

The frequency-dependent absorption coefficient can be calculated similar to Equation 3. To generalize the model, and remove any assumption regarding the microphone and source positions with respect to the reflector, the angle of incidence θ has to be included. Hence, the angle-independent absorption coefficient is [38]:

$$\alpha(\omega, \theta) = (1 - |R(\omega)|^2) \cdot \sin\theta. \tag{6}$$

The absorption coefficients used as features for classification are calculated by averaging $\alpha(\omega, \theta)$ within each frequency band as:

$$\alpha_b = \frac{1}{B_b} \sum_{\omega=\Omega_b - \frac{B_b}{2}}^{\Omega_b + \frac{B_b}{2}} \alpha(\omega, \theta), \tag{7}$$

where Ω_b is the central frequency of the b-th octave band and B_b is the b-th octave band size. Features extracted from recorded RIRs are shown in Figure 5. There, for visualization purposes, dimensionality reduction was applied to the six-dimensional data by using the t-distributed Stochastic Neighbor Embedding [54].

4.2 Classification

This is, to our knowledge, the first attempt to perform a classification of room boundary materials given common acoustic signals, such as RIRs. We propose to use a well-known, simple and effective classification algorithm, based on kNN [6]. However, data estimated by the feature extraction method proposed in Section 4.1 may present some outlying samples. Hence, to reduce their contribution to the classification result, we propose to employ the distance weighted kNN (WKNN) [15].

WKNN is a classification algorithm that implements a vote among neighbors, within a given distance d^{MAX} from the test sample. Closer neighbors will have a greater influence than neighbors which are further away. A weight g_t attributed to the t-th nearest neighbor can be defined by looking at its Euclidean distance d_t from the test sample as [15]:

$$g_t = \frac{1}{d_t}, \qquad d_t \neq 0. \tag{8}$$

$d_t = ||z - x_t||$, where z is the test sample's position, x_t the t-th neighbor's position, and $||\cdot||$ is the ℓ_2 norm operator. To improve the algorithm efficiency, k-dimensional trees are created during the training session [18]. These trees are data structures for organizing points in space, and are used to find the nearest neighbors, by calculating their d_ts. Being $\mathbf{T}$ the set of nearest neightbors, classification can be performed as [20]:

$$y' = \arg\max_{y} \sum_{(x_t, y_t) \in \mathbf{T}} g_t \cdot \delta(y = y_t), \tag{9}$$

where y is the label related to x, and $\delta(y = y_t)$, the Dirac delta function, takes a value of one if $y = y_t$ and zero otherwise.

5 EXPERIMENTS AND RESULTS

Experiments were run to evaluate the proposed audio-visual method. These experiments had as main applications AR and VR. For this reason, the reflector being analyzed to classify its material is always the one closest to the listening position. In fact, we aim to improve the quality of synthesized early reflections, and the first one is, perceptually, the most important one [3].

5.1 Datasets

Audio-visual data was captured in five rooms, having different sizes and reverberation times (RT60s). This data was used to estimate the room boundary and related materials.

5.1.1 Audio-Visual Datasets. RIRs were recorded in five rooms. Depending on the availability during the recording session, we employed either a bi-circular array of microphones [41] or a Soundfield microphone [11]. The bi-circular array circles have radii 85 cm and 106 cm and are composed of 24 omnidirectional microphones each. For the Soundfield microphone, only the W-omni channel is selected for our purposes. Two 360° cameras were, separately, placed at the same position of the microphone array.

The room dimensions are reported in Table 1 (right side). The acoustic properties, such as the RT60s averaged between 125 Hz and 4 kHz, the analyzed reflector materials, and the loudspeaker and microphone positions, are reported on the left side of Table 1. The first room, named as "LR", is a listening environment at the BBC MediaCity, in Salford, UK, where we used the bi-circular array of microphones [28]. The same microphone array was also employed within the second room, named as "UL" [28], that is a lab at the BBC MediaCity. There, to evaluate two reflectors, two microphone array and loudspeaker positions were recorded. "ST" is a large recording studio at the University of Surrey [12]. Also in ST the bi-circular array was used for the recordings. The fourth dataset is a meeting room "MR" at the University of Surrey, where the bi-circular array was employed to record the RIRs. The last dataset is "CY". It is a

[2] It is selected a standard temperature of 20 °C and humidity of 50%.

