# Grounding Acoustic Echoes in Single View Geometry Estimation

**Wajahat Hussain, Javier Civera, Luis Montano**
Robotics, Perception and Real Time Group, I3A
University of Zaragoza, Spain
{hussain, jcivera, montano}@unizar.es

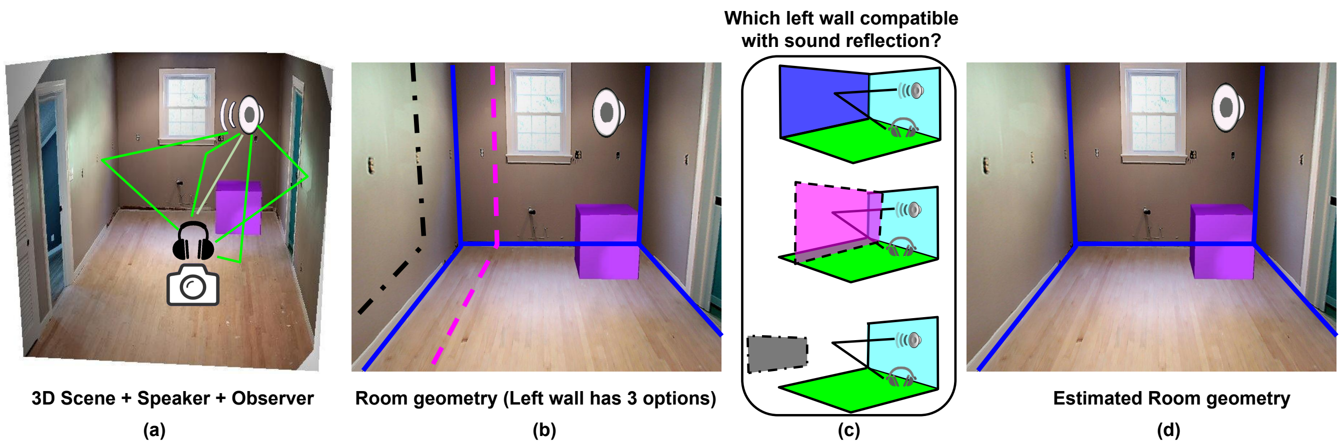


Figure 1: Acoustic echoes serve as a cue for single view geometry.

### Abstract

Extracting the 3D geometry plays an important part in scene understanding. Recently, robust visual descriptors are proposed for extracting the indoor scene layout from a passive agent's perspective, specifically from a single image. Their robustness is mainly due to modelling the physical interaction of the underlying room geometry with the objects and the humans present in the room. In this work we add the physical constraints coming from acoustic echoes, generated by an audio source, to this visual model. Our audio-visual 3D geometry descriptor improves over the state of the art in passive perception models as we show in our experiments.

## 1 Introduction

In order to interact with its surrounding environment, an agent needs first to understand it. Estimating the 3D geometry of the scene forms an important component of this scene understanding. Neverthless, the most studied and used methods for extracting such 3D scene models from visual data are based on the motion of the agent (Simultaneous Localization and Mapping –SLAM– and Structure from Motion –SfM– (Hartley and Zisserman 2000)). This forms a chicken-and-egg problem. Extracting the 3D geometry requires motion, i.e., interaction. And, in order to actively interact with the scene, one needs to understand it first.



Recently, robust learning-based visual descriptors have been proposed for extracting the 3D geometric layout of a scene from a passive agent's perspective, i.e., single image (Hoiem et al. 2009, Saxena et al. 2009). Figure 1d shows the estimated layout geometry of an indoor scene, consisting of the fundamental planes constituting the scene –walls, floor and ceiling. The fact that the majority of the scenes can be simplified into a few fundamental planes (Nedovic et al. 2010) is the motivation for such geometric models.

Furthermore, these data driven approaches can also handle complex scenes with major clutter and active humans. The reconstruction challenge in these complex scenes is two-fold. Firstly, the objects and the humans occlude the geometry. Secondly, a high degree of non-rigid elements –like humans– in the scene is still a challenge for the traditional multi-view geometry approaches. These learning-based approaches utilize the physical constraints offered by the objects and the humans to improve the room geometry. For example, no detected object can exist outside the room walls. These are termed as volumetric constraints (Lee et al. 2010). Similarly, a detected human pose, e.g., sitting, indicates a supporting surface, e.g., chair, occluded by the person. These are affordance constraints (Fouhey et al. 2012). Grounding these volumetric and affordance constraints in the room geometry estimation has shown exciting progress.

In addition to these physical constraints, 3D sound is an additional cue informing about room geometry (Dokmanić et al. 2013; Antonacci et al. 2012). For example, sound echoing in large halls is a common experience. This echoing phenomena exists even in smaller rooms, although not always human-perceivable. In this paper we add 3D sound as a cue for room layout estimation. Our research is motivated by several existing devices that present the combination of audio and visual sensors: mobile phones, laptops, and RGB-D sensors like Kinect.

