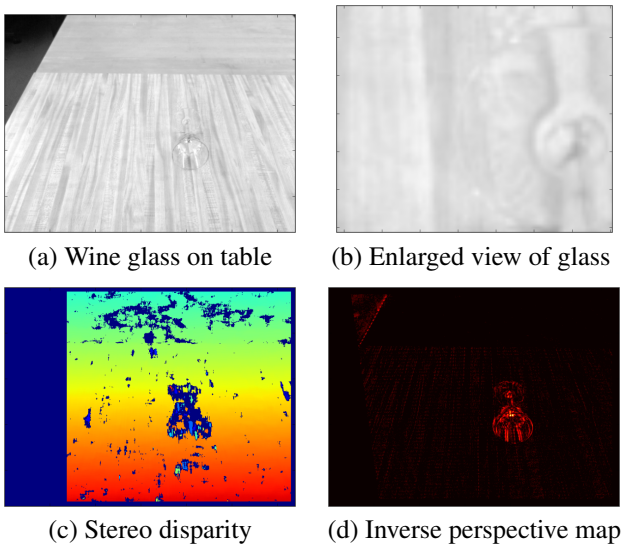
A Novel Stereoscopic Cue for Figure-Ground Segregation of Semi-Transparent Objects

Cody J. Phillips, Konstantinos G. Derpanis and Kostas Daniilidis
GRASP Laboratory, University of Pennsylvania
Philadelphia, PA 19104, USA
{codyp, derpanis, kostas}@cis.upenn.edu

Abstract

The visual perception of semi-transparent objects, such as drinking glasses, is an open challenging problem. Unlike opaque objects, semi-transparent objects violate many of the standard vision assumptions, among them that figure-ground segmentation contains salient boundaries. More specifically, reliable motion and stereo cues for segmenting semi-transparent objects are not present because of the infeasibility of establishing correspondence. This paper describes a new discovery that semi-transparent objects are salient on the plane-parallax image generated by the inverse perspective mapping. A novel cue is introduced that reveals objects extruding from a planar support surface. Points on the support plane are consistent with a planar homography transformation, whereas extruding points from textured surfaces violate this mapping. Furthermore, extruding semi-transparent objects violate the mapping due to the refraction of light and strong specularities. The utility of this new cue is demonstrated in a novel detection and localization approach, where the cue is matched to a database of 3D models of semi-transparent objects. Preliminary empirical results suggest that the presented approach produces a small set of candidate locations for semi-transparent objects and yields accurate localization.



(a) Wine glass on table (b) Enlarged view of glass

(c) Stereo disparity (d) Inverse perspective map

Figure 1. Challenges of perceiving semi-transparent objects. (a) Intensity image of a wine glass resting on its side. Even upon close inspection of the glass, shown in (b), the contours of the glass are weak or absent. A lack of texture causes stereo matching to fail, shown in (c), where failures are denoted in blue. Diffraction, refraction and specularities of the glass cause anomalies in the stereo inverse perspective map, shown in (d).

1. Introduction

1.1. Motivation

Semi-transparent objects are prevalent in many of the environments where one desires to employ autonomous robots. This is especially true in domestic environments, where cups and containers are often made out of semi-transparent materials, such as glass and plastic. Accurate object localization is a prerequisite for tasks, such as grasping, obstacle avoidance and motion planning. Therefore, the ability to perceive semi-transparent objects is vital to successfully perform manipulation tasks in such scenes.

Current sensors and approaches for object detection and localization generally fail when applied to semi-transparent objects. In intensity images, semi-transparent objects have essentially no texture, and exhibit very weak contours, if at all. The lack of contours and color of such objects cause them to be virtually indistinguishable from the background with standard segmentation approaches. Figures 1 (a) and (b) demonstrate how such poorly defined contours make the outline of the wine glass extremely difficult to discern, even upon close inspection. The lack of texture and presence of specularities result in the failure of traditional stereo matching approaches. Figure 1 (c) reveals a hole in the stereo

disparity image, signifying the failure of the stereo computation in the region of the glass. The transmissive property of semi-transparent objects is unfortunate, as it renders ineffective attempts to project texture onto the object surface or use standard time-of-flight light sensing techniques.

