3D Reconstruction in the Presence of Glasses by Acoustic and Stereo Fusion

http://vis.uky.edu/~gravity/Research/audiofusion

Mao Ye
Univ. of Kentucky
Lexington, USA
mao.ye@uky.edu

Yu Zhang
Nanjing Univ.
Nanjing, China
zhangyu606@gmail.com

Ruigang Yang
Univ. of Kentucky
Lexington, USA
r.yang@uky.edu

Dinesh Manocha
Univ. of North Carolina
Chapel Hill, NC, USA
dm@cs.unc.edu

Abstract

We present a practical and inexpensive method to reconstruct 3D scenes that include piece-wise planar transparent objects. Our work is motivated by the need for automatically generating 3D models of interior scenes, in which glass structures are common. These large structures are often invisible to cameras or even our human visual system. Existing 3D reconstruction methods for transparent objects are usually not applicable in such a room-size reconstruction setting. Our approach augments a regular depth camera (e.g., the Microsoft Kinect camera) with a single ultrasonic sensor, which is able to measure distance to any objects, including transparent surfaces. We present a novel sensor fusion algorithm that first segments the depth map into different categories such as opaque/transparent/infinity (e.g., too far to measure) and then updates the depth map based on the segmentation outcome. Our current hardware setup can generate only one additional point measurement per frame, yet our fusion algorithm is able to generate satisfactory reconstruction results based on our probabilistic model. We highlight the performance in many challenging indoor benchmarks.

1. Introduction

Imaging and reconstructing the 3D structure of the scene have been an active area of computer vision research for the past decades. Given the rapid development of range sensors such as ToF (time-of-flight) cameras, laser scanners, and stereo cameras, 3D point clouds with sufficient accuracy can be easily generated for 3D reconstruction. On the other hand, many indoor scenes consist of transparent and refractive objects, which can cause severe artifacts in reconstructed results. Overall, 3D reconstruction of non-diffuse objects remains a challenging problem in the field.

Some techniques have been proposed that can generate impressive results using specialized setup, e.g., laser plus thermal cameras [5], or fluorescent liquid [9]. To the best of our knowledge, however, the reconstruction of building interiors in the presence of large glasses structure has not been addressed. Glasses are commonly used in building interiors and are a key part of modern architectures. Many of these glasses structures are completely transparent. Furthermore, it is not uncommon to have buildings with big or many glass structures, and current methods designed for small objects may not work well in such cases.

We present a simple and inexpensive setup in which an ultrasonic range sensor is added to a Microsoft Kinect camera. The ultrasonic sensor computes a valid depth value even for completely transparent glasses. A user can sweep the camera around to scan the scene of interests, in a fashion similar to KinectFusion [13]. Given the multiple ultrasonic sensor reads, which are registered in the same coordinate frame as the depth readings, we have developed a novel sensor fusion algorithm to combine the sparse range values from the ultrasonic sensor with the depth map based on stereo vision. The main challenge in this fusion algorithm is that the ultrasonic sensor readings tend to be sparse and unevenly distributed, as compared to the depth maps. Assuming piece-wise planar transparent objects, we formulate this fusion problem as a segmentation/labeling problem followed by depth reconstruction. More specifically we define a Bayesian network to optimally infer whether a pixel should be assigned to the depth value generated by the stereo matching, one of the fitted planes from the ultrasonic sensor readings, or infinity (unknown). Given the labeled pixels we then update the depth map to reconstruct the entire scene, including transparent objects.

Acoustic sensors have been used for a few scene reconstruction and recognition applications in different domains. This includes considerable work for 3D scene reconstruction [10], 3D acoustic images for object recognition [21], and augmented scene modeling and visualization of underwater scenes [6]. These environments are quite different from our current work. In general, sound waves and their propagation paths vary considerably based on the frequency. For example, low frequency waves exhibit diffraction and scattering effects. However, ultrasound waves are high frequency waves and at those frequencies it is reason-

able to approximate their paths using geometric acoustics [15, 18]. In this case, the sound propagation paths can be modeled based on ray theory as the sound wavelengths are significantly smaller than the size of the obstacles, and the resulting specular effects can be used to measure the depth values from transparent glasses.

Such an acoustic plus visual sensing scheme has also been explored in the robotics. For example, Yang et al. [26] combine an ultrasonic sensor with a line laser scanner to detect mirrors and glasses for robotic navigation and obstacle avoidance. However, their problem domain is 2D (e.g., a floor plan) and the resulting samples from the range sensor and the laser scanners are dense. In this paper, we instead aim to reconstruct the full 3D model, with rather sparse acoustic range readings.

2. Related Work

Ihrke et al. present an extensive survey concerning the state-of-the-art methods aimed at reconstruction of transparent and refractive objects or phenomena [12]. Interested readers are also referred to an excellent tutorial on reconstructing invisible surfaces at CVPR 2013 [27]. In this paper, we divide existing methods into three categories based on means of acquisition and briefly discuss each of them.

Physical Manipulation It means the target object for 3D reconstruction are physically changed and manipulated. Goesele et al. cover objects of interest with diffuse coatings before recovering the object surface and rendering [7]. Alternatively, color dye [11] or fluorescent liquids [9] can be mixed with the target objects, which enables generating a direct sample of the object geometry without generating a 3D virtual model. These methods can generate impressive results, but the underlying physical manipulation is time consuming and difficult for large scenes.

