2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

2D-3D Object Shape Alignment for Camera-Object Pose Compensation in Object-Visual SLAM

Hanyeol Lee[1], Jae Hyung Jung[2], and Chan Gook Park[1]

***Abstract*— In this study, we propose an object shape alignment method through a robust optimization scheme for 6-degrees-of-freedom (DOF) object pose compensation. Although the pose estimation of the 3D object by the camera has been rapidly improved in recent years with the development of deep learning, the estimate still contains errors due to several factors. To compensate for this, we perform a shape alignment between the 2D segmentation of the object and the projection of the 3D object in the image plane. To avoid convergence to a local minimum in nonlinear optimization, we separate the pose into translation and rotation. This approach derives the optimization of a linear form in terms of a translation with reduced computational cost. For the rotation, the parallel optimization is performed with multiple initial values, reflecting to the uncertainty of an initial value. We formulate an invariant extended Kalman filter (EKF)-based object-visual simultaneous localization and mapping (SLAM) with a camera-object relative pose as the measurement model. To verify the performance of the proposed algorithm, we present the improved results of camera-object relative pose accuracy and localization and mapping accuracy in the several sequences of YCB-video dataset.**

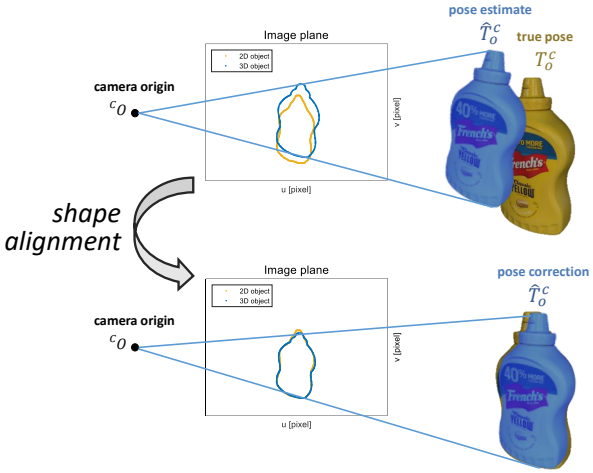


Fig. 1. The shape alignement between a projection of 3D object estimate and 2D segmentation of an object.

I. Introduction

The semantic perception of an object is essential to perform robotic tasks such as object manipulation or exploration to find an object. For this purpose, the problem of classifying objects and estimating their 6-DOF pose in 3D space has been widely studied. Classically, there are methods in the field of computer vision research to estimate the 6-DOF pose of a 3D object based on template/feature matching [1]–[7]. Recently, with the development of deep learning, pose estimation of an object has become one of the major research interests, and learning-based 3D object pose estimators have been actively proposed [8], [9]. With this advance, it is possible to obtain a pose estimate with higher accuracy than a model-based pose estimator. To obtain a probabilistic solution of the object pose expressed in a fixed frame, it is necessary to map the object measurements from the camera frame to a fixed frame. For tasks that require real-time processing, like high-level robotic tasks, object pose estimation should be performed with localization based on the SLAM framework. Recently, object-visual SLAM research has been actively pursued with progress in robotics [10], [11]. However, one of the remaining issues is how to deal with measurement errors obtained from a deep neural network (DNN). If there is an error in the relative pose estimate between the camera and the object, it will be propagated to the localization and mapping errors. Since the probabilistic error distribution of learning-based measurements is difficult to model mathematically, it is also difficult for the estimator to reflect the uncertainty of a measurement. This means that the performance of an estimator, such as optimal information fusion and consistency of estimation, can be degraded by unknown uncertainty. Unfortunately, state-of-the-art 6-DOF pose estimators also have inevitable errors caused by camera resolution, clutter, occlusion, motion blur, and unknown error of a neural network, etc.

In this study, we propose a method to compensate the error of the 6-DOF camera-object relative pose obtained from a learning-based pose estimator. As shown in Fig. 1, the shape alignment of a projected 3D object with a 2D object segmentation in the image plane is performed by the proposed robust optimization scheme. For the convergence of the 6-DOF pose with a high dimension compared to the 2D measurement in the image plane, the optimization framework is composed of two sequential procedures. First,

*This research was supported in part by the Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea (NRF), Unmanned Vehicle Advanced Research Center (UVARC) funded by the Ministry of Science and ICT, the Republic of Korea (No. 2020M3C1C1A01086408), and in part by the National Research Foundation of Korea funded by the Ministry of Science and ICT, the Republic of Korea, under Grant NRF-2022R1A2C2012166.