In this paper, an approach for detecting and localizing semi-transparent objects is presented that is based on the *inverse perspective map* [17, 5]. It is shown that the same properties of semi-transparent objects that present a challenge to their perception can be used as a cue to their presence. Specifically, the tendency of such objects to cause specularity via reflection or to bend light via refraction generates anomalies in stereo parallax. It is demonstrated that these anomalies are often detectable and provide strong evidence for the presence of an object that is warping the light field in the scene. Figure 1 (d) shows an example where such a warping is salient.

In the present work, it is assumed that semi-transparent objects lie on a known supporting plane. This planar assumption allows for an inverse perspective mapping (i.e., a homography) between a stereo camera pair. In brief, all points in a given image view are mapped onto a known ground plane and then reprojected back onto a second image view. The reprojected image is then compared to the true image in the second view to detect intensity differences. This necessitates the assumption that there is sufficient background texture seen through the semi-transparent object for such differences to exist. In general, a point on the supporting surface is seen by the cameras via the two respective light rays that emanate from it. If the path of the the two light rays are unobstructed, their projection onto the camera planes will obey the homography; however, if the rays pass through a semi-transparent object, they will likely bend and project to image locations that violate the homography. This violation is detected by inversely mapping an image view to a second one and comparing their intensity values. In addition to violations due to refraction, specularities will also cause an intensity difference as one camera may receive a high localized intensity not seen from a differing viewpoint. Note that the presented cue does not discriminate between opaque textured objects that are salient in the plane-parallax image from semi-transparent ones. Nonetheless, a salient cue is provided for a problem with no other reliable cues for figure-ground segregation.

1.2. Related work

As compared to the vast amount of work devoted to the analysis of opaque objects (e.g., recognition and segmentation), semi-transparent objects have received relatively less attention in the computer vision literature. Semi-transparent objects are distinct from their opaque counterparts in that they violate many of the fundamental vision assumptions (e.g., Lambertian surface).

In many works, the image of a semi-transparent object has been modelled by a linear combination of layers [1, 22, 15, 6]. Here, the focus is on modeling the attenuation of light through the semi-transparent material at the expense of ignoring refraction and specularities altogether. Another set of work has considered the recovery of shape and pose information from analytically derived reflective and refractive properties [4, 2, 14]. This body of research has assumed highly simplified and controlled environments. Specularities alone have been used in a qualitative manner to recognize objects of known 3D shape [20]. A drawback of this approach is that it ignores potentially rich information in the interior of the object caused by the refraction of the background behind it. Others have eschewed analytic models and instead proposed methods based on learning from an examplar set [7, 19, 18, 10, 16, 13]. Additional key distinctions between approaches include whether the sensor used is passive (e.g., CCD/CMOS camera) or active (e.g., time-of-flight sensor [12]) and whether a single view or multiple views are considered.

Common among the cited approaches is that they ignore potentially useful contextual cues. For the detection of opaque objects, it has been shown that performance is enhanced by introducing geometrical contextual information of the scene layout [11, 3], such as the relationship between objects and their supporting planes.

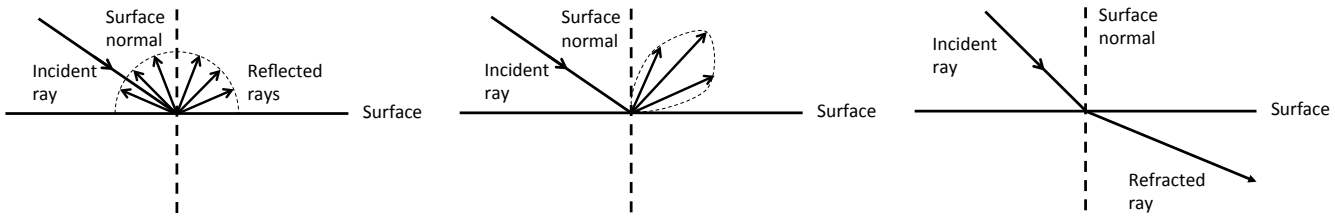
1.3. Contributions

In the light of previous research, the major contributions of the present work are threefold. First, a new discovery is reported that a well-known cue for obstacle detection, the inverse perspective mapping [17, 5], makes an excellent cue for figure-ground segregation of semi-transparent objects. Second, a new representation for model-query image matching in stereo images is proposed, where the model is represented by its expected plane-parallax image based on a known pose. Finally, promising preliminary detection and localization results for semi-transparent objects are presented, in admittedly few images but with very challenging appearance.

