Seeing Glassware: from Edge Detection to Pose Estimation and Shape Recovery

*Cody J. Phillips, *Matthieu Lecce, and Kostas Daniilidis
GRASP Laboratory
University of Pennsylvania
Philadelphia, Pennsylvania 19104
Email: {codyp,mlecce,kostas}@cis.upenn.edu
*these authors contributed equally to this work

***Abstract*—Perception of transparent objects has been an open challenge in robotics despite advances in sensors and data-driven learning approaches. In this paper, we introduce a new approach that combines recent advances in learnt object detectors with perceptual grouping in 2D, and projective geometry of apparent contours in 3D. We train a state of the art structured edge detector on an annotated set of foreground glassware. We assume that we deal with surfaces of revolution (SOR) and apply perceptual symmetry grouping in a 2D spherical transformation of the image to obtain a 2D detection of the glassware object and a hypothesis about its 2D axis. Rather than stopping at a single view detection, we ultimately want to reconstruct the 3D shape of the object and its 3D pose to allow for a robot to grasp it. Using two views allows us to decouple the 3D axis localization from the shape estimation. We develop a parametrization that uniquely relates the shape reconstruction of SOR to given a set of contour points and tangents. Finally, we provide the first annotated dataset for 2D detection, 3D pose and 3D shape of glassware and we show results comparable to category-based detection and localization of opaque objects without any training on the object shape.**

I. Introduction

From just a couple image views of a dinner or kitchen scene, the current state of the art Structure from Motion and object detection techniques allow for a reasonable scene reconstruction along with the localization of most scene objects. However, there is a category of objects that persist as a point of failure: objects made from transparent materials. Such objects are omnipresent in home and commercial kitchens, as well as in chemistry laboratories and manufacturing environments. They remain challenging because they lack salient features, with very little color or texture of their own. Due to their high light transmittence and high specularity, even active light sensors do little to improve our sensing capabalitiy.

With the recent resurgence of infectious diseases, robots could work in portable quarantine labs minimizing the possibility of infection for the lab personnel. A solution for interaction with such transparent objects calls for a reliable perception algorithm as a key component.

While some approaches might detect image areas corresponding to transparent materials or even place a bounding box around a hypothesized object, a robot must know the 3D location and the shape of the object in order to grasp it.

One can observe from a search query of "chemistry equipment" or "glassware" that the majority of such objects are rotationally symmetric. For the purpose of perception, it makes more sense to define all transparent 3D objects which are rotationally symmetric as one class, instead of capturing exemplars for each object class, such as glasses, cups, flasks, tube, bottles, etc. The shape of such objects can be modeled with Surfaces of Revolution (SORs).

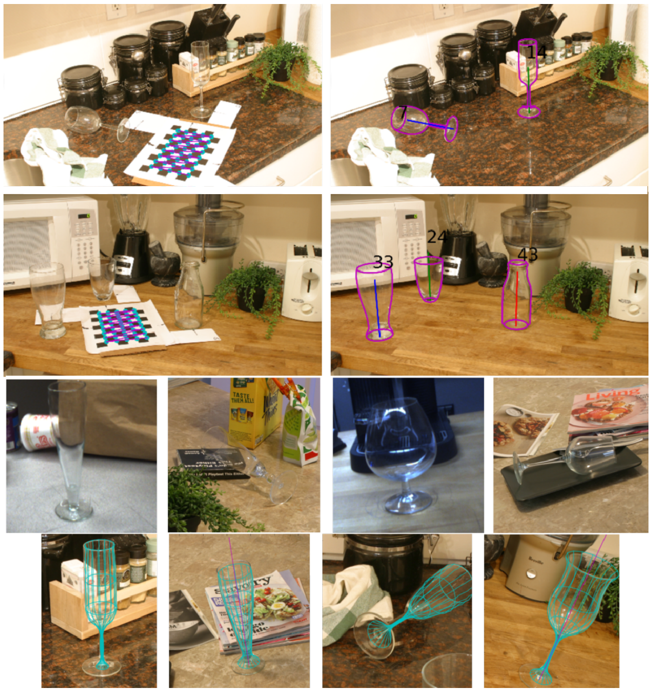

Fig. 1: *Annotated multi-view transparent dataset.* The first two rows each show a single viewpoint, with and without the calibration pattern used to annotate groundtruth camera and object pose. *Example outputs* The last row shows models and poses automatically recovered by the proposed approach.

In this paper, we propose a novel approach for the detection, 3D localization, and shape estimation of the class of transparent rotationally symmetric 3D objects from two calibrated views, where the camera transformation is obtained up to a scale using Structure from Motion on the scene as a whole. The projected contours of an SOR in two views carry information about the shape and pose of the object, but their appearance defeats traditional salient edge detection methods.

To address this challenge, a structured prediction forest is trained from annotated data in order to perform transparent edge detection. Object hypotheses are then generated using geometric properties of SORs projected in two views. Finally, a novel algorithm is proposed to segment, reconstruct and score object hypotheses by efficiently exploring the space of possible SOR shapes to maximize the transparent edge response of the projected contours in two views.

A precisely annotated dataset is introduced to evaluate the visual detection, as well as pose and shape estimation of multiple transparent SORs from two or more views. A set of 34 transparent objects were used in 125 cluttered scenes containing up to 3 transparent SOR instances. Up to 4 calibrated views of each scene were captured, with varying baselines. Fig. 1 shows some views from the dataset, along with annotations.

II. Related Work

The present work lies at the intersection of transparent object modeling, geometric reasoning on the projection of SORs, and boundary detection and grouping.

The challenging problem of detecting, localizing and reconstructing transparent objects has received increasing attention in recent work. Only a few approaches attempted to use visual cues to perform this task. Specific appearance features in RGB images such as transparent edge profile [20] and specularities [21], [11] were modeled to perform localization with promising results on small experiments, but such approaches do not allow for segmentation and reconstruction, only localizing detections with a bounding box. Stereo pairs with a textured background plane were used in [22] to achieve more precise localization, but 3D object shape models are required. Most recent work on transparent objects and glass material steered away from RGB images, either leveraging failure modes of Kinect depth sensors and estimating object pose with a known 3D shape model [28], [16], [17], using time-of-flight cameras [13], sequences with varying illumination [30], or customized structured light sensors [18], [15]. Because of the requirement for prior object models or specific sensors, those approaches do not allow a simple scenario where the only sensor is a stereo camera, or a commercial camera taking a couple pictures of a scene “in the wild”.

