GLASS OBJECT SEGMENTATION BY LABEL TRANSFER ON JOINT DEPTH AND APPEARANCE MANIFOLDS

Tao Wang, Xuming He, Nick Barnes

NICTA & Australian National University, Canberra, ACT, Australia

ABSTRACT

We address the glass object localization problem with a RGB-D camera. Our approach uses a nonparametric, data-driven label transfer scheme for local glass boundary estimation. A weighted voting scheme based on a joint feature manifold is adopted to integrate depth and appearance cues, and we learn a distance metric on the depth-encoded feature manifold. Local boundary evidence is then integrated into a MRF framework for spatially coherent glass object detection and segmentation. The efficacy of our approach is verified on a challenging RGB-D glass dataset where we obtained a clear improvement to the state-of-the-art both in terms of accuracy and speed.

Index Terms— Glass object detection, segmentation, label transfer, adaptive feature learning, MRF inference.

1. INTRODUCTION

Glass object localization has been a challenging problem for the computer vision and robotics community. The appearance of glass objects largely depends on the background and is therefore more difficult to capture by visual features. However, accurate localization of glass objects is a crucial functionality as they are commonly found in various indoor environments.

Most previous work on glass localization focused on the special refractive properties of glass, and their interaction with opaque surfaces in images [1, 2, 3]. Osadchy et al. [4] recognize particular objects from specular reflections which uses knowledge of their 3D shape. On the other hand, McHenry, Ponce and Forsyth [5] design a classifier which attempts to find generic glass/non-glass boundary based on a combination of most commonly used cues, including color and intensity distortion, blurring and specularity. These cues have been further integrated with contours [6] or object categories [7] to infer a coherent object hypothesis.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications, and the Digital Economy, and the Australian Research Council (ARC) through the ICT Centre of Excellence Program. This research was also supported in part by ARC through its Special Research Initiative (SRI) in Bionic Vision Science and Technology grant to Bionic Vision Australia (BVA).

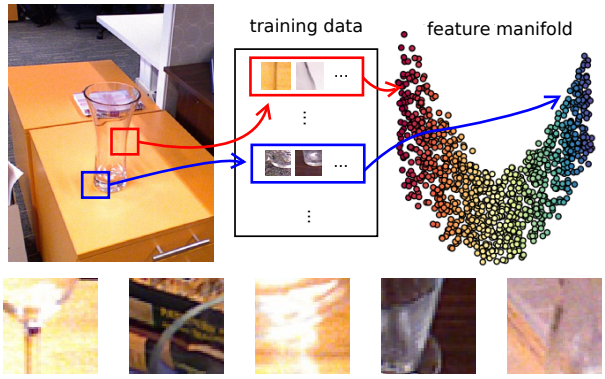

Fig. 1. Top: Illustration of feature manifold based glass boundary classification. We use a learned feature manifold to match every boundary fragment in a test scene (shown as image patches) to training set in order to predict its label. Bottom: Large variation on glass boundaries: patches examples.

Recently, range (depth) cameras have been employed to detect transparent objects, in which the attenuation of signal intensities is exploited. Wallace and Csakany [8] develop a time-of-flight laser sensor based on photon counts. Klank, Carton and Beetz [9] use two images from a time-of-flight camera to detect and reconstruct transparent objects. The popularity of RGB-D sensors (e.g., Kinect) has allowed researchers to utilize both intensity and depth to localize glass objects. Lysenkov, et al. [10] have proposed a model taking into account both silhouette and surface edges, and a CAD-based pose estimation method with a robotic grasping pipeline. Another work [11] exploits the missing-vs-nonmissing pattern in the depth channel which can be used as an effective feature for approximate glass object localization.

Despite those progresses, the generic glass detection and boundary localization is hardly a solved problem. One key reason is the large appearance variations at glass boundaries, as shown in a few examples in Fig. 1 . The state-of-the-art methods, which train generic classifiers for boundaries, produce unreliable predictions (see Fig. 3 for examples). Even with RGB-D cameras, the missing patterns in depth channel can be noisy, or distorted due to local refractive properties [11].

To address this feature variation issue, we propose an im-

age adaptive approach to predicting glass boundaries. In particular, we focus on the scenario that inputs are captured with a RGB-D camera. The main idea of our method is to generate boundary proposals based on a nonparametric feature model. Our model is represented by a joint depth and appearance feature manifold, on which each point is the glass boundary feature of an image patch pair. The boundary label of any pair of neighboring patches is predicted by a weighted voting of its nearest neighbors on the feature manifold. The distance metric on manifold is learned in a supervised manner.

