

LWIR-Based 6D Pose Estimation for Robotic Valve Manipulation in Visually Degraded Environments

Woojin Son ^a, Yeongkyu Kim ^a, Seunghyeon Lee ^a, Taejoo Kim ^a, Dongseok Ryu ^b, Yukyung Choi ^{c,d,*}

^aSejong University, Seoul 05006, Republic of Korea

^bKorea Atomic Energy Research Institute (KAERI), Daejeon, 34057, Korea

^cDept. of Convergence Engineering for Intelligent Drone, Sejong Univ, Seoul 05006, Republic of Korea

^dArtificial Intelligence and Robotics Institute (AIRI), Seoul 05006, Republic of Korea

*Corresponding author: ykchoi@rcv.sejong.ac.kr

***Keywords** : object 6d pose estimation, robotic grasping, LWIR thermal imaging, nuclear accidents

1. Introduction

In severe accidents such as those at nuclear power plants, where human access is restricted, remote robots can perform the initial response [1]. In this process, a robot must not only observe the site but also directly operate equipment, for example by closing the valve of a leaking pipe. Grasping and manipulating a valve presupposes visual perception of the target, and specifically requires its 6D pose, comprising the 3D translation and rotation of the valve. However, such perception is possible only when the target is sufficiently observable in the image. Standard 6D pose estimation techniques take RGB or RGB-D images as input, both of which fundamentally depend on observations in the visible or near-infrared (NIR) bands. At a severe-accident site, airborne particles such as smoke, water vapor, and dust scatter and absorb visible light. Furthermore, if a loss of power leaves no lighting available, RGB images lose key texture and edge information necessary for pose estimation. Active depth sensors likewise rely on projected light reflection, making their measurements equally susceptible to distortion by the same medium. Consequently, under such visually degraded conditions, both modalities fail simultaneously, proving that solutions confined to optical sensing are insufficient.

To overcome the limitations of conventional optical sensing (visible and NIR bands), this study shifts the observation spectral range to long-wave infrared (LWIR). LWIR imaging senses thermal radiation emitted directly by objects, making it completely independent of external illumination. Furthermore, its significantly longer wavelength experiences substantially reduced scattering from airborne particles. As a result, structural geometry is preserved in LWIR images even under conditions where conventional optical sensors fail. Fig.1 demonstrates the same scene captured in both modalities under clear and visually degraded conditions; while the valve's outline is severely degraded in the RGB image under smoke and aerosol coverage, its distinct shape remains clearly visible in the LWIR spectrum.

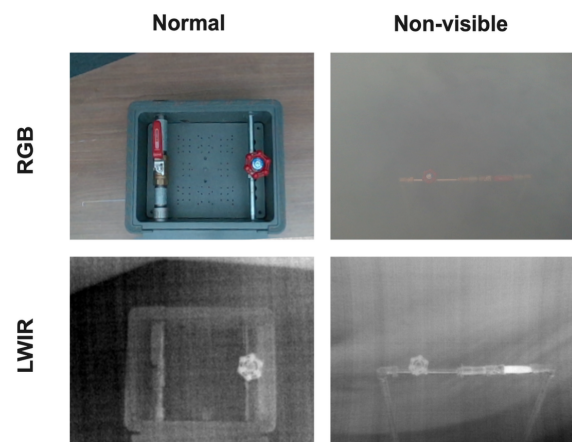


Fig. 1. Comparison of RGB and LWIR images under normal and low-visibility conditions.

Using this preserved shape information, this paper estimates the 6D pose from LWIR images and evaluates performance under both normal and non-visible conditions to demonstrate its effectiveness. Furthermore, using the estimated poses, the robot performs grasping and manipulation tasks on a ball valve and a globe valve to measure task success rates. The analysis reveals that LWIR-based pose estimation suffers only marginal degradation even in non-visible environments where visible-light imaging fails. Crucially, this perception robustness directly translates into high success rates for physical valve manipulation. When grasping and manipulation performance were evaluated separately, performance under the non-visible condition collapsed in the visible-light case, whereas the LWIR pipeline consistently maintained its performance across both stages.

