3D pose estimation and segmentation using specular cues

3D Pose Estimation and Segmentation using Specular Cues

Ju Yong Chang[1,*] Ramesh Raskar[2] Amit Agrawal[1]
[1]Mitsubishi Electric Research Labs (MERL), 201 Broadway, Cambridge, MA
[2]MIT Media Lab, 20 Ames St., Cambridge, MA
juyong.chang@gmail.com

Abstract

We present a system for fast model-based segmentation and 3D pose estimation of specular objects using appearance based specular features. We use observed (a) specular reflection and (b) specular flow as cues, which are matched against similar cues generated from a CAD model of the object in various poses. We avoid estimating 3D geometry or depths, which is difficult and unreliable for specular scenes. In the first method, the environment map of the scene is utilized to generate a database containing synthesized specular reflections of the object for densely sampled 3D poses. This database is compared with captured images of the scene at run time to locate and estimate the 3D pose of the object. In the second method, specular flows are generated for dense 3D poses as illumination invariant features and are matched to the specular flow of the scene.

We incorporate several practical heuristics such as use of saturated/highlight pixels for fast matching and normal selection to minimize the effects of inter-reflections and cluttered backgrounds. Despite its simplicity, our approach is effective in scenes with multiple specular objects, partial occlusions, inter-reflections, cluttered backgrounds and changes in ambient illumination. Experimental results demonstrate the effectiveness of our method for various synthetic and real objects.

1. Introduction

Consider the scene shown in Figure 1 which has multiple specular objects on a cluttered background, resulting in occlusions and inter-reflections. In this paper, we present a simple and yet effective and fast system to locate and estimate the 3D pose of specular objects in such scenes. Assuming a known CAD model of the object, we show that simple appearance/feature matching can surprisingly give fast and reliable segmentation and pose estimates for such challenging scenarios.

Model based 3D pose estimation is a classical vision problem and a variety of solutions based on feature correspondences, texture cues, and range data have been proposed. However, pose estimation and segmentation remains a challenging problem for specular objects. Model based pose estimation has been extensively studied for diffuse objects. Classical approaches attempt to match geometric 3D features on the object to 2D features from images to estimate the object pose, typically ignoring the illumination information. These techniques rely on texture or intensity cues in 2D images or video [15, 18, 21, 22, 23, 28, 32], where it is assumed that the texture is invariant against potential variations of the scene. However, this assumption does not hold if there are severe illumination changes or shadows. Moreover, textureless objects cannot be handled by these approaches.

Figure 1. Localization and pose estimation of specular objects is challenging in typical scenes as shown due to clutter, inter-reflections and partial occlusions.

Knowledge of accurate 3D or depth information can also help in several vision problems such as pose estimation. A straightforward approach for 3D pose estimation would be to estimate 3D/depths in the scene and match it with the object model to estimate the pose. Several range image based methods [4, 8, 10, 11, 12, 29, 30] have been proposed along those lines. However, for specular objects, 3D estimation is challenging, noisy, and non-robust.

Registering real images with synthetic images were proposed by Horn and Bachman [16, 17] for image understanding and automatic terrain classification, where a reflectance model was used to generate synthetic images. In contrast, we simply assume perfect mirror reflection for specular objects, although the actual BRDF may have a diffuse component. We analyze the effect of this assumption on pose

*Currently at Digital Media R&D Center in Samsung Electronics

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

estimation.

Recently, [3] proposed to use multiple monocular cues including polarization and synthesized CAD model images for 3D pose estimation. However, we only use image intensities and propose to use specular flow as another cue. Specular reflection has been used for 3D pose *refinement* [20], starting from a known coarse pose. Their approach uses both texture and specular cues and handles only a single object in the scene. However, absolute pose estimation and segmentation of specular objects is much more difficult than pose refinement. We handle multiple specular objects in cluttered environments. In addition, we specifically use a mirror sphere to obtain the environment map, which is used to render the synthetic images.

Note that the process of pose estimation indirectly gives segmentation (localization of the object). For matching real photos to synthetic images by rendering the specular object requires additional information about the illumination, which is often simplified and represented by the 2D *environment map*. In our second approach, the requirement of the environment map can be removed by matching *specular flows* as an illumination-invariant feature.

As computational power keeps on increasing, one can use simple, brute-force methods for challenging vision problems such as pose estimation. Our goal is to develop a simple, fast and practical system. We report recognition time of few seconds on commodity hardware (without using GPUs) by matching 25000 synthesized images. We propose practical heuristics such as the use of saturated/highlight pixels for fast matching and *normal selection* to minimize the effects of inter-reflections and cluttered backgrounds.

1.1. Benefits and limitations

We demonstrate that the proposed system handles challenging scenes with partial occlusions, background clutters, and inter-reflections. Our method does not require 3D/depths estimation of the target scene and is conceptually simple and easy to implement. The limitations of our approach are as follows.