2. Technical approach

2.1. Opaque vs. semi-transparent objects

Perception of semi-transparent objects is a challenging task because such objects have properties that violate fundamental assumptions made in most vision algorithms about how light interacts with the scene objects. For instance, objects are generally assumed to be diffuse Lambertian surfaces, where light rays that strike the object are isotropically reflected towards all viewpoints (see Fig. 2 (a)). This assumption permits reasoning about light rays and image points using perspective projections. Specularities



(a) Lambertian reflection (b) Specular reflection (c) Refraction

Figure 2. Opaque (Lambertian) vs. specular vs. semi-transparent objects.

and refractions violate the perspective projection of light and require additional mathematical models to be handled properly, such as the Bidirectional Reflectance Distribution Function (BRDF) for specular reflection and Snell's Law for refraction. These models are a function of both the object's shape and material composition, as well as the camera viewpoint, and are thus much harder to employ.

In the case of semi-transparent objects, light is often reflected non-isotropically, leading to specularities arising from the fact that some viewpoints receive the reflected light at high intensity while others do not (see Fig. 2 (b)). Additionally, most light rays are not reflected at all, but rather pass through the object and emerge bent based on the refractive index of the material (see Fig. 2 (c)). Thus, most light rays received at a specific viewpoint originate from behind the object.

2.2. Local binocular semi-transparency cue

Semi-transparent objects refract the light coming from the scene behind and often exhibit specular reflections of light in front. These two properties of specular reflection and refraction violate perspective projection and greatly contribute to the difficulty in perceiving semi-transparent objects. It is therefore interesting that this violation of perspective projection is precisely what can be exploited to increase the saliency of semi-transparent objects in binocular imagery.

When a plane is present within a scene, there exists a homography between each camera's image plane and the scene plane. From these homographies there is also an inverse perspective mapping that maps the (scene) plane's projection in the image plane of one camera onto the image plane of the other. Figure 3 shows these homographies, H_1 and H_2, between the scene and cameras, as well as the inverse mapping, $H_1^{-1}H_2$, that maps the right image plane to the left image plane. This homography can be computed from a stereo image pair via stereo correspondence matching followed by estimating the dominant plane using RANSAC [9].

Given a stereo image pair of a scene plane, such as a table, each point on the plane, X, projects its light intensity onto the two image planes. Assuming the scene plane has

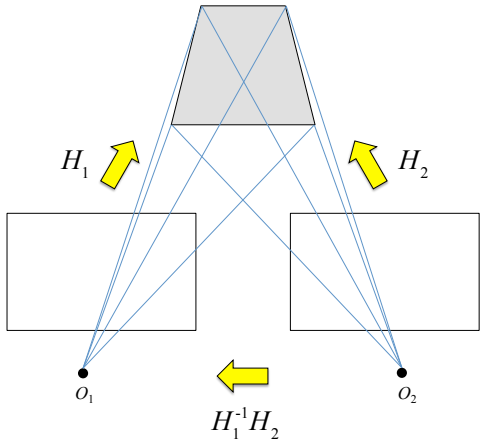


Figure 3. Stereo inverse perspective mapping. Planar homographies H_1 and H_2 relate the scene plane to the two images planes with O_1 and O_2 denoting the respective centers of projection, respectively. The inverse mapping, $H_1^{-1}H_2$, maps the right image plane to the left image plane.

a Lambertian surface, it is expected that the corresponding points on the image plane, X_1 and X_2, have the same intensity. These two image points are related by the *inverse perspective mapping* [17, 5],

$$X_1 = H_1^{-1}H_2X_2. \tag{1}$$

In more general terms, the homography acts as a correspondence function, $f(\cdot)$, that encodes the expected perspective reprojection of X_2 in the other image plane, such that $X_1 = f(X_2)$. The planar homography is used due to the ease of estimation and practicality of assuming a dominant supporting plane.