We then integrate the locally adapted glass boundary predictor into a superpixel-based pairwise MRF [12] for glass object detection and segmentation. The MRF labels every superpixels as glass vs non-glass, in which our boundary prediction is used to modulate the smoothing terms in random fields. As we will show in the experiments, our approach generates more accurate glass boundary predictions, which simplifies the overall model structure and the inference algorithm.

Our work is inspired by the recent progress in nonparametric, data-driven approaches on label transfer and propagation (e.g., [13, 14]). These methods first retrieve a subset of training images based on global image statistics, and use the retrieved images for label transfer on the superpixel level for dense image parsing. In particular, Fathi et al. [15] take a semi-supervised learning approach to learn a metric for label propagation in videos.

Our contributions in this paper are threefold. Firstly, we propose novel features for glass localization and a flexible feature pool for improving performance. Secondly, our work is the first to explore nonparametric label transfer within the context of glass detection, and exploit a joint depth-appearance manifold for transductive learning. Lastly, we integrate our locally adapted glass boundary detector into a MRF framework for glass object detection and segmentation, achieving a clear improvement to the state-of-the-art on a challenging RGB-D glass dataset in terms of accuracy and speed.

2. OUR APPROACH

Our approach first generates candidates for glass boundary and region by over-segment an input image into superpixels. We then estimate the local boundary by a weighted voting scheme on a joint feature manifold. Finally, we use a pairwise MRF to integrate the local estimation and generate spatially coherent glass object hypotheses.

2.1. Superpixels and Features

Superpixels. Our first step is to run SLIC [16] and partition image into superpixels. We choose SLIC as it better follows glass and depth boundaries overall compared to alternatives (e.g., edge detector and triangulation as in [11]).

Boundary features. Suppose we have an input image $\mathbf{I}$ and denote each superpixel with a single letter (e.g., i), then any boundary fragment can be indexed by two letters (e.g., ij, indicating i and j are neighbors and ij is the shared boundary between them). The local boundary feature vector $\boldsymbol{f}_{ij}$ includes: (i) Hue and saturation [5]; (ii) Blurring [5]; (iii) Blending and emission [1]; (iv) Texture distortion [5, 17]; (v) Missing depth [11]. In addition, we add (vi) Color histogram on boundary; (vii) HOG [18] on depth data; and (viii) Range (depth) histogram. Note that the above features are extracted from a pair of windows on either side of a boundary fragment, and we use the non-oriented relative ratios in our feature vector.

We augment the image cues by sampling features on multiple scales and at multiple locations. Specifically, we augment the feature set in the following two aspects:
(A) We run superpixelization at a coarse scale and a fine scale, and perform label transfer separately (see details in Sec. 2.2). Afterwards, we merge the local glass boundary proposals from the coarse into the fine scale. Merging is based on the image spatial location, subject to a fixed pixel error tolerance.
(B) Multi-scale and pattern-based features are extracted for each boundary fragment. The multi-scale extraction involves features within windows at 2-times and 3-times of the default feature window size, while the pattern-based feature sampling further augment the features with randomly selected rectangular patterns, at both sides of a boundary fragment, similar to TextonBoost [19].

2.2. Boundary label transfer

The main challenge of glass localization lies in boundary detection, as the refractive properties of glass lead to large variations in the relative features (i.e., features computed on the difference at both sides of glass boundaries). Instead of building a single classifier in the feature space, we explore the local feature manifold, and label transfer based on local matches on the feature manifold.

More formally, let e_{ij} be a binary variable associated with boundary fragment ij, and $e_{ij} = 1$ if the fragment is part of glass boundary and 0 otherwise. A weighted voting scheme is adopted to estimate $P(e_{ij}|\mathbf{I})$, which we use as a local boundary classifier:

$$\begin{aligned} P(e_{ij}|\mathbf{I}) &\propto \sum_{kl} w_{ij,kl} \cdot \delta(e_{kl} = l_{kl}) \\ &= \sum_{kl} e^{-(\boldsymbol{f}_{ij}-\boldsymbol{f}_{kl})^T \Sigma (\boldsymbol{f}_{ij}-\boldsymbol{f}_{kl})} \cdot \delta(e_{kl} = l_{kl}) \end{aligned} \quad (1)$$