Look at Figure 1a to see an example illustrating the benefits of our approach. The sound generated by the audio source travels different paths before reaching the listener. A few of these paths are shown as rays in Figure 1a. These paths include the direct path, paths with one bounce ($1^{st}$ order) and paths with more than one bounce (higher order). Copies of the same audio signal, or echoes, reach the listener at different times. The $1^{st}$ order echoes inform us about the location of the fundamental planes in the scene. Look at Figure 1b. Three candidates out of the possible left wall hypotheses are shown. By estimating the path travelled by the sound echo which reflected from the left wall, we can localize the left wall as shown in Figure 1c.

Utilizing acoustic echoes in this manner involves two challenges. Firstly, the $1^{st}$ order echoes have to be separated form the higher-order ones (Dokmanić et al. 2013). Secondly, each one of the echoes has to be associated with the correct wall. Only with the correct echo selection and labelling the 3D geometry of the scene can be estimated.

Our main contribution is then grounding these acoustic constraints in the structured prediction-based 3D geometry estimation techniques. Our model jointly estimates the 3D geometry of the scene, selects and labels the acoustic echoes. The input to our method is the single image and the estimated acoustic echoes. Through an extensive evaluation of this algorithm we show that the fusion of audio and visual cues outperform the estimation based on only images.

## 2 Background

The pioneering work for recovering the geometric layout from a single image was from Hoiem et al. (2007). Hoiem et al. divided the scene into 5 dominant planes –floor, sky, left, right and middle walls–; a valid model in most scenarios, both indoors and outdoors. Low-level visual cues such as color, texture and shade were used to train geometric classifiers. The abstract geometry provided by this method is accurate enough to improve the state of the art object detectors (Hoiem, Efros, and Hebert 2008). The grounding of physical rules –e.g., cars need to be supported by the floor below–, helped in removing false detections.

Indoor scenes are more structured than outdoor scenes. This structured nature of the indoor scenes, combined with the low-level visual cues, improved the geometry estimation (Hedau, Hoiem, and Forsyth 2009). Lee et al. (2010) introduced the physcial interaction of the room geometry with the detected objects (Lee et al. 2010). Humans existing and acting in the scene occlude the underlying geometry. Fouhey et al., (2012) transformed the detected human pose into an affordance cue.

3D sound is an addional cue that informs about the volume of the room. Larger halls sound different to smaller rooms. Recently, several algorithms have been proposed for inferring room geometrics from acoustic cues only (Dokmanić et al. 2013; Antonacci et al. 2012; Tervo and Tossavainen 2012). These algorithms do not include neither clutter nor humans in their model. The constraints generated by the objects, detected in the image, are seamlessly incorporated into our audio-visual model. Futhermore, visual data provides a prior on the scene shape which reduces the correspondence search space.

## 3 Overview

Our goal is to ground the physical constraints offered by the acoustic echoes in the passive perception visual model. An overview of the whole algorithm is shown in Figure 2. The inputs of our approach are the image (Figure 2b) and the relative arrival times of the echoed sound signals $\Delta t_i$ at the listener position $L$ (Figure 2c).

Firstly, the image data is used to generate plausible room geometry hypotheses. Indoor scenes usually follow the Manhattan world assumption, meaning that the dominant planes in the scene are aligned along one of three orthogonal directions. These orthogonal directions are given by the vanishing points, which are estimated from the lines in the image (Rother 2002). Given the vanishing points, multiple up-to-scale hypotheses for the room geometry are generated, as shown in Figure 2d. Given the camera height above the ground, the metric parameters of the ground plane can be estimated (Tsai et al. 2011). Having the metric reconstruction of the ground plane, the remaining planes of the room geometry can also be metrically reconstructed.

Learning-based techniques assign a *goodness* score to these hypotheses. Low-level visual cues based on texture and color, object volumetric cues and human affordance cues are used to calculate this score. The aim of the paper is to improve the ranking of the hypotheses by adding the acoustic constraints to the image information.

In a 3D scene, the audio signal generated by the source $S$ travels different paths before reaching the listener $L$ (Figure 1a). As shown in Figure 1c, the paths with one bounce ($1^{st}$ order) help in localizing the fundamental planes in the scene. In practice, we do not have these 3D path rays. What we have is the arrival times $\Delta t_i$ of the $i$ echoes travelling these $\geq$1 bounce paths (Figure 2c). For a known audio signal, these arrival times of echoes can be estimated reliably as shown by (Dokmanić et al. 2013). Solutions also exist where the sound signal is unknown (Gunther and Swindlehurst 1996). Given the source $S$ and the listener $L$ position, the relative arrival time $\Delta t_i$ constraints the layout plane to be tangent to a 3D ellipsoid (Figure 2e), as will be detailed in section 4. Look at Figure 2f. Ideally each plane of the correct hypothesis is the supporting plane of an ellipsoid, i.e., tangent to the ellipsoid in the absence of noise. In Figure 2f, the third hypothesis finds the best support as each of its plane is tangent to an ellipsoid. The remaining room hypotheses are penalized according to their acoustic support. The ellipsoids

