Glass Object Localization by Joint Inference of Boundary and Depth

Tao Wang, Xuming He, and Nick Barnes
NICTA & Australian National University, Canberra, ACT, Australia
E-mail: {tao.wang, xuming.he, nick.barnes}@nicta.com.au

Abstract

We address the problem of localizing glass objects with multi-modal RGB-D camera. Our method integrates the intensity and depth information from a single view point, and build a Markov Random Field that predicts glass boundary and region jointly. Based on the localization, we also reconstruct the depth of the scene and fill in the missing depth values. The efficacy of our algorithm is validated on a new RGB-D Glass dataset of 43 distinct glass objects.

1. Introduction

Semi-transparent objects are commonly found in indoor environments such as household or office scenes, and play a key role in daily human activities. As such, it is important for a vision-based robotic system to be able to localize and interact with them. However, detecting and segmenting such objects from RGB cameras is much more challenging than non-transparent objects due to lack of locally discriminative visual features and homogeneity of surface appearance [11, 4].

Most previous work on glass object localization and recognition has focused on detecting special properties of the glass surfaces and their interaction with the opaque environment in images [13, 1, 12]. In particular, McHenry, Ponce and Forsyth [11] design a classifier which attempts to find a glass/non-glass boundary based on a combination of cues, such as color and intensity distortion, blurring and specularity. In addition, contextual [10] or categorical [4] information is employed to integrate a variety of local features into a coherent surface or object model. Despite those efforts, glass object localization still remains unsatisfactory in practice due to the ambiguity and lack of cues in 2D RGB images.

In this work, we aim to localize semi-transparent surfaces more precisely by exploring multi-mode sensors and incorporating depth information as a novel contextual cue. In particular, we seek to exploit low cost RGB-D consumer cameras, such as the structured-light PrimeSense device (*e.g.* Kinect), to fuse the intensity and depth information from a single view point for indoor environment. While recent work with RGB-D camera is mainly for generic object detection [6, 7, 3], here our goal is joint detection, segmentation and depth inference, which can facilitate many interactive tasks such as manipulation. There has been some work exploiting range devices to detect or reconstruct semi-transparent objects [14, 5]. Unlike those methods, we rely on a single view RGB-D image and combine both intensity and depth cues.

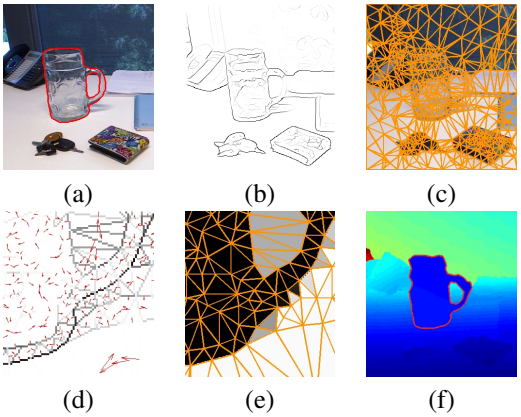

Figure 1. Illustration of the proposed approach. (a) Intensity image with ground truth foreground mask overlaid. (b) Edge detector output. (c) Triangulation result. (d) Boundary classifier output (magnified). (e) Superpixel classifier output (magnified). (f) Reconstructed depth with joint inference result overlaid.

In particular, we exploit the refraction and attenuation that will be experienced by an active signal passing through glass objects. This physical process is difficult to model, but it provides a distinctive missing-vs-nonmissing pattern in the depth channel. We integrate boundary cues from RGB channel with the region cues from depth to build a glass boundary and region detector. In addition, we incorporate the spatial context by constructing a Markov Random Field on a triangularized contour fragments and the corresponding su-

perpixels [2]. A joint inference is designed to predict the glass boundary and region simultaneously. Furthermore, we perform a plane segmentation of the 3D scene in the non-glass region, and fill in the missing depth values caused by glass refraction and other factors. Note that this step would be difficult without the glass boundary/region information. For the glass region, however, due to lack of depth measurement, we approximate its depth by assuming a single cardboard model standing on its (non-transparent) supporting surface.

2. Our Approach

2.1 Boundary and region graph

We begin with detecting edges on the intensity images as proposal for glass boundaries. Depth images are not used here as it is highly noisy at glass boundaries and the missing patterns can either be dilated or corroded depending on the local refractive properties. We use the BGTG boundary detector from [9] to propose glass boundaries (See Figure 1 (b)).