[1]Hanyeol Lee and Chan Gook Park are with the Department of Aerospace Engineering, and Automation and Systems Research Institute, Seoul Natioinal University, Seoul 08826, South Korea han2110@snu.ac.kr; chanpark@snu.ac.kr

[2]Jae Hyung Jung was with the Department of Aerospace Engineering, and Automation and Systems Research Institute, Seoul Natioinal University, Seoul 08826, South Korea, and is corrently with Smart Robotics Lab, School of Computation, Information and Technology, Technical University of Munich, Germany jaehyung.jung@tum.de

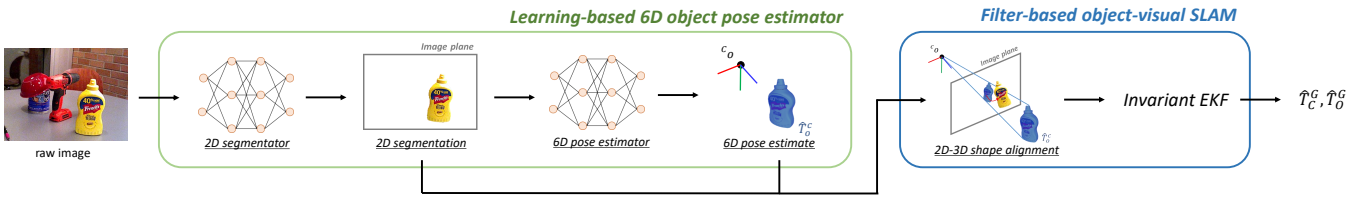


Fig. 2. The overview of our algorithm. From the raw image, 2D segmentation and object pose estimate are extracted by DNN. They are used to obtain more accurate object pose through 2D-3D shape alignment. Finally, object SLAM is performed using the compensated pose.

linear translation compensation, by decoupling translation and rotation, is performed to obtain an accurate initial value for nonlinear optimization. Second, an iterative closest point (ICP) [12] is performed in the 2D image plane for a 6-DOF pose compensation with multiple initial values considering robustness. The obtained measurements are processed in an estimator formulated based on an invariant EKF to perform object SLAM with consistent estimation. We present improved accuracy of the proposed method through a real-world dataset [13].

- We perform to align a projection of 3D object in the image plane to 2D segmentation using the proposed robust optimization scheme. This enable to obtain accurate 6-DOF camera-object relative pose measurements.
- We formulate an invariant EKF-based object-visual SLAM using the camera-object relative pose compensated by a proposed shape alignment.
- We show the improved results of measurement, localization, and mapping by the proposed method on several sequences of the YCB-video dataset.

II. Related Works

The 6-DOF pose estimation of an object had been traditionally studied based on modeling before learning-based methods led the field. There have been studies on template-based methods to derive a pose estimation solution using similarity in a template [1]–[4] and feature-based methods to obtain it through 2D-3D feature correspondence [5]–[7]. Early studies on learning-based object pose estimation have been proposed [4], [6], [14], including poseCNN [13], which is an end-to-end 6-DOF pose estimation algorithm with robustness to occlusion. Relatively accurate 2D mask of an object could be obtained as an improvement of 2D segmentation research [15], and CosyPose [8] provides accurate 6-DOF pose estimate of an object based on 2D mask with high quality. Recently, in the BOP Challenge 2022, [17], GDR-Net [16], and ZebraPose [9] showed high performance in object segmentation. Although the accuracy of object pose estimation has improved recently, there is still an inherent error due to the various factors mentioned in the introduction. Y. Hu *et al*. [18] fuses the corner point of the 3D bounding box (Bbox) and the 2D segmentation of an object using RANSAC and EPnP [19] to obtain the object pose. ContourSLAM [20] formulates the measurement by aligning the 2D segmentation and 3D object model contours using the RGB-D camera with depth information. In this study, we propose the optimization scheme to compensate the relative camera-object pose by aligning a projection of the 3D object model to the 2D object segmentation in the RGB camera. Our work enables object shape alignment even for an RGB camera that cannot obtain direct depth measurements from the sensor.

In object-visual SLAM, the object is often modeled as a 3D Bbox [10], an ellipsoid [21], or segmented based on a CAD model [22] if available. We select the 3D object model considering the more precise representation of the object. An invariant EKF has been proposed for consistent state estimation, which allows linearization to be performed independently with the estimate [23]. It has been actively used in state estimation problems in robotics due to its robustness to linearization points [11]. In this study, we formulate an invariant EKF to jointly estimate the state of the object and the camera with prior knowledge of the object model.

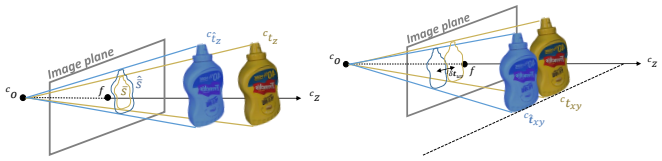
III. 2D-3D Shape Alignment

In general, the 6-DOF pose estimation of an object is a more difficult problem than the 2D segmentation of an object in the image plane, due to its relatively high dimensionality. Motivated by this, our goal is to compensate the 6-DOF pose of an object from DNN using the 2D segmentation of the object. An overview of our algorithm is shown in Fig. 2. DNN extracts a 2D segmentation estimate as output from the raw image and uses it as input to estimate the 6-DOF object pose. In the proposed method, the 2D segmentation output is again used to compensate the object pose from a final output of DNN. Using the compensated pose as a measurement, an invariant EKF-based object-visual SLAM is performed.