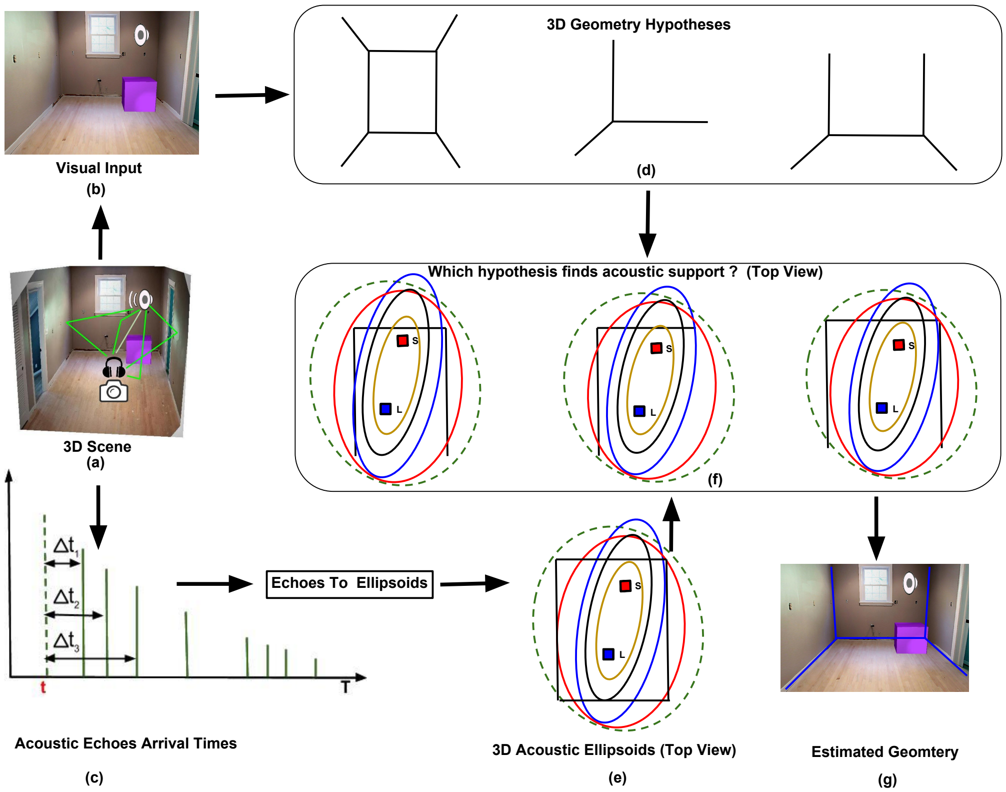


Figure 2: Our approach of grounding acoustic cues in indoor scene geometry estimation. **(e)** Each wall of the ground truth room is the support plane (tangent plane) of $1^{st}$ order echo ellipsoid. **(f)** Only the right most room hypothesis satisfies this constraint.

corresponding to higher order echoes, i.e., $> 1$ bounce paths, do not satisfy this tangent condition, e.g., dashed ellipsoid in Figure 2e. They act as the noise. Our proposal, fusing visual and acoustic data, is able to filter such noise.

## 4 Physically Grounded Scene Geometry

Given an image $I$, a set of room geometry hypotheses $\{r_1, r_2, ..., r_l\}$ is generated. Each room hypothesis $r$ is a set of planes $\{p_1, p_2, ..., p_X\}$, where $1 \leq X \leq 5$ (in a single perpective image at most 5 walls of the room are visible at a time). In the presence of objects or humans, the visual input $I$ also provides a set of detections $\{o_1, o_2, ..., o_M\}$. The acoustic echoes provide a set of 3D ellipsoids $\{e_1, e_2, ..., e_N\}$. We can represent our scene as an indicator vector $\mathbf{s} = (\mathbf{s}_r, \mathbf{s}_o, \mathbf{s}_e)$, where $\mathbf{s}_r = (s_r^1, s_r^2, ..., s_r^l)$, $\mathbf{s}_o = (s_o^1, s_o^2, ..., s_o^M)$, $\mathbf{s}_e = (s_e^1, s_e^2, ..., s_e^N)$. $s_\eta^i = 1$ if $\eta_i$ is the selected item, i.e., room hypothesis, object or ellipsoid, otherwise it is 0. We have to evaluate all the possible instances of the scene configuration $\mathbf{s}$ in order to find the valid one. Each instance of $\mathbf{s}$ contains one room geometry hypothesis, i.e., $\Sigma_i s_r^i = 1$. The selected room $r_i$ is tested for object containment and acoustic violations. Similar to Fouhey et al. (2012), we assume that all the object detections are correct, hence $\Sigma_i s_o^i = M$. Similar to Dokmanic et al. (2013) we assume loudness, meaning that the sound reaches the receiver $L$ after reflecting from all the fundamental planes in the room. Hence, $\Sigma_i s_e^i = X$, where $X$ is number of faces of the selected room geometry $r_i$. The total search space for the scene configuration $\mathbf{s}$ is $l \times X \times N \times 1$. There are $l$ room geometry hypotheses. Each room has X planes. Each plane is tested for tangency against N ellipsoids. All the object detections are valid and considered for each scene configuration.