To facilitate depth reconstruction in non-glass region, we also detect *depth boundary* by computing a local depth orientation map on the smoothed depth image with missing regions filled in by a median filter [6]. The results are supplemented by a Canny edge detector on the intensity image to capture any weak depth discontinuities that co-occur with strong intensity changes. We threshold the boundary map and link the edges so that short isolated edges and noises are removed.

To model the spatial context, we construct a graph on proposed boundaries and planar regions as follows. We first break the linked boundaries into short lines and perform Delaunay triangulation on their end points, which generates two types of nodes and their connectivity: boundary fragment nodes connected with their end points, and triangular superpixel nodes partitioned by boundary fragments. As illustrated in Figure 1 (c), most glass boundaries and depth discontinuities are followed by our partition, and this process partially recovers broken/missing boundary detections.

2.2 A Markov Random Field on Boundaries and Superpixels

We build a Markov Random Field (MRF) model [2] on the boundary fragments and superpixels w.r.t. the graph in Section 2.1, which defines a joint distribution over the glass labeling given an RGB-D image input. Note that our output includes both boundary and region labeling – with which we are able to encode the spatial dependency in a more expressive way. We first introduce the energy function of our model and then describe its components in detail.

Let the boundary fragments be $\mathbf{E} = \{e_{ij}\}$ and its subgraph be G_E. Similarly we have $\mathbf{D} = \{d_i\}$ and G_D for superpixels. We define the state space of d_i as $\{0, 1\}$, indicating glass and non-glass. For boundary variable e_{ij}, we first assign a direction to it and define its left and right side. e_{ij} is 0 if it is not a glass-vs-nonglass boundary, 1 if the glass region lies at left side and -1 otherwise. The energy function we proposes can be written as follows:

$$E = \underbrace{\sum_{e_{ij}\in\mathbf{E}} \phi_E(e_{ij};\mathbf{I})}_{\text{boundary unary}} + \beta \underbrace{\sum_{(ij,kl)\in G_E} \psi_E(e_{ij}, e_{kl};\mathbf{I})}_{\text{boundary pairwise}} + \gamma \underbrace{\sum_{d_i\in\mathbf{D}} \phi_D(d_i;\mathbf{I})}_{\text{superpixel unary}} + \lambda \underbrace{\sum_{(i,j)\in G_D} \psi_D(d_i, d_j, e_{ij};\mathbf{I})}_{\text{superpixel pairwise}} \quad (1)$$

where $\mathbf{I}$ is the input image, and β, γ and λ are weighting coefficients.

Boundary unary potentials. The boundary unary potential is the negative log-probability from a classifier based on local cues:

$$\phi_E(e_{ij};\mathbf{I}) = -\log(P(e_{ij}|\boldsymbol{f_{ij}})) \quad (2)$$

where $\boldsymbol{f_{ij}} \in R^N$ is the local feature vector for boundary fragement e_{ij}. We evaluate two different local classifiers: a Support Vector Machine (SVM) with RBF kernel and a Random Forest (RF) classifier. The classifier input consists of the following features: (i) Hue and Saturation [11]; (ii) Blurring [11]; (iii) Blending and Emission [1]; (iv) Texture Distortion [11, 8]. In addition, we add (v) Missing Depth, which measures the ratio of missing depth readings in the neighborhood of the proposed glass boundary.

The boundary unary potential is illustrated in Figure 1 (d). Each fragment is assigned with a probability for glass object contour (*i.e.* the darker the more possible), and the orientation is marked with red arrows pointing towards detected glass region.

Boundary pairwise potentials. The boundary pairwise potential imposes an direction-sensitive smoothness prior. Note that for each boundary fragment e_{ij} there are three possible states. The model prefers configurations where connected boundary fragments with glass region on the same side. More formally, we define the smoothness prior for two connected boundary fragments e_{ij} and e_{kl} as:

$$\begin{aligned}\psi_E(e_{ij}, e_{kl}) &= 1 - \delta(e_{ij} = e_{kl} \neq 0) \\ &+ C_1\delta(e_{ij} = e_{kl} = 0) + C_2\delta(e_{ij} \neq e_{kl})\end{aligned} \quad (3)$$

where $\delta(\cdot)$ is the indicator function and $C_1 = 0.3 * \delta(\frac{\pi}{2} < \alpha \leq \pi)$, and $C_2 = (1 - \cos\alpha)^3 \delta(\frac{\pi}{2} < \alpha \leq \pi)$. Here α is the angle between two fragments. We prefer configurations when the angle between two neighboring boundary fragments are obtuse, so additional penalties terms for turning off such boundaries (*i.e.* $e_{ij} = e_{kl} = 0$) and incompatible orientation (*i.e.* $e_{ij} \neq e_{kl}$) are added. If the angle is acute, we simply treat all states equally except if the orientation is compatible.