The modeling of projected SORs for pose estimation and reconstruction has been addressed as a single-view problem, where pose and camera calibration can be recovered either from bi-tangent points on the apparent contour [31], [29], imaged cross-sections detected as ellipses [4], [5], [10], or both bi-tangent points and cross-sections [3]. Methods have been proposed for the projective [27] and metric [5] reconstruction of an SOR from a single calibrated view when the pose in known, as well as image contours. The visual detection and localization of SORs in cluttered scenes, before their pose and shape can be recovered with the above approaches, has not been extensively studied. An edge linking approach using SOR geometry is described in [4], but it requires an initial window localization of the object in order to be tractable, and reliable edge detection and low clutter are necessary to prevent the non-linear minimization from converging to a local minimum. In this paper, the problem is related to that of 2D symmetry detection, to which more recent efforts have been devoted [26], [14].

Geometry driven edge grouping is one of the core aspects of this paper. In [8], epipolar matching is performed on curvelets to establish a 3D curve sketch from multiple views. Curves were also used instead of points for SLAM in [24]. Another idea that inspired this work is the use multiple images to cluster a common region while matching boundary orientations [12].

Closest to this work are approaches on SOR reconstruction and pose estimation, using two views and manually segmented contours [23], or automatically segmenting contours in a single view before applying reconstruction [3]. The goal of this work is to jointly segment and reconstruct the object in an effort to achieve more robustness. To the best of our knowledge, the proposed framework is the first attempt to fully automatically detect and reconstruct transparent SORs from two views.

III. Proposed Approach

The goal of the proposed approach is to detect transparent rotationally symmetric objects, and to estimate their pose and shape from two calibrated views of a scene. The 3D shape of a rotationally symmetric object can be modeled with a Surface of Revolution (SOR), obtained by rotating a generatrix around an axis. The projection of an SOR produces apparent contours that are view-dependent, as well as fully or partially visible cross-sections. In the case of a transparent object, these image boundaries appear as faint edges and ridges with a specific image profile, and traditional salient edge detection methods fail to detect those boundaries.

The present work addresses this issue in two ways. First, the concept of transparent edge is learned from data, by training a structured prediction random forest to detect transparent edges and distinguish them from visually salient edges (Sec. III-A). Second, segmentation and reconstruction are performed jointly in two views, so that geometric reasoning and consistency across views drive the edge linking process, allowing for more robustness to edge detection errors. Geometric properties of imaged SORs are used to generate object hypotheses from edge detection score maps in two calibrated views (Sec. III-B). Object hypotheses are then segmented, reconstructed and scored in a process that combines edge linking and geometric reasoning (Sec. III-C). Fig. 2 gives an overview of the proposed detection, shape and pose estimation pipeline.

A. Detecting transparent object contours

The proposed approach detects and reconstructs a transparent SOR by leveraging the properties of its apparent contour in several views. While the apparent contour of a Lambertian object can generally be recovered with a variety of edge and contour detection methods [2], [19], [25], the same methods fail to extract transparent contours. Indeed, transparent edges exhibit an image intensity profile different from general edges,

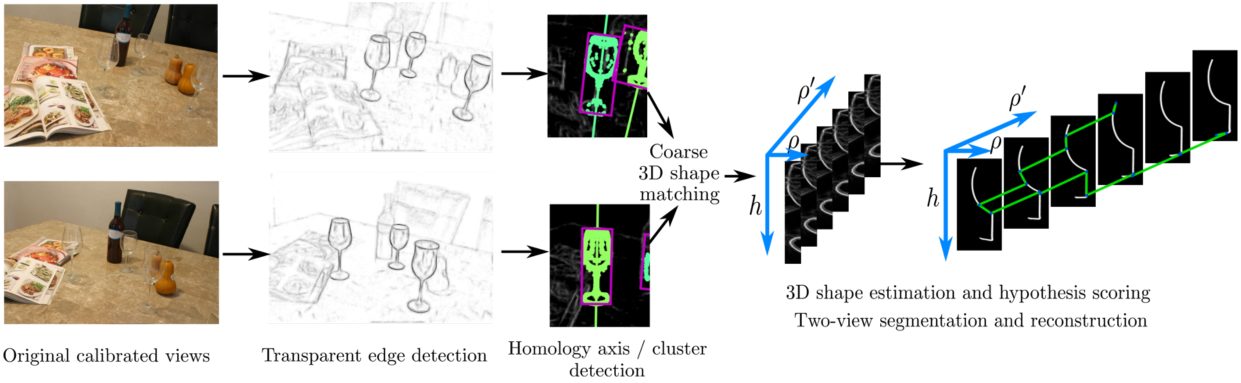
Original calibrated views
Transparent edge detection
Homology axis / cluster detection
Coarse 3D shape matching
ρ'
ρ
h
3D shape estimation and hypothesis scoring
Two-view segmentation and reconstruction

Fig. 2: *Overview of the proposed approach.* A structure prediction random forest trained from annotated data is used to compute transparent edge detection maps. Cylindrical rectifications of the maps are computed and 2D-symmetric curves are extracted and matched across views to generate object hypotheses. Hypotheses that pass consistency checks are then segmented, reconstructed and scored in a process that combines edge linking and geometric reasoning.

closer to a ridge than a step edge. Additionally, it is harder to leverage the differences in image statistics and texture on either side of the contour as in [19], [25], since the background of a glass object is refracted through it, causing the image region inside the object boundaries to be less distinguishable from the outside.

This section describes a way to address these challenges by learning and predicting the appearance of a transparent edge from collected image data. More specifically, transparent edge detection can be treated as a structured prediction problem applied to random forests, the same way salient edge detection is addressed in [6]. Given many examples of transparent objects and their annotated apparent contour and cross-sections, an ensemble of decision trees are used to learn the mapping between an image patch and its corresponding local segmentation mask.