The details of object shape alignment are as follows. The 3D object is projected onto the image plane to align with the 2D segmentation of the object in the image plane. The two 2D point clouds on the image plane are aligned by the optimization. Since this problem optimizes a 6-DOF pose with 2D measurements, a converged solution is highly sensitive to the initial values. To address these limitations, we propose an optimization scheme that consists of linear optimization, decoupling of translation and rotation, and nonlinear optimization with multi-start points considering the uncertainty of an initial value.

A. Linear optimization formulation by decoupling pose

The 6-DOF relative pose of a camera-object is defined by its translation and rotation. The z-axis translation ${}^{C}t_z$ of an object expressed in a camera frame is related to a scale



(a) (b)

Fig. 3. (a) The scale result according to a z-axis translation. (b) The offset result according to a x-axis and y-axis translation.

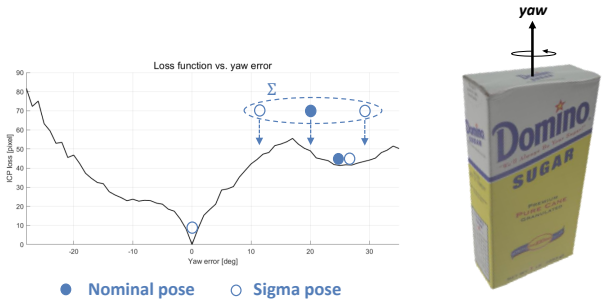


Fig. 4. The loss function and sigma pose definition in terms of yaw error.

of an object as shown in Fig. 3(a). Assuming that there is not rotation error of the object, a scale in an image plane is inversely proportional to ${}^{C}t_z$, which includes the intrinsic parameter, as

$$\bar{s} = \frac{f}{{}^{C}t_z} s \tag{1}$$

where s is a scale of an object, $\bar{s}$ is a projected scale in a image plane, and f is the focal length among the camera intrinsics. By expressing the scale of two point clouds as a ratio, we derive the relationship between scale and z-axis translation,

$$\frac{\hat{\bar{s}}}{\bar{s}} = \frac{\frac{f}{{}^{C}\hat{t}_z} s}{\frac{f}{{}^{C}t_z} s} = 1 + \frac{{}^{C}\delta t_z}{{}^{C}\hat{t}_z} \tag{2}$$

where

$${}^{C}\delta t_z = {}^{C}t_z - {}^{C}\hat{t}_z. \tag{3}$$

${}^{C}\hat{t}_z$ is an estimate of the z-axis translation and ${}^{C}\delta t_z$ is error of the z-axis translation. The scale corresponds to the perimeter in the image plane. To calculate this, the boundaries of the two point clouds are extracted. This also serves to lighten ICP algorithm [12], which will be introduced in Section III. B, matching only the sets of boundary points. Rephrasing (2) with the perimeter calculated from the boundaries instead of the scales yields

$${}^{C}\delta t_z = {}^{C}\hat{t}_z (p_r - 1) \tag{4}$$

where

$$p_r = \hat{\bar{p}} / \bar{p}. \tag{5}$$

$\bar{p}$ is a perimeter of 2D object segmentation in an image plane, $\hat{\bar{p}}$ is a projected perimeter of 3D object, and p_r is a ratio between these perimeters. This results in the optimization problem described in

$${}^{C}\delta t_z^* = \arg\min_{{}^{C}\delta t_z} \| 1 - p_r \| \tag{6}$$

with a closed-form solution as

$${}^{C}\hat{t}_z^* = {}^{C}\hat{t}_z + {}^{C}\delta t_z^*. \tag{7}$$

As shown in Fig. 3(b), the x,y axis translations ${}^{C}t_x$, ${}^{C}t_y$ of an object in the camera frame are converted by the pixel translations ${}^{I}t_u$, ${}^{I}t_v$ in the image plane. Applying this to the two sets of boundary points results in

$${}^{I}\mathbf{t}_{uv} = \begin{bmatrix} {}^{I}t_u \\ {}^{I}t_v \end{bmatrix} = \begin{bmatrix} f({}^{C}t_x / {}^{C}t_z) + c_x \\ f({}^{C}t_y / {}^{C}t_z) + c_y \end{bmatrix} \tag{8}$$

where ${}^{I}t_u$ and ${}^{I}t_v$ are the pixel coordinates in an image plane, respectively, and ${}^{C}t_x$ and ${}^{C}t_y$ are the x and y coordinates of point in a camera frame, respectively. c_x and c_y are the principal point offsets of each axis among camera intrinsics. The equation between a translation error of x, y coordinates ${}^{C}\delta\mathbf{t}_{xy}$ and a pixel error ${}^{I}\delta\mathbf{t}_{uv}$ is as

