Next Viewpoint Recommendation by Pose Ambiguity Minimization for Accurate Object Pose Estimation

Nik Mohd Zarifie Hashim[1,5], Yasutomo Kawanishi[1], Daisuke Deguchi[2], Ichiro Ide[1], Hiroshi Murase[1], Ayako Amma[3] and Norimasa Kobori[4]

[1]*Graduate School of Informatics, Nagoya University, Japan*

[2]*Information Strategy Office, Nagoya University, Japan*

[3]*Toyota Motor Corporation, Japan*

[4]*Toyota Motor Europe, Belgium*

[5]*Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka, Malaysia*

hashimz@murase.is.i.nagoya-u.ac.jp, kawanishi@i.nagoya-u.ac.jp, ddeguchi@nagoya-u.jp, ide@i.nagoya-u.ac.jp, murase@i.nagoya-u.ac.jp, ayako_amma@mail.toyota.co.jp, Norimasa.Kobori@toyota-europe.com

Keywords: 3D Object, Deep Learning, Next Viewpoint, Pose Ambiguity, Pose Estimation.

Abstract: 3D object pose estimation by using a depth sensor is one of the important tasks in activities by robots. To reduce the pose ambiguity of an estimated object pose, several methods for multiple viewpoint pose estimation have been proposed. However, these methods need to select the viewpoints carefully to obtain better results. If the pose of the target object is ambiguous from the current observation, we could not decide where we should move the sensor to set as the next viewpoint. In this paper, we propose a best next viewpoint recommendation method by minimizing the pose ambiguity of the object by making use of the current pose estimation result as a latent variable. We evaluated viewpoints recommended by the proposed method and confirmed that it helps us to gain better pose estimation results than several comparative methods on a synthetic dataset.

1 INTRODUCTION

3D object pose estimation which estimates the three axes rotation of a target object, has recently become one of the focussed topics in the machine vision field for application on tasks in activities by robots. Especially, object picking is an essential task for industrial robots and home helper robots. For example, home helper robots are required to pick an object and pass it to a person.

As a machine vision problem, robots require a sensor for observing their surrounding environment. There are several types of sensors utilized to observe the environment, e.g. light detection and ranging (LiDAR), stereo cameras, RGB and depth (RGB-D) cameras, and so on. All these sensors are being actively improved year by year. In this paper, since they are robust to the object's texture and can obtain much information about the object's shape, we focus on depth images captured by a depth sensor, and use them for estimating the 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. In this approach, the object's pose is described as a relative pose to the sensor.

If an object has a distinct shape to be distinguished from various viewpoints, object pose estimation would be easy. However, most objects have viewpoints where their poses cannot be uniquely distinguished by their appearances because they resemble each other. We call this the "pose ambiguity problem". In single viewpoint pose estimation, this problem leads to inaccurate pose estimation.

Since a robot can move and thus change viewpoints, after the initial observation, it can move to another viewpoint and re-observe the object. By observing an object from multiple viewpoints, the ambiguity could be reduced. In this approach, to estimate the object's pose accurately, it is necessary to choose the best next viewpoint. A better viewpoint helps us to achieve a more accurate object pose estimation as shown in Figure 1. However, if the pose estimation result from the initial viewpoint is ambiguous, the robot could not properly decide in which direction and how far it should move to reach the best

Hashim, N., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H., Amma, A. and Kobori, N.
Next Viewpoint Recommendation by Pose Ambiguity Minimization for Accurate Object Pose Estimation.
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019) - Volume 5: VISAPP*, pages 60-67
ISBN: 978-989-758-354-4

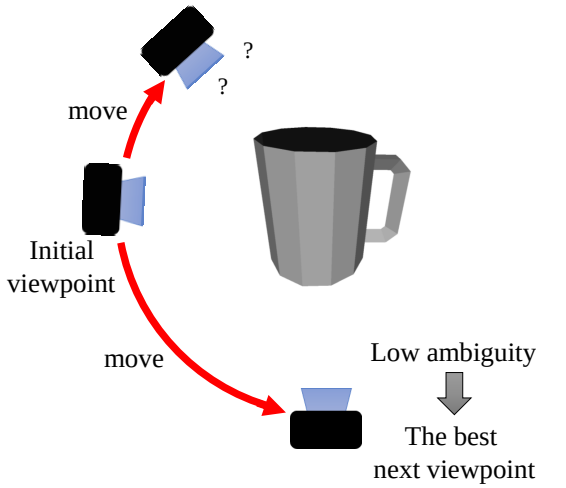

Figure 1: Recommendation of the best next viewpoint.