- We require placing a calibration object (mirror sphere) in the target scene to capture the environment map.
- We require the environment map to be of sufficiently high frequency to induce variations in the synthesized images.
- The specular-flow based method requires specific motion of the environment to induce the specular flow.
- Planar surfaces and surfaces with low curvature cannot be handled.

1.2. Related work

3D pose refinement using specular reflections has been proposed in [20]. Given a short image sequence and initial pose estimates computed by the standard template matching, their approach first separate Lambertian and specular components for each frame and derive environment maps from the estimated specular images. The environment maps are then aligned in conjunction with image textures to increase the accuracy of the pose refinement. Similar to [20], our approach also exploits specular reflections for 3D pose estimation. However, we focus on absolute pose estimation and segmentation of specular objects in a scene rather than pose refinement. Our approach does not require the object to be textured to compute an initial pose estimate and can estimate the absolute 3D pose directly from specular reflection.

Specular surface reconstruction: A wide range of methods derive sparse 3D shape information from the identification and/or tracking of distorted reflections of light sources and special features [5, 6, 26, 31, 34]. Dense measurements can also be obtained based on the general framework of light-path triangulation as shown in [7, 19, 24]. However, these methods usually need to perform accurate calibration and control of environments surrounding the target object and require several input images. On the other hand, our method uses a simple mirror-like sphere as a calibration object and requires just two input images (or a HDR image). In addition, we do not need to estimate the 3D/depths in the scene.

Specular flow [27] refers to the flow induced by a small environmental motion on the image plane for a specular object. It can be utilized for specular surface reconstruction without any environment calibration as shown in [1, 33] for synthetic examples, or for detecting specular surfaces in image [9]. Surface reconstruction using specular flow requires a pair of linear PDE's to be solved using initial conditions, which are not easy to estimate in real situations. In addition, the accuracy of 3D reconstruction using specular flow has not been established for real scenes yet. We show that specular flow can also be used as a cue for 3D pose estimation. Since we avoid 3D reconstruction, we only require to generate the specular flows corresponding to different object poses for subsequent matching.

2. Problem statement

Given a scene consisting of several specular objects, our goal is to simultaneously locate and estimate the absolute 3D pose for a given object using its CAD model. We assume that the target object has perfect *mirror-like BRDF*, although in practice the actual BRDF may differ. The 3D pose is defined by a 3D translation vector (x, y, z) and rotation angles (θ, ϕ, σ). Additionally, it is assumed that the distance between the camera and the target object is approximately known ($z \approx z_0$). This is a reasonable assumption under many controlled applications. All equations in this paper are derived from the assumption of the orthographic projection for notational simplicity. But they can be easily

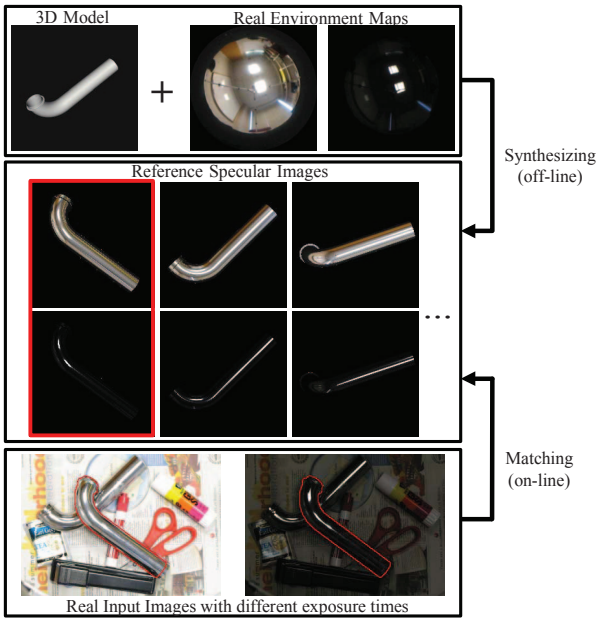


Figure 2. Overview of environment map based pose estimation.

generalized to the perspective case and we actually use the perspective projection for experiments.

We pursue a top-down approach, where we assume known high level information such as the object's geometric and photometric properties and/or illumination information, and then generate and utilize the low level features observable in 2D image to estimate 3D pose. Specifically, we render the given 3D model with the mirror-like BRDF and estimate the best 3D pose and location of the object, which makes the resultant synthetic specular features well matched to the features in the real input image. We use a brute-force matching strategy to obtain a coarse pose, which is further refined using optimization techniques. Now we describe two methods that utilize (a) rendered specular images and (b) specular flows as the cues for matching.