Our evaluating function is $f(I, \mathbf{s})$ as given in equation 1.

$$f(I, \mathbf{s}) = \omega^T f_1(I, \mathbf{s}_r) + \alpha_o f_2(\mathbf{s}_r, \mathbf{s}_o) + \alpha_e f_3(\mathbf{s}_r, \mathbf{s}_e) \quad (1)$$

where $f_1$ measures the fit of the room geometry with respect to the low-level visual features, $f_2$ checks the compatibility of room geometry with the detected objects and humans, and $f_3$ penalizes the room hypothesis not finding support from acoustic echoes. $f_1$ and $f_2$ involve visual data, so their contribution can be summarized in a single function $f_v$. The acoustic contribution of $f_3$ can be summarized in a function $f_a$. $\omega$, $\alpha_o$ and $\alpha_e$ are training parameters.

$$f_v = \omega^T f_1 + \alpha_o f_2 \; , f_a = \alpha_e f_3$$

For each scene configuration $\mathbf{s}$, this function returns a score. The best scene configuration $\mathbf{s}^*$ is the one with the maximum score.

$$\mathbf{s}^* = \operatorname*{argmax}_{\mathbf{s}} f(I, \mathbf{s}) \quad (2)$$

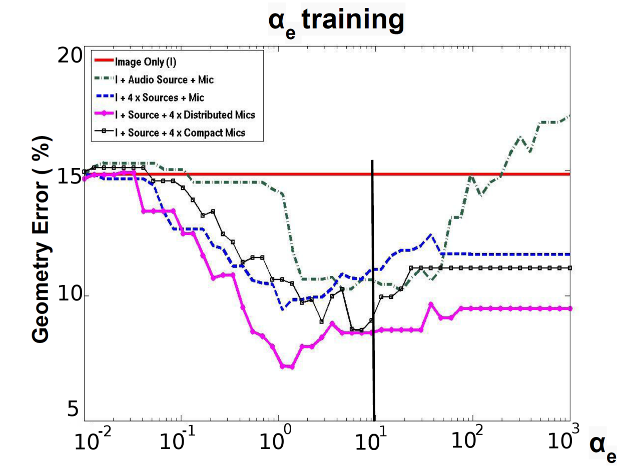


Figure 3: Room geometry labeling error against the acoustic training parameter $\alpha_e$ for several of the cases in our evaluation. Notice that the range of the parameter for which the audiovisual combination improves over only visual data – horizontal red line– is wide and comparable for all the cases. We chose a value of 10. All plots are generated with noisy echoes. Noisy sound source position (10 cm error) is used for last 3 options.

## Scoring Room Geometry with Visual Data

The visual scoring function is given in equation 3.

$$f_v = \omega^T \Psi(I, \mathbf{s}_r) + \alpha_o \phi(\mathbf{s}_r, \mathbf{s}_o) \tag{3}$$

where $\Psi(I, \mathbf{s}_r)$ is the feature vector corresponding to the room hypothesis $r$. Each room hypothesis $r$ is a set of planes, i.e., floor, middle wall, right wall, left wall and ceiling. For each plane, visual features are extracted from its image projected area. These features include color, shade, texture, total lines count, line count $\|$ to plane etc. The first term in equation 3 assigns a score to the room hypothesis $r$ using these low level features. $\phi(\mathbf{s}_r, \mathbf{s}_o)$ measures the compatibility of the room hypothesis $r$ with the detected objects in image space. The rooms not containing the entire object volume are penalized. Look at Figure 6b. The cuboid object is outside the walls of the incorrect room hypothesis (red geometry). $\omega^T$ and $\alpha_o$ are obtained using supervised structured learning. For details check Hedau et al. (2009) and Lee et al. (2010).

## Interaction of Room Geometry with Echoes

Our main contribution is the embedding of acoustic constraints in the visual geometry estimation model of the previous section. The acoustic data consists of the estimated arrival times of $1^{st}$ and higher order echoes. Look at Figure 2c. For each echo we have the time delay $\Delta t_i$ in its arrival after the first copy. Given the relative motion between the source $S$ and the listener $L$ position, the set of these delays $\{\Delta t_1, \Delta t_2, ..., \Delta t_N\}$ is converted into a set of 3D ellipsoids $\{e_1, e_2, ..., e_N\}$. For details on audio source $S$ localization and echo estimation from microphone signals see Antonacci et al. (2012) and Tervo et al. (2012).

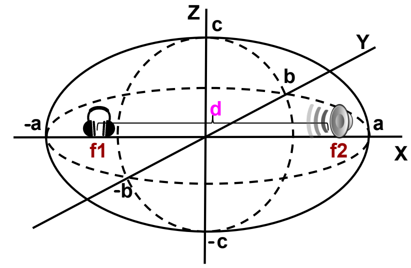


Figure 4: Converting acoustic echo information into a 3D ellipsoid. The Source $S$ and the Listener $L$ form the focal points of ellipsoid. The echo arrival delay time is used to estimate the length of major axis ($a$) and minor axes ($b, c$).