Superpixel unary potentials. This term is similar to the boundary unary term except that features are extracted from triangular superpixels. The result is illustrated in Figure 1 (e).

Superpixel pairwise potentials. This pairwise term specifies valid configurations of a boundary fragment and its neighboring superpixels. Any incompatible state will be penalized. Specifically, for boundary fragment e_{ij} let d_i be the superpixel resides to its left and d_j to the right. We set the pairwise potential as

$$\begin{aligned}\psi_D(d_i, d_j, e_{ij}) = \quad & \delta(d_i \neq d_j, e_{ij} = 0) \\ & - \delta(d_i = 0, d_j \neq 0, e_{ij} = +1) \\ & - \delta(d_i \neq 0, d_j = 0, e_{ij} = -1). \quad (4)\end{aligned}$$

2.3 Joint prediction

We greedily search for the global parameters β,γ and λ using a small held-out validation set, and use 0.25, 50 and 20 in our work. To predict the boundary and region labels jointly, we adopt an alternating inference approach to compute the marginals of the boundary nodes and superpixel nodes. We start with no depth terms and use Loopy Belief Propagation (LBP) [2] to compute an initial guess of the marginals of boundary nodes. In each iteration, we first use mean-field approximation to marginalize out the boundary variables and compute the marginals on depth nodes. Then we update the marginals on boundary nodes in a similar way. This procedure is repeated until no change on the marginals.

Given the segmentation, we can reconstruct the depth of the scene in a post-processing step. First, we perform a plane segmentation of the scene directly in 3D by fitting each superpixel with a plane. Each glass object is modeled as a simple cardboard after its supporting plane is identified. See Figure 1 (f).

3. Experimental Evaluation

3.1 Dataset and Setup

We collect a RGB-D Glass Dataset that contains 171 RGB and depth image pairs of 43 distinct glass objects taken from multiple views and with different levels of background clutter. We manually generated a pixelwise ground-truth segmentation mask for each object. In the experiment that follows, we randomly split the dataset into training and testing subsets, including 92 and 79 RGB-D image pairs respectively.

For the local classifiers on boundary fragments, we extract features from multiple pair of image patches at the two sides (*i.e.* left and right) of the boundary. The locations of those pairs are defined by a triplet $l_i = (d_i, r_{1i}, r_{2i})$, where $d_i \in \{3, 5, 10\}$ is the pixel distance from the patches to the boundary, $r_{1i}, r_{2i} \in \{5, 10, 15, 20\}$ is the length of two adjacent sides. For the Random Forests classifiers, we use a three-fold cross-validation process which resulted in 500 trees with 16 predictors sampled for spliting at each node. The superpixel unary potentials are given by a Support Vector Machine with RBF kernel.

	Intens.+ SVM	Intens.+ Depth	Detached Inference	Joint Inference
Bound	19.52	44.38	54.08	62.27
Region	28.06	55.84	61.85	65.96

Table 1. F-measures at 50% recall for boundary and region accuracy metrics, respectively.

3.2 Results and Comparisons

The experimental results are summarized in Figure 2, which shows the precision-recall curves of our glass detector under two metrics: boundary pixel accuracy and region pixel accuracy. For boundary accuracy, we use the benchmark utility from [9] and the matching procedure. We also report the F measure computed at 50% recall rate in Table 1. Here the F measure is the harmonic mean of the precision and recall rate, *i.e.* $F = 2/(1/Pr + 1/Rc)$. For the local classifier with intensity and depth cues, we report the better performance from two classifiers.

We can see that our method achieves much better performance than the baselines. For the methods that uses features from intensity image only, the performance is poorest due to the challenging nature of our dataset. We have tested the same set of features on the dataset in [11] and achieved similar results as theirs. The performance is greatly improved by using depth cues, and by almost 40% precision on average. For boundary fragments, the Random Forest classifier with features extracted at multiple locations further increases the accuracy, which provides around 20% precision increase at 50% recall.