If $I(\cdot)$ is the intensity function, then the intensity discrepancy function, $D = |I_1(X_1) - I_2(f^{-1}(X_1))|$, is expected to be negligible if the two image points indeed correspond to the same point on a Lambertian surface. A critical source for intensity discrepancies among images captured at different viewpoints, as it relates to semi-transparent objects, is the refraction of light as it travels from the scene plane to the cameras. Since this refraction is not accounted for in the perspective projection, errors in the reprojection are

Algorithm 1: Computing intensity discrepancy.

Input: I_1: View 1 intensity image, I_2: View 2 intensity image, C: Intrinsic stereo parameters
Output: D: Intensity discrepancy, P: Supporting Plane

Step 1: Compute the stereo disparity values between the views.
Step 2: Estimate the dominant plane via RANSAC.
Step 3: Compute the inverse homography $H_1^{-1}H_2$ between views.
Step 4: Remap image in view 2 to view 1 using the homography, (1).
Step 5: Compute absolute difference, D.

Algorithm 2: Localizing semi-transparent objects.

Input: D: Intensity discrepancy, P: Supporting plane, C: Intrinsic stereo parameters, O: 3D object models
Output: S: Similarity map, L: Object localization

Step 1: Establish pose search space using P and C.
Step 2: Sample search space.
Step 3: Create parallax template M from O for each sample point.
Step 4: Compute inner product of each template M and discrepancy image D yielding S.
Step 5: Consider similarity peaks as candidate locations, with top as L.
Step 6: (Optionally) Repeat Steps 1-5 with finer sampling.

introduced as light rays pass through the object and are bent on their path towards the image plane. These errors are then propagated to the inverse mapping between cameras. Generally, the intensity discrepancy image generated by the inverse perspective mapping has a relatively high response on: (i) binocular half-occlusion regions [8], (ii) opaque objects that lie off the scene plane, (iii) specular highlight regions and (iv) semi-transparent objects that distort the view of the background by way of refraction.

2.3. Localizing semi-transparent objects

To detect and localize semi-transparent objects from a stereo pair, the intensity discrepancy (i.e., parallax image) between images from two viewpoints is computed. Given the intensity discrepancy image, a template matching-based approach is employed to perform object localization in 3D space.

The model template is generated by the union of the (binary) silhouettes of the object rendered in the left and right views, with the right silhouette reprojected to the left image plane by the inverse perspective mapping, (1), see Fig. 4 (a). This simulates the stereo parallax and is done to complement the parallax revealed in the intensity discrepancy image by the subtraction of two stereo views. This parallax is seen as objects extend upwards off the plane, resulting in noticeable "double vision" in Fig. 6 (b).

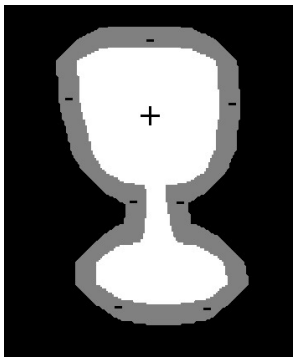


Figure 5. Parallax template as generalized center surround. The positive white center area corresponds to the parallax template mask, M. The negative gray area corresponds to the dilated region surrounding M, represented as dilation$(M) - M$.

The similarity score between the position dependent template and the corresponding intensity discrepancy image region can be computed by an inner product. An issue with using the inner product directly as a similarity score is that it yields high scores for small templates in large regions of high energy. A generalized center-surround is used to penalize such cases. Specifically, the similiarity score consists of the sum of energy over the template (i.e., inner product) less the energy within the dilated area around the template (see Fig. 5). Formally, the similarity score, S, for a given discrepancy image, D, and rendered parallax template mask, M, is computed by

$$S = \sum_{ij} D_{ij} M_{ij} - \alpha \sum_{ij} D_{ij} [\text{dilation}(M) - M]_{ij}, \quad (2)$$

where α is the surround penalty factor. In evaluation, the surround penalty factor, α, was set to 1.0 and 5 iterations of binary dilation were used.