3. Environment map based pose estimation

In this method, we first measure the environmental illumination. In general, this information can be formulated by a 5D plenoptic function [2]. We assume that the target object is sufficiently far away from its surrounding environments, which simplifies it to a 2D *environment map*. Specifically, we put a small mirror-like sphere in the target scene and use its image as the 2D environment map. To handle wide dynamic range, we capture two environment maps, E_L and E_S at large and small exposure times respectively.

3.1. Generating synthetic specular images

Let the object surface $\mathbf{S}(x,y) = (x, y, f(x,y))$ be viewed orthographically from above and illuminated by a far-field environment E as in Figure 3. Let $\hat{\mathbf{r}} = (0,0,1)$ be the viewing direction, $\hat{\mathbf{n}}(x,y)$ the surface normal at surface

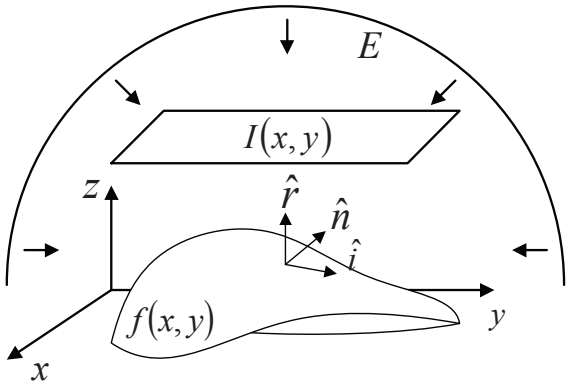


Figure 3. Image formation for a specular object.

point $(x, y, f(x,y))$, and $\hat{\mathbf{i}}(x,y)$ the direction of the incident light ray which can be represented as two spherical angles $\alpha(x,y)$ (elevation) and $\beta(x,y)$ (azimuth). Using the law of the reflection ($\hat{\mathbf{i}} = 2(\hat{\mathbf{n}} \cdot \hat{\mathbf{r}})\hat{\mathbf{n}} - \hat{\mathbf{r}}$), the spherical angles of the incident light ray in terms of surface derivatives are given by $\alpha = \tan^{-1}(\frac{2\|\nabla f\|}{1-\|\nabla f\|^2})$ and $\beta = \tan^{-1}(\frac{f_y}{f_x})$. Then the image of the specular surface is given by

$$\begin{aligned} I(x,y) &= E(\alpha, \beta) \\ &= E(\tan^{-1}(\tfrac{2\|\nabla f\|}{1-\|\nabla f\|^2}), \tan^{-1}(\tfrac{f_y}{f_x})) \end{aligned} . \quad (1)$$

The above equation can be used to generate reference specular images corresponding to pre-defined pose hypotheses, assuming known CAD model. The number of these pose hypotheses should be sufficiently large in order to cover large pose variations. We uniformly sample the rotation angles to generate $25000 - 50000$ reference images by environment mapping [13, 14]. Let $R^L_{\theta,\phi,\sigma}$ and $R^S_{\theta,\phi,\sigma}$ denote the synthetic specular images using E_L and E_S as the environment map respectively.

3.2. Matching for pose estimation

Let I_L and I_S denote the two input images captured using the same exposure times used for environment maps. We compare the input images with the reference specular images corresponding to densely sampled pose hypotheses and then obtain a coarse pose estimate as the best matched pose hypothesis. We propose following two heuristics to improve the speed and reliability of matching.

Using highlights/saturated pixels: In general, the specular image consists of highlights (saturated pixels) due to bright light sources. For mirror-like or metallic objects, the highlight pixels are extracted by applying a simple threshold to the short exposure images. The resulting binary images are referred to as highlight images. Let D_I and D_R refer to the distance transform of the highlight images corresponding to the input and reference highlight images. We use the highlight pixels for fast localization/pose estimation

by minimizing

$$C_{HL}(\theta,\phi,\sigma,x,y) = \frac{\sum_{(u,v)}(D_I(u,v) - D_R^{\theta,\phi,\sigma}(u-x,v-y))^2}{N_{HL}},$$

where (u,v) is pixel coordinate and N_{HL} denotes the number of highlight pixels.

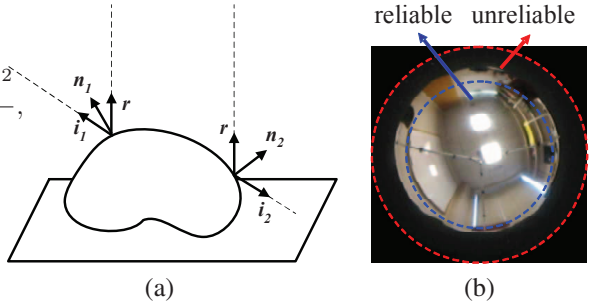
The highlight based cost function has several advantages. Firstly, the highlights are usually sparse in the input image, so they can be used as a strong constraint for restricting the object's location. Secondly, distance transform makes the cost distribution smoother. Finally, since the stencil of the highlights contains a very small number of pixels, computing the above cost function can be done efficiently. In our experiments, the proposed highlights based cost function converges well to a global minimum rapidly.

