論文

Best Next-Viewpoint Recommendation by Selecting Minimum Pose Ambiguity for Category-Level Object Pose Estimation*

Nik Mohd Zarifie HASHIM**,***,†, Yasutomo KAWANISHI†, Daisuke DEGUCHI†, Ichiro IDE†,
Ayako AMMA††, Norimasa KOBORI††† and Hiroshi MURASE†

Object manipulation is one of the essential tasks for a home helper robot, especially in helping a disabled person to complete everyday tasks. For handling various objects in a category, accurate pose estimation of the target objects is required. Since the pose of an object is often ambiguous from an observation, it is important to select a good next-viewpoint to make a better pose estimation. This paper introduces a metric of the object pose ambiguity based on the entropy of the pose estimation result. By using the metric, a best next-viewpoint recommendation method is proposed for accurate category-level object pose estimation. Evaluation is performed with synthetic object images of objects in five categories. It shows the proposed methods is applicable to various kind of object categories.

Key words: best next viewpoint, category-level object pose estimation, entropy, human helper robot, pose ambiguity

1 Introduction

Recently, estimating the rotation of a target object; object pose estimation, has become a focussed topic in the robot vision field. To estimate the target object's pose, a vision sensor is an vital device for a robot to capture the target's information. Stereo cameras, Light Detection And Ranging (LiDAR), RGB and Depth (RGB-D) cameras are popular vision sensors equipped to robots. In this paper, we use depth images captured by a depth sensor to estimate the object pose, because they are robust to texture variations and can capture the shape of the target well.

The template matching approach is one of the earliest pose estimation methods from an input image[1]. Murase and Nayer[2] proposed the Parametric Eigenspace method to lessen the amount of templates in this task. Recently, Ninomiya et al. proposed a method based on the deep feature extraction for embedding the template into a pose manifold [3]. For pose estimation from an image, there is a case that the object pose is ambiguous from the observation as illustrated in **Fig. 1**. On the other hand, since robots have embodyment, they can move to other locations. From the new location, a different observation can be obtained. Based on this, many researchers focus on object pose estimation from multiple viewpoints.

For example, Zeng et al.[4], Erkent et al.[5], Collet and Srinivasa[6], Kanezaki et al.[7], and Vikstén et al.[8], proposed pose estimation methods using observations from multiple viewpoints. However, most of these methods only focus on pose estimation using a given image set and do not consider how to select the viewpoints.

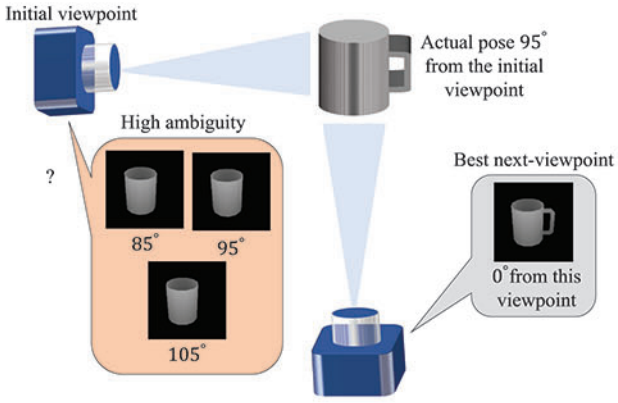

Fig. 1: Pose ambiguity problem from the initial viewpoint

Bajcsy et al. discussed the idea of active perception, initially in the context of the sensor planning problem[9]. Since controlling the camera view is the most crucial issue to recognize objects, recent researches focus on predicting the best next-view[10–12]. The recent active perception works [13][14] propose the best next-viewpoint prediction for pose estimation of multiple target objects based on Hough Forest. However, these methods focus on target objects whose shapes are known. In daily environment, there are variously shaped objects even within an object category. Therefore, pose estimation methods should be robust to the shape variation within the target object category, that is, category-level object pose estimation is desirable.

