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

Mao Ye[1], Yu Zhang[2], Ruigang Yang[1], Dinesh Monacha[3]
[1]University of Kentucky. [2]Nanjing University, China. [3]University of North Carolina at Chapel Hill.

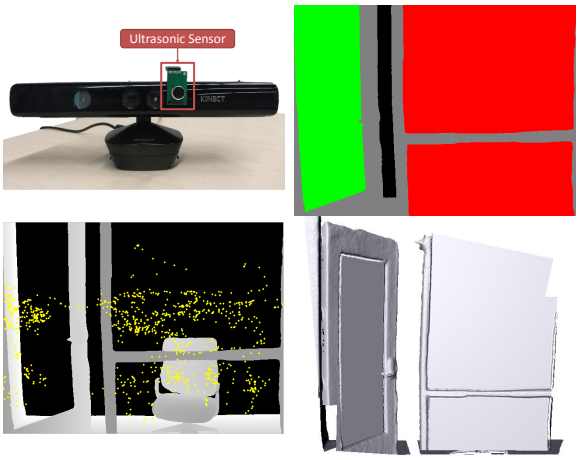


Figure 1: (Top-left): our modified Kinect sensor; (Bottom-left) input depth map, yellow dots are range readings from the ultrasonic sensor; (top-right) color-coded segmentation result; (Bottom-right) final reconstruction results, in which the glass panes are recovered, the chair is behind the door.

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 interior 3D modeling, 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 these completely transparent ones. A user can sweep the camera around to scan the scene of interests, in a fashion similar to KinectFusion [2]. From the multiple ultrasonic sensor 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 are vary sparse and unevenly distributed compared to the depth maps.

Assuming piece-wise planar transparent objects, we formulate this fusion problem as a 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 by the stereo matching, one of the fitted planes from the ultrasonic sensor, or infinity (unknown). From the labeled pixels we then update the depth map in which transparent objects can be reconstructed.

We assume the target transparent objects are piece-wise planar. We first fit multiple planes to the data collected from the ultrasonic sensor using RANSAC. Then we perform segmentation/labeling in the 2D depth map space. 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 $\infty$ label means empty space where no data can be observed from neither sensors. The $\zeta$ label means the first point hit along line of sight is not from a transparent object and the depth data can be observed. By contrast, each $\pi_k$ label defines the pixel from our fitted surface planes $\pi_k$ of the transparent objects.

We define the Bayesian Network in Fig. 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 depth sensor and Ultrasonic depth are represented as $Z_i^t$ and $S_i^t$ respectively. The data captured from depth sensor does not solely depend on labeling, but also is affected by occlusion. Specifically, if a non-transparent object resides behind a transparent one, the depth sensor will very likely capture the geometry of the non-transparent object, while the label of the corresponding pixel should be the transparent object. Therefore, in our model, we use a hidden binary variable $O_i^t$ to explicitly model the phenomena, which takes value 1 if the pixel falls into this situation and 0 otherwise. Our Bayesian Network 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

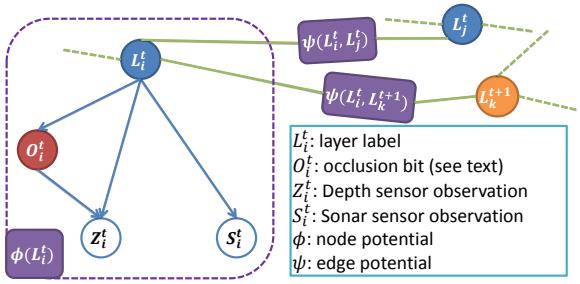


Figure 2: The graphical model.

potential $\phi(L_i^t)$ can be expressed as:

$$\phi(L_i^t) = P(L_i^t) \cdot P(Z_i^t | L_i^t) \cdot P(S_i^t | L_i^t) \tag{1}$$

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

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

The labeling problem can then be cast as a MAP problem that minimizes the following energy function:

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

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. The first term (data cost) and the second term (smoothness cost) are described in detail in the paper. With these terms defined, we use the Graph Cuts algorithm [1] to solve this labeling problem.

The second step of our framework is depth reconstruction for the transparent objects based on the labeling information. 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 the camera center through the pixel and intersecting with the target surface. In order to obtain smooth reconstruction, we adopt the Poisson Blending technique [3] to refine the estimation.

Our reconstruction framework produces promising results on real scenes with various complexities. We hope our work will encourage more exploration in combing aural and visual sensors for 3D reconstruction and beyond.

[1] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[2] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proc. ACM Symp. User Interface Softw. & Tech.*, pages 559–568, 2011.

[3] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 313–318. ACM, 2003.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.