Normal selection for geometrically reliable pixels: To account for inter-reflections and background clutter, we propose a normal selection procedure to use only geometrically reliable pixels by avoiding illumination directions corresponding to small elevation angles. Our geometric stencil selection is as follows. First, we compute the incident light ray direction $\hat{\mathbf{i}}$ for each pixel of the reference image using the law of reflection and known surface normal for the given pose. Then, the *reliability* of the pixel information is defined by considering the illumination direction as shown in Figure 4. Illumination directions corresponding to small elevation angles are usually less reliable because of inter-reflections between the specular object and its neighbors. We use only those specular pixels for which incident light rays have the elevation angles larger than 90 degrees. We define a second cost function based on geometrically reliable pixels

$$C_{GR}(\theta,\phi,\sigma,x,y) = 1 - g(I_L(u,v), R^L_{\theta,\phi,\sigma}(u-x,v-y)),$$

where $g()$ denotes the normalized cross correlation (NCC) function. Although it seems natural to use the object's segmentation mask as the stencil for NCC computation, we found that using only geometrically reliable specular pixels as the stencil produces better results in practice.

Coarse pose estimation: The best match among the reference images is found in two steps. In the first step, for each rotational pose, the best translation in the image plane with its associated cost is obtained using only the highlight pixels based cost function. For this translation optimization, we use the downhill simplex algorithm [25]. As the initial points for the downhill simplex algorithm, we use the three corner points of the input image. Then the translation is refined by performing optimization considering all geometrically reliable pixels. Once we have the optimal translations and cost values for each rotation, we compare these cost values and choose the rotation with the minimum cost. We refer to the obtained pose as *coarse pose*, since it depends on the discretization of the database (number of reference images).



(a) (b)

Figure 4. Reliability computation is illustrated in (a). Illuminations from $\mathbf{i}_1$ and $\mathbf{i}_2$ are considered as reliable and unreliable, respectively. Examples of reliable and unreliable pixels for a mirror sphere are shown in (b).

Pose refinement: Since the above pose estimate is obtained by matching the reference images, it is accurate only up to the discretization of the database. The estimated 3D pose is further refined by optimizing over all five pose parameters using a steepest descent algorithm, where the gradient at each step is computed numerically. We minimize the reliable pixels based cost function and initialize the pose parameters as the coarse pose estimate obtained above.

4. Specular flow based pose estimation

In this method, we utilize *specular flow* [27] as features for matching, which is defined as the optical flow induced by the camera or scene/object motion in the images for specular reflection. While previously specular flow has been used for 3D shape reconstruction [1], we propose to use it for 3D pose estimation. Similarly to [1], we keep the relative pose between the camera and object fixed, and only assume environmental motion. We capture two images of the target scene under pre-defined rotation of environment around known direction (e.g. camera's viewing direction). We use a simple block matching algorithm to obtain the 2D displacement vectors for each pixel.

Note that since the relative pose between the camera and the scene is fixed, optical flow is mainly observed on specular objects due to illumination change. Therefore, in a cluttered scene, this motion cue can be used for strongly constraining the specular object's location, similar to the highlights in the environment map based method.

4.1. Generating reference specular flows

The angular motion of far-field environment can be represented as a vector field $\omega(\alpha,\beta) = (\frac{d\alpha}{dt}, \frac{d\beta}{dt})$ on the unit sphere. This environment motion induces a specular flow $\mathbf{u}(x,y) = (\frac{dx}{dt}, \frac{dy}{dt})$ on the image plane. This flow is related to the environment motion through the Jacobian $\mathbf{J}$ and can be written as

$$\mathbf{u} = \mathbf{J}^{-1}\omega, \tag{2}$$

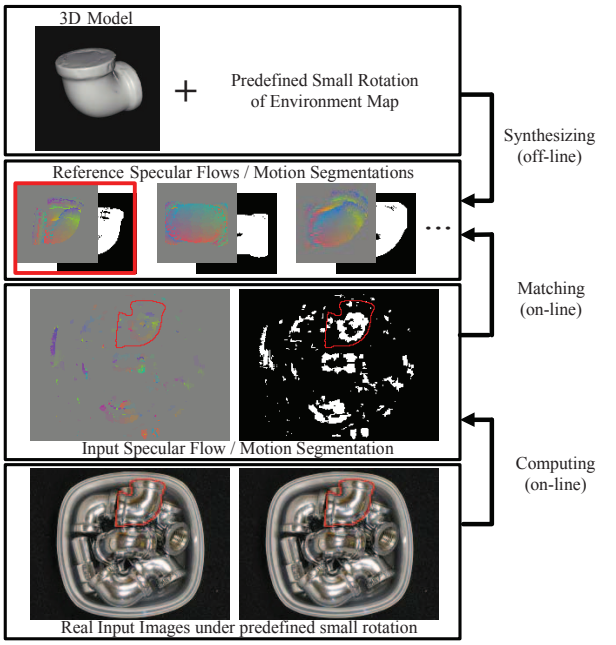


Figure 5. Overview of specular flow based pose estimation.