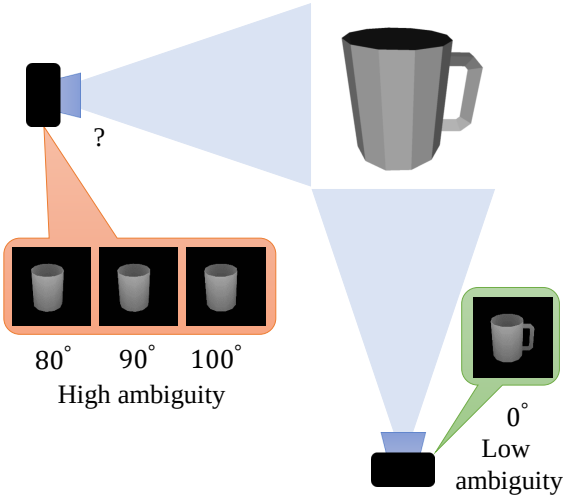

Figure 2: Selection problem in pose estimation from multiple viewpoints.

next viewpoint. The question here is that, how can we know the best next viewpoint from the current observation as illustrated in Figure 2.

Research on pose estimation from multiple viewpoints is actively proposed by the instance-level approach (Doumanoglou et al., 2016) (Sock et al., 2017). Though this approach seems to work, if an application needs to estimate a pose between various object shapes in a same category, they may not perform well. Thus, we focus on the category-level pose estimation.

In this paper, we propose a method for the next viewpoint recommendation for accurate object pose estimation even if there are various shapes in the same category. To evaluate the effectiveness of the recommended viewpoint, we define a metric called **"pose ambiguity"**, which reflects how ambiguous the pose estimation is. By minimizing the pose ambiguity, we will find the next viewpoint which will be the best to estimate the object's pose. This pose estimation is realized by averaging the pose estimation of two viewpoints which are the initial observation and the next viewpoint. To handle the pose estimation ambiguity of the initial observation, we make use of the estimated pose from the initial observation as a latent variable. We introduce an estimation method of the pose ambiguity by marginalizing the latent variable, which can consider all possibilities of the initial estimation result. To make the problem simple and focus on the key idea, in this paper, we limit the movement of the sensor only to the z-axis rotation and consider the non-cascade case for analyzing the next viewpoint. However, the proposed method and discussion could be straightforwardly extended to 3D rotation. We evaluate the effectiveness of the proposed method on a dataset generated from a publicly available 3D object dataset: ShapeNet (Chang et al., 2015).

Our contribution can be summarized as follows:

- We define a metric **"pose ambiguity"** to evaluate the ambiguousness of the pose estimation.
- We propose a next viewpoint recommendation method which finds the best next viewpoint where the pose ambiguity is minimized.
- We show that the proposed method ourperforms two other naive viewpoint recommendation methods, and also that it achieves a better result than the pose estimation result from a single viewpoint.
- We introduce a new paradigm of searching the best next viewpoint for category-level object pose estimation compared to conventional instance-level object pose estimation.

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

2 RELATED WORK

In the past few years, many researchers proposed methods to tackle various difficulties in 3D object pose estimation. Here, we categorize the pose estimation methods into the following two approaches: those from a single viewpoint and those from multiple viewpoints.