Active Illumination Active illumination methods, such as structured light and coded illumination, have been widely investigated. For instance, Ma et al. [17] used coded projection and two cameras to reconstruct transparent objects. These approaches require dedicated setups that are difficult to scale up. Eren developed a novel approach called Scanning-From-Heating [5]. In this setup, a laser beam is shot on the surface of an object and a thermal image is captured based on which the 3D position of the surface point can be triangulated. This method can deal with completely transparent objects, but it is probably too slow to scan any large scale objects.

Passive Methods Passive methods can generate 3D reconstructions directly from captured images, without any physical interaction or active structured illumination on the scene. For example, mirror shapes can be reconstructed by observing the distortions of known patterns (e.g., [8]. Similarly a transparent object can be recovered by observing the disparity in the background through multiple view angles (e.g., [2]), provided that the objects can refract light significantly. If the object shape is known approximately a priori, or partially recovered, some techniques have been proposed to detect non-diffuse objects and fully reconstruct them (e.g., [16, 23]). However, none of these assumptions holds for thin glasses in large indoor scenes.

Our work is also related to **sensor fusion**, which has been actively studied in both robotics and computer vision literature. In the area of 3D reconstruction, photometric stereo and stereo have long been combined to generate highly-detailed metric surface maps (e.g., [4, 19, 28]). Recently, time-of-flight (ToF) sensors and stereo sensors have been used to obtain high accuracy depth maps as the error characteristics of ToF sensor are complementary to passive stereo (e.g., [29]). Fusion can also be performed using multiple samples from the same sensor. Recent methods in this category include [24] and the well known KinectFusion system [13]. Unlike prior multi-modal fusion techniques, the ultrasonic range sensor in our setup provides rather sparse data point sets, as compared to dense depth maps generated using a depth sensor. We therefore developed a novel sensor fusion algorithm to deal with our unique setup.

3. Overview

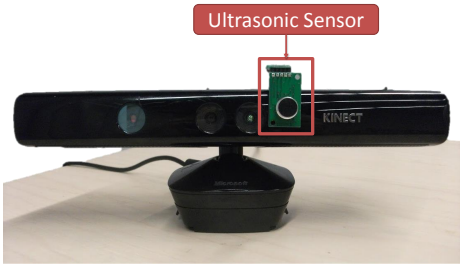


Figure 1. Our modified Kinect camera has an additional ultrasonic range sensor (e.g., a sonar), which can measure distance to any objects, including transparent ones.

In Figure 1, we show our modifications to the Kinect sensor: an ultrasonic range sensor, which can measure transparent objects, is attached. We choose this setup vs a more sophisticated senor array mainly due to its simplicity. The range sensor [1] is operating at 235KHz, with a narrow beam width of just 15 degree. At this high frequency and the narrow beam, the sound wave propagation can be approximated as rays. The sensor also has a USB connector for tethered data collection and control. We mechanically calibrate the range sensor by aligning it to the IR camera as close as possible. In practice, the range sensor has a small field of view, so precise co-axial alignment is not necessary. In our setup we simply assume that the range measurement corresponds to the middle of the depth image. We have developed algorithms to simultaneously capture the Kinect depth map and range sensor readings. During data acquisition, we treat the hybrid camera as a 3D hand-held scanner

to capture the scene of interest. The depth maps are first processed by the Kinect Fusion system [13] to obtain both the scene model and the camera poses. With the camera poses, we can transform the range readings from the ultrasonic sensor to 3D points in the same coordinate system as the reconstructed scene.

We assume the target transparent objects are piece-wise planar. We first fit multiple planes to the data collected from ultrasonic sensors using RANSAC. One simple way to reconstruct the scene geometry is to directly use the fitted surface to replace the empty space. A slightly more advanced method is to use Poisson blending [22] to interpolate a smooth surface that uses the ultrasonic data points as initial seeds. However, both of these naive solutions can generate a lot error, especially in regions where non-transparent objects, behind a transparent object, are captured by the depth sensor. The main reason is the lack of knowledge of the exact location and span of the transparent objects. Therefore, we initially perform a segmentation step in which each pixel is labeled to a certain category. This step is followed by a depth reconstruction procedure based on Poisson blending [22]. Instead of performing segmentation in 3D space, we formulate the segmentation on the 2D image. In other words, we assume that the input to our algorithm includes a depth map (notice that the depth map can be a synthesized one from the KinectFusion result, instead of the raw depth) and a set of planes fitted from the 3D points acquired by the ultrasonic sensor.

4. Segmentation

During the segmentation step, each pixel is labeled as one of the categories in our candidate set $\mathcal{C} = \{\infty, \zeta, \pi_k | k = 1, \cdots K\}$ with $K + 2$ elements. The ∞ label means empty space where no data has been observed from both sensors. The ζ label means that the first point hit along line of sight is not from a transparent object and the depth data can be observed. By contrast, each π_k label defines the pixel label from our fitted surface planes π_k of the transparent objects. It should be emphasized that there is no information about the boundary of these planes a priori.