1) Annotated dataset for transparent edge detection: A set of 34 rotationally symmetric transparent objects were used in a dataset aimed at evaluating transparent edge detection, transparent object detection as well as pose and shape estimation using multiple views. The dataset consists of cluttered scenes containing up to three transparent objects. The scene also contain textured objects so that SfM can provide an estimate of the camera transformation. A total of 100 scenes were captured as a stereo pair from a calibrated Bumblebee camera. In order to evaluate the approach on a wide range of larger baselines, another 25 scenes were captured with a calibrated DSLR camera with fixed focal length and large depth of field, from up to 4 viewpoints. The scenes were split in half for training and evaluation of edge detection and the overall proposed approach. Fig. 1 showcases the variety of object shapes, background and clutter captured by the dataset. A checkerboard was placed at a fixed location in the scenes using fiducial markers in order to estimate the camera poses and object poses. The transparent objects were placed at a known offset with the checkerboard, and each view was captured twice, with and without the checkerboard. Fig. 1 illustrated the process and shows, for a couple views, the checkerboard and the ground truth annotations for each object. The apparent contour was automatically annotated by rendering the contours of the object's 3D model in the image. The 3D model of each object was obtained by spray-painting it, manually segmenting its apparent contour in multiple views and applying the reconstruction process described in [23], which is exact when the object pose in the camera frame is known.

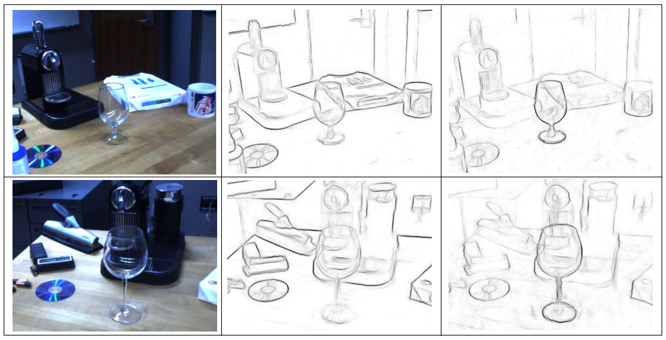

Fig. 3: *Salient VS transparent edge detection.* The second column shows the output of the structured forest edge detector [6], trained on the BSDS500 dataset [1]. The third column shows the output of the same detector, trained on the annotated transparent edge detection dataset. The latter scores transparent edges higher than background edges.

2) Learning a transparent edge detector: We desire a detector that responds to the transparent edges that make up the object's apparent contour and visible cross sections, some portions of which may be seen through the surface of the glass itself. Ground truth labels for these features are generated using the SoRs scaling function in conjunction with the annotations described above. Different labels are given to the outside of the object, the inside of the fully visible cross-section, and the rest of the object interior. Using these labels with the framework described in [6], a structured prediction forest is learned in order to predict, for any color image patch, a patch of segmentation labels of the same size such that transparent edges are located at the boundaries between clusters of different labels. This section provides a very short summary of the structured forest framework, the reader is invited to refer to [6] for more details.

Structured random forests are an extension of the random forest framework for a structured label space. In the present case a label is a patch of the annotated view segmentation mask, typically of size 16-by-16, with three different values given to the background, top cross section and rest of the object interior. In order to address the high dimensionality of such a label space and the difficulty to define an information gain for such labels, the idea proposed in [6] is to map the structured labels to a discrete set of labels, such that similar segmentation masks will be mapped to the same label. The mapping proposed in [6] first encodes label patches in a training set as binary vectors by randomly sampling a small number of pairs of pixel locations and encoding for each patch whether the label values at those locations are the same. Principal Component Analysis (PCA) is then used to further reduce dimensionality, and the resulting vectors are clustered, the cluster numbers being used as discretized labels. Typical parameter values used in [6] are 256 pairs of pixel locations out of 32640 in a 16-by-16 patch, followed by clustering on 5 PCA dimensions. The discretization is randomized and different for the training of every node, resulting in very diverse decision trees better able to cover the large and complex label space. When the maximum tree depth is reached during training, a single patch is stored for prediction. It is obtained applying the above binarization and PCA to all patches in the training subset associated to the leaf and computing the medoid of the obtained vectors. The associated patch is stored for prediction. The same process is used to combine the patches predicted by the trees forming the random forest.

The apparent contour segmentation masks described above were used for training, patches containing a point on the ground truth apparent contour were randomly sampled as positive, all other locations (including any visually salient edges) were randomly sampled as negative, for a total of 10^6 patches, split equally into positive and negative. A typical set of image features were used, namely LUV channels, gradient magnitude at 2 scales and 4 orientations.

A 5-fold cross validation was performed on the annotated dataset, to evaluate edge detection accuracy. Average precision is 0.61 for boundary accuracy, as evaluated in [19]. Fig. 3 shows the interest of fine-tuning the structured prediction forests for transparent edge detection as opposed to visually salient edges.

B. *Generating transparent SOR hypotheses*

The input of the proposed approach consists of two RGB images from two calibrated views with known relative motion R, T. The views are modeled with a pinhole camera model of calibration matrix K, such that the 3D point X projects to $x_i \sim K[R_i|T_i]X, i = 1, 2$ in the two views, and $R = R_1^\top R_2,\ T = T_2 - T_1$.

Once transparent edge response maps are generated, the next step is to generate object hypotheses, using the symmetry properties of SORs. A rotationally symmetric object to be detected can be modeled as a surface of revolution (SOR) defined by rotating a generatrix around a 3D axis, yielding the following possible parametrization:

$$S(h, \theta) = (\rho(h)\cos(\theta), \rho(h)\sin(\theta), h). \quad (1)$$

As described in [27], [5], [23], the cross-sections (visible coaxial circles) of the SOR are projected to ellipses and are view-independent, whereas points on the occluding curve Γ are projected to an apparent contour γ that is view-dependent. Indeed, the 3D curve Γ is view-dependent since changing position of the camera center with respect to the object also changes the locus of points where camera rays are tangent to the object. However complex Γ is, it is symmetric with respect to the plane Π_s containing the camera center and the SOR axis, which means γ is self-invariant through a homology of axis l_s and of vanishing point v_∞ orthogonal to l_s:

$$H_S = I - 2\frac{v_\infty l_s^\top}{v_\infty^\top l_s}, \quad (2)$$

The projected axis l_s is the projection of the plane of symmetry Π_S, and v_∞ is the projection of the direction orthogonal to Π_s. Therefore $K^\top l_s = K^{-1} v_\infty$, or equivalently $l_s = \omega v_\infty$ where $\omega = K^{-\top} K^{-1}$ is referred to as the image of the absolute conic.