Table 1: Acoustic properties of the rooms (left) and evaluation of layout estimation (right).

Data	RT60 (ms)	Material GT	Loudsp. pos. (m)	Mic. pos. (m)	Ground-truth (m^3)	Estimated (m^3)	Vol. err. (%)
MR	270	Acoustic Tile	[2.12, 0.33, 1.00]	[3.31, 0.21, 1.50]	5.61×4.28×2.33	5.52×4.35×2.36	1.3
UL	275	Wood	[4.79, 2.76, 1.07]	[2.52, 2.73, 1.07]	5.57×5.20×2.91	5.92×4.95×2.95	27.0
		Curtains	[1.81, 4.66, 1.07]	[2.62, 3.92, 1.07]			
LR	222	Carpet	[2.81, 0.56, 1.08]	[2.81, 2.55, 1.08]	5.64×5.05×2.90	5.77×5.17×2.98	7.6
ST	913	Wood	[7.12, 12.14, 1.50]	[7.12, 10.14, 1.50]	17.08×14.55×6.50	16.53×14.87×5.70	13.2
CY	688	Concrete	[5.42, 14.99, 1.69]	[3.35, 12.63, 1.70]	10.10×19.0× −	9.61×18.51× −	7.3
AVG	–	–	–	–	–	–	11.3

Table 2: Classification precision for each dataset.

Data	MR	UL	LR	ST	CY	AVG
Precision	62 %	54 %	85 %	98 %	100 %	80 %

courtyard placed within the Center for Vision, Speech and Signal Processing, at the University of Surrey. This is a peculiar dataset, different from the others since it does not have any ceiling. Here, the Soundfield microphone was used.

RIRs were recorded by employing the swept-sine method [16]: the five rooms were excited through a 10 s swept-sine signal between 20 Hz and 24 kHz, with a sampling frequency of 48 kHz, and the recorded signals were then deconvolved to obtain RIRs. To perform this kind of measurements the background noise level was always kept 30 dB below the produced sound's. From the RIRs recorded at each omnidirectional microphone available, by using a single loudspeaker position, the first arriving reflection is processed by our feature extraction method, as described in Section 4.1 (i.e. we have one test sample for each omnidirectional microphone available during the recording sessions). The obtained samples were then used as test data for the classification algorithm in Section 4.2.

5.1.2 Training Dataset. The training dataset for classification was created by combining two types of data. The first one is composed of about 250 samples. It was obtained by selecting the absorption coefficients of materials provided in the literature [13, 55]. These were calculated by employing traditional controlled experiments, e.g. by using a reflective tube within an anechoic chamber [13]. However, in our proposed method, we follow a more generalizable approach: we calculate the new samples' absorption coefficients from recorded RIRs, by using the proposed method in Section 4.1. Thus, there is need for enriching the training data with samples calculated with the same procedure. Therefore, we recorded RIRs in a small room, at the University of Surrey, originally built as RF anechoic chamber. There, different material samples were acoustically excited, separately, to record their reflections. Then, we applied the proposed feature extraction method.

The classes represented in the whole training dataset are: "Curtains", "Wood", "Concrete", "Plasterboard", "Glass", "Acoustic Tile", "Mineral Wool", "Linoleoum", and "Carpet". They were defined by fusing together those sets of similar materials that are available in the Google VR SDK to synthesize sound [19] (e.g. different type of concrete in [19] are considered here as part of the general class "concrete"). However, the five rooms, that we employed for testing, contained a limited number of materials, that did not span the whole range of classes available in the training dataset that we generated. Therefore, five are the classes tested in this paper: "Carpet", "Concrete", "Wood", "Acoustic Tile", and "Curtains".