4.1. Probabilistic Model

We define the Bayesian network in Figure 2 to describe our labeling process. Here $L_i^t \in \mathcal{C}$ denotes which category that node i^t (defined as pixel i at frame t) belongs to and is our target. Our observations from the kinect sensor and the ultrasonic sensor are represented as Z_i^t and S_i^t, respectively. The data captured from depth sensor does not solely depend on labeling, and is also affected by occlusion. Specifically, if a non-transparent object resides behind a transparent one, the depth sensor will very likely capture the depth values corresponding to the non-transparent object, while the label of the corresponding pixel should be that of the transparent

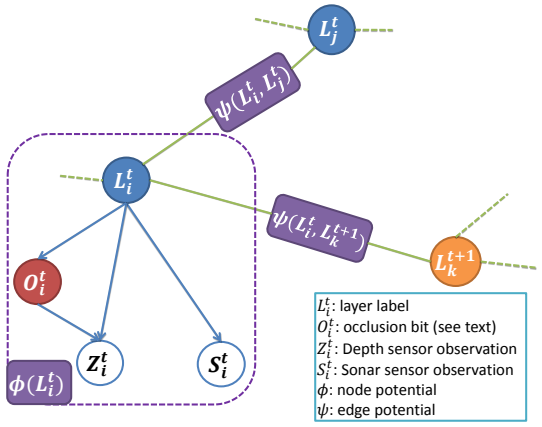


Figure 2. The graphical model shows the conditional dependence structure between random variables and the smoothing term on the right side defines the constrains between neighboring labels while the data term on the left side describes the sensor measurement process.

object. Therefore, in our model, we use a hidden binary variable O_i^t to explicitly model this phenomena, which takes value 1 if the pixel falls into this situation and 0 otherwise. Our Bayesian Network also takes into consideration both the spatial connectivity $\psi(L_i^t, L_j^t)$ for pair (i, j) at frame t and temporal consistency $\psi(L_i^t, L_k^{t+1})$ between node i^t and its correspondence i^{t+1} in the next frame.

Based on this graphical model, the node potential $\phi(L_i^t)$ can be expressed as:

$$\phi(L_i^t) = P(L_i^t) \cdot P(Z_i^t | L_i^t) \cdot P(S_i^t | L_i^t). \quad (1)$$

The introduction of the hidden variable O makes it more intuitive to model the probability $P(Z|L)$:

$$\begin{aligned} P(Z_i^t | L_i^t) = & P(Z_i^t | O_i^t = 0, L_i^t) \cdot P(O_i^t = 0 | L_i^t) \\ & + P(Z_i^t | O_i^t = 1, L_i^t) \cdot P(O_i^t = 1 | L_i^t). \end{aligned} \quad (2)$$

The labeling problem is cast as a MAP (Maximum a posteriori) problem that minimizes the following energy function:

$$E = -\sum_t \sum_i \log(\phi(L_i^t)) - \sum_{<i,j,f,g>} \log(\psi(L_i^f, L_j^g)), \quad (3)$$

where the quadruple $< i, j, f, g >$ defines a pair of pixels (i, j) that are either spatially ($f = g$) or temporally ($f \neq g$) connected and forms an edge in the graph as shown in Figure 2. The first term (data cost) and the second term (smoothness cost) are described in detail in Section 4.2 and Section 4.3, respectively. With these terms defined, we use the Graph Cuts algorithm [3] to solve this labeling problem.

4.2. Data Terms

The probabilities in Eq. 1 and Eq. 2 are formulated according to the characteristic of depth and ultrasonic sensors.

In the following, we will define the terms in Eq. 1 from the right to the left.

The probability $P(S_i^t|L_i^t)$ relates the ultrasonic sensor measurement to the scene geometry. Since the ultrasonic sensor can correctly measure the scene depth, we can model this probability as a Gaussian when the ultrasonic measurement is not zero. However, due to the sparsity of ultrasonic measurements, a majority of pixels will have no corresponding depth value computed by ultrasonic sensor. In this case, we simply acknowledge such a situation by assigning a high probability value c_h. In contrast, it is unlikely that the ultrasonic sensor will provide any measurements for empty space, so a small constant value c_l is used for $P(S_i^t = d|L_i^t = \infty)$. Table 1 provides our formulation for $P(S_i^t|L_i^t)$. Here z_i^t is the value captured by the depth sensor. Therefore, $P(S_i^t = d|L_i^t = \zeta) = N(d|z_i^t, \sigma_s^2)$ measures the discrepancy of ultrasonic data with respect to the depth sensor observation with a Gaussian. Similar $P(S_i^t = d|L_i^t = \pi_k) = N(d|T(\pi_k, i, k), \sigma_s^2)$ measures the discrepancy of ultrasonic data with the value $T(\pi_k, i, t)$ which is the intersection of fitted surface π_k with the ray from camera center to pixel i^f. The variance in both terms σ_s^2 encodes the ultrasonic sensor noises, which can be approximated as Gaussian noises [25]. The value of σ_s^2 can either be given by the manufacturer or evaluated through residue computation in surface fitting.

The probability $P(Z_i^t|L_i^t)$ computation is slightly more complicated, as shown in Eq. 2. The first term $P(Z_i^t|O, L_i^t)$ can be defined with similar consideration as above, as shown in Table 2. The terms $P(Z_i^t|O_i^t = 1, L_i^t \in \{\infty, \zeta\})$ are undefined, because the two conditions contradict with each other. Namely, transparent objects occluding non-transparent ones ($O_i^t = 1$) means that the label can only be one of the fitted surfaces ($L_i^t = \pi_k$ for some k), instead of $\{\infty, \zeta\}$. Consequently, the corresponding prior $P(O_i^t = 1|L_i^t \in \{\infty, \zeta\})$ will be zero and cancels out these terms. One term that might seem counter-intuitive at the first glance is $P(Z_i^t = d|O_i^t = 1, L_i^t = \pi_k) = N(d|z_i^t, \sigma_k^2)$.

S_i^t \ L_i^t	∞	ζ	π_k
0	c_h	c_h	c_h
d	c_l	$N(d\|z_i^t, \sigma_s^2)$	$N(d\|T(\pi_k, i, \mathrm{t}), \sigma_s^2)$

Table 1. The probabilistic modeling of ultrasonic measurements, namely $P(S_i^t|L_i^t)$

.