where the Jacobian can be expressed in terms of surface shape as

$$\mathbf{J} \triangleq \frac{\partial(\alpha,\beta)}{\partial(x,y)} = \begin{pmatrix} \frac{f_x f_{xx}+f_y f_{xy}}{\|\nabla f\|\cdot(1+\|\nabla f\|^2)} & \frac{f_x f_{xy}+f_y f_{yy}}{\|\nabla f\|\cdot(1+\|\nabla f\|^2)} \\ \frac{f_x f_{xy}-f_y f_{xx}}{2\|\nabla f\|^2} & \frac{f_x f_{yy}-f_y f_{xy}}{2\|\nabla f\|^2} \end{pmatrix}. \quad (3)$$

This equation can be used for generating reference specular flows corresponding to densely sampled pose hypotheses. Let the reference specular flow image synthesized from orientation (θ, ϕ, σ) be denoted by $R_{\theta,\phi,\sigma}$.

The specular flow does not depend on the illumination information but only on the motion and the object's shape and pose. Therefore, under the assumption that the motion and the object's shape are given, it can be used as the *illumination-invariant feature* for pose estimation. Note that the determinant of $\mathbf{J}$ can be written as $\det(\mathbf{J}) = \frac{2K(1+\|\nabla f\|^2)}{\|\nabla f\|}$, where K is the Gaussian curvature of the surface. Thus, planar surfaces and surfaces with low curvatures cannot be handled by this method. In addition, for the specular flow to be estimated reliably, the environment should have sufficiently high frequency variations.

4.2. Matching specular flows

Similarly to the previous method, the reference specular flows are compared with the input specular flow I to estimate a coarse pose as follows:

Using motion segmentation for fast matching: We define a motion segmentation image as the binary image indicating the presence/absence of specular flow. As discussed, since the relative pose of the camera and object is fixed, the motion segmentation image gives strong cues for location of specular objects, similar to saturated pixels in environment map based approach. Thus, a fast location search can be done using motion segmentation image.

Again, let D_I and D_R denote the distance transformation of motion segmentation images for scene specular flow and the reference specular flow $R_{\theta,\phi,\sigma}$ respectively. A motion segmentation based cost function is defined as

$$C_{MS}(\theta,\phi,\sigma,x,y) = \frac{\sum_{(u,v)}(D_I(u,v) - D_R^{\theta,\phi,\sigma}(u-x,v-y))^2}{N_{MS}},$$

where the summation is carried out for motion segmentation pixels of $R_{\theta,\phi,\sigma}$ and N_{MS} denotes the number of such pixels.

Using specular flow: We define a matching error between the input specular flow $I(u,v)$ with the translated reference specular flow $R_{\theta,\phi,\sigma}(u-x, v-y)$. In reality, specular flow contains many outliers, so simple cost functions such as sum of squared differences (SSD) does not work well. Instead, we use a robust cost function based on the number of inlier pixels. First, we define the inlier pixels as ones where the difference between the input specular flow vector $I(u,v)$ and the reference specular flow vector $R_{\theta,\phi,\sigma}(u-x, v-y)$ is less than a small threshold (1.0 in our experiments). Then, the matching cost function C_{SF} is defined as

$$C_{SF}(\theta,\phi,\sigma,x,y) = -|\mathbf{M}|,$$

where $\mathbf{M}$ is the set of inlier pixels.

Coarse pose estimation: First, translation (x,y) is optimized for each rotation (θ,ϕ,σ) by using the downhill simplex algorithm and motion segmentation based cost function, initialized by three corner points of the input image. Then translation is refined by minimizing C_{SF}. Finally, by comparing best costs from all translation optimized poses, the best rotation values are chosen.

Pose refinement: Using the above coarse pose as the initial starting pose, we refine all five DOF parameters for pose by minimizing the SSD cost function.

5. Experimental results

In this section, we present the results of both approaches on various synthetic and real objects. All experiments have been performed on a standard PC with 2.66 GHz Intel quadcore CPU and 3.25 GB RAM. Before pose estimation, reference synthetic images in Section 3 or reference specular flows in Section 4 are synthesized (using OpenGL) and stored *off-line*. The resolutions of the input images and the reference images are 400×300 and 100×100, respectively.