Each point on the surface plane is considered to be a potential object position. The search space is sampled by projecting each pixel location to its corresponding three-dimensional point on the plane. While general pose estimation has six degrees of freedom, the assumption that the objects rests on a known supporting surface reduces the degrees of freedom in the object's pose to five. Additionally, the assumption of rotationally symmetric upright objects decreases the search space further to two degrees. This two-dimensional search space is defined as the set of three-dimensional points that lay on the supporting plane. A coarse sampling of every ten pixels is used for an initial

(a) Parallax template (b) Templates over search space (c) Search space similarity scores

Figure 4. Parallax templates and search space. (a) Summary for the creation of the parallax template. Conceptually, the model is rendered in both the left and right views, and then the right view is remapped by the planar homography to the left view. The union of the binary masks form the parallax template. (b) An example sampling of the search space. A binary template is rendered at each sampled point in the search space. Its inner product with the intensity discrepancy image yields the similarity for that location. (c) Example score peak at wine glass location. The search space is determined by projecting each pixel location of the intensity image to the supporting plane. Each pixel location in this image corresponds to a intensity image point and its projection on the supporting plane. The color represents the similarity of the templates centered at these points. Regions in red, yellow and blue, signify high, intermediate and low response, respectively.

localization that is then refined in a local window at a higher sampling density. An example of the search space sampling is illustrated in Fig. 4 (b).

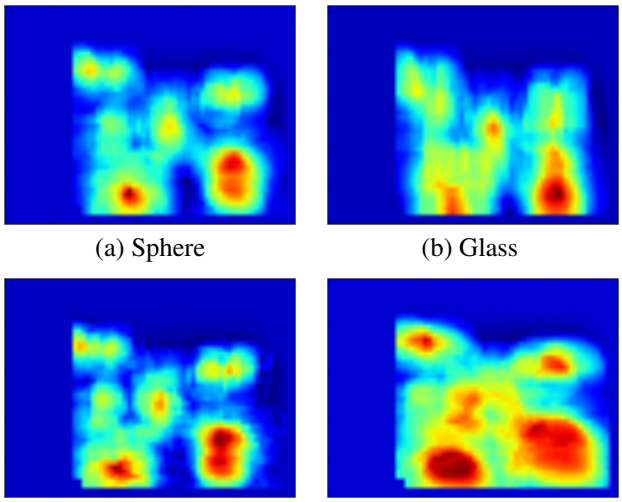
To produce candidate locations for the object's location, several peaks are extracted from the similarity image; Fig. 4 (c) shows an example similarity image for a wine glass input template. The location of the peak in the similarity image directly corresponds to the pixel location in the intensity image and the metric location on the supporting plane in the world. The quality of a candidate is ranked by the height of its peak. Finally, candidate locations are identified via an iterative non-maximum suppression procedure that terminates once a local peak is encountered that has a value below a predefined threshold.

To recapitulate, the two main aspects of the proposed approach are summarized in Algorithms 1 and 2.

3. Empirical evaluation

To evaluate the presented approach, a test scene was created consisting of the following five semi-transparent glass objects: wine glass, sphere, drinking glass, mug and bowl. In addition, a textbook was used to represent an opaque distractor. A PointGrey BumbleBee2 stereo camera was used to capture the binocular imagery. The left view of the scene is shown in Fig. 6 (a) and the inverse perspective mapping discrepancy image is shown in Fig. 6 (b). To generate the parallax templates, five models that were roughly similar in shape and size to the objects in the test set were selected from the Princeton Shape Benchmark [21] and Google's 3D Warehouse (http://sketchup.google.com/3dwarehouse).

Object localizations were evaluated based on two criteria: (i) the rank of the best candidate and (ii) the metric position error from the ground truth of this candidate. To determine the rank, the candidates were ordered by their similarity value. To measure the metric error, ground truth virtual model positions were established manually and the distance between the ground truth and recovered positions were computed in both pixel and scene space.