In this paper, we focus on how to select the next viewpoint given an initial viewpoint for object pose estimation from multiple viewpoints. We propose a best next-viewpoint recommendation method based on a novel metric "pose ambiguity", which reflects the difficulty to estimate the pose correctly from the given inputs as illus-

* Received October 14, 2020
Accepted February 15, 2021
** Correspondence to: hashimz@murase.is.i.nagoya-u.ac.jp
*** Centre for Telecommunication Research & Innovation (CeTRI), Fakulti Kejuruteraan Elektronik Dan Kejuruteraan Komputer, Universiti Teknikal Malaysia Melaka (Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia)
† Nagoya University (1 Furo-cho, Chikusa Ward, Nagoya, Aichi 464-8601, Japan)
†† Toyota Motor Corporation (Kirigabora-543 Nishihirosechō Toyota, Aichi 470-0309, Japan)
††† Woven Planet Group (Nihonbashi Muromachi Mitsui Tower, 16-20F 3-2-1 Nihonbashi Muromachi, Chuo-ku, Tokyo, 103-0022, Japan)

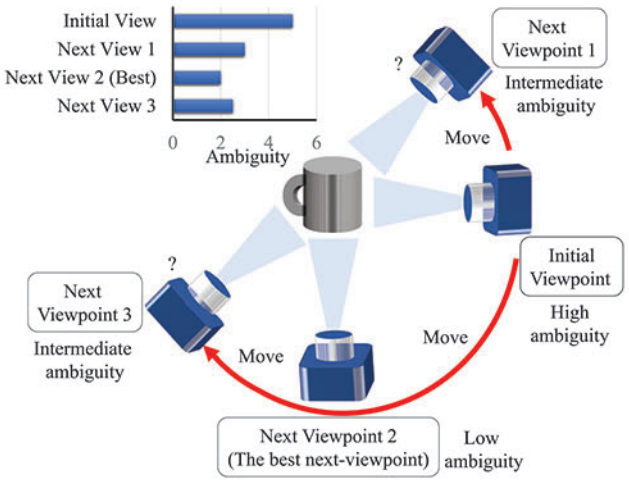

Fig. 2: Selection of the next viewpoint from the initial viewpoint based on the ambiguity scale. The bar graph indicates the ambiguity value for the three cases of the next viewpoint as illustrated

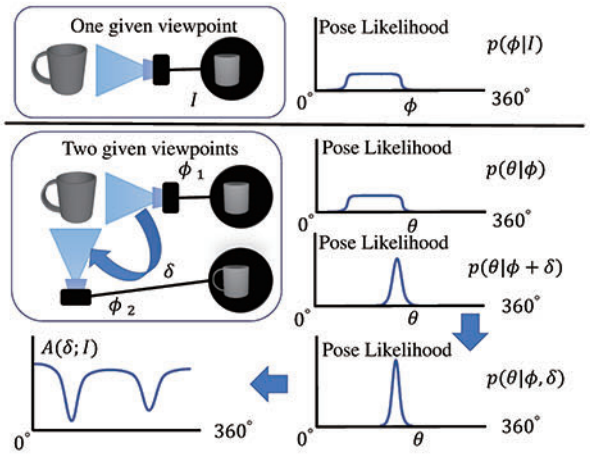

Fig. 3: Illustration of the idea on the one given viewpoint (top), and two given viewpoints (bottom) for the next viewpoint recommendation

trated in **Fig. 2**.

The best next-viewpoint can be discovered by electing the viewpoint where the pose ambiguity is the smallest given observations from the initial and the next viewpoints. However, since the estimated initial viewpoint is also ambiguous, we consider the initial viewpoint as a latent variable to keep all possibilities of the object pose from the initial viewpoint.

In this paper, to focus on the essential part of the problem setting and idea, we explain and evaluate the method by limiting to a single axis rotation. However, the extension to 3D rotation could be straightforwardly explainable. With the utilization of object images rendered from the public 3D object dataset, ShapeNet[15], we conducted several assessments on the proposed method's effectiveness.

Our contributions are summarized as follows:

- Definition of a new metric called *"pose ambiguity"* to examine the ambiguousness for the pose estimation task.
- Introduction of a new standard for finding the best next viewpoint for category-level object pose estimation.
- The proposed method outperforms two other naive viewpoint recommendation methods in several pose estimation analysis, and also it achieves a better result than the result from a single viewpoint.

Note that this paper is an extended version of our previous conference paper[16]. Compared to it, firstly, we added an evaluation on various object categories to confirm the performance of category-level object pose estimation. The evaluation is performed with objects in five categories. Secondly, we revised the formulations of the proposed method to make the concept of the paper clearer.