For the environment map based method using 25000 reference images, on average our approach takes 2.26 and 0.44 seconds for coarse pose estimation and fine pose refinement respectively. The corresponding numbers for specular flow based approach are 32.99 and 0.49 seconds respectively. Note that the computation time is dominated

	x (mm)	y (mm)	θ (°)	ϕ (°)	σ (°)
Env. Map	0.29	0.39	2.12	1.64	1.69
Spec. Flow	0.46	0.39	3.64	2.85	2.06

Table 1. Average pose errors for successful pose estimates (maximum error less than 1 mm and 10° for translation and rotation respectively).

by coarse pose estimation which utilizes brute-force matching and the downhill simplex algorithm. We believe that this process can be parallelized by using modern graphics processing units (GPUs) similarly to [12], and could be reduced.

5.1. Synthetic objects

For the quantitative evaluation of our methods, we performed experiments using a synthetic object, but with a real environment map captured using a mirror ball. The reference synthetic images are generated using the known CAD model and the environment map, assuming perfect mirror-like BRDF. To evaluate the pose estimation accuracy, we generate a noisy input image using a random pose and added Gaussian noise. For the specular flow based method, two input images are generated by assuming small rotation (5°) of the environment map. The input specular flows are similarly generated using a random pose for the object.

We compare the results with ground truth pose parameters. The resultant pose is regarded as *success* if the maximum pose errors are less than 1 mm and 10° for translation and rotation angles respectively. The average pose errors for successful pose estimates after 50 trials for both methods are shown in Table 1. The environment map based method appears to be more accurate than the specular flow based method. Horn and Bachman [16, 17] argues that using pixel intensities instead of first extracting features is better for matching real images to synthetic images. Our results also show that the environment map based method is more accurate than the specular flow based method.

5.2. Robustness analysis

Note that we made simplifying assumptions on BRDF (mirror-like) as well as assumed that exact geometry and environment map is available. In practice, however, the BRDF could consist of diffuse component as well as a specular lobe. In addition, the CAD model may not be accurate and the changes in ambient illumination could effect the environment map. We investigate how these deviations degrade the performance of our system.

We first generate reference synthetic images/specular flows from noiseless 3D model assuming perfect specular reflection and known environment map. We then apply both techniques to input images synthesized by assuming variations in (a) object geometry (b) object reflectance, and (c) environment map. The resultant success rates and mean/variance of rotational pose errors are shown in Table 2. For the geometric and photometric variations, we added noise to the normal vectors of the 3D model and a small diffuse component to the specular reflection, respectively. The environment map variations are summarized in Figure 6. Note that these variations are handled well by the environment map approach. We can see that both methods are robust to variations except the geometric variation in the specular flow based method. This is because the specular flow depends on the first and second derivatives of the surface function, so it is highly sensitive to noise in the 3D model.

Env. Map	(%)	θ (°)		ϕ (°)		σ (°)	
Geometric	94	2.59	4.73	1.80	2.08	1.95	2.25
Photometric	94	2.59	5.26	2.15	4.33	2.07	3.41
Environment	95	2.36	4.29	1.86	2.68	1.87	2.51
Spec. Flow	(%)	θ (°)		ϕ (°)		σ (°)	
Geometric	44	5.01	6.86	2.66	2.72	3.38	3.36
Photometric	80	3.28	5.43	2.40	2.59	2.01	3.22

Table 2. Success rates and rotational pose errors (mean/variance) for various deviations from the ideal case for environment map based method and specular flow based method, respectively.

(a) (b) (c) (d)

