Viewpoint Recommendation for Object Pose Estimation via Pose Ambiguity Minimization

Nik Mohd Zarifie HASHIM[1,3,a)] Yasutomo KAWANISHI[1,b)]
Daisuke DEGUCHI[2,c)] Ichiro IDE[1,d)] Hiroshi MURASE[1,e)]

Abstract

Recently, helper robots become popular in our social life, especially for helping the elderly and the disabled people to perform their daily tasks at home. To handling objects, object pose estimation from a depth image is an essential task of the helper robots. However, an object's pose is often ambiguous from an observation from only a single viewpoint. If we can observe the object from additional viewpoints, the pose estimation result will be better. Thus, we propose a next viewpoint recommendation method based on pose ambiguity minimization. We confirmed and showed the proposed method outperformed other comparative methods on synthetic object images.

1. Introduction

Object pose estimation has recently become one of the focussed topics in the machine vision field for application on tasks such as object picking by a robot. Primarily, object picking is an essential task for home helper robots and industrial robots. For observing the surroundings of a robot, it is usually equipped with several sensors such as RGB cameras and Depth cameras. In this paper, we focus on depth images captured by a depth camera and utilize them for estimating an object's pose.

Among techniques for pose estimation from depth images, the simplest approach is estimating an object's pose from a single depth image captured from a certain viewpoint. If an object has a distinct shape to be distinguished from various viewpoints, the object's pose estimation would be easy. However, most objects have good and bad viewpoints for their pose estimation. We define "pose ambiguity" as how difficult to estimate the object's pose is. High pose ambiguity leads to inaccurate pose estimation.

To the extent of our knowledge, the techniques for object pose estimation are divided into two; estimating an object's pose from a single image known as the single viewpoint pose estimation, and from images captured from multiple viewpoints known as the multiple viewpoints pose estimation.

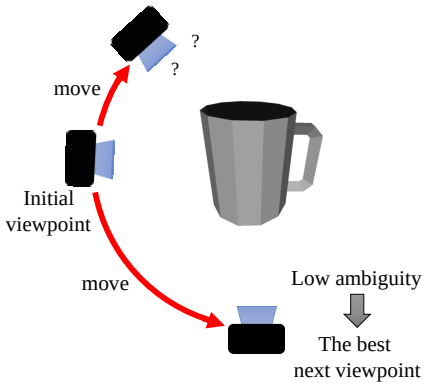

Fig. 1 Recommendation of the best next viewpoint.

Chin et al. [1] proposed a template matching method for the single viewpoint pose estimation. To reduce the number of templates, Murase and Nayer [2] proposed the Parametric Eigenspace method. However, in their work, images with similar appearances may be embedded to similar points in a low-dimensional subspace, which makes it difficult to distinguish a pose accurately. Recently, Ninomiya et al. [3] proposed a supervised feature extraction method for embedding images into a deep feature manifold. They modified DCNNs [4] for object pose estimation, named Pose-CyclicR-Net, which can accurately handle an object's rotation by describing the rotation angle using trigonometric functions.

In general, object pose estimation from a single viewpoint faces the problem of inaccurate pose estimation due to the pose ambiguity issue; namely, an object may have some poses which look similar and hard to be distinguished.

There are several work for the multiple viewpoint pose estimation, such as those by Zeng et al. [5] and Kanezaki et al. [6]. As such, there are several work for object pose estimation from multiple viewpoints, however these methods do not consider which viewpoint is effective for the estimation. Recently, some work focuses on predicting next-best-view for object pose estimation. Doumanoglou et al. [7] and Sock et al. [8] proposed next-best-view prediction methods for multiple object pose estimation based on Hough Forest [9]. However, we acknowledge that [7] and [8] could not be applied for the category-level object pose estimation since