The remaining of this paper is structured as follows: In Chapter 2, the proposed method will be introduced in detail. Chapter 3 will introduce the evaluation with discussion on the results. Finally, we conclude the paper in Chapter 4.

2 Best Next-Viewpoint Recommendation

2.1 Overview

In this research, we propose a novel problem setting, which is the best next-viewpoint recommendation[16]. We define a metric called "pose ambiguity" given two different viewpoints; the initial viewpoint ϕ and the rotation angle δ to the next viewpoint from ϕ are considered as its parameters. In this chapter, we propose a method to recommend the best next-viewpoint by selecting the next viewpoint whose "pose ambiguity" will be the minimum. By a standard pose estimation method, because of the shape of the target object, the pose estimation result from the current observation I may be ambiguous from the initial viewpoint ϕ. Assuming the current viewpoint as a latent variable, the defined pose ambiguity function is decomposed into "pose ambiguity under given two viewpoints" and "viewpoint ambiguity under a given observation" as shown in **Fig. 3**. The minimum value of the pose ambiguity infers the best next viewpoint as $\phi + \delta$. Once we can obtain the rotation angle δ, we can re-observe the object from the recommended viewpoint and estimate the pose by averaging these two viewpoints. We will present the process in details in the subsequent sections.

2.2 Minimum Pose Ambiguity Selection Framework

This framework measures the defined pose ambiguity in a quantitative way. For the framework, we first need to define the pose ambiguity. Here, we define it as the difficulty to estimate the pose of an object in a category from a viewpoint. If the possibility of the estimated object pose θ is widely distributed, the result can be considered as ambiguous. Therefore, we define the pose ambiguity $A(\delta; I)$ based on the pose likelihood distribution $p(\theta|I, \delta)$ given the initial observation I and rotation angle δ. Here, we introduce a mapping function G from the pose likelihood distribution to the pose ambiguity. For example, G can be defined by the entropy of $p(\theta|I, \delta)$ as

$$A(\delta; I) = G(p(\theta|I, \delta)) = -\int p(\theta|I, \delta) \log p(\theta|I, \delta) \mathrm{d}\theta. \quad (1)$$

Here, the pose likelihood distribution given an image I is observed from the initial viewpoint, and then this pose likelihood dis-

tribution will yield the rotation angle δ 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 will tell us the best next-viewpoint for accurate pose estimation from the two viewpoints. By using the formulation, we find the best next-viewpoint by finding the minimum entropy as

$$\widehat{\delta} = \arg\min_{\delta} A(\delta; I). \tag{2}$$

To calculate the ambiguity, we further decompose the pose likelihood distribution using a latent variable ϕ, which represents the initial viewpoint as follows:

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

The first term $p(\theta|\phi,\delta)$ gives the pose likelihood distribution given two viewpoints ϕ and $\phi+\delta$, and the remaining $p(\phi|I)$ indicates the viewpoint likelihood given an observation I. In the following sections, we explain more details on the two distributions.

2.3 Viewpoint Likelihood Distribution

To measure the object pose, we need to define the origin of the object rotation. Then, we can estimate the object pose as a relative rotation angle of the viewpoint from the origin. However, it is difficult to estimate a fixed rotation angle from the observation, the viewpoint likelihood distribution given an observation is used. In the ideal case, if we have a pose estimator, which can output the distribution itself, such as a discrete pose classifier, we can use the output distribution directly. On the other hand, if we take a regression-based approach for the pose estimation, such as Pose-CyclicR-Net proposed by Ninomiya et al.[3], we may only obtain an estimation result such as

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

where I represents a given observation 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 the object category, by applying pose estimation to those images, we can obtain many pose estimation results ϕ_i. From these results and corresponding ground-truth poses, we can obtain a large number of pairs of an estimation result and its ground truth.

By applying density estimation to these data, we can obtain a conditional distribution as $p(\phi|f(i)) = p(\phi_{\mathrm{gt}}|\phi_{\mathrm{est}})$, where ϕ_{gt} represents the ground truth and ϕ_{est} the estimation result. By using this conditional distribution, we can obtain the viewpoint likelihood distribution as

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

for the given regression-based object pose estimator.