The 3D ellipsoid model has 5 parameters $(a, b, c, f_1, f_2)$ as shown in Figure 4. $f_1$ and $f_2$ are the focal points. $a$, $b$ and $c$ are the lengths of the major and minor axes respectively. These parameters are calculated using equations 4, 5 and 6.

$$t = \frac{d}{v} \tag{4}$$

where $t$ is the time in which the direct copy of the audio signal reached $L$ from $S$, $d$ is the distance between $S$ and $L$, $v = 343ms^{-1}$ is the speed of sound.

$$a_i = \frac{v(t + \Delta t_i)}{2}\ , b_i = \sqrt{a_i^2 - \frac{d}{2}^2}\ , c_i = b_i \tag{5}$$

$$f_2 = [\frac{d}{2}, 0, 0]\ , f_1 = -f_2 \tag{6}$$

The ellipsoids generated with equations 5 and 6 are in a local frame. They are transformed into the observer coordinate frame using two transformations $\mathbf{T}_m$ and $\mathbf{T}_c$. $\mathbf{T}_m$ is the motion from the local frame to the observer's microphone one. $\mathbf{T}_m$ is given by equations 7 to 9 .

$$\mathbf{T}_m = [\mathbf{R}_m, \mathbf{t}_m] \tag{7}$$
$$\mathbf{R}_m = \mathrm{AlignVectors}([1, 0, 0], [S - L]) \tag{8}$$
$$\mathbf{t}_m = \frac{S + L}{2} \tag{9}$$

where $\mathbf{R}_m$ is the 3x3 rotation matrix, $\mathbf{t}_m$ is the 3x1 translation vector. In local coordinates, the major axis of the ellipsoid is along X axis, i.e., [1,0,0] (Figure 4). In microphone coordinates, the major axis of the ellipsoid is along the axis pointing from $L$ to $S$ (**dir** $= [S - L]$). $\mathbf{R}_m$ aligns [1,0,0] with **dir**. $\mathbf{t}_m$ sets $S$ and $L$ as the focal points of ellipsoid instead of $[\frac{\pm d}{2}, 0, 0]$. $\mathbf{T}_c$ is the calibration between the observer's microphone and the camera (Legg and Bradley 2013).

Now that we have the set of ellipsoids $\mathbf{s}_e$, we can measure the acoustic support $f_a = \alpha_e \chi(\mathbf{s}_r, \mathbf{s}_e)$ for each room hypothesis using algorithm 1. The value of the acoustic weight is set $\alpha_e = 10$. Figure 3 shows the insensitivity of geometric labelling error to this parameter. Notice the logarithmic scale in the $\alpha_e$ axis and the wide range where the fusion improves over the image-only understanding.

**Algorithm 1** Acoustic Penalty Algorithm

INPUT: $\mathbf{s}_r$ {room hypotheses}, $\mathbf{s}_e$ {acoustic ellipsoids}, $vp$ {vanishing points}, $h$ {camera height} K {camera intrinsics}
OUTPUT: $\chi(\mathbf{s}_r, \mathbf{s}_e)$ {rooms acoustic penalty}, $\mathbf{s}'_e$ {selected ellipsoids}

$\mathbf{R} = [vp_x, vp_y, vp_z]$
**for** $i = 1$ to $l$ **do**
    $[\chi(s_r^i, \mathbf{s}_e), \mathbf{s}_e'^i] = \min_{\mathbf{R}} AcousticPenalty(s_r^i, \mathbf{s}_e, \mathbf{R}, h)$
**end for**

**FUNCTION** $AcousticPenalty(s_r^i, \mathbf{s}_e, \mathbf{R}, h)$
$\{p_1, p_2, ..., p_X\} = get\_room\_planes(r_i, \mathbf{R}, h, \mathrm{K})$ {See Tsai et al. (2011)}
$\mathbf{s}_e'^i = \emptyset$
**for** $j = 1$ to $X$ **do**
    $\{k_1, .., k_N\} = get\_\mathbf{s}_e\_support\_planes(\mathbf{s}_e, p_j)$ {See Figure 5 for support planes.}
    $\{u_1, .., u_N\} = get\_distance(\{k_1, .., k_N\}, p_j)$
    $u_* = minimum(\{u_1, u_2, ..., u_N\})$
    $d_j = u_*$
    $\mathbf{s}_e'^i = \mathbf{s}_e'^i \cup s_e^*$
    $\mathbf{s}_e = \mathbf{s}_e \setminus \mathbf{s}_e'^i$
**end for**
$\chi(s_r^i, \mathbf{s}_e) = \Sigma_a d_a$
**return** $\chi(s_r^i, \mathbf{s}_e)$, $\mathbf{s}_e'^i$