[1] Graduate School of Informatics, Nagoya University, Japan
[2] Information Strategy Office, Nagoya University, Japan
[3] Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka, Malaysia
a) hashimz@murase.m.is.nagoya-u.ac.jp
b) kawanishi@i.nagoya-u.ac.jp
c) ddeguchi@nagoya-u.jp
d) ide@i.nagoya-u.ac.jp
e) murase@i.nagoya-u.ac.jp

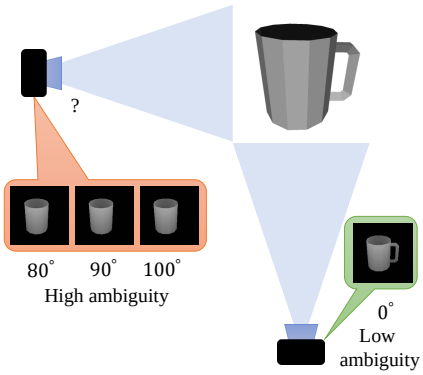

Fig. 2 Viewpoint selection problem in pose estimation from multiple viewpoints.

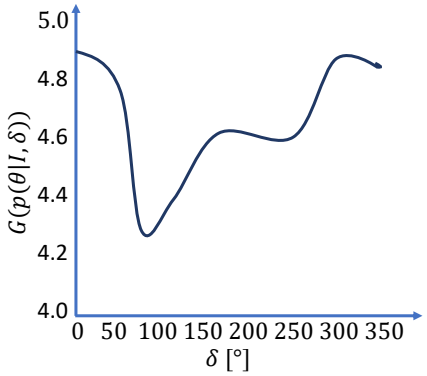

Fig. 3 Pose ambiguity minimization (Input image with 90° rotation angle).

they are designed only for instance-level object pose estimation. As the pose estimation on category-level has not been studied in the past, we initiated the study with the proposed method.

In this paper, we propose a method for estimating the next viewpoint, where the pose ambiguity will be minimized. In the multiple viewpoint pose estimation, to estimate the object's pose accurately, it is necessary to choose the best set of viewpoints. Here, given an observation of an object, we consider how to select one more viewpoint to observe the object. We call the viewpoint the next best viewpoint. A better viewpoint helps us to obtain a more accurate object pose, as shown in Figure 1. It is easy to select the best viewpoint when we know the current viewpoint and the shape of the object accurately. However, if the pose estimation result from the current viewpoint is ambiguous, it is difficult to determine in which direction and how far the robot should move to reach the best next viewpoint. The question here is that, how can we know the best next viewpoint from the current observation as illustrated in Figure 2.

In this paper, we propose a method of viewpoint recommendation for accurate object pose estimation. To evaluate the effectiveness of a candidate viewpoint, we define a metric called "pose ambiguity", which reflects how ambiguous the pose estimation is. By finding the viewpoint where the pose ambiguity is the minimum, we can obtain the next viewpoint, which will be the best viewpoint to estimate the object's pose by combining with the current observation.

To make the problem simple and focus on the fundamental idea, in this paper, we limit the movement of the depth camera only to rotation around the z-axis of the target object. However, the proposed method and discussion could be straightforwardly extended to 3D rotation. We evaluate the effectiveness of the proposed method on dataset generated from a subset of publicly available 3D object dataset: ShapeNet [10].

Our contribution can be summarized as follows:

- We define a metric "pose ambiguity" to evaluate how difficult the pose estimation is.
- We propose a next viewpoint recommendation method which finds the best next viewpoint where the pose ambiguity is minimized.

The remaining of this paper is structured as follows: In Section 2, the proposed method will be introduced. After that, in Section 3, we will explain the evaluation setting. In Section 4, we will discuss the evaluation results. Finally, we conclude our paper in Section 5.

2. NEXT VIEWPOINT RECOMMENDATION

2.1 Overview

We define a metric called "pose ambiguity" given two different viewpoints which should be minimized. Since the current viewpoint may be ambiguous, by handling the current view point ϕ and the angle to the next best viewpoint from the current viewpoint δ as latent variables, the pose ambiguity function is decomposed into pose ambiguity under given two viewpoints and viewpoint ambiguity under a given observation.