2.4 Pose Likelihood Distribution

Here, we explain the pose likelihood distribution given two viewpoints ϕ and $\phi+\delta$, where ϕ represents the initial viewpoint and δ the rotation angle to the next viewpoint. The likelihood distribution can simply be decomposed into two pose likelihoods as

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

where $p(\theta|\phi)$ and $p(\theta|\phi+\delta)$ represent 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 in the same way as in Section 2.3.

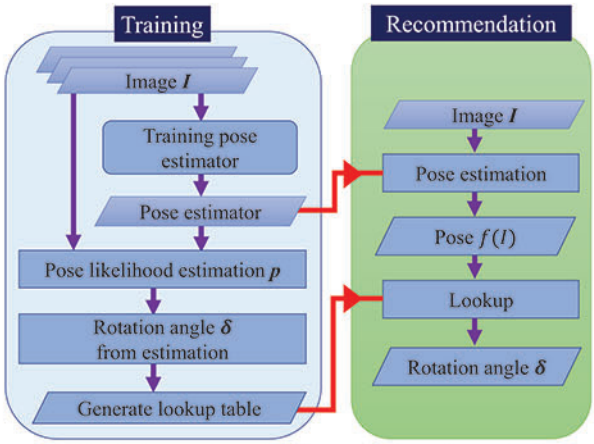

Fig. 4: Process flow of the best next-viewpoint method

2.5 Pose Estimation

After finding the best δ by Equation (2), we can finally estimate the object pose from the two viewpoints; the initial viewpoint ϕ and the next viewpoint $\phi+\widehat{\delta}$. Here, I_1 is the image observed from the initial viewpoint ϕ. After rotating $\widehat{\delta}$, we obtain I_2, which is the image observed from the next viewpoint.

We estimate the pose θ_e from the two viewpoints as the average of pose estimation results ϕ_1 and ϕ_2 from I_1 and I_2, respectively, considering the rotation angle $\widehat{\delta}$ as

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

where $\phi_1 = f(I_1)$ is the pose estimation from the initial viewpoint and $\phi_2 = f(I_2)$ is that from the next viewpoint. Since $\widehat{\delta}$ is selected in terms of the minimum pose ambiguity given an initial viewpoint and the rotation angle $\widehat{\delta}$, the averaged pose θ_e will be optimal.

The next viewpoint recommendation idea, which consists of training and recommendation phases, is shown in **Fig. 4**. In practice, the pose likelihood distribution $p(\theta|I,\delta)$ is implemented by a lookup table. This lookup table is pre-computed in the training phase. In the recommendation phase, after the current pose is estimated from the initial viewpoint $p(\phi|I)$, we can obtain $\widehat{\delta}$ by referring to the lookup table.

3 Experiments

3.1 Dataset

To show the effectiveness of the proposed viewpoint recommendation method, we performed a simulation-based evaluation. For the simulation, we selected 125 3D models in five object categories "Airplane", "Car", "Chair", "Mug", and "Toilet" from the ShapeNet dataset[15]. Concretely, we put a 3D model in a virtual

Fig. 5: Example of images from the "Airplane", "Car", "Chair", "Mug", and "Toilet" class in the ShapeNet dataset[15)]

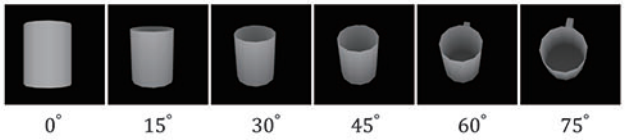

Fig. 6: Example of "Mug" images observed from different elevation angles

environment and observed it using a virtual depth sensor. By rotating the sensor around the z-axis of the 3D model, 360 depth images in the range of $[0°, 360°)$ are obtained for each model as shown in **Fig. 5**. To focus on the essential part of the proposed algorithm, we estimated the pose in the single axis rotation setting. Additionally, in the simulation, we changed the elevation angle of the virtual sensor as 0°, 15°, 30°, 45°, 60°, and 75° which is elevated upright from the z-plane as shown in **Fig. 6**.

In total, 125 objects were observed from each elevation angle with a total of 45000 images. We used the rendered images for training and testing in the evaluation. We divided these 125 objects in a category from the synthetic datasets into five folds for evaluating the proposed pose estimation method compared to other methods in a five-fold cross-validation setup. For each fold, images of 25 objects were used for testing and the remaining objects for training the model.