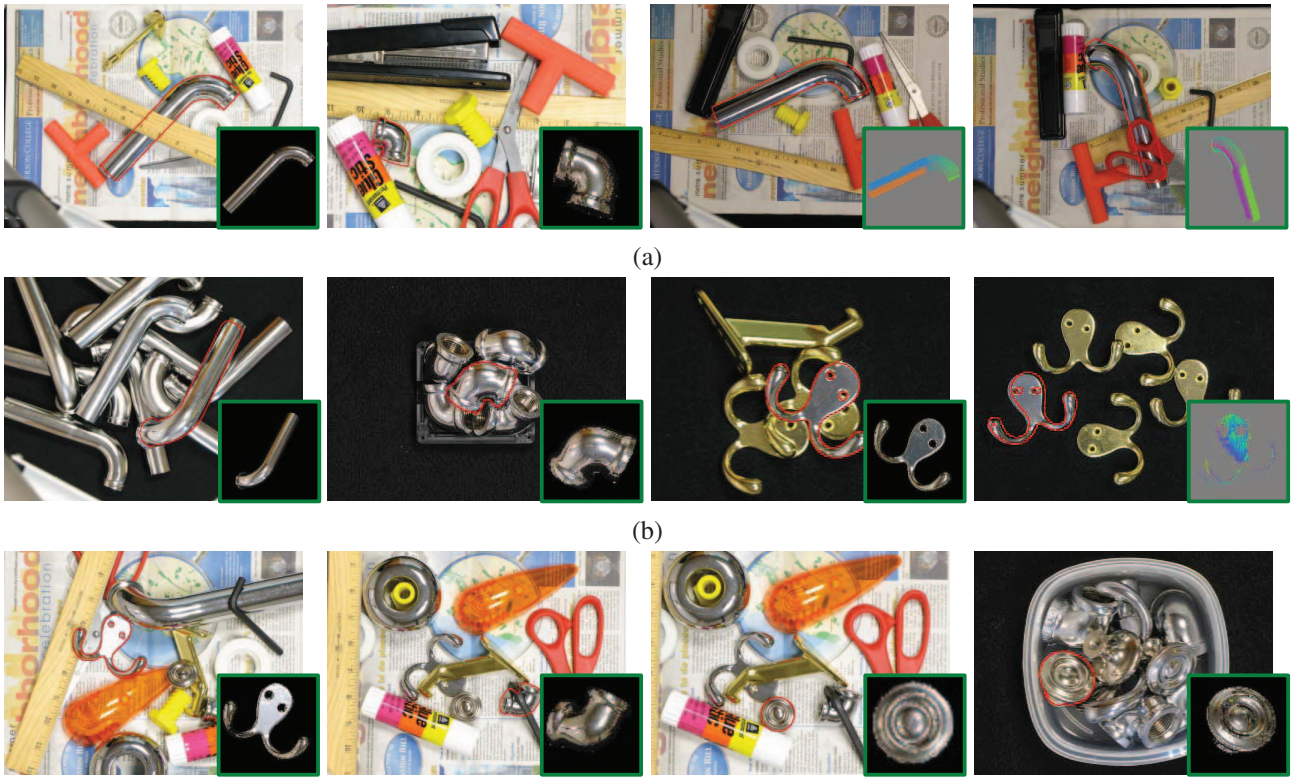
Figure 6. Variations of the environment map. (a) Using half of original exposure time. (b) Adding people and opening door. (c) Opening window. (d) Translating mirror to side for environment map capture.

5.3. Real objects

We performed real experiments using specular objects in cluttered environments along with multiple diffuse/specular objects. To obtain the CAD model of the specular object, we spray-paint the object to make it diffuse and use a commercial laser scanner. Note that this will result in an imperfect CAD model. We applied our techniques to various real scenes including partial occlusions, cluttered backgrounds, and multiple specular objects with inter-reflections. Since the ground truth 3D pose is not available, qualitative evaluation is performed by overlaying the rendered image synthesized using the estimated pose on the input images.

Examples of real experimental results using both methods are illustrated in Figures 2, 5, and 7. Figures 2 and 7 show that the environment map based method works well as the rendered silhouettes are well matched with the input images. In Figure 5, the computed input specular flow seems to be noisy and sparse because of textureless regions in the input image. Nonetheless, the proposed specular flow based method detects the target object with correct pose.

Cluttered background: Object detection and pose es-



(a) (b) (c)

Figure 7. The environment map based method and the specular flow based method are applied to various real scenes containing (a) cluttered background, (b) multiple specular objects of same kind, and (c) specular objects of different kind. Rendered synthetic silhouettes in estimated pose are overlaid on input images. The corresponding reference specular images or specular flows are displayed in corner.

(a) (b)

Figure 8. Failure cases for pose estimation approaches. (a) Since discriminative features are not visible on the end of pipe, estimated pose has large out of plane rotation error. (b) Specular reflections on a different object having similar shape lead to an incorrect match.

timation is usually difficult in cluttered backgrounds due to several local minima/longer convergence time. We handle this effectively using the highlights/motion segmentation information. Notice that this sparse binary information efficiently constrains the pose of the specular object in Figures 2 and 5. Other examples in Figure 7(a) show that the proposed approach is robust to the cluttered backgrounds.

Multiple specular objects of same kind make the segmentation and 3D pose estimation problem challenging due to severe inter-reflections between them. Our approach works reasonably well for this case as illustrated in Figure 7(b). We found that our technique for obtaining geometrically reliable pixels in Section 3.2 plays an important role in this case by excluding inter-reflection pixels.

Mixed specular objects of different kind can be also handled by our approach as illustrated in Figure 7(c). Complex specular reflections from different kind of objects are observed in this case. This makes the pose discrimination more ambiguous especially for the sparse information such as highlights or motion segmentation. We can see that our approach resolves this difficulty and detects different kind of objects in the same scene.

Failure cases: Figure 8 illustrates typical failure cases. In Figure 8(a), the estimated pose has out of plane rotational error along the main axis of the long pipe. Typically, in-plane rotation and translation has better estimation accuracy compared to out of plane rotation and depth estimate. In Figure 8(b), pose estimate is located on different specular object due to similar specular reflections.