where $\boldsymbol{f}_{ij}$ and $\boldsymbol{f}_{kl}$ are local feature vectors for boundary fragments ij and kl, Σ is a diagonal matrix with diagonal elements being the distance between $\boldsymbol{f}_{ij}$ and $\boldsymbol{f}_{kl}$, $\delta(\cdot)$ is an indicator function, and l_{kl} is the ground-truth label of e_{kl}. Here we sum up weighted votes from every training boundary fragment e_{kl}. The weight $w_{ij,kl} = \exp\big(-(\boldsymbol{f}_{ij} - \boldsymbol{f}_{kl})^T \Sigma (\boldsymbol{f}_{ij} -$

$\boldsymbol{f}_{kl})\big)$ is based on a distance metric learned on the feature manifold. In this work, we only estimate $P(e_{ij}|\mathbf{I})$ with k-nearest neighbours, i.e., the local feature manifold, and set $k = 10$ in our experiments.

The weight $w_{ij,kl}$ is learned with manually labeled samples, by adopting the strategy proposed in [15] which casts a distance metric learning problem as a binary classification task. We define a target metric as $w_{ij,kl} = 1$ if $l_{ij} = l_{kl}$, and $w_{ij,kl} = 0$ otherwise. Learning of Σ is performed with linear regression on training data. Intuitively, we prefer the similarity weight $w_{ij,kl}$ to be high if both fragments are part of glass boundary, or both are not.

2.3. Object model and inference

Our glass object model follows a pairwise Markov random field [12] formulation with unary and pairwise terms on superpixel nodes. Denote the set of all image sites (i.e., superpixels) as $\mathcal{S}$. Let $\mathcal{G}$ be the neighbourhood graph on $\mathcal{S}$ based on the spatial relationship. Denote $\mathbf{D} = \{d_i\}$ as a set of binary variables associated with superpixels, and we assume binary state space $\{0, 1\}$ for d_i, with 1 indicating glass regions. Our energy function can be written as follows:

$$E(\mathbf{D}) = \sum_{i \in \mathcal{S}} \phi_D(d_i; \mathbf{I}) + \beta \sum_{(i,j) \in \mathcal{N}} \psi_D(d_i, d_j; \mathbf{I}) \quad (2)$$

where β is the weighting coefficient between unary and pairwise terms, and $\mathcal{N}$ is the neighborhood. The unary term $\phi_D(d_i; \mathbf{I})$ is the negative log-likelihood given by a local SVM classifier:

$$\phi_D(d_i; \mathbf{I}) = -\log(P(d_i|\boldsymbol{g_i})) \quad (3)$$

where $\boldsymbol{g_i}$ is features extracted for superpixel d_i. The features we use for superpixels only include (i), (v), (vii), and (viii) of those used for boundary (see Sec. 2.1 for all boundary features). We also extract multi-scale image features for each superpixel.

For the pairwise term $\psi_D(d_i, d_j; \mathbf{I})$, we utilize $P(e_{ij}|\mathbf{I})$ estimated by boundary label transfer to apply the boundary-superpixel compatibility constraint. We set penalty terms for incompatibility between a boundary fragment e_{ij} and its neighbouring superpixels d_i and d_j as:

$$\begin{aligned} &\psi_D(d_i, d_j; \mathbf{I}) = \delta(d_i \neq d_j) P(e_{ij} = 0|\mathbf{I}) \\ &+ \alpha\delta(d_i = d_j) P(e_{ij} \neq 0|\mathbf{I}) \end{aligned} \quad (4)$$

where $P(e_{ij}|\mathbf{I})$ is estimated by the locally adapted k-nearest neighbour voting described in Sec. 2.1 . In the experiments that follows, we use Loopy Belief Propagation (LBP) [12] to compute the marginals for MRF inference. Model parameters α and β were learned through cross-validation.

3. EXPERIMENTAL EVALUATION

3.1. Data Specifications and Setup

We test our approach on the RGB-D glass dataset used in [11], which contains 171 RGB-D image pairs with 43 distinct glass objects[1] . We follow the training/test data split in [11]. As shown in Fig. 3, the dataset were collected in various scene categories and many of the glass objects are very challenging for localization due to background clutter.