$${}^{C}\delta\mathbf{t}_{xy} = \begin{bmatrix} {}^{C}\delta t_x \\ {}^{C}\delta t_y \end{bmatrix} = \begin{bmatrix} {}^{C}\hat{t}_z {}^{I}\delta t_u / f \\ {}^{C}\hat{t}_z {}^{I}\delta t_v / f \end{bmatrix} = \frac{{}^{C}\hat{t}_z}{f} {}^{I}\delta\mathbf{t}_{uv} \tag{9}$$

where

$${}^{C}\delta\mathbf{t}_{xy} = {}^{C}\mathbf{t}_{xy} - {}^{C}\hat{\mathbf{t}}_{xy} \text{ and } {}^{I}\delta\mathbf{t}_{uv} = {}^{I}\mathbf{t}_{uv} - {}^{I}\hat{\mathbf{t}}_{uv}. \tag{10}$$

To apply this equation to correct the pose of an object, ${}^{I}\delta\mathbf{t}_{uv}$ is calculated as the offset between the centroids of the set of boundary points in the image plane. Similar to the z-axis optimization problem, the optimization problem expressed in

$${}^{C}\delta\mathbf{t}_{xy}^* = \arg\min_{{}^{C}\delta\mathbf{t}_{xy}} \| {}^{I}\delta\mathbf{t}_{uv} \| \tag{11}$$

has a closed-form solution as

$${}^{C}\hat{\mathbf{t}}_{xy}^* = {}^{C}\hat{\mathbf{t}}_{xy} + {}^{C}\delta\mathbf{t}_{xy}^*, \tag{12}$$

assuming that ${}^{C}\delta t_z$ and roatation errors are small. The translation part in the assumption is reasonable due to the compensation of ${}^{C}\hat{t}_z$ from (7) and the rotation part is described in Section III. B. This linear optimization form enables to obtain a precise initial translation estimates as only a sum and product with little computation. This is the first step in the proposed optimization scheme providing the accurate initial values of the nonlinear optimization for 6-DOF pose in the next subsection.

B. 2D-ICP in terms of 6-DOF pose

We construct 2D-ICP in the image plane to estimate the camera-object pose. However, unlike the corrected initial value for translation in Section III.A, the initial value for rotation is a raw measurement from the DNN-based object pose estimator. In the problem of estimating 6-DOF pose with a loss in the 2D image plane, an inaccurate initial value greatly increases the risk of convergence to a local minimum more than in general 3D-ICP for 6-DOF pose. To address this, we formulate the optimization approach with multi-start points about each rotation. Fig. 4 shows the 2D-ICP loss as a function of the object's yaw axis error. To converge to the

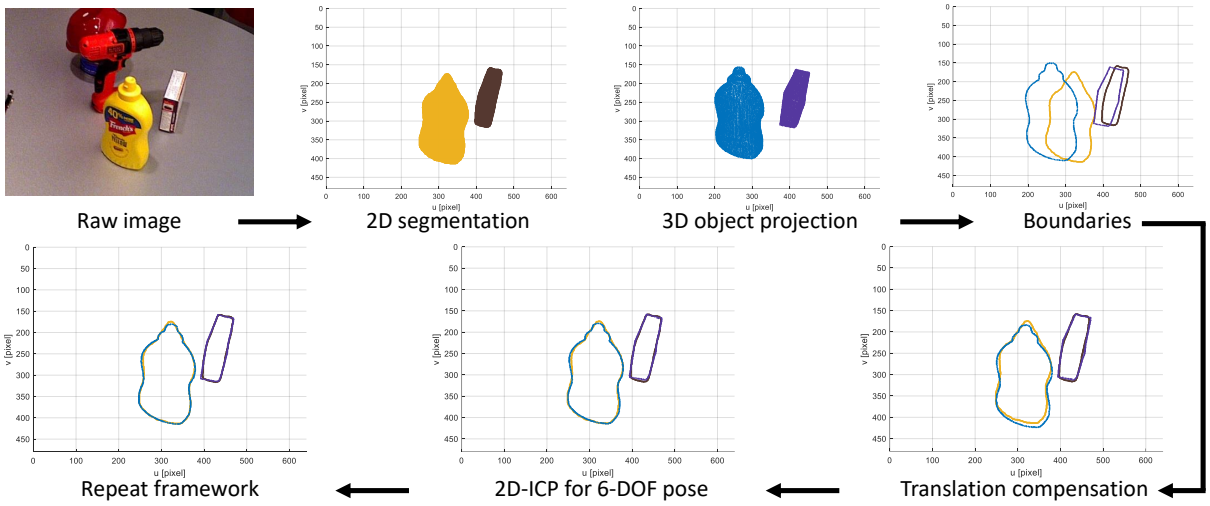


Fig. 5. A visualization of the 2D-3D shape alignment process. 2D segmentation and 3D object projection are obtained from the DNN and the object model. Their boundaries are used for linear optimization of translation and 2D-ICP is performed using accurate initial values. In order to satisfy the assumption for pose decoupling, the entire framework is iterated.