The prior $P(O_i^t|L_i^t)$ represents the probability of certain non-transparent objects residing behind the transparent foreground. As mentioned above, the invalid terms are set to zero, which leads to the alternatives being one: $P(O_i^t = 0|L_i^t \in \{\infty, \zeta\}) = 1$. Currently all other prior terms are uniformly set as $\frac{1}{2}$ based on parameter tuning. Advance image analysis techniques can possibly be used to infer per-pixel priors, which we consider as a future work.

L_i^t	∞		ζ		π_k	
Z_i^t \ O_i^t	0	1	0	1	0	1
0	c_h	–	c_l	–	c_h	c_l
d	c_l	–	$N(d\|z_i^t, \sigma_k^2)$	–	$N(d\|T(\pi_k, i, \mathrm{t}), \sigma_k^2)$	$N(d\|z_i^t, \sigma_k^2)$

Table 2. The probabilistic modeling of depth sensor measurements, namely $P(Z_i^t|O_i^t, L_i^t)$

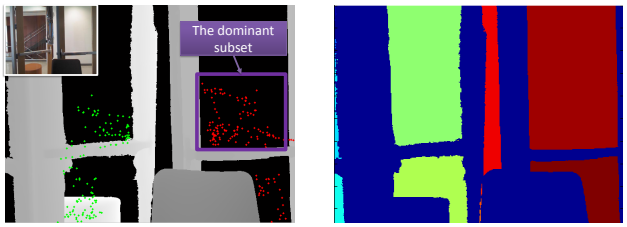
The label prior $P(L_i^t)$ encodes the belief of a pixel taking a certain label before performing any inference through our Bayesian network. A non-information prior can be used, similar to the other prior above. However, one major limitation with this formulation is lack of knowledge about the empty space, as no measurement can be used to favor or oppose a pixel being labeled as empty space. Consequently, the zero region (defined as the union of pixels with zero depth value in the depth image) might tend to be labeled as one of the surfaces instead of the empty space. In practice, we notice that most of transparent surfaces in our daily environments are bounded by certain non-transparent objects, for example windows are usually bounded by frames. If a certain region actually corresponds to the empty space, no single ultrasonic measurement will be captured in this region. The idea is to favor empty space labels in such regions, while favoring some particular surfaces if there are ultrasonic measurement associated to this surface fall into this region.

Towards this end, we segment the zero region into individual connected components $\{Q_j^t\}$ as illustrated in Figure 3(b) and identify which label is most likely to be dominant in each component. For each component, we locate the set of Ultrasonic points whose projections are inside this region, as marked by the purple rectangle in Figure 3(a). If there are sufficient number of points inside the region (e.g. more than 10 points), we can identify the dominant subset that is associated to a single surface, and can be classified as the dominant label. In the example shown in Figure 3(a), all points inside the rectangle are associated to the same surface, therefore the subset is the same as the entire set. We denote the label of this dominant surface as l_j^t for component Q_j^t. Distance transform is applied to each pixel in Q_j^t using this subset of ultrasonic points as seeds to calculate the affinity score $\alpha_i^t, i \in Q_j^t$. Next, a prior is computed for the pixels inside the region ($i \in Q_j^t$) as follows to favor the dominant surface:

$$p_i^t = \max\left\{0.5, \exp\left(-\alpha_i^t / \max_i\{\alpha_i^t\}\right)\right\}, \quad (4)$$

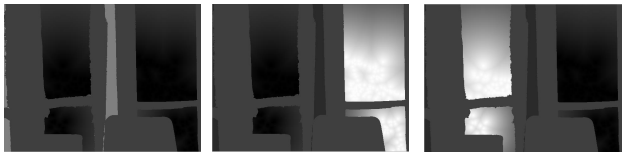
$$P(L_i^t = l) = \begin{cases} p_i^t & \text{if } l = \pi_{l_j^t} \\ (1 - p_i^t)/(K+1) & \text{otherwise} \end{cases}. \quad (5)$$

For the component without sufficient ultrasonic measurement support, our prior model favors the ∞ over other la-



(a) Depth map with projected sonar points

(b) Connected components generated from zero region



(c) The prior of label being ∞ (left), π_1(middle) and π_2 (right)

(d) Labeling result with non-informative prior

(e) Labeling result with our prior model

Figure 3. An example demonstrating our prior model. In this example π_1 is the surface fitted from the entire red points in (a) and π_2 is the surface fitted from the entire green points in (a).

bels by setting:

$$P(L_i^t = \infty) = \begin{cases} c_\infty & \text{if } l = \infty \\ (1 - c_\infty)/(K+1) & \text{otherwise} \end{cases}. \quad (6)$$

where c_∞ is set to 0.5 in our experiments. For the region with positive depth values, we simply use uniform priors $1/(K+2)$. Figure 3(c) illustrates the per pixel prior of the two surfaces in this example. The advantage of using our prior model over the non-informative prior is shown in Figure 3(d-e). Throughout this paper, we represent our labeling result using different colors. Black indicates the ∞ and gray corresponds to ζ. Other colors indicates one of the fitted surfaces π_k.

4.3. Smoothness Term

The smoothness term ψ in Equation 3 considers both the spatial and temporal consistency. The smoothness cost $E_s(L_i^f, L_j^g) = -\log \psi(L_i^f, L_j^g)$ for a identified pair is defined as:

$$E_s(L_i^f, L_j^g) = \begin{cases} \omega(i,j,f) \cdot D(L_i^f, L_j^f) \cdot c_s & \text{if } f = g \\ \delta(L_i^f \neq L_j^g) \cdot c_s & \text{if } f \neq g \end{cases}. \quad (7)$$

where c_s leverages the relatively importance of this smoothness term over the data term and $\delta(\cdot)$ is an indicator function. The terms $\omega(i,j,f)$ and $D(L_i^f, L_j^f)$ are defined in Equation 8 and Equation 9, respectively.