2.2 Pose Ambiguity Minimization Framework

In this framework, the method measures the pose ambiguity in a quantitative way. We define the pose ambiguity G as a functional of the pose likelihood distribution $p(\theta)$. For example, G can be defined by the entropy of $p(\theta)$ as

$$G(p) = \int -p(\theta) \log p(\theta) d\theta. \quad (1)$$

Here, we evaluate the pose likelihood distribution under an image observed from the initial viewpoint, and then yield the rotation angle to the best next viewpoint. Therefore, we define the pose likelihood distribution as a conditional distribution $p(\theta|I,\delta)$ when an image I from the current viewpoint and a rotation angle δ are given. By using the formulation, we find the best viewpoint by minimizing the entropy as

$$\hat{\delta} = \arg\min_{\delta} G(p(\theta|I,\delta)). \quad (2)$$

An example of $G(p(\theta|I,\delta))$ in terms of all δ is illustrated in Figure 3.

To handle the ambiguity of the initial viewpoint, we further decompose the pose likelihood distribution as follows:

$$p(\theta|I,\delta) = \int p(\theta|\phi,\delta) p(\phi|I) d\phi. \quad (3)$$

The first term $p(\theta|\phi,\delta)$ indicates the pose likelihood distribution under two given viewpoints ϕ and $\phi+\delta$, and the

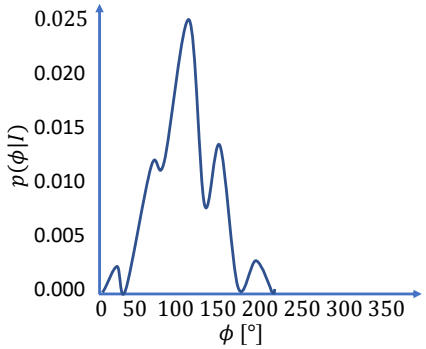

Fig. 4 Viewpoint likelihood distribution $p(\phi|I)$ (Input image with 90° rotation angle).

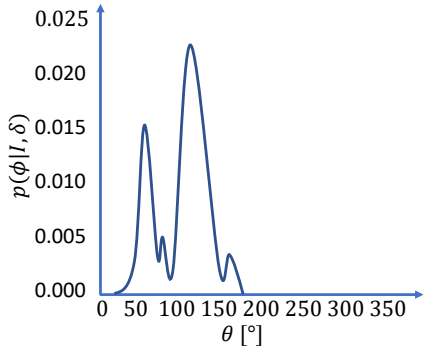

Fig. 5 Pose likelihood distribution given two viewpoints (Input image with 90° rotation angle).

rest part $p(\phi|I)$ indicates the viewpoint likelihood under a given observation. In the following sections, we explain more details on the two distributions.

2.3 Estimation of Viewpoint Likelihood Distribution $p(\phi|I)$

Since the viewpoint of an observation is difficult to obtain, the viewpoint likelihood distribution can be considered as a relative pose estimation from the initial viewpoint. We may only obtain an estimation result if we take a regression-based approach for the pose estimation, such as Pose-CyclicRNet [3],

$$\phi = f(I), \tag{4}$$

where I represents a given image and f the pose estimator. Since we have many images I_i of various objects in a class, by applying pose estimation for many images, we can obtain many pose estimation results ϕ_i. From these pose estimation results and their groundtruth, we can obtain a huge number of pairs of an estimation result and a ground truth . By applying density estimation to the data, we can obtain a conditional distribution as $p(\phi|f(I_i)) = p(\phi_{\mathrm{gt}}|\phi_{\mathrm{est}})$, where ϕ_{gt} represents the ground truth and ϕ_{est} the estimation result.