2.1 Object Pose Estimation from a Single Viewpoint

The template matching approach is one of the earliest pose estimate method from a single viewpoint (Chin and Dyer, 1986). This method utilizes many templates of the target object captured from various viewpoints beforehand, and the pose estimation result is taken from the best matched template. To reduce the number of templates, Murase and Nayer (Murase and Nayar, 1995) proposed the Parametric Eigenspace method. This method represents an object's pose variation on a manifold in a low-dimensional subspace obtained by Principal Component Analysis (PCA). By interpolating the pose (template) of the target object on the manifold, we can achieve accurate object pose estimation even with few templates. Since PCA focuses on the appearance variation of templates, some poses with similar appearances may be mapped to similar points in the low-dimensional subspace, which is difficult to distinguish. This diminishes the accuracy of the pose estimation. Moreover, the method is based on PCA, which is an unsupervised learning method, so it does not fully utilize the pose information for estimating the object's poses.

Recently, Ninomiya et al. (Ninomiya et al., 2017) proposed a supervised feature extraction method for embedding templates into a pose manifold. They focused on Deep Convolutional Neural Networks (DCNNs) (Krizhevsky et al., 2012) which is one of the deep-learning models, as a supervised learning method for manifold embedding. They modified DCNNs for object pose estimation, named Pose-CyclicR-Net, which can correctly handle object rotation by describing the rotation angle using trigonometric functions. By introducing the Pose-CyclicR-Net based manifold embedding, which is called Deep Manifold Embedding, the method estimates the object's pose from a single viewpoint.

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

2.2 Object Pose Estimation from Multiple Viewpoints

To avoid the pose ambiguity issue, several methods focus on object pose estimation from multiple viewpoints. Collet and Srinivasa (Collet and Srinivasa, 2010) proposed a multi-view object pose estimation method based on multi-step optimization and global refinement. Erkent et al. (Erkent et al., 2016) tackle object pose estimation in cluttered scenes. This is a multi-view approach based on probabilistic, appearance-based pose estimation. Vikstén et al. (Vikstén et al., 2006) proposed a method combining several pose estimation algorithms and information from several viewpoints. Zeng et al. (Zeng et al., 2017b) proposed a self-supervised approach for object pose estimation in the Amazon Picking Challenge (Zeng et al., 2017a) scenario. Kanezaki et al. (Kanezaki et al., 2018) proposed the RotationNet, which takes multi-view images of an object as input and jointly estimates its pose and object category. As such, there are various methods for object pose estimation from multiple viewpoints, but these methods do not consider which viewpoint is effective for the estimation. Unlike others, in this paper we propose an idea of estimating the current viewpoint which will increase the pose estimation later.

Recently, some researches focus on predicting Next-Best-View for object pose estimation. Doumanoglou et al. (Doumanoglou et al., 2016) and Sock et al. (Sock et al., 2017) proposed next-best-view prediction methods for multiple object pose estimation based on Hough Forest (Gall and Lempitsky, 2009). We expect that this approach will be the next interesting topic. This idea will allow us to support an application in which various instances in a specific object category need to be considered as the target object. However, we acknowledge that these methods could not be applied for the category-level object pose estimation since 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 our proposed method.

3 NEXT VIEWPOINT RECOMMENDATION

3.1 Overview

In this paper, we propose a novel next viewpoint recommendation method based on pose ambiguity minimization; We define a metric called "pose ambiguity" given two different viewpoints which should be minimized. Since the initial viewpoint may be ambiguous, by handling the current viewpoint as a latent variable, the pose ambiguity function is decomposed into "pose ambiguity under given two viewpoints" and "viewpoint ambiguity under a given observation". Figure 3 illustrates the angle distribution for the "pose ambiguity" G. Here, the minimum value of G at y-axis infers the best next viewpoint as δ [°]

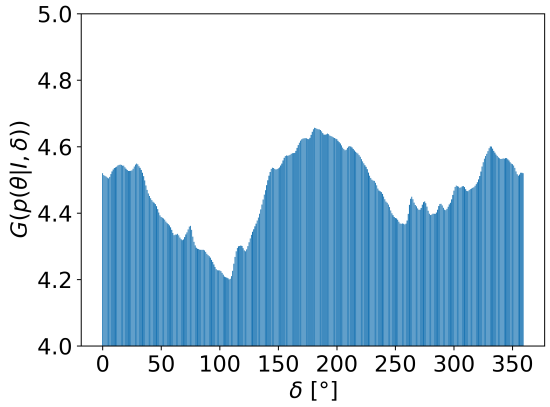

Figure 3: Pose ambiguity minimization (Input image with 90° rotation angle).

from x-axis. This δ will be utilized to estimate all the object poses. We calculate the pose estimation for two viewpoints by averaging them as it provides us a reliable pose between the two different angles. We will introduce the details of the process in the following sections.