## Acoustic Penalty Algorithm

The single bounce echoes are reflected from the planes of the room. Therefore, all the planes of the correct room hypothesis must support (tangent to) the one bounce path rays (Figure 1c) and the corresponding ellipsoids (Figure 2f). In practice, due to noise in the estimated parameters, e.g., room orientation coming from vanishing points (vp), sound source localization etc., the planes are not exactly tangent. Look at Figure 5. Dashed ellipsoid corresponds to the echo reflected from the right wall. This rightwall-ellipsoid correspondence is performed by finding the closest ellipsoid. The right wall is moved so that it becomes tangent to the closest ellipsoid. The vertical dashed lines in Figure 5 show the amount of displacement of the right wall for each ellipsoid. The ellipsoid which requires minimum right wall displacement d is selected. This displacement is the acoustic penalty for the right wall, i.e., $d = u_5$ (Figure 5). Similarly, the penalties for the remaining planes of the room hypothesis are estimated. The cumulative penalty for a given room hypothesis is $\Sigma_a d_a$. Intuitively, this acoustic penalty should be less for the correct room hypthesis as compared to any random room hypothesis.

Experiments have shown that this wall-ellipsoid correspondence and the acoustic penalty computation is sensitive to the errors in the estimated parameters, e.g., room orientation, sound source localization (which affects the 3D ellipsoids parameters). We use particle optimization (Birge 2003) to overcome this problem. Starting with the intial room orientation $\mathbf{R}$, we generate multiple particles for room orientation within $\pm$ 5° along each axis. The ellipsoid correspondence and the acoustic penalty is computed for each orientation particle using the alogrithm 1. The orientation particle with minimum acoustic penalty at step $t$, is selected as initial seed for step $t+1$. The process is repeated until no significant change in penalty occurs. For a given orientation particle, a wrong wall-ellipsoid correspondence may reduce the penalty. However, it is unlikely that the same orientation generates low penalty, incorrect correspondences for the remaining walls of the room. Our experiments show that this algorithm can handle noisy vanishing points, noisy higher order echoes and the sound source localization error (up to 10 cm).

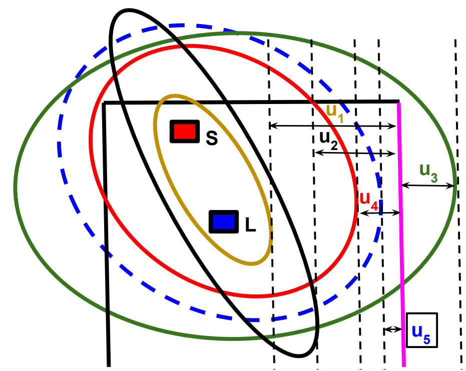


Figure 5: Solving right wall-ellipsoid correspondence. Step 1, move the rigth-wall to make it supporting plane for each ellipsoid (dashed lines). Step 2, select ellipsoid which requires minimum motion. Dashed ellipsoid is the right correspondence. $u_5$ is the penalty for this correspondence.

# 5 Experimental Evaluation

| | | Clean echoes | | Noisy Echoes | |
|---|---|---|---|---|---|
| | | No pert. | Pert. | No pert. | Pert. |
| Pixel Err. | I | 14.8 | 14.8 | 14.8 | 14.8 |
| | I+S | 17 | 12.4 | 14.5 | 10.6 |
| | I+O | 11.5 | 11.5 | 11.5 | 11.5 |
| | I+O+S | 17 | **8.7** | 14.3 | **9.1** |
| Label Err. | I | 65 | 65 | 68 | 68 |
| | I+S | 57 | 42 | 65 | 55 |
| | I+O | 60 | 60 | 67 | 67 |
| | I+O+S | 60 | **35** | 67 | **51** |

Table 1: Experimental results for the fusion of images and single sound source. *Pixel Err.* is % of incorrect pixel labels. *Label Err.* is the % echo-wall correspondence error. *I* is Hedau et al. (2009). *I+O* is Lee et al. (2010).

The input to our algorithm is a single image and the estimated acoustic echoes inside a 3D scene. There is no benchmark dataset available in this regard. In order to generate

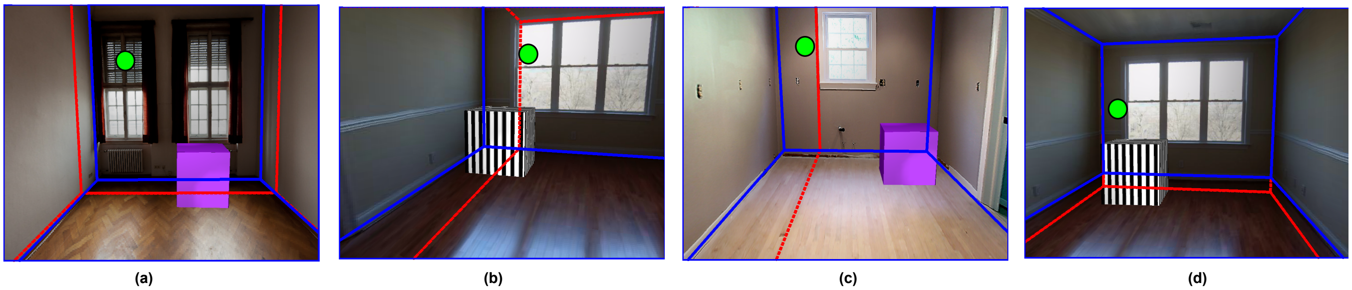


Figure 6: Qualitative Results. Image-only (Hedau, Hoiem, and Forsyth 2009) (red dashed line). Image + echoes (solid blue line). The sound source is shown as a green circle. Notice how the addition of acoustic contraints improves the wall boundaries.