In the spatial domain, 4-connectivity is used to associate pixels and each frame forms a regular grid sub-graph. The weight between a spatial pair $\omega(i,j,f)$ is determined purely based on the depth information, because the captured color information in transparent region is a mixture of reflection and the scene behind the transparent object. The weight represents the degree of inherent similarity between these two nodes and is defined as follows.

$$\omega(i,j,f) = \max\left\{m_\omega, \exp\left(-\frac{\|z_i^f - z_j^f\|^2}{\sigma_d^2}\right)\right\}, \quad (8)$$

where $m_\omega = 0.01$ servers as safeguard to ensure minimum smoothness and σ_d^2 models the degree of scene discontinuity that could be adapted according to the scene. The function $D(L_i^f, L_j^f)$ measures the label assignment cost for these two nodes and is also based on depth values:

$$D(L_i^f, L_j^f) = \delta(L_i^f \neq L_j^f) \cdot \left| v(L_i^f) - v(L_j^f) \right|, \quad (9)$$

$$\text{where} \quad v(L_i^f) = \begin{cases} 0, & \text{if } L_i^f = \infty \\ z_i^f, & \text{if } L_i^f = \zeta \\ T(\pi_k, i, f), & \text{if } L_i^f = \pi_k \end{cases}. \quad (10)$$

In the temporal domain, we use the relative transformation between frames to locate correspondences and rely on visibility test to remove the outliers. Specifically, the pixel i^f, whose depth value is positive, is transformed to the space of another frame g and compared with the target value in the depth map at frame g. If the pixel is outside the image space or the target value is zero, or the projected depth value is larger than the target value over a certain threshold (5 cm in our current settings), then no correspondence is associated. Otherwise, we build an edge between the two nodes in our graph and assign the cost according to Equation 7.

5. Scene Reconstruction

The second step of our framework is depth reconstruction for the transparent objects based on the labeling information inferred in the previous stage. Pixels that are labeled as one of the transparent surfaces require the depth values being re-estimated. For each of these pixels, we calculate an initial value by casting a ray from camera center through the pixel and intersecting with the target surface. In order to obtain smooth reconstruction, we adopt the Poisson blending technique [22] to refine the estimation. Due to the severe noises at the object boundary in the captured depth maps, we further include boundary pixels into the blending region, as shown in Figure 4. We also need to take into account

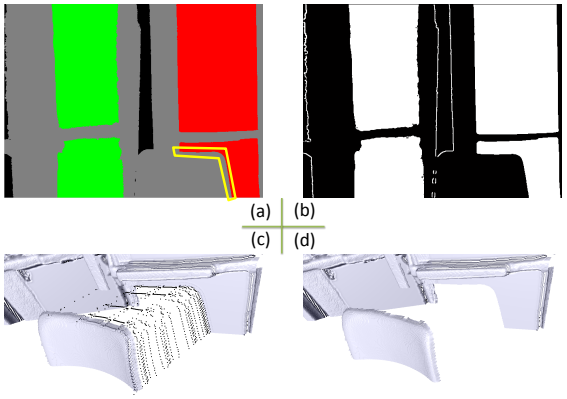


Figure 4. A basic mask for Poisson blending can be obtained from our labeling result in (a). In this case, the mask will contain the red and the green region. To reduce the noises in the object boundary caused by the depth sensor, we extend this mask by dilating the mask for 5 pixels. The boundary along empty space is also included and the resulting mask is shown in (b). (c) and (d), which demonstrates the necessity of adopting the adaptive differential operators during Poisson blending.

the non-uniform smoothness across the entire scene. Take Figure 4(a) as an example, enforcing smoothness along the object boundary inside the yellow mask will produce points along the boundary that is floating between two objects. To address this issue, we incorporate the adaptive differential operators described in [28] when constructing the Laplacian operator. The adaptive weights are calculated based on the depth map filled with the intersection values. Figure 4 shows the effectiveness of the adaptive weighting.

6. Experiments

We validate our framework on real world scenarios with various complexities as shown in Figure 6 and Figure 7. The first case consists of an opaque object (flower) behind a single piece of glass and is the simplest scenario in our benchmarks. In the second scene, the opaque object behind the glass crosses the boundary between two glass pane and is more challenging. The third one goes a step further with an extra foreground object (the chair) occluding part of the glasses. The fourth case is a fish tank where the four sides are all transparent. Column (b) shows the original scan from KinectFusion [20], which completely misses the transparent objects, as expected. Our approach successfully reconstructs all these scenes by taking into account the ultrasonic measurements. For the first cases, the labeling was performed simultaneously on five frames, and the reconstructed meshes shown in Figure 6, are generated from these five frames. Due to space consideration, only two frames with distinctive views are shown per case in column (a). For the last case, the KinectFusion can easily get lost when trying to perform a surrounding scan. Due to this limitation, in this particular dataset, we only use one single frame from the front view. In the case of five frames, the entire labeling procedure takes less than 15 seconds and the depth reconstruction step takes less than 5 seconds. The settings of the parameters used in these experiments are shown in Table 3 (the depth and ultrasonic measurements are in the unit of meter).

c_l	c_h	σ_s^2	σ_d^2	c_s
0.1	0.8	4e-4	1e-4	3000

Table 3. The parameters used in our experiments.