3.2 Evaluation Method

3.2.1 Pose Estimation Method

For the proposed method, any regression-based pose estimation method can be used. Since this part is not the core of the proposed method, we simply use a network architecture similar to the Pose-CyclicR-Net proposed by Ninomiya et al.[3)] as the pose estimator. Since we assume 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.

3.2.2 Evaluation Criteria

We evaluate how the recommended viewpoints are appropriate for the pose estimation by using several criteria. Basically, we evaluate it by comparing the pose estimation results using the initial viewpoint and the recommended viewpoint. One criterion is the Mean Absolute Error (MAE) of the pose estimation results to the ground truth. The pose estimation results are obtained by using a pair of the initial viewpoint and the recommended viewpoint. By

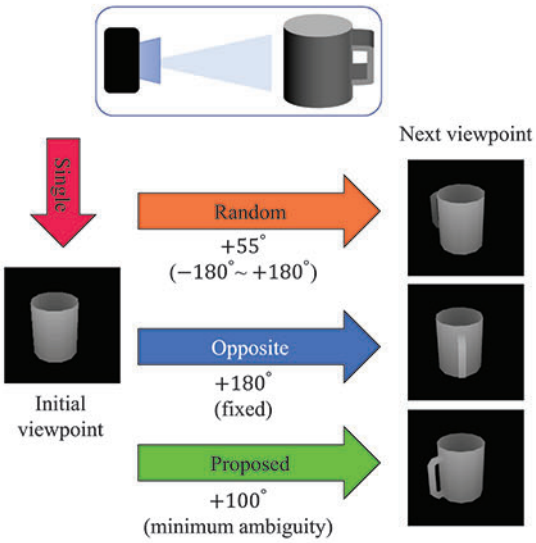

Fig. 7: Example of images observed from the estimated viewpoints by the proposed method and comparative methods

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 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

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

The other criterion is Pose Estimation Accuracy (PEA), which is 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 the ground truth θ_g^i, $F(\cdot)$ is a function which returns 1 if the condition in the function holds and 0 vice versa.

Standard deviation over the five-fold cross-validation is also evaluated.

3.2.3 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 work that could be directly compared with the proposed next best-viewpoint recommendation method for the category-level object pose estimation task except for Sock et al.'s work. They proposed several methods; one is "Random" which randomly selects the next viewpoint, and the other is "Furthest" (in this paper, we called it "Opposite" instead) which recommend the completely opposite side. We use these methods as comparative methods. We also prepared a pose estimation method from a single viewpoint, which just applies a Pose-CyclicR-Net-like network to the input image from the initial viewpoint as a baseline method.

Table 1: Comparison of MAE for the five categories when the elevation angle is 0° by five-fold cross validation

Target Object	Single	Random	Opposite	Proposed
"Airplane"	12.86°	11.76°	11.79°	**11.45°**
"Car"	8.73°	7.75°	8.12°	**7.09°**
"Chair"	10.88°	8.24°	**7.30°**	8.77°
"Mug"	13.07°	11.34°	11.03°	**9.58°**
"Toilet"	10.12°	8.54°	7.96°	**7.66°**

Table 2: Comparison of MAE for different elevation angles of "Mug" by five-fold cross validation

Elevation Angle	Single	Random	Opposite	Proposed
0°	12.55°	10.67°	10.14°	**9.34°**
15°	12.63°	10.79°	10.36°	**9.51°**
30°	12.11°	10.25°	9.97°	**8.81°**
45°	8.98°	7.46°	7.12°	**6.76°**
60°	6.05°	5.04°	5.07°	**4.68°**
75°	4.80°	4.01°	**3.83°**	4.44°

Table 3: Comparison using Partial-AUC (pAUC) of Pose Estimation Accuracy by changing the error threshold for for the five categories when the elevation angle is 0° by five-fold cross validation

Target Object	Single	Random	Opposite	Proposed
"Airplane"	89.01	89.01	89.03	**89.65**
"Car"	92.35	92.00	91.52	**92.58**
"Chair"	88.76	91.27	**92.19**	90.72
"Mug"	87.12	88.35	88.46	**90.24**
"Toilet"	90.72	91.20	91.55	**92.23**