2. Methodology

The system used for the evaluation is shown in Fig. 2 and consists of three stages: perception, planning, and execution. Two target objects are used: a ball valve with a lever handle and a globe valve with a handwheel. The RGB and LWIR images each pass through an identically structured perception stage, after which the planning and

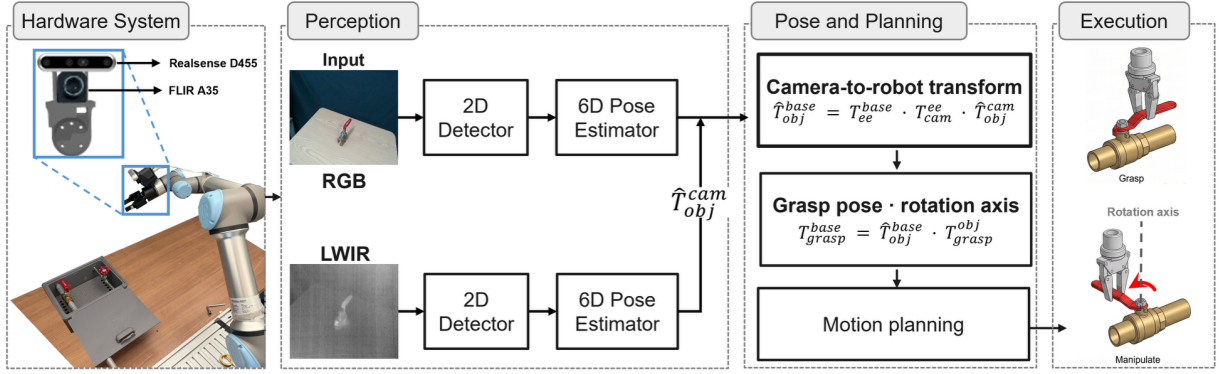


Fig. 2. Overview of the system, comprising three stages: perception, pose transformation and motion planning, and execution.

execution stages are applied in the same way to both conditions.

2.1. Perception

The perception stage estimates the 6D pose of the target object from a single image. A 2D detector [2] first identifies the bounding box of the target object in the input image, and the cropped region-of-interest is provided to the 6D pose estimator [3]. The estimator predicts the rigid-body transformation of the object frame relative to the camera frame, $\hat{T}_{obj}^{cam} = [\hat{R}|\hat{t}]$, which serves as the input to the subsequent planning stage. For both RGB and LWIR modalities, the pipelines employ network modules of identical architecture, trained via the same procedure, and differ only in the spectral band of the input image. Both setups operate on a single image captured by a monocular camera.

2.2. Pose Transformation and Motion Planning

Since the pose estimated by the perception stage is expressed in the camera frame, it must be transformed to the robot base frame before execution. Because the camera is mounted on the robot's end-effector, the transformation matrix obtained by hand-eye calibration and the forward kinematics at the time of image capture are used together to transform the pose as follows:

$$\hat{T}_{obj}^{base} = T_{ee}^{base} \cdot T_{cam}^{ee} \cdot \hat{T}_{obj}^{cam} \quad (1)$$

Here, T_{ee}^{base} is the end-effector pose computed by forward kinematics, T_{cam}^{ee} is the transformation between the camera and the end-effector obtained by hand-eye calibration, and $\hat{T}_{obj}^{base}$ is the object pose expressed in the robot base frame.

The grasp pose and rotation axis are predefined in the object frame so that the same motion is possible regardless of how the object is placed. Denoting the grasp pose defined in the object frame as T_{grasp}^{obj} , the grasp pose the robot must reach is given as follows.

$$T_{grasp}^{base} = \hat{T}_{obj}^{base} \cdot T_{grasp}^{obj} \quad (2)$$

Here, T_{grasp}^{base} is the target pose that the gripper reference frame must reach. The rotation axis is represented by a point on the axis and a direction vector; the transformation in Eq. (1) is applied to the point as is, while only the rotational component is applied to the direction vector, transferring it to the robot base frame. Using the resulting grasp pose as the target, a collision-aware motion planner generates the joint-space path [4].

2.3. Execution

Execution is performed in two stages: grasping and manipulation. In the grasping stage, the gripper is moved to the planned grasp pose to seize the valve handle; in the manipulation stage, the seized handle is rotated about the rotation axis by the target angle. The manipulation scheme is applied in two ways according to the positional relationship between the rotation axis and the grasp point.

For the ball valve, the lever handle is grasped, so the rotation axis lies outside the grasp point. In this case the grasp point must move along a circular arc centered on the rotation axis; the arc is therefore discretized into several waypoints, and a joint solution is computed at each waypoint to generate the path. For the globe valve, the handwheel is grasped, so the rotation axis coincides with the tool axis of the gripper. In this case only the end joint is rotated by the target angle, manipulating the handle in place.