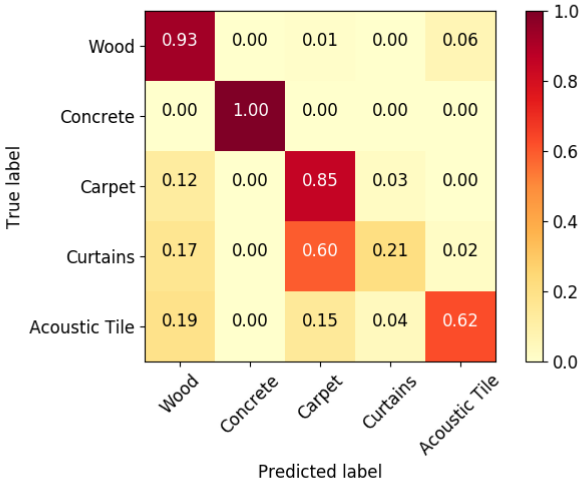

Figure 6: Normalized Confusion Matrix.

5.2 Classification Evaluation Metrics

The material classification is evaluated by calculating [48]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{10}$$

where TP stands for true positives and FP for false positives. This is calculated for each dataset and reported in percentage. Furthermore, we evaluate the algorithm accuracy in recognizing different materials by calculating a confusion matrix C. This is constructed such that $C_{l,p}$ is equal to the number of observations known to be in group l but predicted to be in group p. Since the number of tested samples is unbalanced, we report the normalized confusion matrix, calculated by looking at the class support size [40].

5.3 Geometry Estimation Result

The right side of Table 1 shows the ground-truth and estimated dimensions with volume error against the ground-truth for the test scenes. The estimated room dimensions of MR, LR and CY are very close to the ground-truth (less than 10%) due to sufficient features in the scene. UL shows large errors because of depth estimation error on the featureless dark wall behind the TV and transparent

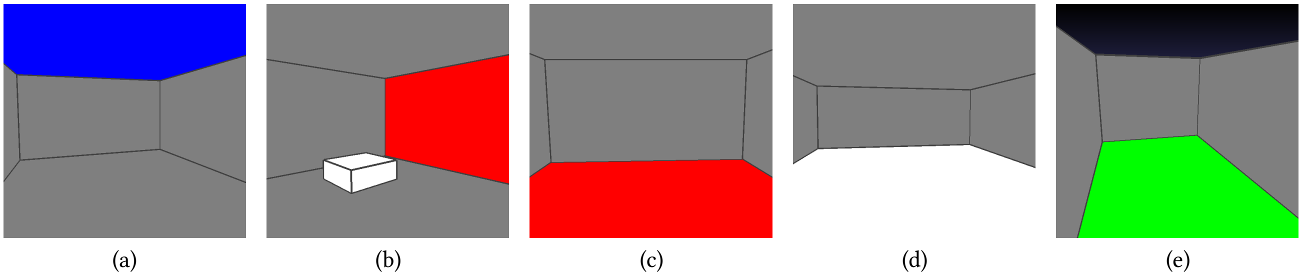

Figure 7: Estimated room layouts with material estimated for the main acoustic reflectors. (a) represents MR, with the ceiling being classified as Acoustic Tile; (b) is UL, with a wall classified as carpet, and a table as wooden; (c) is LR, with the floor being recognized as made of carpet; (d) is ST, with the floor classified as wooden; (e) is CY, with the floor material estimated to be concrete and the ceiling labeled in dark violet because missing (i.e. the courtyard is open air).

windows. In the ST scene, the height of the space was incorrectly estimated due to the rails on the ceiling and saturated lights. CY does not have ceiling and the error relates to the area of side walls.

5.4 Classification Results

Results, related to the proposed classification method, are reported in Table 2. This table highlights the high performance of our proposed method, with a precision, averaged over the five datasets, that is 80 %. This precision percentage is better than the one that can be achieved by using vision-based material recognition methods. Considering cross-dataset experiments, in fact, they usually provide an accuracy of about 40 % [4]. Our method works perfectly in CY and has a precision of 98 % in ST. In LR, it provides good performance, with a precision that is more than 80 %. However, the proposed method faces some issues in MR and UL, with a precision of 62 % and 54 %, respectively. These two rooms are the only rooms, among the tested ones, being furnished. They were both set up to reproduce living room-like environments. This generates scattered energy that interferes with the analyzed reflections. Furthermore, in both these rooms, the microphone array was placed next to sofas.