The MRF model further improves the performance, particularly in maintaining high precision into high recall regime. We observe a 10% precision gap between

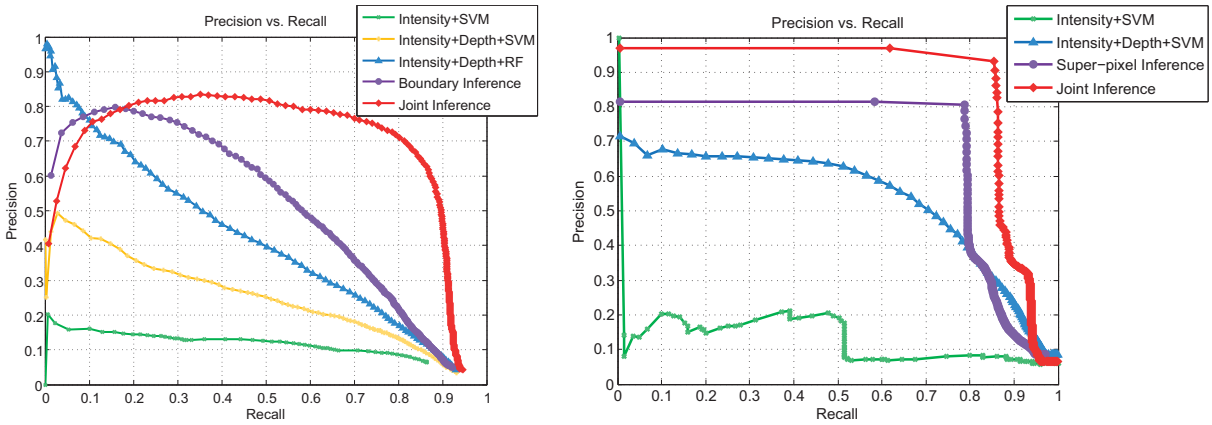

Figure 2. The precision-recall curves based on boundary matching (left panel) and pixel-wise region matching (right panel).

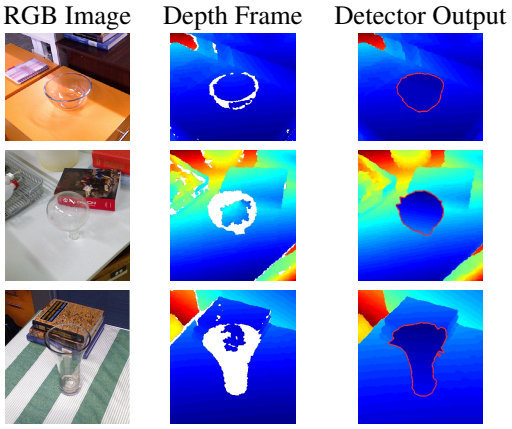

Figure 3. Examples of glass detection results on our new RGB-D Glass dataset. Note that missing areas are shown in white, and depth readings are recovered by a piece-wise planar model.

local classifier performance and its results from the MRF. Joint inference is the most effective method of all. The precision for both boundary fragments and pixel-wise matching sustained at a high level until around 80% recall. This would make our method practical for achieving higher-level scene understanding. We present some examples in Figure 3, which also shows reconstructed depth with missing values filled.

4 Conclusion

In this paper, we have proposed a novel approach to glass localization with consumer RGB-D cameras. By setting up an MRF which jointly encodes boundary fragment and superpixel properties and constraints, we proposed a global optimization procedure for glass detection, segmentation and recovery of the noisy depth maps. We validated the efficacy of this approach on our new RGB-D Glass dataset, which shows the superior performance of our method.

References

[1] E. Adelson and P. Anandan. Ordinal characteristics of transparency. In *AAAI*, 1990.

[2] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[3] B. Frank, R. Schmedding, C. Stachniss, M. Teschner, and W. Burgard. Learning the elasticity parameters of deformable objects with a manipulation robot. In *IROS*, 2010.

[4] M. Fritz, M. Black, G. Bradski, and T. Darrell. An additive latent feature model for transparent object recognition. 2009.

[5] U. Klank, D. Carton, and M. Beetz. Transparent object detection and reconstruction on a mobile platform. In *ICRA*, 2011.

[6] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011.

[7] K. Lai, L. Bo, X. Ren, and D. Fox. Sparse distance learning for object recognition combining rgb and depth information. In *ICRA*, 2011.

[8] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[9] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell*, 26(5):530–549, 2004.

[10] K. McHenry and J. Ponce. A geodesic active contour framework for finding glass. In *CVPR*, 2006.

[11] P. J. McHenry, K. and D. Forsyth. Finding glass. In *CVPR*, 2005.

[12] H. Murase. Surface shape reconstruction of an undulating transparent object. In *ICCV*, 1990.

[13] M. Osadchy, D. Jacobs, and R. Ramamoorthi. Using specularities for recognition. In *ICCV*, 2003.

[14] A. Wallace, P. Csakany, G. Buller, A. Walker, and S. Edinburgh. 3d imaging of transparent objects. In *BMVC*, 2000.