6. Discussions

We demonstrated that specular reflections can be used for localization and pose estimation of specular objects using a known CAD model. We showed that simple feature/intensity matching can surprisingly handle textureless and highly specular objects in challenging scenes with clut-

tered background, inter-reflections and partial occlusions. Our approach uses monocular camera images and does not require 3D scanning of the target scene. The proposed approach uses simple matching cost functions and optimization algorithms, and is fast and easy to implement. Fast nearest neighbor search using k-d trees and dimensionality reduction algorithms can further reduce the computational cost of our approach.

Apparent limitation of our approach is that both proposed methods require specific assumptions such as sparse highlights have to be observed in the input image or motion has to be restricted to only the small environmental motion. Removing these assumptions and extending our approach to handle more general (e.g. partially diffuse and specular) or challenging (e.g. translucent) objects is an interesting future work.

Acknowledgements We thank the anonymous reviewers and several members of MERL for helpful suggestions. We also thank Jay Thornton, Keisuke Kojima, John Barnwell and Haruhisa Okuda, Mitsubishi Electric, Japan for help and support.

References

[1] Y. Adato, Y. Vasilyev, O. Ben-Shahar, and T. Zickler. Toward a theory of shape from specular flow. In *Proc. Int'l Conf. Computer Vision*, 2007.

[2] E. Adelson and J. Bergen. *Computational Models for Visual Processing*. MIT Press, 1991.

[3] B. Barrois and C. Wohler. 3d pose estimation based on multiple monocular cues. In *BenCOS*, 2007.

[4] P. Besl and N. McKay. A method for registration of 3d shapes. *IEEE Trans. Pattern Anal. Machine Intell.*, 1992.

[5] A. Blake. Specular stereo. In *Proc. Int'l Joint Conf. on Artificial Intelligence*, volume 2, pages 973–976, 1985.

[6] A. Blake and G. Brelstaff. Geometry from specularities. In *Proc. Int'l Conf. Computer Vision*, pages 394–403, 1988.

[7] T. Bonfort, P. Sturm, and P. Gargallo. General specular surface triangulation. In *Proc. Asian Conf. Computer Vision*, pages 872–881, 2006.

[8] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *Robotics and Automation*, 1991.

[9] A. DelPozo and S. Savarese. Detecting specular surfaces on natural images. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.

[10] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy. Geometrically stable sampling for the icp algorithm. In *3DIM*, 2003.

[11] N. Gelfand, N. Mitra, L. Guibas, and H. Pottmann. Robust global registration. In *Eurographics Symposium on Geometry Processing*, 2005.

[12] M. Germann, M. D. Breitenstein, I. K. Park, and H. Pfister. Automatic pose estimation for range images on the gpu. In *3DIM*, 2007.

[13] N. Greene. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications*, 6(11):21–29, 1986.

[14] P. Haeberli and M. Segal. Texture mapping as a fundamental drawing primitive. In *Fourth Eurographics Workshop on Rendering*, pages 259–266, 1993.

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[16] B. Horn and B. Bachman. Using synthetic images to register real images with surface models. *Communications of the A.C.M.*, 21(11):914–924, 1978.

[17] B. Horn and B. Bachman. Registering real images using synthetic images. In *Artificial Intelligence: An MIT Perspective*, volume 2, pages 129–159, 1979.

[18] M. J. Jones and P. Viola. Fast multi-view face detection. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2003.

[19] K. N. Kutulakos and E. Steger. A theory of refractive and specular 3d shape by light-path triangulation. In *Proc. Int'l Conf. Computer Vision*, pages 1448–1455, 2005.

[20] P. Lagger, M. Salzmann, V. Lepetit, and P. Fua. 3d pose refinement from reflections. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.

[21] V. Lepetit and P. Fua. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005.

[22] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[23] D. G. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Trans. Pattern Anal. Machine Intell.*, 13:441–450, 1991.

[24] D. Nehab, T. Weyrich, and S. Rusinkiewicz. Dense 3d reconstruction from specularity consistency. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.

[25] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[26] M. Oren and S. K. Nayar. A theory of specular surface geometry. *Int'l J. Computer Vision*, 24(2):105–124, 1997.

[27] S. Roth and M. Black. Specular flow and the recovery of surface structure. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 1869–1876, 2006.

[28] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2003.

[29] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. *ACM Transactions on Graphics (TOG)*, 21(3):438–446, 2002.

[30] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3DIM*, 2001.

[31] S. Savarese, M. Chen, and P. Perona. Local shape from mirror reflections. *Int'l J. Computer Vision*, 64(1):31–67, 2005.

[32] C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. In *Proc. Conf. Computer Vision and Pattern Recognition*, 1996.

[33] Y. Vasilyev, Y. Adato, T. Zickler, and O. Ben-Shahar. Dense specular shape from multiple specular flows. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2008.

[34] P. Zisserman, A. Giblin, and A. Blake. The information available to a moving observer from specularities. *Image and Vision Computing*, 7(1):38–42, 1989.