3.2 Pose Ambiguity Minimization Framework

This framework will measure the pose ambiguity in a quantitative way. First of all, what is pose ambiguity? Here, we define it as the difficulty to estimate the pose of an object from a viewpoint. Given a pose likelihood distribution over the possible poses, if the distribution is largely diverse, the pose will be difficult to be estimated correctly. Thus, we define the pose ambiguity G as a function of the pose likelihood distribution $p(\theta)$. For example, G can be defined by the Entropy of $p(\theta)$ as

$$G(p) = \int -p(\theta)\log p(\theta)d\theta. \tag{1}$$

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.

The minimum value of the pose ambiguity in G will tell us the best next viewpoint for accurate pose estimation using the two viewpoints. By using the formulation, we find the best viewpoint by minimizing the entropy as

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

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

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

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

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

Since the absolute viewpoint of an observation is difficult to obtain, the viewpoint likelihood distribution can be considered as a relative pose estimation from the initial viewpoint. In the ideal case, if we have a discrete pose classifier in hand for the pose estimation, we may obtain not only the estimation result (pose) but also the likelihood for all possible poses. On the other hand, if we take a regression-based approach for the pose estimation, such as Pose-CyclicRNet proposed by Ninomiya et al. (Ninomiya et al., 2017), we may only obtain an estimation result such as

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

where I represents a given image and f the pose estimator. For such a regression-based pose estimator, how can we obtain the viewpoint likelihood distribution? 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 ground truth, 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 $p(\phi|f(i)) = p(\phi_{\text{gt}}|\phi_{\text{est}})$, 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.

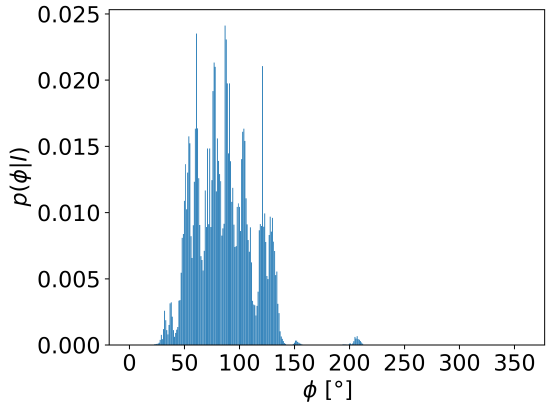

Figure 4: Viewpoint likelihood distribution $p(\phi|I)$ (Input image with 90° rotation angle).

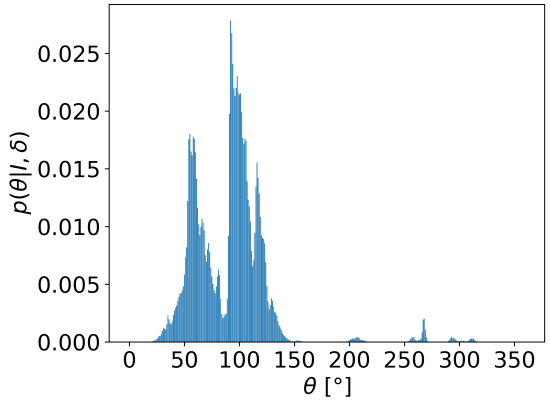

Figure 5: Pose likelihood distribution given two viewpoints $p(\theta|I,\delta)$ (Input image with 90° rotation angle).