Projected axis image search Given the edge detection score map of a view, object hypotheses can be generated by searching for clusters of points with high edge detection response that are self-invariant with respect to a homology of parameters l_s, v_∞ to be determined. Since the views are calibrated and K is known, there are only two parameters to estimate since the choice of an axis l_s determines the position of the corresponding vanishing point $v_\infty = \omega^{-1} l_\infty$. A RANSAC procedure is performed over the edge response map, where a pair of image points x, x' can be hypothesized to be the projection of two symmetric points on the occluding curve Γ of an SOR. The relationship $x' = H_S x$ determines a hypothesized axis l_s and corresponding vanishing point v_∞. This hypothesis can be scored locally by using H_S to examine the edge responses of homologous points. The RANSAC procedure returns an axis l_s and a set of pairs of point that are coherent with l_s and could belong to the contours of an SOR of projected axis l_s. In order to detect multiple instances and speed up the search, the RANSAC procedure is made local by sampling seeds in the image, and for each seed only considering points within a certain distance of the seed. Seeds are randomly sampled according to a distribution that's obtained by applying a mean filter of large size to the edge map: regions with higher average edge response have a higher probability to contain sampled seeds. For each seed, an axis and cluster of inliers are obtained. The local axes and clusters obtained this way are then merged to form complete SOR localization hypotheses. Two axes l_s, l_s' and their corresponding clusters are merged when the corresponding planes of symmetry Π_s, Π_s' are close, which means their normals $K^\top l_s, K^\top l'_s$ are almost collinear, and their inner product is above a certain threshold. For the experiments on the proposed dataset, the following parameter values were used: a mean filter of size 100×100 pixels to obtain the seed sampling distribution; a set of 30 seeds; a seed radius of 150 pixels; a total of 1000 sampled pairs for each seed; a threshold of 0.98 on the inner product of plane normals for axis merging.

The result of this stage is a set of SOR object detection hypotheses in a single view. A hypothesis is scored by summing the edge response of the corresponding inliers, and an oriented bounding box along the direction of axis captures the extent of the object hypothesis. The detection performance of this stage is evaluated in Sec. IV-A. The rest of the pipeline matches hypotheses across two calibrated views to recover 3D localization of multiple SORs from two views, and reconstructs their shapes.

Two-view coarse shape matching and 3D localization For each pair of self-invariant clusters detected in two views, it must be determined whether they could originate from the same object. First, the two cluster axes are backprojected and intersected, to form a 3D SOR axis hypothesis. Then, for each of the two clusters, the points with minimum and maximum height along the projected 2D axis are used to estimate the 3D height range along the 3D axis $(h_{min,i}, h_{max,i}), i = 1, 2$ of a potential object projecting to the symmetric cluster. Insufficient overlap of the heights results in discarding the pair of clusters. The hypotheses are further selected by roughly segmenting the apparent contours in each view and evaluating if their shape is consistent with a single SOR shape. For each of the clusters, an edge linking algorithm is applied to the edge response map in the cluster bounding box, to obtain a coarse estimate of γ. The rough segmentation is performed by a predictive filter dynamic program described in Appx. A. The small-size dynamic program takes a finds an optimal smooth symmetric curve with high edge response and symmetric edge orientations. The curve and the 3D axis are used to perform single-view SOR reconstruction with known pose, as described in previous work [5], [23]. The construction used in this step is the one from [23], and leverages the fact that for a point X at height h and radius r on the occluding curve Γ and its corresponding image point $x \in \gamma$, the imaged cross-section at height h and radius r (appearing as an ellipse in the image) is tangent to γ at x, as shown in Fig. 4. The tangent orientation at x is therefore used to recover $h, r = \rho(h)$ and reconstruct the generatrix ρ point by point. Edge linking followed by this single-view reconstruction yields an imperfect generatrix, as will be discussed in Sec. III-C. However the coarse shape estimates can be used to further filter hypotheses: for two generatrices $\rho_1(h), \rho_2(h)$ recovered from a pair of clusters in two views, residuals $\delta_\rho(h) = |\rho_2(h) - \rho_1(h)|$ for h in the intersected height range capture the discrepancy between the two reconstructions. The count of inliers such that $\delta_\rho(h)$ is used to score the localization hypothesis. The maximum radius of the coarsely reconstructed generatrices, along with the intersected height range, are used to defined a 3D cylinder bounding the object hypothesis.

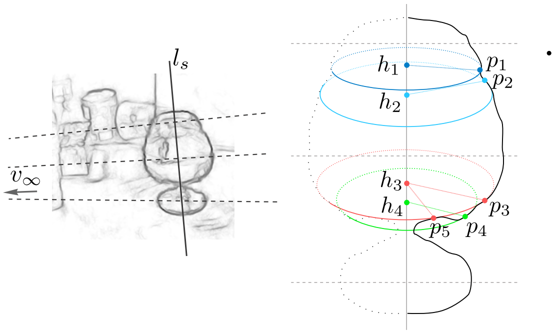

Fig. 4: **SOR reconstruction issues with noisy data.** *Left:* Edge response map and a detected axis l_s. *Right* Reconstruction process on the symmetric curve obtained by edge linking (homology is rectified to put v_∞ at infinity). Points p_i are reconstructed using cross-section tangency.