By using the conditional distribution, we can obtain the viewpoint likelihood distribution as,

$$p(\phi|I) = p(\phi|f(I)) \tag{5}$$

for a regression-based object pose estimator. This viewpoint likelihood distribution is illustrated in Figure 4.

2.4 Estimation of Pose Likelihood Distribution $p(\theta|\phi,\delta)$

The likelihood represents how accurately the objects' pose can be estimated given the two viewpoints; ϕ and $\phi+\delta$, where ϕ represents the current viewpoint and δ the rotation angle to the next viewpoint. The pose likelihood distribution given two viewpoints is illustrated in Figure 5. Here, we simply decompose the likelihood distribution into two pose likelihoods as

$$p(\theta|\phi,\delta) = p(\theta|\phi)p(\theta|\phi+\delta), \tag{6}$$

where $p(\theta|\phi)$ and $p(\theta|\phi+\delta)$ denote the pose likelihood distributions given a viewpoint ϕ and $\phi+\delta$, respectively. This equation holds by assuming $p(\theta)$, which is the pose likelihood without any information, follows a uniform distribution.

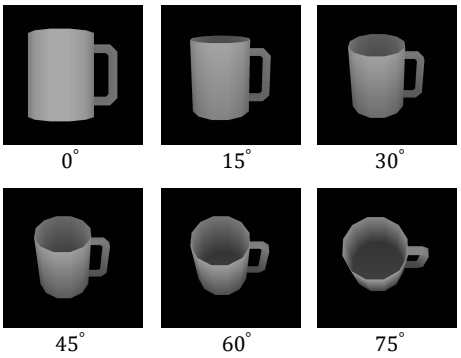

Fig. 6 Example of images observed from different elevation angles.

2.5 Pose Estimation θ_e

Finally we can estimate the object's pose from two viewpoints: the initial viewpoint and the next viewpoint. Here, I_1 is the image observed from the initial viewpoint. After rotating the depth camera δ degrees, we obtain I_2, which is the image observed from the next viewpoint.

We estimate the pose for these two viewpoints θ_e as the average of pose estimation results from I_1 and I_2 (by considering the rotation angle δ) as

$$\theta_e = \frac{\phi_1 + \phi_2 - \delta}{2}, \tag{7}$$

where $\phi_1 = f(I_1)$ is the pose estimation from the initial viewpoint and $\phi_2 = f(I_2)$ that from the next viewpoint.

3. Evaluations

3.1 Dataset

To show the effectiveness of the proposed viewpoint recommendation method, we performed a simulation-based evaluation. For the simulation, we use 135 3D models of "Mug" class in the ShapeNet dataset [10]. Rendering them by rotating around the z-axis, we obtain 360 depth images in the range of $[0°, 360°)$ for each model. We apply the rendering from several elevation angles of the virtual depth camera for each model. Images of 100 objects are randomly selected for the training set and the rest are used for the testing set.

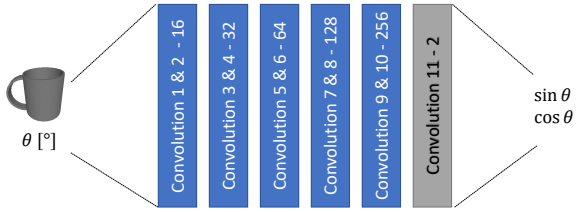

Fig. 7 Network architecture.

3.2 Pose Estimation Method

We prepared a network architecture similar to the Pose-CyclicR-Net proposed by Ninomiya et al. [3] as the pose estimator. The modified network architecture in the proposed method is shown in Figure 7. Since we assumed that the object pose variation is limited to a single axis rotation, we modify the network output to a pair of trigonometric functions $(\cos\theta, \sin\theta)$ instead of the original quaternion. We train the pose estimator using the training images.