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

Here we explain the pose likelihood distribution given two viewpoints ϕ and $\phi+\delta$, where ϕ represents the current viewpoint and δ the rotation angle to the next viewpoint. The likelihood represents how accurately the objects' pose can be estimated given the two viewpoints. 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. Each likelihood distribution given a viewpoint can also be calculated by applying density estimation for the pairs of a pose estimation result and the ground truth similarly as to the method described in Section 3.3.

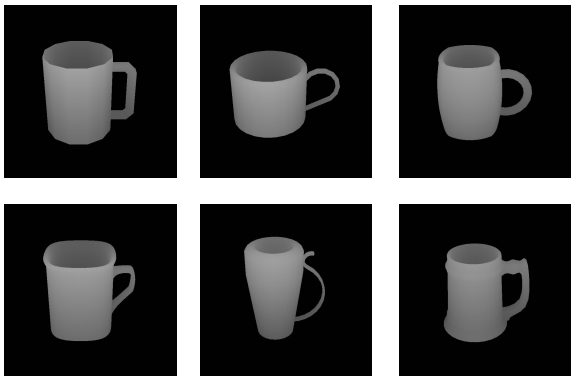

Figure 6: Example images from the "Mug" class in the ShapeNet dataset (Chang et al., 2015).

3.5 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 δ [°], 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.

4 EXPERIMENTS

4.1 Dataset

To show the effectiveness of the proposed viewpoint recommendation method, we performed a simulation-based evaluation. For the simulation, we used a set of 3D models in the ShapeNet (Chang et al., 2015). We collected 135 models in the "Mug" class. Concretely, we put a 3D model in a virtual environment and observed it using a virtual depth sensor. By rotating the sensor around the z-axis of the 3D model, we obtained 360 depth images in the range of $[0°, 360°)$ for each model as shown in Figure 6.

Additionally, in the simulation, we changed the elevation angle of the virtual sensor as 0°, 15°, 30°,

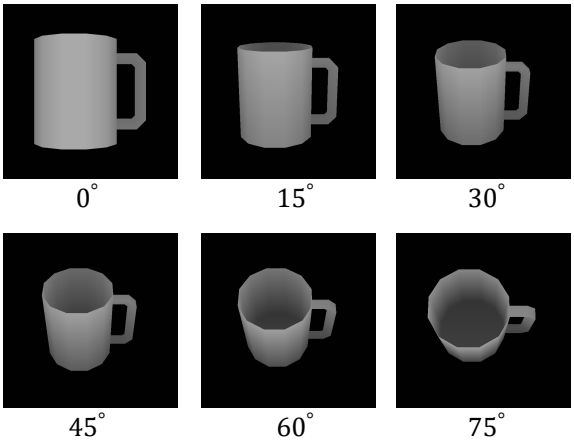

Figure 7: Example of images observed from different elevation angles.

45°, 60°, and 75° which is elevated upright from the z-plane as shown in Figure 7. In total, 135 objects were observed from each elevating angle with a total of 48,600 images. We used the rendered images for training and testing in the evaluation. These images are divided into two folds: a training set and a testing set. Images of 100 objects are randomly selected for the training set and the rest are used for the testing set. We conducted an experiment with the synthetic mug dataset, which we consider is adequate to address the ambiguity issue for category-level pose estimation.

4.2 Evaluation Method

4.2.1 Pose Estimation Method

We prepared a network architecture similar to the Pose-CyclicR-Net proposed by Ninomiya et al. (Ninomiya et al., 2017) as the pose estimator. The original network architecture is shown in Figure 8. Since we assumed that the object pose variation is limited to a single axis rotation, we modified the network output to a pair of trigonometric functions $(\cos\theta, \sin\theta)$ instead of the original quaternion. We trained the pose estimator using the training images.

4.2.2 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

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

The other criterion is Pose Estimation Accuracy (PEA). This criterion is be defined as

$$\mathrm{PEA}(\tau) = \frac{1}{N}\sum_{i=1}^{N} F(d(\theta_e^i, \theta_g^i) < \tau), \tag{10}$$

where τ represents a threshold error which reflects the difference of pose estimation result θ_e^i and ground truth θ_g^i , $F(\cdot)$ is a function which returns 1 if the condition in the function holds and 0 vice versa.