By looking, at Figure 6, we can then analyze the proposed method from the performance in classifying materials. In general, performance is high, with carpet (i.e. in LR), concrete (i.e. in CY) and wood (i.e. in ST and UL) estimated with a precision always above 85 %. However, acoustic tile, that was the ceiling material in MR, is recognized with lower precision. This is due to the reasons already discussed, related to the microphone array being placed next to a sofa, which generates scattered energy. It is interesting, finally, to note the recognition accuracy of curtains. Although the precision it is reported to be pretty low (i.e. slightly above 20 %), the proposed method mainly confuses these samples for carpet ones. This is understandable, since carpet and curtains have similar absorption coefficients [13] (see also Figure 3). This issue could be addressed in future work, by adding, some priors to strengthen the model. For instance, it could be assumed that carpets are found on the floor, whereas curtains are typically parallel to the walls.

5.5 Discussion

Some of the main areas of application for the proposed audio-visual method are VR and AR. For the first time in the literature, we propose a method to determine the material of the acoustic reflector that is closest to the listening position. To do so, we do not aim to calculate the exact value of its absorption coefficient. Instead, we aim to estimate it accurately enough to match with the respective absorption coefficient value, measured in controlled environments [13, 55]. This material classification allows a perceptually plausible synthesis of the most salient early reflection. Nevertheless, in the future, it could be extended to every reflector in the room, enabling appropriate synthesis of the late reverberation.

Results related to the proposed audio-visual method, to reconstruct the room layout and labeling the main reflector material, are graphically visualized in Figure 7. There, the material that has been classified with the highest probability is used to color the acoustic reflector, that was localized by using the vision-based method. To partially overcome the lack of materials available in the tested rooms, in UL (i.e. Figure 7(b)), two reflectors were classified, by using two microphones and two loudspeakers. In this way, we have been able to include in the analysis an additional class, i.e. "curtains". However, as already discussed, the classification algorithm confuses it for a carpet. The second material in UL referred to a tea table, and the classification algorithm correctly labeled it as "wood". The other rooms' reflectors are all correctly classified: LR's floor (Figure 7(c)) is classified as carpet; MR's ceiling is classified as Acoustic Tile (Figure 7(a)); ST's floor is estimated as being wooden (Figure 7(d)); and CY's floor is classified as concrete (Figure 7(e)).

6 CONCLUSIONS

A novel audio-visual pipeline, that locates and classifies the materials of acoustic reflectors, has been proposed. Experiments were performed on five rooms, showing the 80 % precision of the novel material classification algorithm and the importance of its integration within the vision-based room reconstruction method.

Future work may look at improving both the feature extraction, perhaps by looking at using the angle of incidence information. In addition, the core of the classification algorithm may be improved. Moreover, The method may be extended to recognize materials of other reflectors in the room. Furthermore, the reconstructed geometry may be applied to VR toolkits to produce synthetic spatial sound. The test dataset will be also enlarged by adding new rooms, to increase the number of tested materials.

ACKNOWLEDGMENTS

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, along with the terms for data access, are available from: http://dx.doi.org/10.15126/surreydata.00812228.

REFERENCES

[1] Y. El Baba, A. Walther, and E. A. P. Habets. 2018. 3D Room Geometry Inference Based on Room Impulse Response Stacks. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26, 5 (2018), 857–872.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (2017).

[3] S. Bech. 1998. Spatial aspects of reproduced sound in small rooms. *J. Acoustical Society of America* 103, 1 (1998), 434–445.

[4] S. Bell, P. Upchurch, N. Snavely, and K. Bala. 2015. Material Recognition in the Wild With the Materials in Context Database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] L. P. Berg and J. M. Vance. 2017. Industry use of virtual reality in product design and manufacturing: a survey. *Virtual Reality* 21, 1 (2017), 1–17.

[6] N. Bhatia and Vandana. 2010. Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security* 8, 2 (2010), 302–305.