appropriate minimum, it is assumed that there is uncertainty Σ in the initial value of rotation. The sigma pose ϕ_σ is then given by the nominal pose ϕ and the standard deviation σ about each axis as

$$\begin{aligned} \phi_\sigma &= \phi + n\sigma_\phi \text{ and } \phi - n\sigma_\phi \\ \theta_\sigma &= \theta + n\sigma_\theta \text{ and } \theta - n\sigma_\theta \\ \psi_\sigma &= \psi + n\sigma_\psi \text{ and } \psi - n\sigma_\psi. \end{aligned} \tag{13}$$

The initial poses ξ are composed of a nominal pose and $2n$ sigma poses about each rotation axis, and the number of initial values is $(2n+1)^3$. We set $n=1$ and process them in parallel considering the computational complexity. The optimization of an object pose is performed with multiple initial rotation values as

$$\hat{T}_O^{C^*} = \arg\min_{\hat{T}_O^C} \min_{\xi_i} \sum_{{}^I\mathbf{p},{}^O\mathbf{p}} e \tag{14}$$

where

$$e = \|{}^I\mathbf{p} - h({}^O\mathbf{p})\|^2 \text{ and } h({}^O\mathbf{p}) = \pi(R_O^C {}^O\mathbf{p} + {}^C\mathbf{t}_{co}). \tag{15}$$

ξ_i is a component in ξ and π is the projection function of point from a camera frame to the image plane. ${}^O\mathbf{p}$ is a projected boundary point in the image plane obtained from a point cloud of the object model and ${}^I\mathbf{p}$ is a boundary point in the image plane from 2D segmentation. R_O^C and ${}^C\mathbf{t}_{co}$ are the rotation and the translation from T_O^C. (14) is solved iteratively by the Gauss-Newton method including a regularization term with respect to a displacement to an initial value, and the details are as shown in

$$\xi_{ik+1} = \xi_{ik} - (J^T J + \mu I)^{-1} J^T e \tag{16}$$

where

$$J = -2({}^I\mathbf{p} - h({}^O\mathbf{p}))^T \frac{\partial {}^I\mathbf{p}}{\partial {}^I\tilde{\mathbf{p}}} \frac{\partial {}^I\tilde{\mathbf{p}}}{\partial {}^C\tilde{\mathbf{p}}} \frac{\partial {}^C\tilde{\mathbf{p}}}{\partial \xi_i}. \tag{17}$$

To satisfy the assumption of a small rotation error in Section III. A, the entire framework in Section III is carried out m times, where m is a hyperparameter set to 2. Lastly, we set the convergence criterion of the optimization based on the magnitude of the final loss as

$$\hat{T}_O^C = \hat{T}_O^{C^*}, \text{ if } \overline{\sqrt{e}} < \lambda. \tag{18}$$

$\overline{\sqrt{e}}$ is a mean of the distance of matched points and λ is set to 10 pixel. If the shape alignment does not converged, we use the raw T_O^C obtained from the object pose estimator. The overall shape alignment process is shown in Fig. 5.

IV. Object-visual SLAM

In this section, we describe how to formulate object-visual SLAM with camera-object pose measurements using an invariant EKF [11] based on the optimal error formulation for $SE(3)$. Our state is described by

$$X = (T_C^G, T_{O_1}^G, ..., T_{O_i}^G). \tag{19}$$

i is the number of mapped objects. The system model for the invariant EKF can be obtained from various sensors, such as IMU, odometer, etc., as

$$\frac{d}{dt} T_C^G(t) = T_C^G(t) \exp((\mathbf{u}(t) + \mathbf{n}_\omega(t))^\wedge) \tag{20}$$

$\mathbf{u}(t)$ is an odometry measurement and $\mathbf{n}_\omega(t)$ is its noise. $\exp((\cdot)^\wedge)$ is an exponential mapping from $\mathfrak{se}(3)$ to $SE(3)$ in Lie theory. We express the camera-object relative pose measurement model for filter update as

$$Y_j(t_k) = T_G^C(t_k) T_O^G(t_k) \exp(\mathbf{n}_v(t_k)^\wedge). \tag{21}$$

$Y_j(t_k)$ is a camera-object relative pose measurement at time t_k and $\mathbf{n}_v(t_k)$ is an measurment noise and $\mathbf{n}_v \sim N(0, \sigma_v^2)$. We set $\mathbf{n}_v$ adaptively depending on whether the shape alignment converges or not as

$$\sigma_v = \begin{cases} k\sigma_v, & \text{if (17) is converged} \\ \sigma_v, & \text{otherwise} \end{cases}, \tag{22}$$

and k is set to 0.5.