Table 4: Comparison using Partial-AUC (pAUC) of Pose Estimation Accuracy by changing the error threshold τ from 0° to 100° by five-fold cross validation

Elevation Angle	Single	Random	Opposite	Proposed
0°	87.73	89.09	89.52	**90.68**
15°	87.79	88.91	89.16	**90.53**
30°	88.23	89.44	89.74	**91.11**
45°	91.01	92.13	92.50	**92.95**
60°	93.68	94.49	94.44	**94.92**
75°	94.85	95.51	**95.64**	95.14

Fig. 7 shows an example of outputs of each viewpoint recommendation method for a given "Mug" image.

3.3 Results

3.3.1 Mean Absolute Error (MAE)

We conducted pose estimation error analysis with various object categories; "Airplane" "Car", "Chair", "Mug", and "Toilet". **Table 1** tabulates the MAE comparison between all five object categories when the elevation angle is 0°. We see that the proposed method almost outperforms the comparative methods for different target objects. Only in the "Chair" category, the proposed method ranks the third. Overall, this result shows that the proposed method is capable to produce a good estimation for various categories.

Then, to investigate the relation between the elevation angle with the pose estimation method, we choose "Mug" as the target object and show the results in **Table 2**. For most of the elevation angles, the proposed method almost outperforms the comparative methods. The "Opposite" method outperforms the proposed method only in the case of 75° elevation angle. For this elevation angle, the proposed method is not the best but still could be considered comparable to the comparative methods. However, since in our work, the main priority is the pose estimation from 0° to 45° considering the ambiguity problem, these cases become less critical.

These results clearly show that the proposed method is effective and gives a better way (next viewpoint) for object pose estimation. We successfully managed to reduce the pose ambiguity in the difficult observation which has been mentioned earlier in Fig. 2. We can see that estimating an object's pose from two viewpoints yields a better result than that from a single viewpoint. By comparing with the other pose recommendation methods, the proposed method achieves better results by carefully selecting the best viewpoint for object pose estimation.

3.3.2 Pose Estimation Accuracy (PEA)

In general, smaller error is the main priority for pose estimation analysis with the comparative methods. We analyze the pose estimation accuracy by changing the error threshold τ in Equation (10) in the case of elevation angle 0°. To see the performance of the proposed method with the various categories, we calculated partial-AUCs (pAUCs) as summarized in **Table 3**. The proposed method achieved the highest pAUCs than the comparative methods for the four object categories.

Using "Mug" as the target object category, an evaluation for PEA is conducted with the elevation angles, from 0° to 75°. In five out of six elevation angles, the proposed method achieved the most accurate results compared to the other methods as illustrated in **Table 4**. For the remaining elevation angle, 75°, the proposed method was not the best but could be considered comparable to the comparative methods. This is because object poses are easily distinguishable from the initial viewpoints. In these cases, viewpoint selection becomes less important.

3.3.3 Example of Pose Ambiguity Minimization Output Result

In **Table 5**, estimation results of different inital viewpoints are shown. We divided the initial viewpoints into four groups as shown in **Fig. 8**; Viewpoint group 1 (0° – 44 ° and 315° – 359°), viewpoint group 2 (45° – 134 °), viewpoint group 3 (135° – 224°), and viewpoint group 4 (225° – 314°). The proposed method outperforms the comparative methods in most viewpoint groups.

For qualitative study, we compare the initial viewpoint image with the best next-viewpoint using the "Mug" images. Since the proposed method could suggest and select the best next-viewpoint, even though we have an ambigious image as the initial viewpoint, the proposed method could still estimate the object's pose accurately. **Table 6** provides the output examples from a less ambigious

Table 5: Comparison of MAE for "Mug" for different viewpoint areas (0° elevation angle)

Viewpoint group	Single	Random	Opposite	Proposed
1	9.03°	7.31°	6.30°	**6.00°**
2	17.66°	11.43°	**9.23°**	9.97°
3	7.13°	6.77°	6.30°	**6.01°**
4	6.97°	6.71°	9.22°	**5.41°**