[7] M. Billinghurst, A. Clark, and G. Lee. 2015. A Survey of Augmented Reality. *Foundations and Trends®in Human-Computer Interaction* 8, 2–3 (2015), 73–272.

[8] N. Bonneel, C. Suied, I. Viaud-Delmon, and G. Drettakis. 2010. Bimodal Perception of Audio-visual Material Properties for Virtual Environments. *ACM Transacions on Applied Perception* 7, 1 (2010), 1:1–1:16.

[9] K. Chen, Y.-K. Lai, and S.-M. Hu. 2015. 3D indoor scene modeling from RGB-D data: a survey. *Computational Visual Media* 1, 4 (2015), 267–278.

[10] S. Choi, Q.-Y. Zhou, and V. Koltun. 2015. Robust Reconstruction of Indoor Scenes. In *Proc. CVPR.*

[11] P. Coleman, A. Franck, D. Menzies, and P. J. B. Jackson. Berlin, Germany, 2017. Object-based reverberation encoding from first-order Ambisonic RIRs. In *Proc. of the 142th AES Conv.*

[12] P. Coleman, L. Remaggi, and P. J. B. Jackson. 2015. S3A room impulse responses. http://dx.doi.org/10.15126/surreydata.00808465. (2015).

[13] T. Cox and P. D'Antonio. 2016. *Acoustic absorbers and diffusers, third edition: theory, design and application.* CRC Press - Taylor & Francis Group.

[14] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli. 2013. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences* 110, 30 (2013), 12186–12191.

[15] S. A. Dudani. 1976. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6, 4 (1976), 325–327.

[16] A. Farina. 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. of the 108th Audio Engineering Society Convention.*

[17] Forge. 2018. Forge AR/VR Toolkit. http://forgetoolkit.com/. (2018).

[18] J. H. Friedman, J. L. Bentley, and R. A. Finkel. 1977. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. Math. Software* 3, 3 (1977), 209–226.

[19] Google. 2018. Google VR SDK. https://developers.google.com/vr/. (2018).

[20] J. Gou, L. Du, Y. Zhang, and T. Xiong. 2012. A New Distance-weighted k-nearest Neighbor Classifier. *J. of Information and Comput. Science* 9 (2012), 1429–1436.

[21] V. Hulusic, C. Harvey, K. Debattista, N. Tsingos, S. Walker, D. Howard, and A. Chalmers. 2012. Acoustic Rendering and Auditory-Visual Cross-Modal Perception and Interaction. *J. Computer Graphics Forum* 31, 1 (2012), 102–131.

[22] D. Hun, L. Bo, and X. Ren. 2011. Toward Robust Material Recognition for Everyday Objects. In *Proc. of the British Machine Vision Conference (BMVC).* 48.1–48.11.

[23] M. W. Hussain, J. Civera, and L. Montano. 2014. Grounding Acoustic Echoes in Single View Geometry Estimation.. In *Proc. AAAI.* 2760–2766.

[24] M. Hvilshøj, S. Bøgh, O. S. Nielsen, and O. Madsen. 2012. Autonomous individual mobile manipulation (AIMM): past, present and future. *Industrial Robot: An International Journal* 39, 2 (2012), 120–135.

[25] H. Kim, T. d. Campos, and A. Hilton. 2016. Room Layout Estimation with Object and Material Attributes Information Using a Spherical Camera. In *Fourth International Conference on 3D Vision (3DV).* 519–527.

[26] H. Kim, R. J. Hughes, L. Remaggi, P. J. B. Jackson, A. Hilton, T. J. Cox, and B. Shirley. Berlin, Germany, 2017. Acoustic Room Modelling Using a Spherical Camera for Reverberant Spatial Audio Objects. In *Audio Engineering Society Convention 142.* http://www.aes.org/e-lib/browse.cfm?elib=18583

[27] H. Kim, L. Remaggi, P. J. B. Jackson, F. M. Fazi, and A. Hilton. 2017. 3D Room Geometry Reconstruction Using Audio-Visual Sensors.. In *Proc. 3DV.* 621–629.