TABLE I

ROOT MEAN SQUARE ERROR (RMSE) OF MEASUREMENT, LOCALIZATION, AND MAPPING IN YCB-VIDEO DATASET

	Measurement				Localization				Mapping			
	CosyPose		Proposed		CosyPose		Proposed		CosyPose		Proposed	
	rot[deg]	trans[cm]	rot[deg]	trans[cm]	att[deg]	pos[cm]	att[deg]	pos[cm]	att[deg]	pos[cm]	att[deg]	pos[cm]
seq.00	4.12	**1.12**	**3.20**	2.03	8.41	5.77	**6.66**	**5.32**	14.72	14.43	**10.28**	**12.26**
seq.00+rot.	**7.52**	**1.12**	8.81	2.92	14.42	**4.69**	**10.67**	5.71	17.05	**14.71**	**13.40**	16.87
seq.00+trans.	4.12	9.96	**3.26**	**5.55**	8.31	6.80	**5.79**	**6.44**	16.71	17.81	**8.73**	**11.87**
seq.00+total	7.52	9.96	**5.99**	**6.86**	14.15	**6.49**	**11.41**	6.91	19.41	17.81	**12.52**	**16.55**
seq.25	2.39	**0.91**	**2.16**	2.37	4.76	6.03	**3.11**	**5.98**	0.97	1.42	**0.82**	**1.37**
seq.25+rot.	5.83	**0.91**	**4.92**	2.69	8.79	6.66	**5.91**	**6.35**	2.32	3.21	**1.59**	**1.99**
seq.25+trans.	2.39	8.91	**2.14**	**6.54**	4.41	8.08	**3.28**	**6.57**	**0.73**	1.78	0.95	**1.60**
seq.25+total	5.83	8.91	**5.17**	**6.85**	8.35	9.41	**6.17**	**7.52**	2.11	3.48	**1.73**	**2.28**
seq.72	4.44	**2.51**	**4.00**	2.57	8.95	**8.21**	**6.81**	8.90	2.23	1.99	**1.80**	**1.05**
seq.72+rot.	6.87	**2.51**	**6.26**	2.70	13.93	**8.52**	**9.82**	9.50	2.86	2.99	**2.10**	**2.16**
seq.72+trans.	4.44	9.11	**4.02**	**6.51**	9.51	**7.82**	**7.29**	8.70	2.50	2.18	**2.02**	**1.46**
seq.72+total	6.87	9.11	**6.26**	**6.49**	14.39	**8.96**	**10.63**	10.00	3.03	2.86	**2.38**	**2.29**
seq.76	5.37	**2.44**	**4.90**	2.74	9.15	**5.64**	**6.98**	5.81	3.11	2.90	**2.60**	**2.50**
seq.76+rot.	7.76	**2.44**	**7.02**	2.81	13.08	7.34	**10.00**	**7.08**	3.58	3.51	**2.95**	**2.62**
seq.76+trans.	5.37	9.49	**4.91**	**7.76**	9.97	6.04	**6.40**	**5.77**	3.33	3.17	**2.25**	**2.62**
seq.76+total	7.76	9.49	**7.13**	**7.58**	13.87	**7.03**	**10.55**	7.27	3.84	3.84	**3.11**	**3.55**

V. EXPERIMENTS

To validate the proposed method, we evaluate the accuracy of three parts in the several sequences of YCB-Video dataset [13]. We compute an average of the rotation/translation root mean square errors (RMSE) from the camera-objects relative pose errors for the validity of a shape alignment method, and the localization/mapping RMSE of object-visual SLAM for the accuracy of an entire algorithm. In detail, localization refers to the camera's pose estimation and mapping refers to the object's pose estimation. The mapping RMSE represents the average of the pose RMSEs of the estimated objects.

In the proposed algorithm, we adopt CosyPose [8] without the refine module as a learning-based camera-object pose estimator, considering both a low computational burden and an accuracy of a pose estimation. For an evaluation of the proposed shape alignment, we selected our algorithm using the raw measurement obtained by CosyPose as a comparison. To demonstrate the robustness of the proposed algorithm in a challenging condition where convergence is difficult due to initial errors, we perform an analysis in a situation where the measurement error is relatively large. We apply rotation error, translation error, and errors of both rotation and translation to the relative pose obtained from cosypose, which correspond to seq.+rot., seq.+trans., and seq.+total, respectively, in Table I. We implement this by introducing a white zero-mean Gaussian error to the relative pose measurement and setting the standard deviations for the introduced rotation and translation to σ_{rot} = 3deg and σ_{trans} = 5cm. We perform object-visual SLAM using keyframes of 1 Hz, considering that the motion in the dataset is small. It also assumes a non-symmetric, non-occluded object, and the uncertainty Σ for multiple initial values is set to 1^2 deg for each rotation axis, considering an error distribution of the object pose from CosyPose as shown in Fig. 6, which expresses each element of ${}^{C}\delta\phi$. The distributions are biased from the zero-mean and slightly different from the Gaussian distribution.