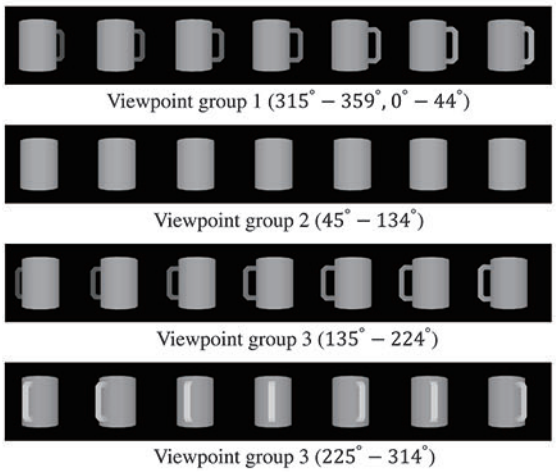

Fig. 8: Image examples for the four group area using the "Mug" object category

initial viewpoint for the proposed method and comparative methods. We can see that the proposed method achieves better pose estimation results than the comparative methods.

3.3.4 Statistical Analysis of Pose Estimation Errors

For the statistical evaluation of the results introduced above, we show the boxplot of the pose estimation errors in **Fig. 9**. This boxplot graph indicates the minimum, lower quartile, median, upper quartile, and maximum values. The median values of the results of the proposed method shows the lowest value for "Airplane" and "Mug". For "Car" and "Toilet" the proposed method performs compatatively with the "Opposite" method. However, for "Chair", the proposed method ranked the third. The outlier's absence shows that the proposed method delivers a promising approach of estimating the pose from an ambiguous viewpoint.

By focussing on the elevation angle at 0° with "Mug" images, the pose estimation error and the pAUCs of PEA for each fold is illustrated in **Table 7**. The standard deviation of the proposed method is 3.92 which shows that the proposed method is the second stable among all the methods. For the PEA analysis, for the fold numbers 2 to 5, the proposed mwthod gained the highest estimation accuracy. The standard deviation of the proposed method, for the PEA analysis, is 3.05 which also shows that the proposed method is the most stable method among all the methods.

4 Conclusion

We proposed a new framework to recommend the best next viewpoint for an accurate pose estimation for category-level object pose estimation. We also proposed a new method for finding the minimum pose ambiguity. We showed that the proposed method statistically outperforms three comparative methods in the category-level pose estimation through five-fold cross-validation. By using the recommended viewpoints, a high pose estimation accuracy was stably achievable. This approach could help the development of the human helper robot field.

For future improvement, we are planning to extend the proposed method to multi-dimensional axis rotation. Expanding the best next-viewpoints recommendation to multi-step viewpoints recommendation has also been projected as our upcoming task. In this paper, we used a simple averaging for the two pose estimation results ϕ_1 and ϕ_2 and considering the rotation angle $\widehat{\delta}$ as shown in Equation (9), but there is room for futher improvement. Handling real depth data with heavy noise is one of future work.

Table 6: Comparison of MAE for "Mug" when the elevation angle is 0° for viewpoint areas. The value in bracket represent the different between image initial viewpoint and the pose estimation result

Image	Single	Random	Opposite	Proposed
 0.0°	342.0° (18.0°)	356.0° (3.5°)	357.5° (2.5°)	0.0° (**0.0°**)
 30.0°	40.0° (10.0°)	39.5° (9.5°)	41.5° (11.5°)	32.0° (**2.0°**)
 210.0°	219.0° (9.0°)	223.0° (13.0°)	220.5° (10.5°)	207.5° (**2.5°**)
 330.0°	298.0° (32.0°)	292.0° (38.0°)	298.5° (31.5°)	326.5° (**3.5°**)

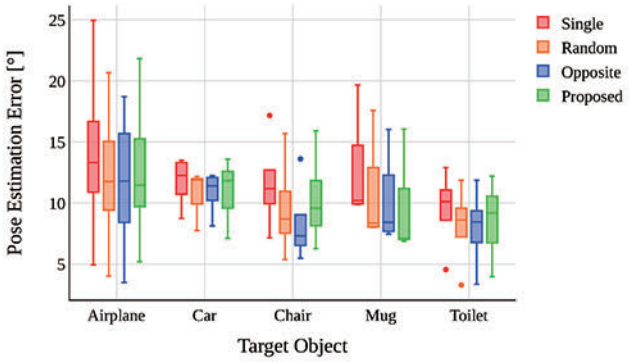

Fig. 9: Boxplot for all target object for 0° elevation angle

Acknowledgements

The authors would like to thank the Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Education (MOE) Malaysia for the financial support under the scholarship of Skim Latihan Akademik IPTA (SLAI). Parts of this research were supported by MEXT Grant-in-Aid for Scientific Research (17H00745).