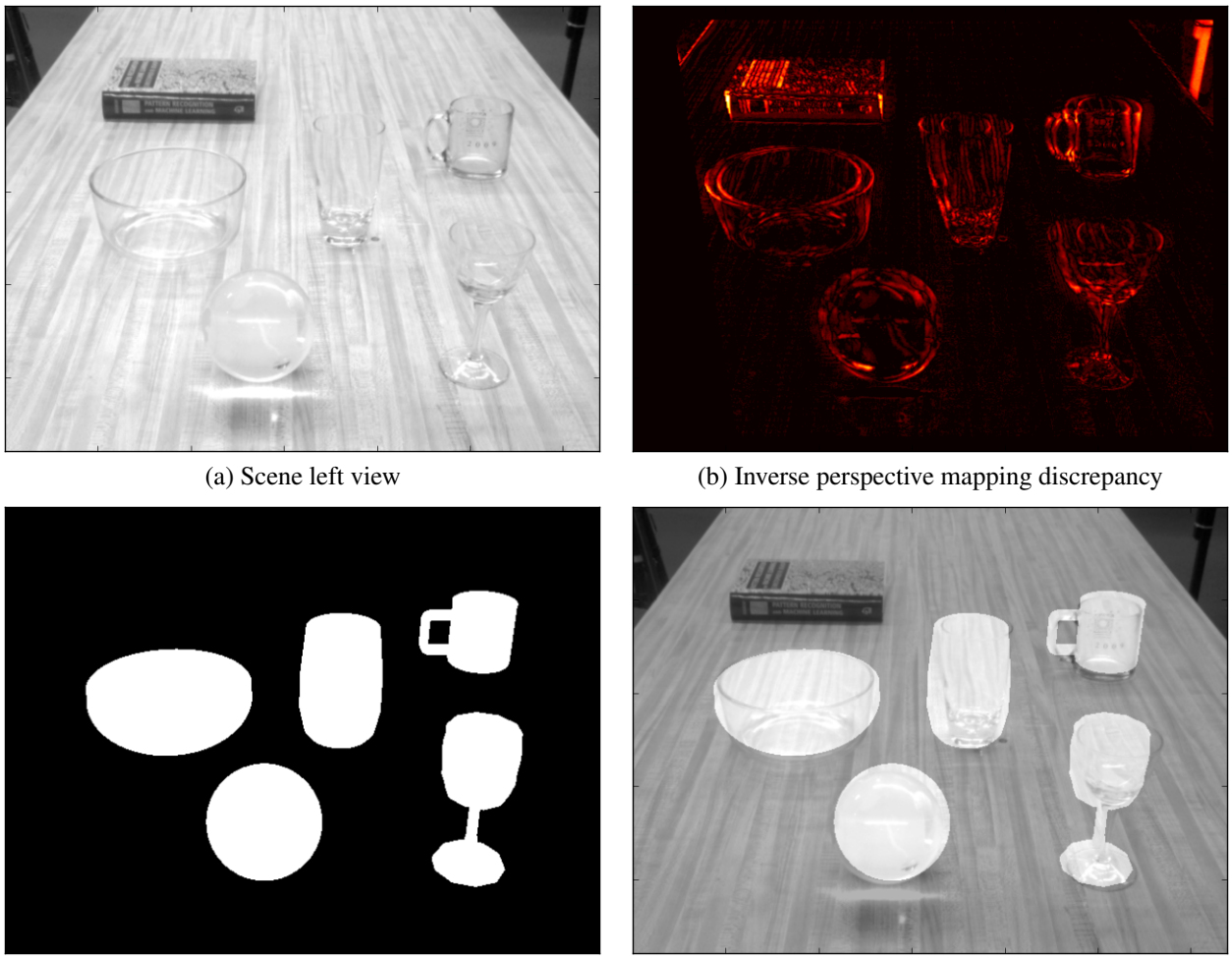


(a) Sphere (b) Glass

(c) Mug (d) Bowl

Figure 7. Similarity images and candidate peaks. The similarity images are created by taking the inner product of the center-surround parallax template and the intensity discrepancy image at the corresponding image locations. Red signifies areas of high energy and peak locations correspond to object location candidates.