Notice the opaque objects behind the transparent foreground have been removed during our scene reconstruction step (Section 5) and are replaced by the target transparent objects. If those objects need to be kept in the final result, as shown in Figure 6, the following procedure should be applied. First we need to identify depth pixels in the zero region of the depth map, which is the region labeled as one of the transparent surfaces. Then these pixels can be back-projected to 3D space to recover the geometry of the opaque objects that are behind the transparent foreground. Finally the new geometry is merged with the geometry from the fused depth map.

In the first case, the scanned ultrasonic data only cover the transparent object in the center, excluding the one to the left and two to the right. With our assumption that the target transparent surface to be reconstructed should be supported by a certain amount of ultrasonic measurements, those excluded pieces of transparent surfaces should be labeled as ∞ and no depth value will not be reconstructed, as shown in the first row of Figure 6. By contrast, the flower region (in image space) is successfully labeled as the transparent surface and the corresponding geometry is appropriately reconstructed as shown in (c). In the second case, a small piece at the top right corner is missing in the mesh shown in column (c). It is due to the fact that the region is not inside the camera view-region in those input frames, therefore, no labeling and reconstruction is preformed for this region. More frames can be used if additional coverage is desired. The third case is more challenging due to the existence of opaque objects on both side of the glasses. Our algorithm successfully differentiates these two situations based on the depth and ultrasonic measurements. The hole region caused by the chair occluding the glasses is not an artifact or failure, instead it reflects the accuracy of the observation, namely it is a true occlusion and none of the camera observations can see this region. In Figure 7, there are objects residing inside the fish tank. With sufficient ultrasonic measurement supports, our approach can still recover the scene information and reconstruct the geometry.

Since there is not much prior in reconstructing such large indoor scenes with glass, we compare our algorithm with two baseline methods that will represent the basic attempt

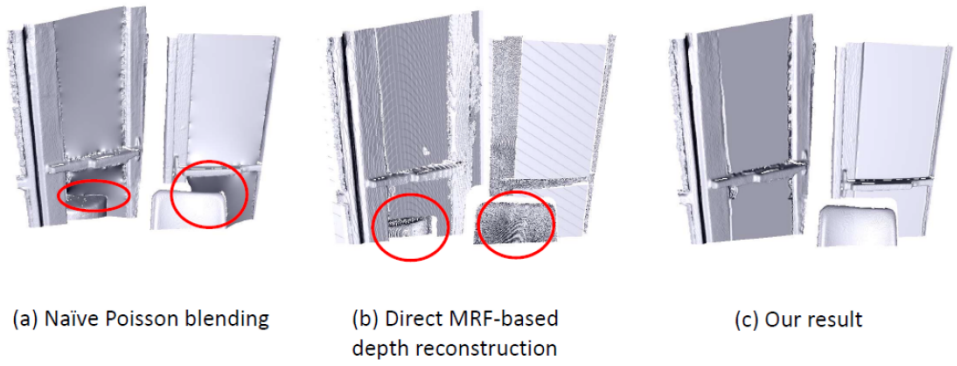


Figure 5. Comparison of our result with two other alternatives. Artifacts produced by these alternatives are highlighted in the red circles.