4.2.3 Comparative Methods

We compared the pose estimation results by the proposed method and several other baseline methods. To the best of our knowledge, there is no existing method that could be directly compared with our proposed method as the category-level next best viewpoint estimation study is just initiated by us. Thus as a baseline, we used pose estimation from a single viewpoint, which just applies Pose-CyclicR-Net-like Network to the input image. We also compared with several viewpoint recommendation methods.

We adapted two other baseline methods from (Sock et al., 2017) which are “Random” and “Furthest”. The first baseline viewpoint recommendation method is recommending the next viewpoint randomly, named “Random”. Because of the randomness, we selected ten viewpoints randomly and averaged the estimation results. The second baseline viewpoint recommendation method is recommending the next viewpoint by simply selecting the opposite side or the furthest point from the initial viewpoint, named “Opposite” (equivalent to “Furthest”).

4.3 Results

4.3.1 Mean Absolute Error (MAE)

The experimental results are summarized in Table 1. Here, for all elevation angles, the proposed method

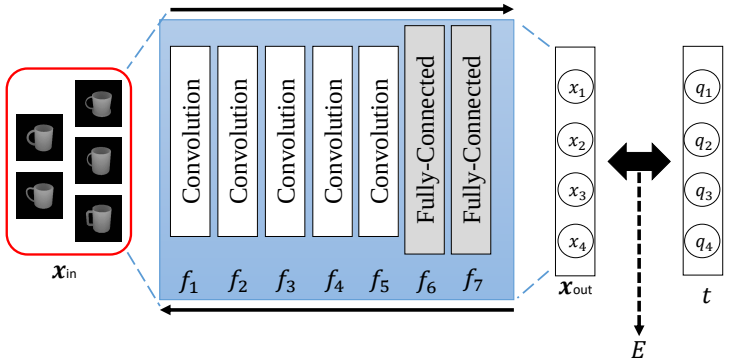

Figure 8: Original Pose-CyclicR-Net (Ninomiya et al., 2017).

Table 1: Comparison of overall Pose Estimation Accuracy.

Elevation angle	Single	Random	Opposite	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°**

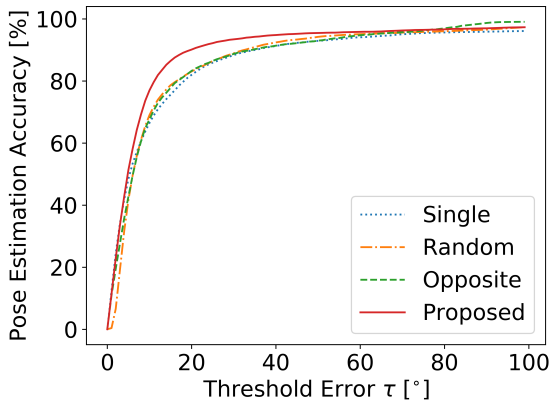

Figure 9: Pose Estimation Accuracy within threshold error 0° – 100° (elevation angle = 15°).

outperformed all other comparative methods. This result clearly shows that the proposed method is promising and gave a better way (next viewpoint) for object pose estimation. We successfully managed to reduce the pose ambiguity in the difficult observation viewpoint which has been mentioned earlier in Figure 1 at the beginning of this paper. We can see that estimating object's pose from two viewpoints yields better results than from a single viewpoint. By comparing with the other pose recommendation methods, the proposed method achieved better results by carefully selecting the best viewpoint for object pose estimation. By reducing the pose ambiguity, the proposed method achieved the lowest pose estimation error.

Table 2: Comparison using Partial-AUC (pAUC) between τ from 0° to 60°.