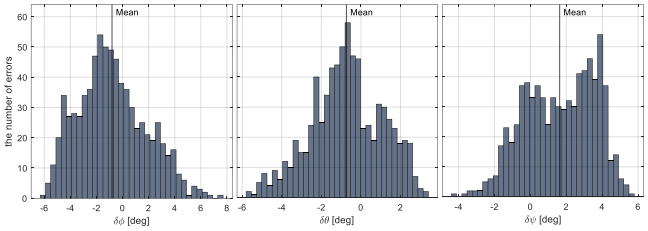
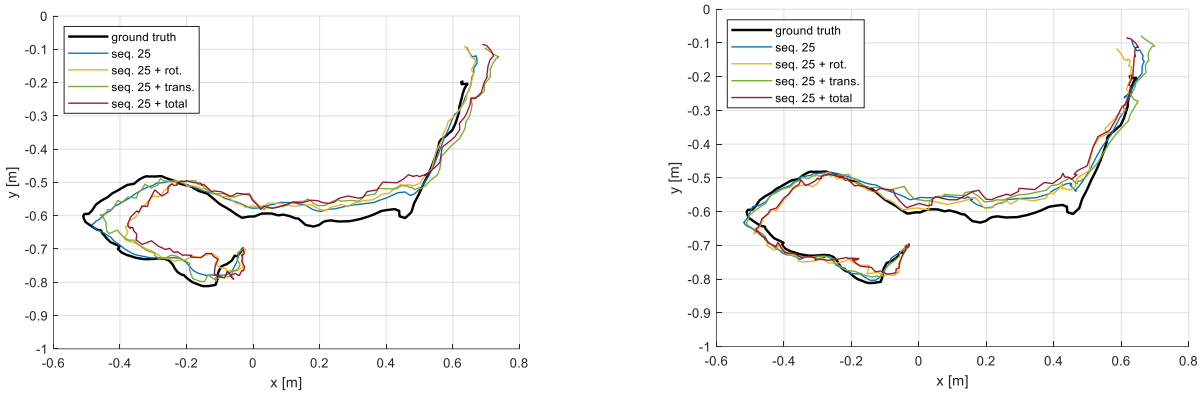


Fig. 6. The rotation error distributions of the single object pose estimate from CosyPose in seq.00 of the YCB-Video dataset.

As shown in Table I, we can see that in the original sequences, the rotation error is improved by using the proposed method compared to CosyPose. The translation error shows a different tendency from the rotation error, which is due to the error contained in the 2D segmentation measurements rather than a flaw in the proposed optimization scheme. The projection of the 3D object and the 2D segmentation have different shapes. For example, in Fig. 5, the shapes between the tops from the segmentation (yellow) and the projection (blue) of the object model of the same bottle are different. This results in inaccurate centroids and perimeters used for translation



(a) CosyPose (b) Proposed

Fig. 7. The estimated trajectories using the CosyPose and the proposed method are shown in (a) and (b) in the top view, respectively.

TABLE II
OBJECT POSE RMSE ACCORDING TO THE COMPENSATION METHODS

rot[deg]/trans[cm]	CosyPose	2D-ICP*	Proposed
seq.00+trans.	4.12/9.96	4.84/9.53	**3.26/5.55**
seq.25+trans.	2.39/8.91	4.22/8.35	**2.14/6.54**
seq.72+trans.	4.44/9.11	5.47/8.34	**4.02/6.51**
seq.76+trans.	5.37/9.49	6.21/8.66	**4.91/7.76**

* 2D-ICP means the compensation without the linear optimization.

compensation. We predict that these shape differences will be reduced when the accuracy of the 2D segmentation algorithm is improved according to the progress of deep learning. In the case where the rotation and translation measurements have an initial error, the RMSE of the measurements is reduced in both cases. Instead of converging to an inappropriate local minimum depending on the initial pose error, the proposed algorithm effectively performs 2D-3D shape alignment due to the robust design of the optimization scheme. Even when the initial rotation error and the initial translation error are combined, each initial error has little effect on the convergence of the other. For example, the rotation RMSE for seq.25+rot. and seq.25+total have similar values, as do the translation RMSEs for seq.25+trans. and seq.25+total. These results are interpreted to reflect the decoupling of translation and rotation to make the optimization robust to an initial error of their opposite. For a thorough validation of the proposed optimization scheme, the results of 2D-ICP without decoupling pose are shown in Table II. The proposed optimization scheme consistently yields lower error levels for several sequences than using only 2D-ICP. It shows that the linear optimization with decoupling has an effective role in compensating the pose by providing an accurate initial value.