3.3 Evaluation Criteria

We evaluated how the recommended viewpoints are appropriate for the pose estimation by using several criteria. One criterion is the Mean Absolute Error (MAE) of the pose estimation results with the ground truth. The pose estimation results are obtained by using a pair of the initial viewpoint and the recommended viewpoint. By considering the circularity of angles, the error can be calculated as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} d(\theta_e^i, \theta_g^i), \tag{8}$$

where N represents the number of images, θ_e^i and θ_g^i are the pose estimation result and the ground truth, respectively. $d(\theta_e^i, \theta_g)^i$ is the absolute difference of the poses considering the circularity defined as

$$d(\theta_e, \theta_g) = \begin{cases} |\theta_e - \theta_g| & \text{if } |\theta_e - \theta_g| > 180^\circ, \\ 180^\circ - |\theta_e - \theta_g| & \text{otherwise.} \end{cases} \tag{9}$$

3.3.1 Comparative Methods

We compared the pose estimation results by the proposed method and several other baseline methods. As a baseline, we use pose estimation from a single viewpoint which just applies Pose-CyclicR-Net-like Network to the input image. We adapt two other baseline methods from [8] which are "Random" and "Furthest".

4. Results

The experimental results are summarized in Table 1. Here, for all elevation angles, the proposed method outperformed all other comparative methods. This result clearly shows that the proposed method is promising and achieves better object pose estimation results. We successfully managed to reduce the pose ambiguity in the difficult observation viewpoint which has been mentioned in this paper.

5. Conclusion

We proposed a new idea for the best next viewpoint recommendation for an accurate pose estimation by minimizing the pose ambiguity by considering the current viewpoint and rotation angle as latent variables. We showed that the proposed method outperforms the other three comparison methods, and confirmed that a reliable and high pose estimation accuracy is achievable by the best next viewpoint.

Table 1 Comparison of overall Pose Estimation Accuracy in MAE.

Elevation angle	Single	Random	Furthest	Proposed
0°	18.36°	16.34°	14.91°	**14.18°**
15°	15.55°	14.01°	13.35°	**11.58°**
30°	15.40°	13.87°	12.75°	**11.96°**
45°	10.71°	9.32°	8.81°	**8.28°**
60°	8.15°	7.27°	7.15°	**6.31°**
75°	7.36°	6.57°	6.36°	**5.15°**

ACKNOWLEDGEMENT

The authors would like to thank Toyota Motor Corporation, Minister of Higher Education of Government of Malaysia (MOHE) and Universiti Teknikal Malaysia Melaka (UTeM). Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research (17H00745).

References

[1] R. T. Chin and C. R. Dyer, Model-based recognition in robot vision, ACM Computing Surveys, vol. 18, no. 1, pp. 67–108, 1986.
[2] H. Murase and S. K. Nayar, Visual learning and recognition of 3-D objects from appearance, International Journal of Computer Vision, vol. 14, no. 1, pp. 5–24, 1995.
[3] H. Ninomiya, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, N. Kobori, and Y. Nakano, Deep manifold embedding for 3D object pose estimation, Proceedings of the 12th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, vol. 5, pp. 173–178, 2017.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, Proceedings of the 25th International Conference on Neural Information Processing Systems, pp. 1097–1105, 2012.
[5] A. Zeng, S. Song, K. T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. Chavan Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching, Computing Research Repository, arXiv/1710.01330, 2017.
[6] A. Kanezaki, Y. Matsushita,and Y. Nishida, RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints, Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5010–5019, 2018.
[7] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T. K. Kim, Recovering 6D object pose and predicting next-best-view in the crowd, Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3583–3592, 2016.
[8] J. Sock, S. H. Kasaei, L. S. Lopes, and T. K. Kim, Multi-view 6D object pose estimation and camera motion planning using RGBD images, Proceedings of the 2017 IEEE Conference on Computer Vision Workshops, pp. 2228–2235, 2017.
[9] J. Gall, and V. Lempitsky, Class-specific Hough forests for object detection, Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1022–1029, 2009.
[10] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, ShapeNet: An information-rich 3D model repository, Computing Research Repositor, arXiv:1512.03012, 2015.