C. Multiple view segmentation and reconstruction

Once object hypotheses are formulated as 3D cylinders, the edge information from both views is used to reconstruct the 3D shape of each hypothesis. Previous work on SOR reconstruction takes as an input a 3D axis pose and an apparent contour in a single view, as was just described in the coarse shape matching step. The input contour is either manually segmented [23], or obtained through edge linking [4]. Fig. 4 illustrates accuracy and inconsistency issues that arise when attempting reconstruction on a 2D curve segmented from noisy edge data. Two examples are illustrated to provide some intuition on those issues: 1) a "wiggly" section of the curve, where tangents to the 2D curve vary rapidly, cause two points p_1, p_2 close to each other to be reconstructed at incorrect heights and radii, with a large height gap $|h_2 - h1|$; 2) some sections of the curve contain inconsistencies, for instance the tangent cross-section at p_4 is not contained inside the area enclosed by the curve, which is inconsistent with the fact that the apparent contour γ of an SOR is the envelope of the tangent cross-sections at all of its points, as stated in [27]. Intuitively, the same cross-section at height h_3 is tangent to the curve at p_3, p_5, but given the axis pose (object seen from above), p_4 is reconstructed at height $h_4 < h_3$. This means p_5 should not be visible and should be instead occluded by a surface that contains the green cross-section p_4. The 2D segmented curve is not a proper SOR apparent contour for the given axis pose.

The proposed algorithm is designed to address these issues by integrating 3D geometric reasoning into the edge linking process, instead of committing to segmenting an image curve before reconstruction. The objective is to find a generatrix $\rho(h)$ that has good edge support when projected into the image. The core of the algorithm is a dynamic program that recovers $\rho(h)$ by going through heights h monotonically, thus preventing inconsistencies described above. Similarly to previous work, the algorithm works with a single view and a hypothesized 3D SOR axis. However, it is formulated in a

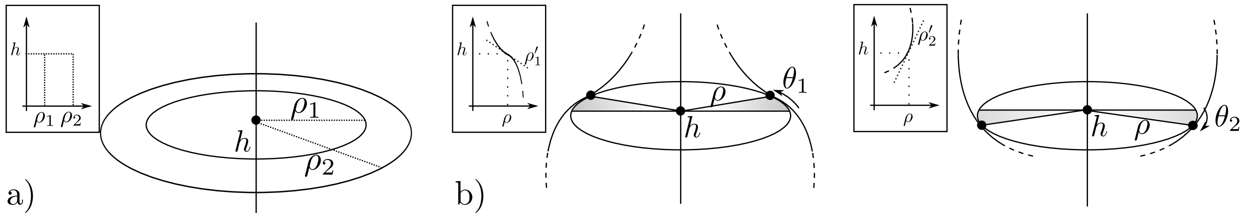

Fig. 5: 2D point supporting a hypothesis for radius ρ and and tangent ρ' at a given height h.

way that also allows to aggregate information from two or more views, and perform reconstruction and segmentation in all views simultaneously. This provides more robustness to edge localization errors in a single view. The space of all scaling functions ρ within the hypothesis cylinder is explored through dynamic programming, in order to recover a smooth 3D shape that gets most support from the edge detection maps.

Given a 3D cylinder hypothesis of axis A (given as an origin at the base of the cylinder, and a direction), height H and radius R, the edge map for a given view is rectified in order to make the projected axis l_s vertical through the image center: the homography described in Eq. (2) becomes a simple 2D symmetry since v_∞, orthogonal to l_s, is at infinity. In this section, the left and right parts of γ with respect to l_s are considered seperately, as well as the corresponding parts of Γ on either side of Π_s for a given view. The algorithm considers the left and right apparent contours in both views as four observations of the object to be combined to estimate the scaling function $\rho(h)$, where h is the height along the 3D axis. For simplicity, Γ and γ will refer to the right side of the 3D occluding curve and 2D contour for one view.

1) Axis-based 2D-3D mapping: The space of all SORs within a 3D cylinder of axis A, height H and radius R can be seen as the set P_A of all possible $\mathcal{C}^1$ (smooth) scaling functions ρ from $[0, H]$ to $[0, R]$. A possible generatrix can be evaluated by projecting the corresponding SOR and evaluating the edge responses along the apparent contour γ. An efficient way of exploring this space can be designed by asking two questions. First, for a possible $\rho(h)$ value at height h, what image evidence (edge response) supports $\rho(h)$? Second, for a given point x in the image space, what are the possible ρ functions such that $x \in \gamma$?

At a given height h and for a possible value $\rho(h)$, the point that would be on the occluding curve Γ at height h belongs to the cross-section circle at height h and radius $\rho(h)$. This circle projects to an ellipse in the image space, as shown in Fig. 5 a. The slope of the generatrix $\rho'(h) = \frac{d\rho}{dh}(h)$, determines exactly which point on the cross-section circle is on Γ: for instance, in Fig. 5 b, assuming the object is seen from above (l_∞ appears above the object in the image), a negative $\rho'(h)$ (object getting thinner as h increases) will cause the occluding 3D point to lie further away from the camera than the cross-section center ($\theta < 0$), whereas a positive $\rho'(h)$ will have the opposite effect ($\theta > 0$). For a given height h, a possible radius $\rho(h)$ and tangent $\rho'(h)$, there is a unique point on Γ at height h, and a unique corresponding point on the apparent contour γ [3].

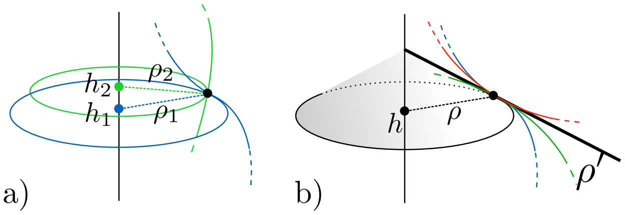

Fig. 6: Pre-image of a single point in the image space by γ_A.

This enables the definition of the function, $\gamma_A(h, \rho(h), \rho'(h))$ mapping h and possible $\rho(h), \rho'(h)$ to a unique point in the image space and edge score map supporting the choice of $\rho(h), \rho'(h)$. Note that γ_A is defined for the axis A of the object to segment.

For a given point x in the image space, the question of which scaling functions would produce a contour γ such that x lies on γ can be answered by finding the pre-image of x by γ_A. The function γ_A is not one-to-one since many apparent contours could go through x and the corresponding point $X \in \Gamma$ would be at a different depth, as shown in Fig. 6 a. However, the restriction of γ_A for a constant ρ' is invertible, as illustrated in Fig. 6 b and stated in the following theorem.