the test data, we used the publicly available *GSound* system (Schissler and Manocha 2011). *GSound* is a realistic sound rendering system in a 3D environment. We used the 3D scene models to render sound as shown in Figure 1a. The 3D models contained objects (cuboids) in addition to sound source and the room geometry. Realistic rendering using *GSound* generated the observer perspective image (Figure 1b) and the room impulse response (RIR) at observer's position. The peaks in RIR represent the acoustic echoes in a 3D scene (Figure 2c). As Tervo et al. (2012), we have 27 echoes from each RIR. 21 out of these 27 echoes are higher-order noisy echoes. Our test set contains 17 scene renderings.

Table 1 shows the quantitative results of our evaluation of the fusion algorithm. For evaluating our geometry estimation, we use the standard metric of incorrect pixel label percentage (*Pixel Err.* in the tables), e.g., floor pixels labelled as middle wall or ceiling pixels labelled as left wall etc. We also provide the percentage of erroneous correspondences between the sound ellipsoids and the walls (*Label Err* in the tables). For both, the lower the better. The results are the average numbers over the total number of scenes in our dataset. In this table we report results for the cases of clean, $1^{st}$ order sound echoes, and noisy, $1^{st}$ + higher order echoes, separately. We also compare the results with and without the perturbation model from algorithm 1. For each of this situations, we evaluate 4 different algorithms for the data, shown as rows in the table: *I* standing for visual input only, *I+S* standing for visual and acoustic data, *I+O* standing for visual input only but with object information, and *I+O+S* standing for the fusion of visual data, objects and sound.

Our results show that the perturbation model in the algorithm 1 (*Pert.*) is essential to obtain accurate results from the sensor fusion. Notice the improvement when a perturbation model is applied in the rows where the sensor fusion is involved (*I+S* and *I+O+S*); when compared to a more naive estimation without such optimization (*No Pert.*). Secondly, with this perturbation model the fusion of the two modalities improves over the state of the art results using single image. The best results are boldfaced.

Figure 6 shows some qualitative results from our experiments. The green dot is the source position, the red dashed line stands for the single image layout results, and the solid blue line is the result after the inclusion of the audio cues. Notice the improvement, and the high accuracy of the blue layout estimation.

We have evaluated our algorithm in a large number of cases and configurations (look at Figure 7). Table 2 presents a summary of the results of this extensive evaluation.

Specifically, we evaluated the cases of multiple sources –second row in the table, as *4S*– and multiple microphones in two different configurations: in a wide-baseline configuration, distributed within the room *–4M (dist.)* in the table, $3^{rd}$ row–; and in a short-baseline configuration as in Kinect-like devices – *4M (short-b.)* in the table, $4^{th}$ row. It can be seen that, in general, the results of the cases with multiple sources and multiple receivers outperform the single-source single-receiver one.

For each of those configurations, we show in columns the results of clean $1^{st}$ order echoes (F) and noisy echoes, i.e., $1^{st}$ + higher order echoes (H). The third column, titled as *O+H*, shows the results when object reasoning is fused with the sound case. The fourth column, titled as *N+H*, summarizes our analysis of the performance of the algorithm with noise. Specifically, we added a noise of $10cm$ in the relative position of the source and the receiver. Notice that the performance of the algorithm is only slightly degraded. Finally, the last column, *–O+N+H–*, the combined case of the latest two columns. Notice that, even for this noisy case, our algorithm improves over the single image case by $4\%$.

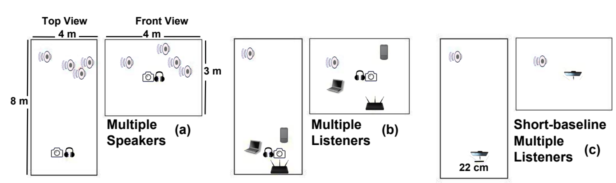


Figure 7: Schematic view of the geometric configuration of the multisource and multisensor experiments.