We use SLIC [16] to generate superpixels, with initial region sizes 10 and 30 px. The pixel error tolerance for merging the boundary proposals from the coarse superpixel layer is set to 5 px. For local boundaries, we extract features on 3 different scales, and each scale consists of 50 randomly selected rectangular patterns on both side of the detected boundary, resulting in 300 feature windows. The local superpixel feature set is also generated at 3 scales, and we use SVM with RBF kernel for the unary potential in our MRF. The model parameters α and β chosen by cross validation were 0.5 and 0.25 respectively.

3.2. Results and Discussion

The quantitative and qualitative results using our method are shown in Fig. 2 and Fig. 3, respectively. We compare our approach with [11], referred as "Joint". We also show the performance based on the boundary classifier output, and see why our method is capable of producing superior results with a simpler MRF model. These local boundary classifier outputs are referred to as "Unary" in the figures.

The overall precision and recall on the RGB-D glass dataset is shown in Fig. 2 . The left and middle plots present the precision-recall figures under two metrics: boundary pixel accuracy and region pixel accuracy. For boundary accuracy, we use the benchmark utility from [20] and follow the matching procedure. As both the method from [11] and our method are capable of recovering major glass surface (as a result of using depth features), region pixel accuracy can be less sensitive to noise at glass boundaries as it measures pixel-wise accuracy over the entire image. Therefore we additionally present another region pixel accuracy based result in the right plot which only considers pixels within 10 px of ground-truth glass boundaries. This metric directly reflects the region recovery quality near glass boundaries, which is vital to accurately recovering the shape of glass objects. We achieved superior results on both glass boundary detection and final inference results. While joint inference is able to boost the the performance of noisy unary responses, having cleaner boundary proposals will allow us to adopt simple and more efficient inference algorithms.

Fig. 3 presents some hard examples for comparison for both methods. Note that noisy boundary estimate is the main

[1]The dataset can be accessed from `http://users.cecs.anu.edu.au/~taowang`.

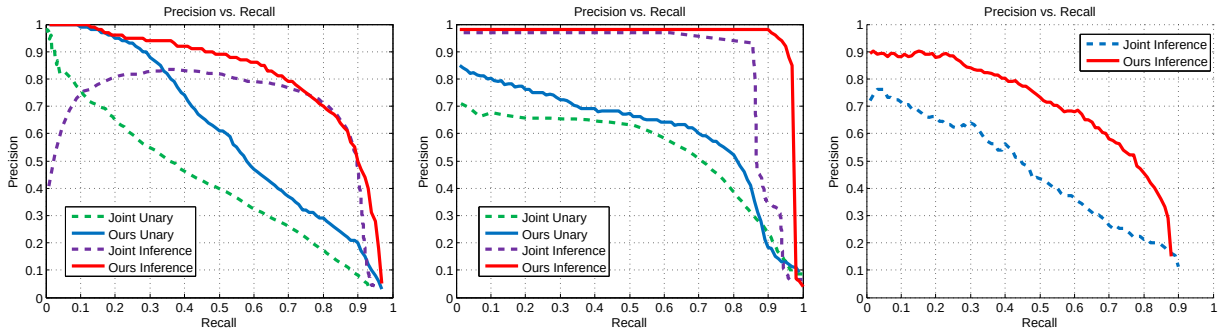

Fig. 2. The overall precision and recall on RGB-D glass dataset for various methods. **Left**: Performance based on boundary pixel accuracy. **Middle**: Performance based on region pixel accuracy on the whole dataset. **Right**: Performance based on region pixel accuracy in the glass boundary neighbourhoods (i.e., regions within 10 px of ground-truth glass boundaires).