The empirical results are summarized in Table 1. In terms of position error, all objects were successfully localized, with a maximum error of approximately nine pixels in the image and 11 mm on the table. These localizations are shown in Fig. 6 (c) and (d). Of the five objects, the correct wine glass and sphere locations were found as the top candi-



(a) Scene left view

(b) Inverse perspective mapping discrepancy

(c) Recovered object positions

(d) Objects overlayed on input image

Figure 6. Detection and localization results. (a) The left stereo image of the scene with five semi-transparent glass objects and a textbook. (b) The intensity discrepancy from the inverse perspective mapping. (c) The resulting object localizations returned by the proposed approach. (d) The object masks superimposed over the input image at their respective detection positions.

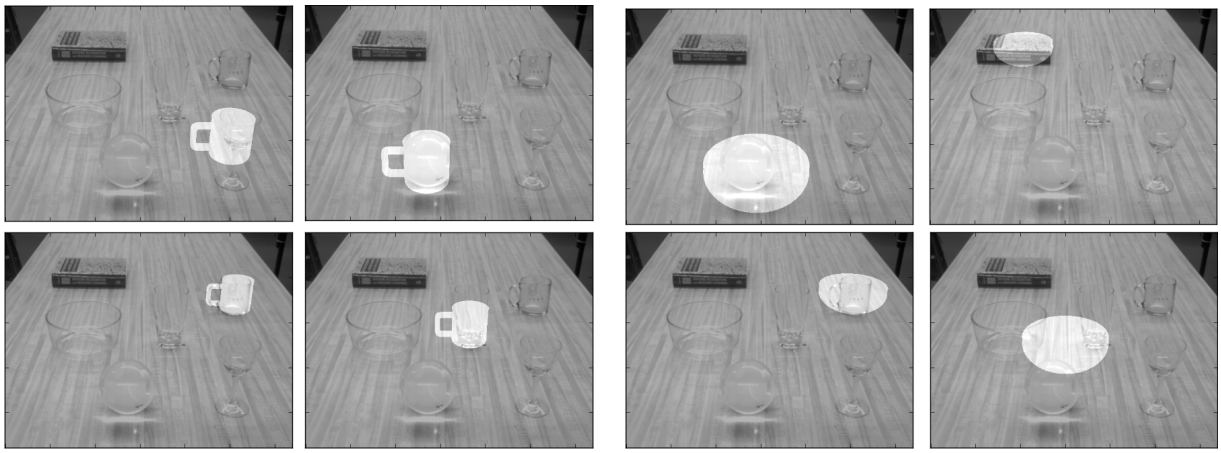
dates, the correct glass and mug locations were ranked third and the bowl was ranked eighth. The sphere and the wine glass were a source of confusion when searching for the glass and mug. The similarity images in Fig. 7 (b) and (c) reveal that their locations resulted in score peaks higher than the correct localization. These candidates are shown for the mug in Fig. 8 (a). The similarity images for the sphere and wine glass (see Fig. 7 (a) and (b), respectively), illustrate that these objects have very high peaks and are thus likely to be selected as prime candidates. The bowl's best candidate was ranked eighth. Its similarity image shows (see Fig. 7 (d)) many high, yet flat peak regions that caused confusion. These regions are caused by the large size of the template, enabling it to encompass the energy of any object in the scene, see Fig. 8 (b).

The localization of an object requires an evaluation of 1000-3000 templates for a 640×480 image. OpenGL is utilized to render the model template at each sample location. The search step (coarse search with refinement of best candidate) takes under 10 seconds per object with unoptimized Python code.

4. Discussion and summary

The concept of using inverse projective mapping to extract a salient cue indicating the presence of semi-transparent objects was introduced. In addition, the utility of this cue has been demonstrated in a novel approach to perform model-query image matching in stereo imagery.

