Depth Estimation of Transparent Objects

Yunho Choi
2017-15648
Department of Mechanical Engineering

Abstract

Research on object manipulation using robots has been steadily processing. Dealing with a transparent object, however, is a challenge for computer vision process due to the difficulties of obtaining semantic and geometric information caused by its reflective and see-through nature. In order to tackle this issue, we first modified a prior method Dex-NerF[1] to enhance time efficiency. Then, we introduce TeRF(Thermal Radiance Field for Rendering Depth in Transparent Objects), performing depth estimation by applying thermal images to radiance field synthesis for robust object depth estimation of multiple transparent objects.

1 Introduction

Object manipulation using a robot equipped with sensor system is being used in various places. Generally, grasping operation of a robot firstly estimate the depth of objects based on the information of RGBD sensors. However, transparent objects such as a plastic bottle or glass rarely appear clearly on RGBD sensors.

It is challenging to apply major tasks of computer vision techniques on thermal objects. Because of their reflective and see-through nature, conventional techniques fail to precisely recognize geometry or classify what an object is.

To tackle this problem, we utilized thermal images. Different from RGB images where sensors sense rays in visible section, thermal cameras capture infrared rays. The infrared ray cannot penetrate the surface of transparent objects as well as non-transparent objects', so thermal images contain more accurate geometric information of transparent objects than RGB-colored images.

By applying this concept to Plenoxel[2], which is currently one of the popular methods to render novel views by multiple view from different locations and directions, to estimate 3D geometry of an object. We are inspired by [1], which estimates depth using NeRF[3].

2 Related Work

2.1 Detecting transparent Objects

The most recent approaches detecting and recognizing transparent objects are data driven[1]. That is, large datasets are needed to train neural networks. ClearGrasp[4] is one of typical example of this method. The network proposed is trained by the set of RGB image, input depth image, and true depth image.

Dex-NerF[1] however, based on neural radiance fields(NeRF)[3], doesn't need any set of paired datasets. Rather it directly utilize RGB images captured from various views of the transparent object, along with locations and directions, to train neural radience field. Since Dex-NerF use the Vanilla-NerF, training neural radiance field takes a considerable amount of time, so it is not desirable in real-time situation where we have to immediately obtain 3-dimentional geometric information of objects.

2.2 Radiance Fields

Figure 1: Neural Radiance Fields

Recently, implicit neural representations have led to significant progress in 3D object shape representation and encoding the geometry and appearance of 3D scenes. Mildenhall et al. [3] presented Neural Radiance Fields (NeRF), a neural network whose input is a 3D coordinate with an associated view direction, and output is the volume density and view-dependent emitted radiance at that coordinate. Due to its view-dependent emitted radiance prediction, NeRF can be used to represent non-Lambertian effects such as specularities and reflections, and therefore capture the geometry of transparent objects.

However, NeRF is slow to train and has low data efficiency. Keil et al. proposed Plenoxel[2], representing a scene as a sparse 3D grid with spherical harmonics. This representation can be optimized from calibrated images via gradient methods and regularization without any neural components. Different to Dex-NerF[1], we use Plenoxel as the background radiance field generator for TeRF.

3 Method And Experiment

3.1 Depth Estimation from Radiance Fields

Radiance fields learns a neural scene representation that maps a 5D coordinate containing a spatial location (x, y, z) and viewing direction (θ, ϕ) to the volume density σ and RGB color $\mathbf{c}$. The expected color $C(\mathbf{r})$ of the camera ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ between near and far scene bounds t_n and t_f is:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt \tag{1}$$

where $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$ is the probability that the camera ray travels from near bound t_n to point t without hitting any surface. The discrete computation of the expected color $\hat{C}(\mathbf{r})$ as:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i \tag{2}$$

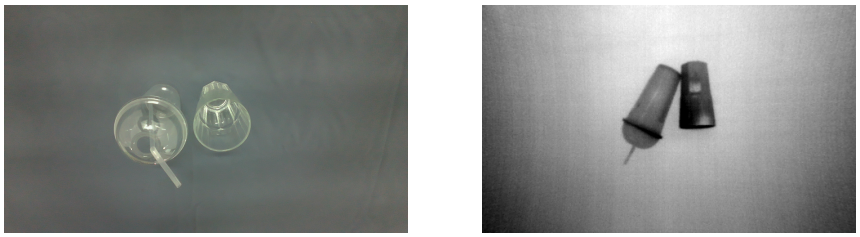
where $T(t) = \exp\left(-\Sigma_{j=1}^{i-1}\sigma_j\delta_j\right)$ and $\delta_i = t_{i+1} - t_i$ is the distance between consecutive samples on the ray r.