[28] H. Kim, L. Remaggi, P. J. B. Jackson, and A. Hilton. 2017. S3A audio-visual captures. https://doi.org/10.15126/surreydata.00812228. (2017).

[29] H. Kuttruff. 2009. *Room Acoustics - Fifth edition.* Spon press.

[30] S.-W. Kwon, F. Bosche, C. Kim, C. Haas, and K. Liapi. 2004. Fitting range data to primitives for rapid local 3D modeling using sparse range point clouds. *Automation in Construction* 13, 1 (2004), 67–81.

[31] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz. 2010. Exploring features in a Bayesian framework for material recognition. In *Proc. CVPR.* 239–246.

[32] H. Liu, X. Song, J. Bimbo, L. Seneviratne, and K. Althoefer. 2012. Surface material recognition through haptic exploration using an intelligent contact sensing finger. In *IEEE/RSJ International Conference on Intelligent Robots and Systems.* 52–57.

[33] M. Long. 2014. *Architectural Acoustics - 2nd Edition.* Academic Press.

[34] E. Lopez-Caudana, O. Quiroz, A. Rodríguez, L. Yépez, and D. Ibarra. 2017. Classification of materials by acoustic signal processing in real time for NAO robots. *International Journal of Advanced Robotic Systems* 14, 4 (2017).

[35] D. Markovic, K. Kowalczyk, F. Antonacci, C. Hofmann, A. Sarti, and W. Kellermann. 2014. Estimation of Acoustic Reflection Coefficients Through Pseudospectrum Matching. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 22, 1 (2014), 125–137.

[36] S. Mori, S. Ikeda, and H. Saito. 2017. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications* 9, 1 (2017), 1–17.

[37] V. Murino and A. Fusiello. 2004. Augmented Scene Modeling and Visualization by Optical and Acoustic Sensor Integration. *IEEE Transactions on Visualization and Computer Graphics* 10, 6 (2004), 625–636.

[38] C. Nocke. 2000. In-situ acoustic impedance measurement using a free-field transfer function method. *Applied Acoustics* 59, 3 (2000), 253–264.

[39] Oculus. 2018. Oculus SDK. https://developer.oculus.com/. (2018).

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[41] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang. 2017. Acoustic Reflector Localization: Novel Image Source Reversion and Direct Localization Methods. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25, 2 (2017), 296–309.

[42] L. Remaggi, H. Kim, P. J. B. Jackson, F. Fazi, and A. Hilton. 2018. Acoustic Reflector Localization and Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[43] Z. Ren, H. Yeh, R. Klatzky, and M. C. Lin. 2013. Auditory Perception of Geometry-Invariant Material Properties. *IEEE Transactions on Visualization and Computer Graphics* 19, 4 (2013), 557–566.

[44] F. Rumsey. 2002. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *J. Audio Engineering Society* 50, 9 (2002), 651–666.

[45] D. Scharstein and R. Szeliski. 2002. A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms. *International Journal of Computer Vision* 47, 1 (2002), 7–42.

[46] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *Proc. CVPR.* 519–528.

[47] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556 (2014).

[48] M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437.

[49] S. Song, S. Lichtenberg, and J. Xiao. 2015. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *Proceedings of CVPR.*

[50] D. G. Stork and M. E. Hennecke. 2013. *Speechreading by humans and machines: models, systems, and applications.* Springer.

[51] H. Su, H. Fan, and L. Guibas. 2017. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. In *Proc. CVPR.*

[52] TheStonefox. 2018. VRTK: Virtual Reality Toolkit. https://vrtoolkit.readme.io/. (2018).

[53] M. Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (2014), 189–195.

[54] L. van der Maaten. 2014. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* 15 (2014), 3221–3245.

[55] M. Vorländer. 2007. *Auralization: Fundamentals of Acoustics, Modelling, Simulations, Algorithms, and Acoustic Virtual Reality.* Berlin, Germany: Springer-Verlag.

[56] J. Xu, B. Stenger, T. Kerola, and T. Tung. 2017. Pano2CAD: Room Layout from a Single Panorama Image. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV).* 354–362.

[57] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. 2016. Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction without 3D Supervision. In *Proc. NIPS.*

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