The localization and mapping RMSEs in Table I show that the attitude and position of a camera and an object are estimated proportionally to the accuracy of the measurements. Overall, the attitude and position RMSE of the proposed method are improved compared to the case of using the raw measurement of CosyPose. In Fig. 7, for all conditions in

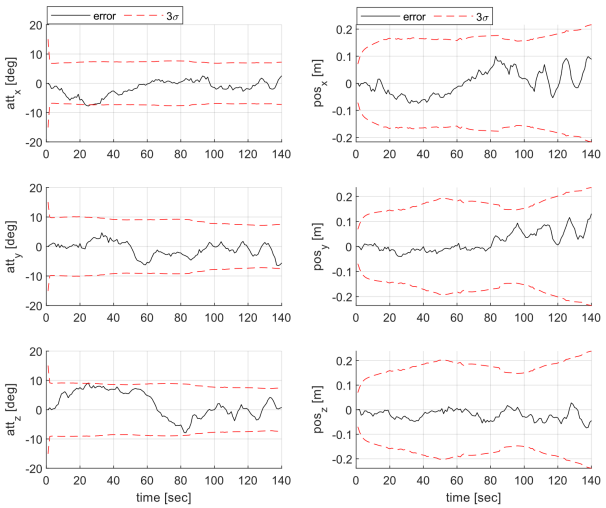


Fig. 8. Localization error and an estimated 3σ of the proposed algorithm in seq.25+total.

seq.25, the proposed method shows overall results close to the ground truth compared to CosyPose. For seq.25+total, which has the worst relative pose, the localization error and the 3σ from the estimated error-covariance are plotted in Fig. 8. Even in the most challenging case, the estimation performance satisfies the consistency of the filter in one-shot.

VI. CONCLUSION

In this study, we have proposed a shape alignment method between a projection of a 3D object and a 2D segmentation of an object. To compensate the 6-DOF pose of an object from 2D measurements in the image plane, we design a robust optimization scheme by decoupling translation and rotation and introducing multiple initial values for nonlinear optimization based on uncertainty. We formulate an invariant EKF-based object-visual SLAM. The results show the validity of the proposed method in correcting the raw measurements and improving the localization and mapping results. Our future work includes the extension of the proposed pose compensation concept to occluded objects.

REFERENCES

[1] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, V. Lepetit, "Gradient response maps for real-time detection of texture-less objects," *IEEE Transactions on Pattern Analasis and Machine Intelligence*, vol. 34, no. 5, pp. 876–888, 2012.

[2] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scene," *in Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 548–562, 2012.

[3] R. Rios-Cabrera and T. Tuytelaars, "Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable Approach," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2048–2055, 2013.

[4] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, N. Navab, "SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1521–1529, 2017.

[5] D. G. Lowe, "Object recognition from local scaleinvariant features," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1150–1157, 1999.

[6] B. Tekin, S. N. Sinha, P. Fua, "Real-time seamless single shot 6D object pose prediction," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 292–301, 2018.

[7] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, K. Daniilidis, "6-DOF object pose from semantic keypoints," in *in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2011–2018, 2017.

[8] Y. Labbé, J. Carpentier, M. Aubry, J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," *in Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 574–591, 2020.

[9] Y. Su, M. Saleh, T. Fetzer, J. Rambach, N. Navab, B. Busam, D. Stricker, F. Tombari, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6748, 2022.

[10] S. Yang, S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transaction on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.

[11] J. H. Jung, C. G. Park, "Gaussian Mixture Midway-Merge for Object SLAM With Pose Ambiguity," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 400-407, 2023.

[12] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analasis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[13] Y. Xiang, T. Schmidt, V. Narayanan, D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv:1711.00199*, 2018.

[14] M. Rad, V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3828–3836, 2017.

[15] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask r-cnn," *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

[16] G. Wang, F. Manhardt, F. Tombari, X. Ji, "GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16611–16621, 2021.

[17] M. Sundermeyer, T. Hodan, Y. Labbe, G. Wang, E. Brachmann, B. Drost, C. Rother, J. Matas, "BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects," *arXiv:2302.13075*, 2023.

[18] Y. Hu, J. Hugonot, P. Fua and M. Salzmann, "Segmentation-driven 6D Object Pose Estimation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3385–3394, 2019.

[19] V. Lepetit, F. Moreno-Noguwe and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *International Journal of Computer Vision*, vol. 81, pp. 155–166, 2009.

[20] S. Lin, J. Wang, M. Xu, H. Zhao, Z. Chen, "Contour-SLAM: A Robust Object-Level SLAM Based on Contour Alignment," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-12, 2023.

[21] L. Nicholson, M. Milford, N. Sünderhauf, "QuadricSLAM: Dual Quadrics From Object Detections as Landmarks in Object-Oriented SLAM," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1-8, 2019.

[22] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, A. J. Davison, "SLAM++: Simultaneous Localisation and Mapping at the Level of Objects," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1352–1359, 2013.

[23] A. Barrau, S. Bonnabel, "The invariant extended Kalman filter as a stable observer," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1797-1812, 2016.