3. Experiments

This chapter analyzes the effectiveness and performance of LWIR-based 6D pose estimation in non-visible environments. By comparing against RGB-based results under identical conditions, we examine how shifting the observation band affects perception performance. We then examine whether the perception performance thus secured carries over to the robot's actual valve-manipulation task, analyzed separately for the grasping and manipulation stages.

3.1. Experimental Setup

Modality	Metric	Normal	Non-visible	All
RGB	ADD	84.46	32.17	62.20
	ADD-S	100.00	59.44	82.74
LWIR	ADD	74.61	67.13	71.43
	ADD-S	95.34	95.80	95.54

Table I. 6D pose estimation accuracy under different environmental conditions (%)

The experiments were conducted on a system composed of a 6-DOF robot arm and a parallel gripper, with the RGB and LWIR cameras mounted on the robot's end-effector. Model training was performed on an NVIDIA A100 GPU, and inference was performed on an NVIDIA Jetson AGX Orin mounted on the robot. The perception stage, including detection and pose estimation, ran at about 10 fps.

Two environmental conditions were distinguished: normal and non-visible. The non-visible environment was created with a fog machine and steam. This condition was applied identically to both the 6D pose estimation evaluation and the valve-manipulation experiments. The manipulation experiments were performed 20 times per object and condition. The valve was kept fixed while the robot's viewpoint was varied at every trial.

6D pose estimation performance was evaluated with ADD[5] and ADD-S. ADD is the mean distance between corresponding points when the 3D model points are transformed by the estimated pose and by the ground-truth pose. ADD-S uses the distance to the nearest point instead of the correspondence, so it can measure the error even when the shape is symmetric and the pose is not uniquely determined. For both metrics, a trial was judged an accurate estimate when the value was within 10% of the object diameter, and this ratio was used as the accuracy. Manipulation performance was evaluated by the grasp success rate and the manipulation success rate; since manipulation is performed only for trials in which grasping succeeded, the manipulation success rate does not exceed the grasp success rate.

3.2. 6D Pose Estimation Results

Table I shows the 6D pose estimation accuracy measured under the two environmental conditions. Under the normal condition RGB showed higher accuracy, but under the non-visible condition RGB dropped by 52.29% in ADD and 40.56% in ADD-S, whereas LWIR maintained nearly the same values as the normal condition for both metrics. As a result, LWIR leads in the overall accuracy combining the two conditions. RGB achieves higher accuracy under the normal condition because the visible-light image also carries texture and edge information that is advantageous for pose estimation. However, this advantage holds only when the observation condition is secured. The gap

Object	Modality	Normal		Non-visible	
		Grasp	Manip.	Grasp	Manip.
Ball valve	RGB	95	90	45	15
	LWIR	75	50	85	60
Globe valve	RGB	100	65	20	5
	LWIR	80	60	75	65
Average	RGB	98	78	32	10
	LWIR	78	55	80	62

Table II. Valve grasping and turning success rates under different environmental conditions (%)

between the two metrics illustrates this difference. For LWIR, ADD-S remains largely consistent across the two conditions, and ADD drops by 7.48%. In contrast, RGB exhibits a severe performance drop, with its ADD falling sharply from 84.46% to 32.17%, which is insufficient for reliable robot manipulation. When texture and edges are lost to scattering, the cues on which RGB relies disappear with them, whereas LWIR preserves the object's shape through thermal-radiation contrast and thus delivers consistent performance regardless of the observation condition. Owing to this consistency, pose estimation at a level usable by the robot is secured even in the non-visible environment, which is the prerequisite for carrying over to actual manipulation.

3.3. Valve Manipulation Results

This section evaluates how directly perception performance translates into actual task success rates. Table II presents the results of the robot grasping and manipulating the two valve types using the estimated poses. Under the normal condition, RGB achieved higher success rates in both grasping and manipulation owing to its superior perception accuracy. Under the non-visible condition, however, performance for RGB dropped drastically across both tasks, with the reduction in task performance far exceeding that of perception accuracy. This indicates that perception degradation was amplified at the execution stage. By contrast, LWIR exhibited higher robustness, maintaining relatively consistent perception accuracy and task success rates compared to RGB under the same condition.