While the present formulation of the detection and localization approach yields confusions in discriminating between similar semi-transparent objects, the inverse projective mapping technique yields a highly informative cue for



(a) Mug detection candidates (b) Bowl detection candidates

Figure 8. Confusion among semi-transparent objects. Candidates are ordered left-to-right and top-to-bottom by decreasing similarity peak score. (a) There were only five peaks detected for the mug, with the third (lower left) candidate being the correct match. All mug candidates were centered over semi-transparent objects. (b) Due to its large size, the bowl template easily encompasses any area of high energy, and can even straddle a gap to cover energy from three separate models, as shown in the lower right.

Object	Error (pixels)	Error (mm)	Rank
wine glass	9.2	10.4	1
sphere	6.3	6.1	1
glass	6.3	8.7	3
mug	4.2	11.1	3
bowl	1.4	2.7	8

Table 1. Per object localization performance metrics. Error values are between the position of the manually placed virtual model and the estimated object location. The rank represents the position of the best candidate, determined by similarity score value.

the detection of semi-transparent objects. Future work involves improving the similarity metric to better detect and localize specific instances of semi-transparent objects.

While this paper uses plane-parallax for its simplicity and ease of estimation, the use of inverse projective mapping to detect semi-transparency extends to any correspondence function when used in conjunction with perspective projection. Such a correspondence function about the background surface is possible even if the support surface is not planar because its 2.5D depth function can be captured with a depth sensor.

In summary, this paper has presented a novel cue for detecting the presence of semi-transparent objects and an approach for detecting and localizing objects in stereo imagery. The approach is founded on the inverse perspective mapping. While this mapping has previously been proposed for detecting opaque objects, its application to semi-transparent objects, as done in this paper, has not been considered. Preliminary results suggest that the approach can achieve accurate detection and localization of semi-transparent objects. Finally, it is anticipated that the proposed cue will prove useful in a variety of applications beyond the particular detection and localization approach considered here.

Acknowledgements

Research presented here has been partially supported by the Army Research Laboratory through ARL RCTA W911NF-10-2-0016, by the National Science Foundation NSF-OIA-1028009, and by Willow Garage through the lease of a PR2.

References

[1] E. Adelson and P. Anandan. Ordinal characteristics of transparency. In *AAAI Workshop on Qualitative Vision*, pages 77–81, 1990.

[2] S. Agarwal, S. Mallick, D. Kriegman, and S. Belongie. On refractive optical flow. In *European Conference on Computer Vision*, pages II: 483–494, 2004.

[3] S. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

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

[5] M. Bertozzi, A. Broggi, and A. Fascioli. Stereo inverse perspective mapping: Theory and applications. *Image and Vision Computing*, 16(8):585–590, 1998.

[6] K. Derpanis and R. Wildes. The structure of multiplicative motions in natural imagery. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence, 32(7):1310–1316, 2010.

[7] A. Efros, V. Isler, J. Shi, and M. Visontai. Seeing through water. In *Advances in Neural Information Processing Systems*, 2004.

[8] G. Egnal and R. Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1127–1133, 2002.

[9] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

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

[11] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *IEEE International Conference on Computer Vision*, pages 1849–1856, 2009.

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

[13] V. Kompella and P. Sturm. Detection and avoidance of semi-transparent obstacles using a collective-reward based approach. In *IEEE International Conference on Robotics and Automation*, 2011.

[14] K. Kutulakos and E. Steger. A theory of refractive and specular 3D shape by light-path triangulation. *International Journal of Computer Vision*, 76(1):13–29, 2008.

[15] A. Levin, A. Zomet, and Y. Weiss. Separating reflections from a single image using local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 306–313, 2004.

[16] C. Liu, L. Sharan, E. Adelson, and R. Rosenholtz. Exploring features in a Bayesian framework for material recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 239–246, 2010.

[17] H. Mallot, H. Bulthoff, J. Little, and S. Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological Cybernetics*, 64:177–185, 1991.

[18] K. McHenry and J. Ponce. A geodesic active contour framework for finding glass. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 1038–1044, 2006.

[19] K. McHenry, J. Ponce, and D. Forsyth. Finding glass. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 973–979, 2005.

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

[21] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The Princeton shape benchmark. In *International Conference on Shape Modeling and Applications*, 2004.

[22] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 246–253, 2000.