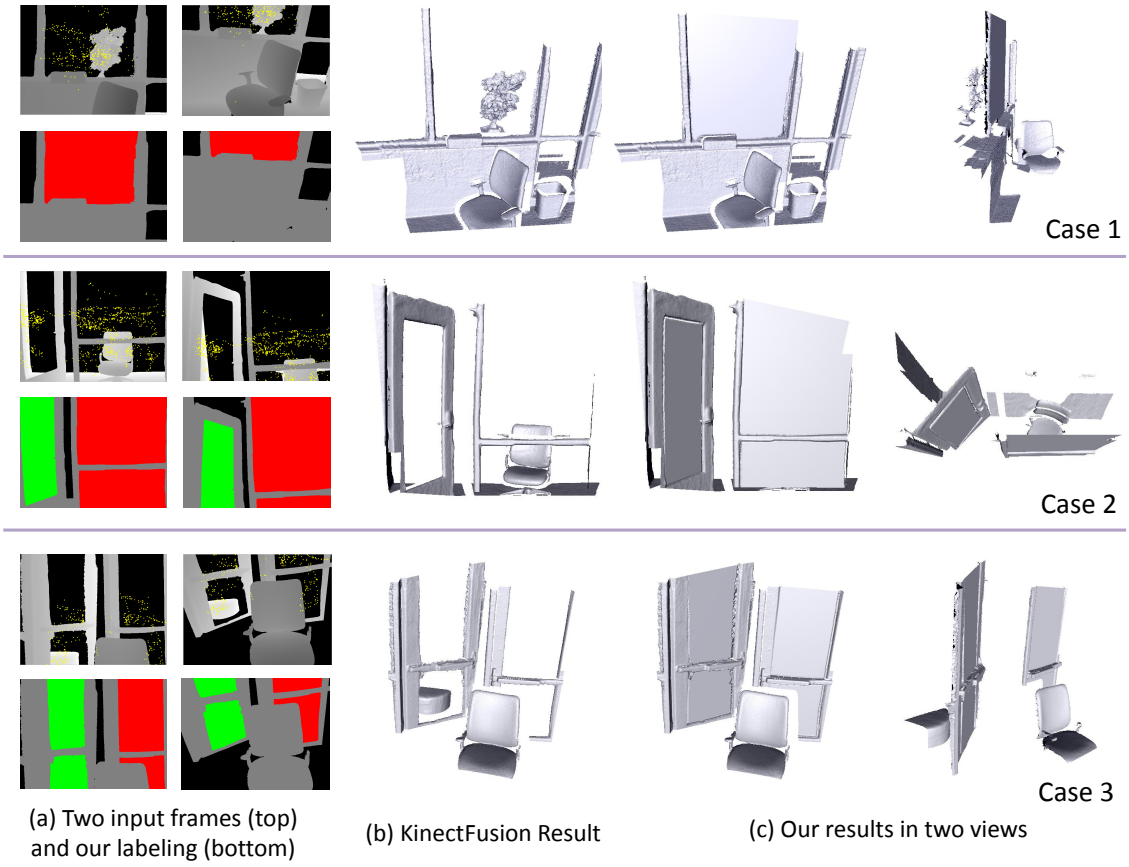


Figure 6. Our results on various real scenes.

to reconstruct such scenes, as shown in Figure 5. The first one is naïve Poisson blending without inferring the label for each pixel. This method can only handle the simplest case where no occlusion exists. In this case, due to lack of knowledge about the labeling, there exists severe artifacts around the boundary. More importantly, it cannot recover the transparent object when there are opaque objects behind it. The second alternative is based on MRF (Markov Random Field) but estimates depth from a set of candidates directly, similar to the classic stereo reconstruction. In this case, we only use depth information, but no color cues. It suffers from the same limitation in terms of handling occlusions. In addition, due to the quantization of depth space, there are usually layering artifacts. A larger set of depth layers candidates can be used to reduce the artifacts, but will significantly increase the computational cost.

6.1. Limitations

Our current approach suffers from certain limitations. In particular, there is no measurement data in empty space.

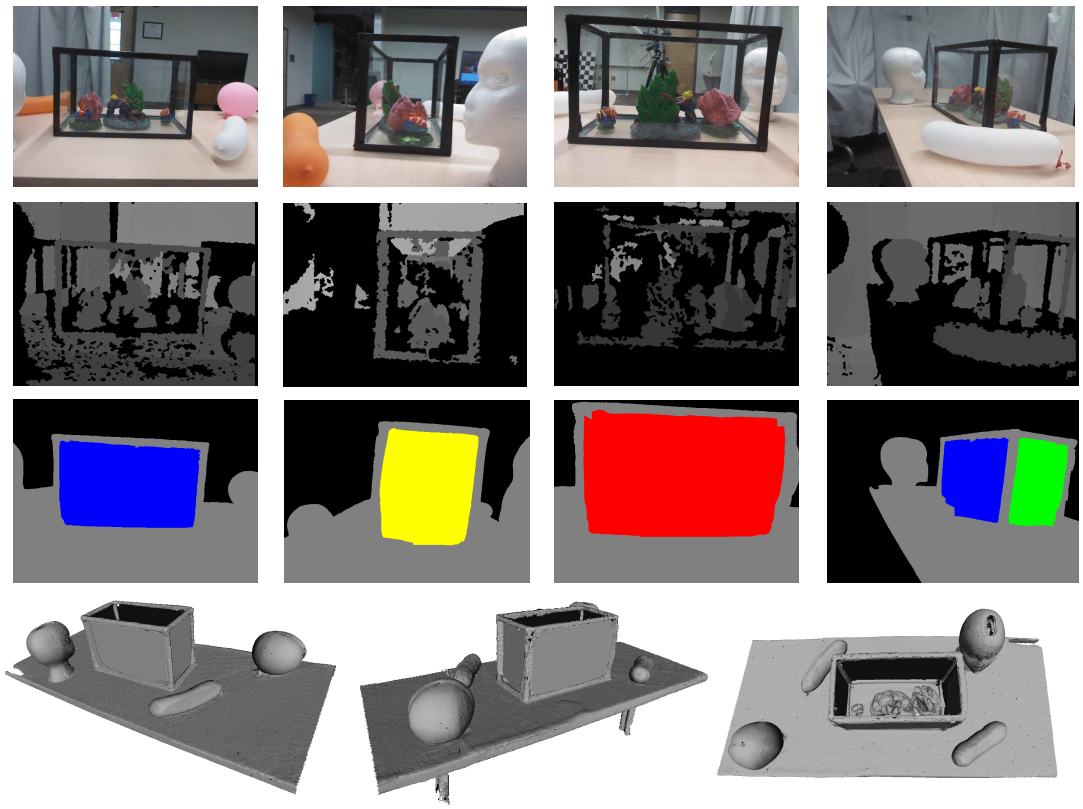


Figure 7. Our results on fish tank scene. We scanned the fish tank separately from four different sides, the third row shows the labeling. The final row shows the final merged result.

Currently we use our prior models described in Section 4.1 to guide our inference. However, if the transparent object is directly connected to the empty space (the case of no boundary or frame), in our current framework, no information can be effectively used to identify the empty space and the labeling result might be incorrect. One possibility is to rely on the ultrasonic sensor to identify the empty space. In the empty space, the ultrasonic sensor will get zero reading as opposed to positive readings when it hits a solid surface. This cue can be embedded in our graphical model, specifically in the probability $P(S_i^t|L_i^t)$ to guide the inference. However, the current ultrasonic sensor [1] has a very limited effective cone-shape region for measurement. Zero readings can be produced if the viewing angle of the sensor is beyond the limit and does not necessarily reflect empty space. This behavior can possibly be modeled in our measurement probability and we can investigate as part of future work. Our current approach also shares the limitation of KinectFusion. In an entirely glass environment, KinectFusion will lose its tracking capability, which can affect the overall scene reconstruction algorithm. We are also unable to perform extra large scale reconstruction due to the reconstruction volume limit imposed by the KinectFusion implementation in the Kinect SDK.