Theorem: Consider the space P_A of scaling functions and the corresponding SORs within a cylinder in the camera frame. For a given value of $\rho' = \tau$, the function $\gamma_{A,\tau}(h, \rho) = \gamma_A(h, \rho, \tau)$ from $[0, H] \times [0, R]$ to the image space is one-to-one.

Proof: Consider a point x in the image space reachable by $\gamma_{A,\tau}$. Because of the definition of $\gamma_{A,\tau}$, any input h, ρ mapped to x means that one or several scaling functions $\rho \in P_A$ such that $\rho(h) = \rho$ and $\rho'(h) = \tau$ create surfaces that are tangent to the ray λx from the projection center through x. Since for any such surface, $\rho'(h) = \tau$, this means the surface is tangent to a cone $C(h)$ of axis A and generatrix defined by τ and the locus of tangency is the cross-section circle at height h, r. This means that no matter where on this circle the ray λx is tangent to the surface, the ray is also tangent to the cone $C(h)$. Considering the family of cones $C(h)$ for any h along the axis, there is only one such cone that is tangent to the ray λx, and it defines h and r. This means h, r are unique for a given x.

2) Joint segmentation and reconstruction: Given a coarse SOR localization hypothesis, segmenting and reconstructing the object can be interpreted as exploring the space P_A of all possible scaling functions within the cylinder. This space can be discretized by considering an array of uniform values of h

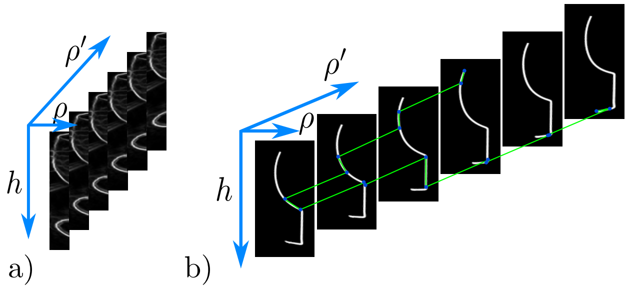

Fig. 7: Dynamic programming for joint segmentation and reconstruction from an edge map.

Algorithm 1: Volume DP segmentation and reconstruction

```
Input: Edge map(s) $\mathcal{E} = \{E_i\}$ from calibrated view(s)
       3D SOR axis $A$ in camera frame
Output: Scaling function $\rho(h)$
$H \leftarrow \{h_{\min} + k * h_{\text{res}},\ k = 1 \dots (h_{\max} - h_{\min})/h_{\text{res}}\}$
$P \leftarrow \{k * \rho_{\text{res}},\ k = 1 \dots \rho_{\max}/\rho_{\text{res}}\}$
$P' \leftarrow \{\cos(k * \pi / n_{\rho'}),\ k = 1 \dots n_{\rho'}\}$
aggregateScores($H, P, P', A, \mathcal{E}$)
$H_\Delta \leftarrow \{h_{\min} + k * \Delta h,\ k = 1 \dots (h_{\max} - h_{\min})/\Delta h\}$
for $(h, \rho, \rho') \in H \times P \times P'$ do
    $h_{\text{prev}} = h - \Delta h$
    // Use only neighboring tangent angles
    $N(\rho') \leftarrow \{\rho'_{\text{prev}} \in P \text{ s.t. } |\theta(\rho') - \theta(\rho'_{\text{prev}})| \leq \pi/8$
    for $\rho'_{prev} \in N(\rho')$ do
        $\rho_{\text{prev}} \leftarrow \rho - \rho'_{\text{prev}} \Delta h$
        score($\rho'_{\text{prev}}$) $\leftarrow$ OPT($h_{\text{prev}}, \rho_{\text{prev}}, \rho'_{\text{prev}}$) + lineSum($h_{\text{prev}}, \rho_{\text{prev}}, \rho'_{\text{prev}}$)
    end
    OPT($h, \rho, \rho'$) = $\max_{\rho'_{\text{prev}} \in N(\rho')}$ score($\rho'_{\text{prev}}$)
end
backTrack($\max_{\rho, \rho'}$ OPT($h_{\max}, \rho, \rho'$))
```

and r within the cylinder, and performing a cut in this array, choosing one radius value $\rho(h)$ for each height, and enforcing smoothness properties. If γ_A were invertible each location x in the image space could be mapped to a single $h(x), r(x)$ value in the array, meaning that the edge response at x supports the choice of $\rho = r(x)$ for the height $h = h(x)$. While this is not achievable because of the ambiguity illustrated in Fig. 6 a, the invertibility of $\gamma_{A,\tau}$ enables the mapping from an image point x to a unique h and r for a given tangent orientation at x. In practice, the mapping $\gamma_{A,\tau}$ is computed using the following geometric reasoning: for a fixed $\rho' = \tau$ and for a given image point x, the ray from the camera center through x is tangent to the SOR, and therefore orthogonal to the surface normal. The surface normal is at an angle of $\pi/2 - \tau$ with the SOR axis. This means that the plane orthogonal to the ray, at distance λ of the camera center such that $X = \lambda x$, intersects the SOR axis at a point X_{axis} such that the angle between $\overrightarrow{XX_{\text{axis}}}$ and the axis is $\pi/2 - \tau$. This enables to solve for λ, and X is projected on the SOR axis to find $h, \rho = \gamma^{-1}_{A,\tau}(x)$. This procedure is vectorized and applied to all image locations.

The tangent angle $\arccos(\rho')$ is sampled into $n_{\rho'} = 40$ values, and the corresponding mappings between the image space and the generatrix space are stored as lookup tables: mapped edge responses form a volume spanned by values of (h, ρ, ρ') as shown in Fig. 7. Responses from the left and right folds of each of the clusters from the two views are aggregated in this way, a volume location (h, ρ, ρ') is therefore supported by edge responses at *four* image locations. Finding the generatrix getting most support from the evidence corresponds to finding a path through the volume with highest score, with the constraint that ρ and ρ' be consistent with $\rho' = d\rho/dh$.