Elevation angle	Single	Random	Opposite	Proposed
0°	75.61	76.75	78.19	**80.80**
15°	79.13	79.14	79.16	**84.05**
30°	79.79	79.69	80.23	**83.75**
45°	84.94	85.46	86.07	**87.65**
60°	88.31	88.42	88.57	**90.49**
75°	89.47	89.42	89.91	**92.03**

4.3.2 Pose Estimation Accuracy (PEA)

In Figure 9, we plotted the pose estimation accuracy by changing the threshold error τ in Equation 10 in the case of elevation angle = 15°. We confirmed that the proposed method outperforms all the comparison methods when the threshold error is within $0° < \tau < 60°$. When $60° < \tau < 100°$, the "Opposite" method outperformed the proposed method. However, a large threshold error value will not critically influence the pose estimation accuracy, so we consider the results when $\tau > 60°$ are not significant for this purpose. We also calculated partial-AUCs (pAUCs) for all curves as summarized in Table 2. With all six elevation angles, we showed that the proposed method achieves the most accurate pose estimation compared to the other methods.

5 CONCLUSION

We proposed a new idea to estimate the best next viewpoint for an accurate pose estimation and a new framework for minimizing the pose ambiguity. We showed that the proposed method outperforms three baseline methods by utilizing the latent variable which provides us with a new next viewpoint in the category-level pose estimation. Therefore, by having this next

viewpoint, a reliable and high pose estimation accuracy is achievable.

For future improvement, we are looking forward to evaluating the proposed method with multi-dimension elevation angles as to compare with our current single elevation angle implementation in this paper. The improvement for obtaining a higher pose estimation accuracy by expanding the two viewpoints into several points has also been projected as our upcoming task. To overcome the non-cascade next viewpoint in specific cases, we also consider to have "several" best next viewpoint where the multi-dimensional elevation angles are utilized. This approach may be utilized with a different class of multiple objects for having a wider scope of application and could help the development of the human helper robot field.

ACKNOWLEDGEMENT

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

REFERENCES

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. (2015). ShapeNet: An information-rich 3D model repository. *arXiv:1512.03012*.

Chin, R. T. and Dyer, C. R. (1986). Model-based recognition in robot vision. *ACM Computing Surveys*, 18(1):67–108.

Collet, A. and Srinivasa, S. S. (2010). Efficient multi-view object recognition and full pose estimation. In *Proceedings of the 2010 IEEE International Conference on Robotics and Automation*, pages 2050–2055.

Doumanoglou, A., Kouskouridas, R., Malassiotis, S., and Kim, T.-K. (2016). Recovering 6D object pose and predicting next-best-view in the crowd. In *Proceedings of the 2016 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3583–3592.

Erkent, Ö., Shukla, D., and Piater, J. (2016). Integration of probabilistic pose estimates from multiple views. In *Proceedings of the 2016 European Conference on Computer Vision*, volume 7, pages 154–170.

Gall, J. and Lempitsky, V. (2009). Class-specific Hough forests for object detection. In *Proceedings of the 2009 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1022–1029.

Kanezaki, A., Matsushita, Y., and Nishida, Y. (2018). RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the 2018 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 5010–5019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 1097–1105.

Murase, H. and Nayar, S. K. (1995). Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14(1):5–24.

Ninomiya, H., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H., Kobori, N., and Nakano, Y. (2017). Deep manifold embedding for 3D object pose estimation. In *Proceedings of the 12th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 173–178.

Sock, J., Kasaei, S. H., Lopes, L. S., and Kim, T. K. (2017). Multi-view 6D object pose estimation and camera motion planning using RGBD images. In *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops*, pages 2228–2235.

Vikstén, F., Söderberg, R., Nordberg, K., and Perwass, C. (2006). Increasing pose estimation performance using multi-cue integration. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, pages 3760–3767.

Zeng, A., Song, S., Yu, K., Donlon, E., Hogan, F. R., Bauzá, M., Ma, D., Taylor, O., Liu, M., Romo, E., Fazeli, N., Alet, F., Dafle, N. C., Holladay, R., Morona, I., Nair, P. Q., Green, D., Taylor, I., Liu, W., Funkhouser, T. A., and Rodriguez, A. (2017a). Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. arXiv/1710.01330.

Zeng, A., Yu, K.-T., Song, S., Suo, D., Walker, E., Rodriguez, A., and Xiao, J. (2017b). Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge. In *Proceedings of the 2017 IEEE International Conference on Robotics and Automation*, pages 1386–1383.