Table 7: Comparison of the overall Pose Estimation Accuracy (PEA) and Partial-AUC (pAUC) for each fold (0° elevation angle)

Testing	Single		Random		Opposite		Proposed	
Fold Number	PEA	pAUC	PEA	pAUC	PEA	pAUC	PEA	pAUC
Fold 1	19.68°	82.23	17.59°	82.91	**16.03°**	84.29	16.06°	**85.53**
Fold 2	13.07°	87.12	11.34°	88.35	11.03°	88.46	**9.58°**	**90.24**
Fold 3	10.20°	89.36	8.05°	91.45	7.76°	91.73	**6.85°**	**92.59**
Fold 4	9.92°	90.09	8.00°	91.59	7.45°	92.04	**7.10°**	**92.59**
Fold 5	9.88°	89.87	8.35°	91.15	8.41°	91.09	**7.09°**	**92.42**
Average	12.55°	87.23	10.67°	89.09	10.14°	89.52	**9.34°**	**90.68**
σ	4.20°	3.29	4.11°	3.70	**3.59°**	3.25	3.92°	**3.05**

Reference

1) RT. Chin and CR. Dyer: Model-based recognition in robot vision, ACM Computing Surveys, **18**, 1, (1986) 67.
2) H. Murase and SK. Nayar: Visual learning and recognition of 3-D objects from appearance, Int. J. of Computer Vision, **14**, 1, (1995) 5.
3) H. Ninomiya, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, N. Kobori and Y. Nakano: Deep manifold embedding for 3D object pose estimation, Proc. 12th Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications, (2017) 173.
4) A. Zeng, KT. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez and J. Xiao: Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge, Proc. 2017 IEEE Int. Conf. on Robotics and Automation, (2017) 1386.
5) Ö. Erkent, D. Shukla and J. Piater: Integration of probabilistic pose estimates from multiple views, Proc. 14th European Conf. on Computer Vision, **7**, (2016) 154.
6) A. Collet and SS. Srinivasa: Efficient multi-view object recognition and full pose estimation, Proc. 2010 IEEE Int. Conf. on Robotics and Automation, (2010) 2050.
7) A. Kanezaki, Y. Matsushita and Y. Nishida: RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints, Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition, (2018) 5010.
8) F. Vikstén, R. Söderberg, K. Nordberg and C. Perwass: Increasing pose estimation performance using multi-cue integration, Proc. 2006 IEEE Int. Conf. on Robotics and Automation, (2006) 3760.
9) R. Bajsy: Active perception vs. passive perception, Proc. 1985 IEEE Workshop on Computer Vision: Representation and Control, (1985) 55.
10) D. Wilkes, SJ. Dickinson and JK. Tsotsos: A quantitative analysis of view degeneracy and its use for active focal length control, Proc. 5th IEEE Int. Conf. on Computer Vision, (1995) 938.
11) R. Pito: A sensor-based solution to the "next best view" problem, Proc. 13th Int. Conf. on Pattern Recognition, (1996) 941.
12) JE. Banta, LR. Wong, C. Dumont and MA. Abidi: A next-best-view system for autonomous 3-D object reconstruction, IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans, **30**, 5, (2000) 589.
13) A. Doumanoglou, R. Kouskouridas, S. Malassiotis and TK. Kim: Recovering 6D object pose and predicting next-best-view in the crowd, Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, (2016) 3583.
14) J. Sock, S. H. Kasaei, LS. Lopes and TK. Kim, Multi-view 6D object pose estimation and camera motion planning using RGBD images, Proc. 2017 IEEE Int.Conf. on Computer Vision Workshops, (2017) 2228.
15) AX. 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 Resipotory arXiv preprint, 1512.03012, (2015).
16) NMZ. Hashim, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, A. Amma and N. Kobori: Next viewpoint recommendation by pose ambiguity minimization for accurate object pose estimation, Proc. 14th Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications, (2019) 60.