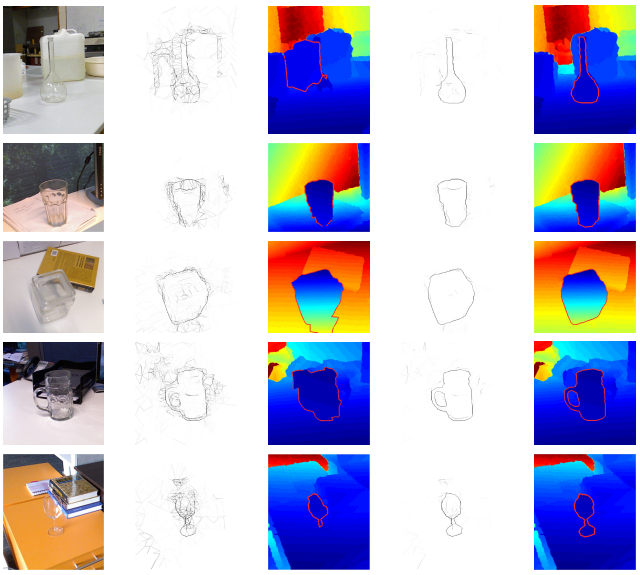

Fig. 3. Hard examples of glass detection results on the RGB-D glass dataset. Column **(a)**: RGB image frame. **(b)**: Unary response from local glass boundary classifiers in [11]. **(c)**: Joint inference and depth recovery results in [11]. **(d)**: Glass boundary proposal results with locally adapted label transfer. **(e)**: Inference and depth recovery results with the proposed method. Note that missing depth readings are recovered by a piece-wise planar model for glass region [11] and smoothed out using a median filter elsewhere.

reason for failure of the joint inference method. The proposed method, on the other hand, showed very reliable and accurate prediction results. Our method has eliminated some circumstances where predictions on the boundary nodes and superpixel nodes are inconsistent (e.g., the second example in Fig. 3). As we can see, the success of the proposed method is primarily due to cleaner glass boundary proposals based on the learned feature manifold. Even sophisticated inference is unlikely to recover glass boundary if the initial estimates are too

	Local (s)	Inference (s)	Total (s)
Joint	0.257	14.542	14.799
Ours	0.928	0.898	1.826

Table 1. Per-image runtime statistics for method in [11] and the proposed method. On average the proposed method is about 8 times faster. See text for details.

weak or severely contaminated by their neighbors.

Finally, we compare the runtime of both methods with our mixed MATLAB and C (mex) implementation. The runtime was broken down into two major components: local boundary estimation and inference. The local part shall include preprocessing, feature extraction and local classification. The proposed method takes longer as we need to extract more features. The inference part for the method in [11] requires up to 20 runs for LBP or mean-field approximations, while ours requires only once. The post-processing (i.e., plane segmentation and depth recovery) takes only a fraction of the total runtime, and therefore is not timed. We report the average runtime per image on an Intel i3 laptop with 4GB RAM in Table 1 . Note that with a native implementation, our method may be further accelerated for real-time applications.

4. CONCLUSION

In this paper, we explored a joint feature based label transfer approach to glass object localization. We propose a novel depth and appearance feature representation for glass boundary and surface detection, and learn a distance metric on the relative feature manifold for glass boundary label transfer. By integrating our glass boundary proposals into a pairwise MRF model, we obtained a significant improvement to the state-of-the-art on challenging examples in a RGB-D glass dataset. Our method can be used as a starting point for more sophisticated algorithms that involves glass surface reconstruction. We would also like to further explore depth-encoded feature manifold for learning with weakly labeled data.

5. REFERENCES

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

[2] H. Murase, "Surface shape reconstruction of an undulating transparent object," in *ICCV*, 1990.

[3] C.J. Phillips, K.G. Derpanis, and K. Daniilidis, "A novel stereoscopic cue for figure-ground segregation of semi-transparent objects," in *ICCV Workshops*, 2011.

[4] M. Osadchy, D. Jacobs, and R. Ramamoorthi, "Using specularities for recognition," in *ICCV*, 2003.

[5] K. McHenry, J. Ponce, and D. Forsyth, "Finding glass," in *CVPR*, 2005.

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

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

[8] A.M. Wallace, P. Csakany, G. Buller, A. Walker, and S. Edinburgh, "3d imaging of transparent objects," in *BMVC*, 2000.

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

[10] Ilya Lysenkov, Victor Eruhimov, and Gary Bradski, "Recognition and pose estimation of rigid transparent objects with a kinect sensor," in *RSS*, 2012.

[11] T. Wang, X. He, and N. Barnes, "Glass object localization by joint inference of boundary and depth," in *ICPR*, 2012.

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

[13] J. Tighe and S. Lazebnik, "Superparsing: scalable nonparametric image parsing with superpixels," in *ECCV*, 2010.

[14] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. PAMI*, vol. 33, no. 12, pp. 2368–2382, 2011.

[15] A. Fathi, M.F. Balcan, X. Ren, and J.M. Rehg, "Combining self training and active learning for video segmentation," in *BMVC*, 2011.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," *École Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep*, 2010.

[17] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *IJCV*, 2001.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[19] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006.

[20] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. PAMI*, vol. 26, no. 5, pp. 530–549, 2004.