Plenoxel reconstruction just as same as NeRF's converts σ_i to an occupancy probability α_i. It then applies the transformation $w_i = \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j)$. The original depth estimation suggested by Wang et al. [3] compute the depth at pixel coordinate $[u, v]$ as $D[u, v] = \sum_{i=1}^{N} w_i\delta_i$. However, this results in noisy depth maps.

Inspired by Ichnowski et al. [1], we consider transparency-aware method that searches for the first sample along the ray for which $\sigma_i > m$ for some fixed threshold m. The depth is then set to the distance of that sample δ_i. For our experiment, we set $m = 15$.

(a) Input RGB image (b) Depth map (c) Drawback of RGB image

Figure 2: **Depth Estimation using only RGB images** (a) We use 40 input RGB images from different views to train Plenoxel. (b) The acquired depth map. (c) The expansion of the red rectangle area of (b) We can see that the depth of the surface of the bottle is poorly estimated.



(a) RGB images (b) Thermal images

Figure 3: RGB images and thermal images regarding on transparent objects

Figure 2 shows the result of applying transparency-aware method on Plenoxel to obtain depth map. Here we can see that the acquired depth map(Figure 2(b)) accurately estimate the depth in regard to objects' boundaries. The performance is similar to that of Dex-NeRF[1], but our method took only 10-15 minutes whereas the original took 10-16 hours thanks to the efficiency of Plenoxel[2].

Nevertheless, we can find drawback from the result shown in Figure 2(c). The depth at the center of the bottle is not properly estimated as we can see some holes in that area. We argue that this is due to the reflectiveness and transparency of the bottle's surface, hindering radiance fields from properly learn 3D geometry.

3.2 Using Thermal Images as Input

In order to solve the issue mentioned in 3.1, we substitute RGB images with Thermal images. As we can see in Figure 3, thermal images are more superior to representing accurate geometry of transparent objects; that is, features of edges and surface of transparent objects are more conspicuous on thermal images.

We first directly replaced RGB images with thermal images and train Plenoxel without any modification of its structure or parameter values. Because thermal images show the temperature, and because the temperature difference between the object and the background is small, the performance deteriorates. Therefore, to increase the temperature difference, we normalized them based on the section where the temperature is concentrated. We performed depth rendering by optimizing temperature and density using camera pose and thermal images.

We coupled a thermal camera to robotic arm to acquire data sets of transparent objects in various environments, such as one or more objects in a low-light environment. Figure 4 shows the results

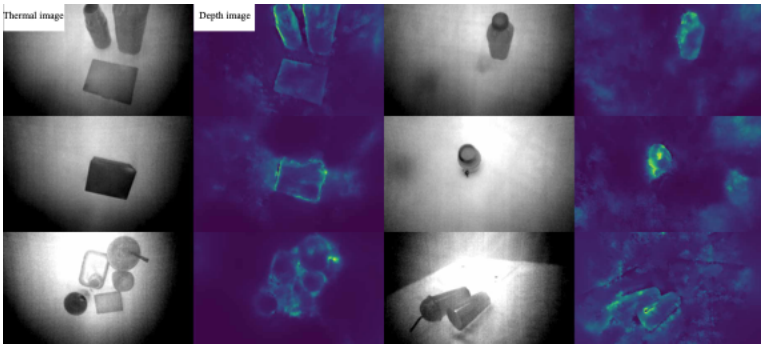


Figure 4: TeRF estimates depth for various transparent objects made of plastic, glass. The left images show the normalized thermal image, and the right images show the depth rendering from the image view.

of depth estimation using TeRF on 5 datasets. We can see that depth has been estimated in some extent, yet the quality is not sufficient to fulfill our goal. The research on TeRF is still on progress. Nevertheless, we have shown the possibility that thermal images can become a key to estimate depth of an object where RGB images are no longer useful.

4 Conclusion

We have introduced two new methods to estimate depth of transparent objects. By modifying Dex-NeRF with Plenoxel, we increase the time efficiency significantly. TeRF is now in early stage, so further development is needed to improve the quality of depth estimation using thermal images. In addition, camera calibration should also be considered in future to properly match the acquired depth to real world coordinates. Though we mainly focused on detecting transparent objects, TeRF can also be applied in circumstances where there is scarce light(RGB images hardly show information of objects) to estimate objects' depth.

Not only depth estimation, but also classification using thermal images is under research procedure to fulfill our goal to perform robust manipulation of robots.

References

[1] J. Ichnowski*, Y. Avigal*, J. Kerr, and K. Goldberg, "Dex-NeRF: Using a neural radiance field to grasp transparent objects," in *Conference on Robot Learning (CoRL)*, 2020.

[2] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5501–5510.

[3] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "Nerf–: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[4] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642.