## 6 Conclusion & Future Work

In this paper we present a model that adds the information coming from acoustic echoes to passive perception visual models. Our proposal is based on the ranking of sev-

| | | F | H. | O+H | N+H | O+N+H |
|---|---|---|---|---|---|---|
| Pixel Err. | 1S+1M | 12.4 | 10.6 | 9.1 | – | – |
| | 4S | 5.9 | 8.8 | 8.7 | 11 | 10.4 |
| | 4M (dist.) | 5.8 | 7.6 | 7.6 | 8.6 | 8.6 |
| | 4M (short-b) | 11.9 | 11.7 | 7.8 | 9 | 8.4 |
| Label Err. | 1S+1M | 42 | 56 | 51 | – | – |
| | 4S | 5 | 21 | 21 | – | – |
| | 4M (dist.) | 9 | 19 | 19 | – | – |
| | 4M (short-b) | 19 | 20 | 19 | – | – |

Table 2: Summary of the evaluation in different cases. S: sound source, M: mic, F: $1^{st}$ order echoes, H: $1^{st}$ + higher order echoes, O: object, N: noise of 10 cm in sound source position. *Pixel Err.* is % of incorrect pixel labels. *Label Err.* is the % echo-wall correspondence error.

eral room hypotheses. The scoring function is the weighted sum of an image-based and a sound-based term. We have performed an extensive evaluation of the algorithm in several cases: single-sensor single-source, multi-sensor single-source, single-sensor multi-source, and erroneous source position. The results show that the combination of the two inputs consistently outperforms the results of the state-of-the-art vision-only approaches.

Up to our knowledge, this paper is the first one addressing scene understanding from audiovisual cues. Future work includes extending this model to absorbent scenes where the walls absorb the sound. There, the visual texture of these absorbing walls can act as a cue.

We believe that the fusion of acoustic and visual data can lead to a wide array of long-term research lines, including

- **Overcoming Range Limitations of Depth Sensors** Active sensors like Kinect are usually limited in the range they can measure. Our audiovisual model, making use of *RGB* data, does not have such range limitation and hence could be used in large rooms without degradation.
- **Audiovisual Cocktail Party Problem** The geometric contraints provided by our method can be used to separate multiple speakers which are active simultaneously. This would improve automatic speech recognition in crowded environments.
- **Audio Augmentation** Our model can be extended for augmenting a scene with the audio, e.g., talking character in augmented reality games etc., or removing the audio-visual print (diminished reality) in case of phobia or dislike, e.g., removing the image and the sound of a barking pet dog etc.

## 7 Acknowledgements

The research here was funded by the project DPI2012-32168 from the Spanish government (Ministerio de Economía y Competitividad).

## References

Antonacci, F.; Filos, J.; Thomas, M. R. P.; Habets, E. A. P.; Sarti, A.; Naylor, P. A.; and Tubaro, S. 2012. Inference of Room Geometry From Acoustic Impulse Responses. *Audio, Speech, and Language Processing, IEEE Transactions on* 20(10):2683–2695.

Birge, B. 2003. PSOt-a particle swarm optimization toolbox for use with matlab. In *Proceedings of the 2003 IEEE Swarm Intelligence Symposium (SIS'03)*, 182–186. IEEE.

Dokmanić, I.; Parhizkar, R.; Walther, A.; Lu, Y. M.; and Vetterli, M. 2013. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*.

Fouhey, D. F.; Delaitre, V.; Gupta, A.; Efros, A. A.; Laptev, I.; and Sivic, J. 2012. People watching: Human actions as a cue for single view geometry. In *Computer Vision–ECCV 2012*. Springer. 732–745.

Gunther, J., and Swindlehurst, A. 1996. Algorithms for blind equalization with multiple antennas based on frequency domain subspaces. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, volume 5, 2419–2422.

Hartley, R., and Zisserman, A. 2000. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press.

Hedau, V.; Hoiem, D.; and Forsyth, D. 2009. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, 1849–1856. IEEE.

Hoiem, D.; Efros, A. A.; and Hebert, M. 2007. Recovering surface layout from an image. *International Journal of Computer Vision* 75(1):151–172.

Hoiem, D.; Efros, A. A.; and Hebert, M. 2008. Putting objects in perspective. *International Journal of Computer Vision* 80(1):3–15.

Lee, D. C.; Gupta, A.; Hebert, M.; and Kanade, T. 2010. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, volume 1, 3.

Legg, M., and Bradley, S. 2013. A combined microphone and camera calibration technique with application to acoustic imaging. *IEEE transactions on image processing* 22(10):4028–4039.

Nedovic, V.; Smeulders, A. W.; Redert, A.; and Geusebroek, J.-M. 2010. Stages as models of scene geometry. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32(9):1673–1687.

Rother, C. 2002. A new approach to vanishing point detection in architectural environments. *Image and Vision Computing* 20(9):647–655.

Saxena, A.; Sun, M.; and Ng, A. Y. 2009. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(5):824–840.

Schissler, C., and Manocha, D. 2011. Gsound: Interactive sound propagation for games. In *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society.

Tervo, S., and Tossavainen, T. 2012. 3d room geometry estimation from measured impulse responses. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 513–516. IEEE.

Tsai, G.; Xu, C.; Liu, J.; and Kuipers, B. 2011. Real-time indoor scene understanding using bayesian filtering with motion cues. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 121–128. IEEE.