The change in task success rate corresponds to the nature of the perception degradation observed in Section 3.2. Grasping requires the gripper to reach the handle and is therefore affected by position error, whereas manipulation requires alignment with the rotation axis and is therefore affected by orientation error. For RGB, position and orientation collapse together under the non-visible condition, so both stages break down; for LWIR, position is retained, so grasping held at the normal-condition level, and the increase in orientation error was small enough that manipulation followed.

Meanwhile, cases where grasping succeeded but manipulation failed were observed under every condition.

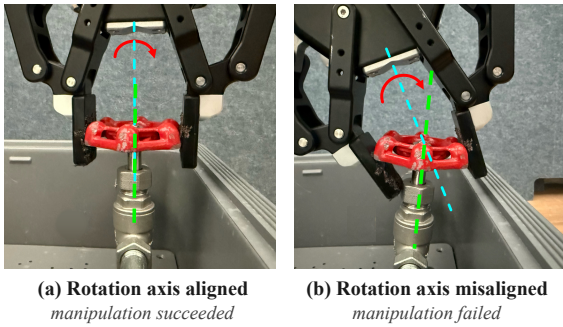


Fig. 3 Comparison of successful (left) and failed (right) rotation-axis alignment. In the latter, the handle was grasped successfully, but the gripper axis was misaligned with the rotation axis and the turning motion could not be performed.

This is because the two motions have different tolerances to pose error. Grasping succeeds provided that the handle is positioned within the gripper opening, thereby offering a tolerance to positioning errors. Manipulation, however, requires the gripper to move about the actual rotation axis, and in the open-loop scheme where the pose is estimated only once, an initial orientation error accumulates over the entire manipulation trajectory. For the ball valve, the grasp point moves along an arc outside the rotation axis, so an error in the axis position is amplified in proportion to the travel distance. For the globe valve, the rotation axis must coincide with the gripper axis, so if the two axes are misaligned when manipulation is attempted, the handle slips or spins idly. Fig. 3 compares a case where the two axes are aligned with a case where they are misaligned; on the right, grasping succeeded but the misaligned axis prevented manipulation. For these reasons, the manipulation success rate falls short of the grasp success rate under every condition. Regarding the proportion of grasp-successful trials completed through manipulation, RGB dropped from 79.6% under normal conditions to 31.3% under non-visible conditions. In contrast, LWIR remained stable, shifting slightly from 70.5% to 77.5%. This minor variation falls within expected statistical variance.

4. Conclusions

At a nuclear power plant severe-accident site, a robot must know the target's 6D pose to manipulate a valve, but once visibility is lost, perception that relies on visible light is disabled. This study verified whether 6D pose estimation with the observation band shifted to LWIR is effective in non-visible environments, and compared, under conditions identical to RGB, whether that perception performance carries over to actual valve manipulation. Under the non-visible condition, RGB's ADD dropped by 52.29% while LWIR dropped by only 7.48%, and this robustness did not stop at perception but carried over to manipulation, maintaining 80% grasping and 62% manipulation. Under the same condition, RGB collapsed to 32% and 10%, with both stages breaking down together.

That the manipulation success rate fell short of the grasp success rate under every condition shows that the bottleneck of valve manipulation lies not in approach but in aligning the rotation axis. Since RGB is superior under the normal condition, LWIR is not always ahead; however, at accident sites where visibility cannot be guaranteed, the fact that its performance is maintained matters more.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144385), by the Institute of Information & Communications Technology Planning & Evaluation (IITP) - ICAN (ICT Challenge and Advanced Network of HRD) grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00436528), and by the Technology Innovation Program (RS-2024-00442513, Development of robotic manipulation task learning based on Foundation model to understand and reason about task situations) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

REFERENCES

- [1] K. Nagatani, S. Kiribayashi, Y. Okada, K. Otake, K. Yoshida, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi, M. Fukushima, and S. Kawatsuma, Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots, *Journal of Field Robotics*, Vol. 30, No. 1, pp. 44-63, 2013.
- [2] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, YOLOX: Exceeding YOLO Series in 2021, *arXiv preprint arXiv:2107.08430*, 2021.
- [3] X. Liu, R. Zhang, C. Zhang, G. Wang, J. Tang, Z. Li, and X. Ji, GDRNPP: A Geometry-Guided and Fully Learning-Based Object Pose Estimator, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 47, No. 7, pp. 5742-5759, 2025.
- [4] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. Van Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, N. Ratliff, and D. Fox, cuRobo: Parallelized Collision-Free Robot Motion Generation, *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.
- [5] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes, *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 548-562, 2012.