A dynamic program (Alg. 1) is used to optimize the path, traversing the volume from top to bottom as the height h along the axis varies. The height and radius are sampled in the volume at resolutions $h_{\text{res}}, \rho_{\text{res}}$ (in the experiments, both resolutions are set to 1mm) within the bounds of the object localization hypothesis $(h_{\min}, h_{\max}, \rho_{\max})$. For instance, a height range $h_{\max} - h_{\min} = 150\text{mm}$ and $\rho_{\max} = 25\text{mm}$ will be stored as an array of $150 \times 25 \times 40$ nodes for the given resolutions and number of orientations. In order to enforce the constraint between ρ and ρ', heights are discretized with a step Δh (in the experiments, $\Delta h = 1\text{cm}$), such that the path constraint can be expressed as $\rho = \rho_{\text{prev}} + \rho'_{\text{prev}} * \Delta h$, where $\rho_{\text{prev}}, \rho'_{\text{prev}}$ are the previous values chosen by a given path. The resulting generatrix is therefore piecewise linear, but the relatively small step $\Delta = 1\text{cm}$ resulted in good shape estimation results.

The aggregateScores function mentioned in Alg. 1 aggregates edge scores from left and right folds of the of the clusters in both views, using a $\gamma_{A.\tau}$ mapping for each volume slice of corresponding $\rho' = \tau$. The lineSum(h, ρ, ρ') function sums values along the 2D segment $(h, \rho), (h + \Delta h, \rho + \rho' \Delta h)$ in the slice corresponding to ρ', and corresponds to evaluating the edge support of a small segment of generatrix of constant ρ'. When OPT is computed, the argmax ρ' value is stored for the backTrack function to recover the optimal generatrix.

An example of the optimal path for a given volume is shown in Fig. 6 b.

IV. Experiments

A. Single view detection and localization

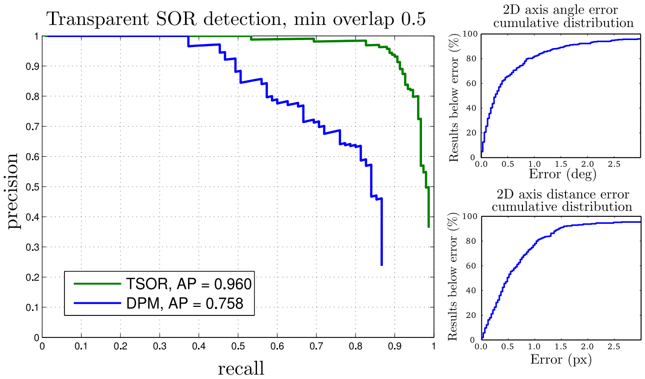

Fig. 8: **2D detection and localization results.** *Left:* PR curve for transparent SOR detection in a single view, with comparison to a DPM model trained on the dataset. *Right:* 2D axis error for correctly detected instances.

Using a single view as an input, the transparent edge detection followed by RANSAC axis detection (Sec. III-A and III-B) can be used for the detection of transparent SORs. The performance is evaluated similarly to object class detection in the PASCAL VOC challenge [7], with a required bounding box overlap of 0.5. The dataset introduced in this paper contains scenes where one or multiple (up to three) SOR instances are present in any given view, therefore this evaluation includes multiple instance detection. A DPM model [9] with 3 components was trained on ground truth bounding boxes in the training set, and a comparison is shown in Fig. 8. Additionally, axis angle and distance error distributions were computed for the boxes that have sufficient overlap. The distributions suggest that the specific symmetry properties of imaged SORs enable the approach to achieve reliable 2D localization that is more informative than a bounding box hypothesis.

B. 3D localization from two calibrated views

The matching step described in Sec. III-B is evaluated as a 3D detection and localization task, where one or multiple SOR instances are imaged in two views. A correct detection requires an overlap of at least 50% between the hypothesized bounding cylinder and the ground truth cylinder encasing the SOR instance. In order to isolate the performance of the matching step and capture only errors occurring in that step, the output 2D bounding boxes of the single view detection task were examined, and the matching task was to detect any 3D instance for which the two corresponding 2D boxes were correctly detected in both views. Similarly to the 2D task evaluation, Fig. 9 shows precision and recall for the 3D detection, as well as distributions of 3D axis pose error for correctly detected hypotheses. The reliability of the pose estimation pictured in those distributions validates the proposed pipeline where an axis pose is first hypothesized and used to reconstruct the object.

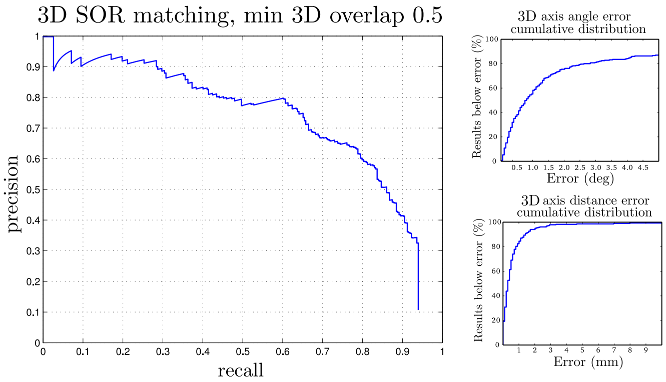

Fig. 9: **3D localization results.** *Left:* PR curve for 3D localization from matching in two views. *Right:* 3D axis error for correctly detected instances.

C. Shape reconstruction

The shape reconstruction is evaluated by computing shape residuals $\delta_\rho(h) = \rho(h) - \rho_{\text{GT}}(h)$ between all points of the reconstructed and ground truth scaling functions, and computing a cumulative distribution of those residuals (Fig. 10). A baseline in the spirit of previous work was implemented for comparison: symmetric curve segmentation as described in Appendix A is performed in a single view, followed by reconstruction with known axis pose as described in [23], [5]. Two observations can be made: 1) the volume DP yields more accurate results than the baseline, as it is designed with noisy edge responses in mind, 2) results obtained with the proposed Volume DP formulation are also computed with evidence from a single view, and the results from two views only provide a marginal improvement. It can be hypothesized that using two views helps in a few cases where edge data is missing or noisier in one of the two views.