7. Conclusion and Future work

To conclude, we have developed a simple yet effective system for the challenging task of transparent object reconstruction using a combination of depth and ultrasound acoustic sensors, for which little research has been done. Our algorithm produces promising results on real scenes with various complexities.

We believe that there is considerable potential in terms of combining different acoustic and visual sensors for scene reconstruction and understanding. There are many interesting directions for future exploration. For example, instead of using a single sensor, an array of aural sensors can be used to provide more precise reading in depth and expand the usable range. The continuous echo signal can be analyzed instead of treating the aural sensors as a black box. Another direction for future work is to integrate advance image analysis techniques to detect transparent objects such as in [14] and then use that information as prior to guide our inference.

Acknowledgement This work is supported in part by ARO Contract W911NF-14-1- 0437, US National Science Foundation grants IIS-123154, IIS-1208420, and CNS-1305286 and Natural Science Foundation of China grant 61332017, 61305011, 61371166, 61422107. Ruigang Yang is the corresponding author for this paper.

References

[1] http://www.robot-electronics.co.uk/htm/srf235tech.htm.

[2] M. Ben-Ezra and S. K. Nayar. What does motion reveal about transparency? In *International Conference on Computer Vision*, pages 1025–1032, 2003.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[4] J. E. Cryer, P.-S. Tsai, and M. Shah. Integration of shape from shading and stereo. *Pattern recognition*, 28(7):1033–1043, 1995.

[5] G. Eren, O. Aubreton, F. Meriaudeau, L. S. Secades, D. Fofi, A. T. Naskali, F. Truchetet, and A. Ercil. Scanning from heating: 3d shape estimation of transparent objects from local surface heating. *Opt. Express*, 17:11457–11468, 2009.

[6] A. Fusiello and V. Murino. Augmented scene modeling and visualization by optical and acoustic sensor integration. *IEEE Transactions on Visualization and Computer Graphics*, 10(6):625–636, Nov. 2004.

[7] M. Goesele, H. P. A. Lensch, J. Lang, C. Fuchs, and H.-P. Seidel. Disco: Acquisition of translucent objects. *ACM Transactions on Graphics*, 23(3):835–844, 2004.

[8] M. T. Hendrik, H. Lensch, M. Goesele, and H. peter Seidel. Shape from distortion: 3d range scanning of mirroring objects. In *SIGGRAPH*, page 248, 2002.

[9] M. B. Hullin, M. Fuchs, I. Ihrke, H.-P. Seidel, and H. P. A. Lensch. Fluorescent immersion range scanning. *ACM Transactions on Graphics*, 27(3):87:1–87:10, 2008.

[10] N. Hurtós, X. Cufí Solè, and J. Salvi. Integration of optical and acoustic sensors for d underwater scene reconstruction. *Instrumentation viewpoint*, (8):43, 2009.

[11] I. Ihrke, B. Goidluecke, and M. Magnor. Reconstructing the geometry of flowing water. In *International Conference on Computer Vision*, pages 1055–1060. IEEE, 2005.

[12] I. Ihrke, K. N. Kutulakos, H. P. Lensch, M. Magnor, and W. Heidrich. State of the art in transparent and specular object reconstruction. In *EUROGRAPHICS 2008 STAR–STATE OF THE ART REPORT*. Citeseer, 2008.

[13] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, 2011.

[14] U. Klank, D. Carton, and M. Beetz. Transparent object detection and reconstruction on a mobile platform. In *IEEE International Conference on Robotics and Automation*, pages 5971–5978. IEEE, 2011.

[15] H. Kuttruff. *Acoustics: An Introduction*. CRC Press, 2007.

[16] I. Lysenkov, V. Eruhimov, and G. Bradski. Recognition and pose estimation of rigid transparent objects with a kinect sensor. *Robotics*, page 273, 2013.

[17] C. Ma, X. Lin, J. Suo, Q. Dai, and G. Wetzstein. Transparent object reconstruction via coded transport of intensity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3238–3245. IEEE, 2014.

[18] D. Manocha. Interactive sound rendering. In *SIGGRAPH 2009 Courses*, pages 82–127. ACM, 2009.

[19] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics*, 24(3):536–543, 2005.

[20] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011.

[21] R. C. Patel and A. R. Greig. Segmentation of 3d acoustic images for object recognition purposes. In *OCEANS'98 Conference Proceedings*, volume 1, pages 577–581. IEEE, 1998.

[22] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. In *ACM Transactions on Graphics*, volume 22, pages 313–318. ACM, 2003.

[23] G. Saygili, L. van der Maaten, and E. AHendriks. Hybrid kinect depth map refinement for transparent objects. In *International Conference on Pattern Recognition*, pages 2751–2756. IEEE, 2014.

[24] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–350. IEEE, 2009.

[25] L. Svilainis and V. Dumbrava. The time-of-flight estimation accuracy versus digitization parameters. *Ultragarsas (Ultrasound)*.

[26] S.-W. Yang and C.-C. Wang. Dealing with laser scanner failure: Mirrors and windows. In *IEEE International Conference on Robotics and Automation*, pages 3009–3015. IEEE, 2008.

[27] J. Yu. 3d reconstruction of the ïnvisibles¨. CVPR 2013 Tutorial, http://www.eecis.udel.edu/ yu/CVPR2013_Tutorial/, 2013.

[28] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu. Edge-preserving photometric stereo via depth fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2479, 2012.

[29] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1400–1414, 2011.