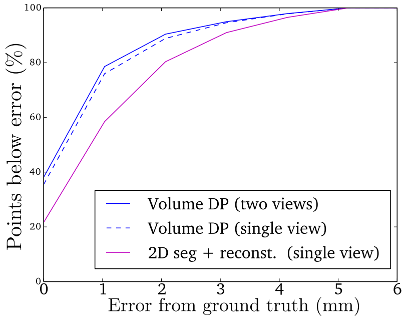

Fig. 10: **Shape reconstruction results.** Generatrix shape error cumulative distribution, for the proposed Volume DP and baseline.

V. Conclusion

A full framework and an annotated dataset were proposed to address the problem of detecting, localizing and segmenting glassware in two calibrated views. It was shown that targeted learning of transparent edges combined with shape-based reasoning are a good way to capture the most informative part of an imaged transparent SOR, namely its contour. Future work will explore extensions to other classes of transparent objects, and to the case of uncalibrated views.

Acknowledgement

We are grateful for support through the following grants: NSF-DGE-0966142 (IGERT), NSF-IIP-1439681 (I/UCRC), NSF-IIS-1426840, , ARL RCTA W911NF-10-2-0016, and a grant from GlaxoSmithKlein.

Appendix A

Symmetric edge linking The dynamic program runs vertically along a cluster's axis, operating over image coordinates u, v. The score of the best path that ends at point (u, v) is represented as $O(u, v) = \max_{u^- \in (u-5 \ldots u+5)} [O(u^-, v^-)] + C_{\text{edge}}$. The edge score is $C_{\text{edge}} = \sqrt{E_l * E_r}$, where $v^- = v - 1$ is the previous pixel height along the axis, E_l, E_r are the left and right fold edge responses, and E is a combination of the left and right edge responses.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011. 3

[2] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):679–698, 1986. 2

[3] C. Colombo, D. Comanducci, and A. Del Bimbo. Camera calibration with two arbitrary coaxial circles. In *ECCV*, pages 265–276, 2006. 2, 6

[4] C. Colombo, D. Comanducci, A. Del Bimbo, and F. Pernici. Accurate automatic localization of surfaces of revolution for self-calibration and metric reconstruction. In *CVPR Workshop, 2004. CVPRW'04*, pages 55–55. 2, 5

[5] C. Colombo, A. Del Bimbo, and F. Pernici. Metric 3d reconstruction and texture acquisition of surfaces of revolution from a single uncalibrated view. *PAMI*, pages 99–114, 2005. 2, 4, 5, 8

[6] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1841–1848. IEEE, 2013. 3, 4

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 8

[8] R. Fabbri and B. Kimia. 3d curve sketch: Flexible curve-based stereo reconstruction and calibration. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1538–1545. IEEE, 2010. 2

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010. 8

[10] D. Fioravanti, C. Colombo, and B. Allotta. Self-calibrated visual servoing with respect to axial-symmetric 3d objects. *Robotics and Autonomous Systems*, pages 451–459, 2009. 2

[11] M. Fritz, G. Bradski, S. Karayev, T. Darrell, and M. J. Black. An additive latent feature model for transparent object recognition. *Advances in Neural Information Processing Systems*, pages 558–566, 2009. 2

[12] D. Glasner, S. N. Vitaladevuni, and R. Basri. Contour-based joint clustering of multiple segmentations. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2385–2392. IEEE, 2011. 2

[13] U. Klank, D. Carton, and M. Beetz. Transparent object detection and reconstruction on a mobile platform. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 5971–5978. IEEE, 2011. 2

[14] T. S. H. Lee, S. Fidler, and S. Dickinson. Detecting curved symmetric parts using a deformable disc model. In *ICCV*, 2013. 2

[15] D. Liu, X. Chen, and Y.-H. Yang. Frequency-based 3d reconstruction of transparent and specular objects. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 660–667. IEEE, 2014. 2

[16] I. Lysenkov, V. Eruhimov, and G. Bradski. Recognition and pose estimation of rigid transparent objects with a kinect sensor. In *Proceedings of Robotics: Science and Systems (RSS)*, Sydney, Australia, July 2012. 2

[17] I. Lysenkov and V. Rabaud. Pose estimation of rigid transparent objects in transparent clutter. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 162–169. IEEE, 2013. 2

[18] C. Ma, X. Lin, J. Suo, Q. Dai, and G. Wetzstein. Transparent object reconstruction via coded transport of intensity. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3238–3245. IEEE, 2014. 2

[19] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *PAMI*, 26(5):530–549, May 2004. 2, 3, 4

[20] K. McHenry, J. Ponce, and D. Forsyth. Finding glass. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 973–979. IEEE, 2005. 2

[21] M. Osadchy, D. Jacobs, and R. Ramamoorthi. Using specularities for recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1512–1519. IEEE, 2003. 2

[22] C. J. Phillips, K. G. Derpanis, and K. Daniilidis. A novel stereoscopic cue for figure-ground segregation of semi-transparent objects. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1100–1107. IEEE, 2011. 2

[23] C. J. Phillips, M. Lecce, C. Davis, and K. Daniilidis. Grasping surfaces of revolution: Simultaneous pose and shape recovery from two views. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1352–1359. IEEE, 2015. 2, 3, 4, 5, 8

[24] D. Rao, S.-J. Chung, and S. Hutchinson. Curveslam: An approach for vision-based navigation without point features. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 4198–4204. IEEE, 2012. 2

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000. 2, 3

[26] S. Tsogkas and I. Kokkinos. Learning-based symmetry detection in natural images. In *ECCV*, pages 41–54. Springer, 2012. 2

[27] S. Utcke and A. Zisserman. Projective reconstruction of surfaces of revolution. In *PR*, pages 265–272. 2003. 2, 4, 5

[28] T. Wang, X. He, and N. Barnes. Glass object localization by joint inference of boundary and depth. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3783–3786. IEEE, 2012. 2

[29] K.-Y. Wong, P. R. S. Mendonca, and R. Cipolla. Camera calibration from surfaces of revolution. *PAMI*, pages 147–161, 2003. 2

[30] S.-K. Yeung, T.-P. Wu, C.-K. Tang, T. Chan, and S. Osher. Normal estimation of a transparent object using a video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1–1, 2014. 2

[31] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, and N. Pillow. 3d object recognition using invariance. In *AI*, pages 239–288. 1995. 2