Stereo Vision-Based Depth Estimation for Ocean Environmental Monitoring and Wave-Height Analysis

Jun-Seok Jin[1], Gi-Sang Yang[2] and Chang-won Kim[1*]

[1] Control and Instrumentation Engineering, Changwon National University, Changwon 51140, Republic of Korea
[2]Department of AI Convergence Engineering, Changwon National University, Changwon 51140, Republic of Korea
**Corresponding Author E-mail: cwkm22@changwon.ac.kr*

ABSTRACT

Ocean waves are an important factor that requires continuous observation for marine environmental monitoring and coastal safety management. Recently, remote sensing technologies such as X-band marine radar have been used to estimate wave-height, but they require specialized equipment and separate signal processing procedures. Therefore, this study proposes a stereo vision-based method for three-dimensional reconstruction of the sea surface and estimation of wave-height using optical cameras. In the proposed method, depth information of the sea surface is estimated from stereo images acquired by two cameras using WAFT-Stereo and converted into a three-dimensional point cloud using the camera intrinsic parameters. An area of interest is then defined from the reconstructed sea surface point cloud, and a reference sea surface is determined through RANSAC-based plane estimation. The coordinate system is subsequently aligned with the reference plane to calculate the relative sea surface elevation at each location, and the elevation values within a specified spatial region are extracted over time to analyze wave-induced surface variations. Based on these measurements, wave height is estimated. For quantitative validation of the estimation results, reference wave-height data are obtained using a self-developed measurement sensor in a field environment, and the proposed image-based estimation results are compared with the reference data to evaluate the performance of the proposed method. This study presents a method for spatially acquiring the three-dimensional shape of the sea surface using stereo cameras without separate large-scale marine observation equipment and for quantitatively analyzing wave height.

Keywords: Stereo Vision/ Ocean Wave/ Wave-Height Estimation/ 3D Sea Surface Reconstruction

1. INTRODUCTION

Ocean waves are important characteristics of the marine environment and are closely related to coastal safety, marine transportation, offshore structures, and environmental monitoring. Continuous observation of wave conditions is therefore essential for understanding changes in the sea surface and for supporting reliable marine monitoring systems. In particular, wave-height such as wave height, wave period, wavelength, and wave direction provide important information for characterizing ocean wave conditions. Various measurement and remote sensing techniques have been developed to observe ocean waves. Conventional measurement systems, such as wave gauges and buoy-based instruments, can provide accurate measurements at specific locations but have limitations in spatial coverage. Recently, remote sensing techniques such as X-band marine radar have been widely used to observe sea surface conditions over relatively large areas [1]. These systems can provide spatially distributed wave information without direct contact with the water surface. However, radar-based observation generally requires specialized equipment, system configuration, and additional signal processing for extracting wave-height, which can increase the cost and complexity of the observation system. Optical cameras provide a low-cost and flexible alternative for remote observation of the sea surface. In particular, stereo vision enables the three-dimensional structure of an observed surface to be reconstructed from images acquired from two viewpoints. By applying stereo depth estimation to synchronized images of the sea surface, the spatial variation of the water surface can be obtained without direct physical contact with the measurement target [2,3,4]. This makes stereo vision a promising approach for spatially continuous ocean wave monitoring using relatively simple observation equipment. However, applying stereo vision

to ocean surfaces remains challenging. Unlike typical terrestrial scenes, the water surface often contains weak and repetitive visual features, and its appearance changes continuously due to wave motion and illumination conditions. These characteristics can lead to ambiguous correspondences between stereo images and unstable depth estimation. Therefore, reliable reconstruction of the three-dimensional sea surface is an important prerequisite for accurate wave-height estimation. In this study, a stereo vision-based method is developed to reconstruct the three-dimensional sea surface and estimate wave-height. Stereo images acquired from two cameras are processed using a stereo depth estimation model, and the resulting depth information is transformed into a three-dimensional point cloud. The reconstructed sea surface is then used to estimate a reference plane based on RANSAC [5], and the coordinate system is aligned with the estimated plane to obtain relative sea surface elevation. By extracting the elevation within a predefined spatial region over time, wave-induced surface variations are obtained and used to estimate wave-height. For quantitative evaluation, a self-developed measurement sensor is used to acquire reference wave-height data in a field environment, and the estimated results are compared with the reference measurements. The objective of this study is to establish a simple and spatially continuous camera-based approach for quantitative ocean wave monitoring.

2. METHODOLOGY

2.1 Overall System

The overall architecture of the proposed stereo vision-based wave-height estimation system is depicted in Figure 1 The pipeline consists of four key processing modules: (1) stereo image alignment and vertical offset correction, (2) dense disparity estimation via WAFT[6]-Stereo and 3D point cloud reconstruction, (3) RANSAC-based reference plane fitting and coordinate alignment, and (4) time-series surface elevation processing, ground truth (GT) sensor preprocessing, and spectral analysis.

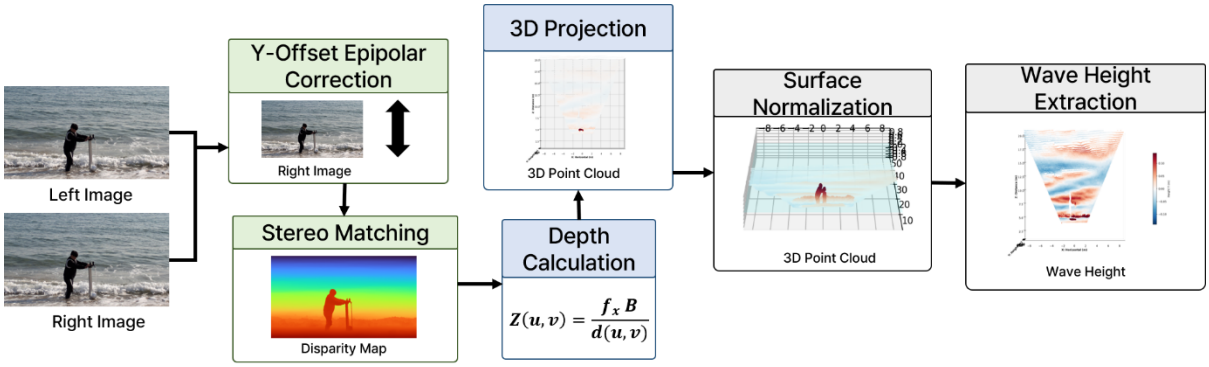


Figure 1. Overall system for stereo vision-based ocean wave-height estimation.

2.2 Stereo Image Preprocessing & Y-Offset Alignment

Vertical misalignments between paired cameras distort epipolar lines and degrade matching accuracy on low-texture water surfaces. Using ORB [7] feature matching between left (I_L) and right (I_R) images, vertical shifts $\Delta y_i = y_i^R - y_i^L$ are extracted. A robust offset Δy_{final} is determined from inliers bounded by the median:

$$y_{offset} = \Delta y_{final} = \frac{1}{|K|}\sum_{k\in K}\Delta y_k \ , where\ |\Delta y_k - median(\Delta y)| < 3.0 pixel \quad (1)$$

An affine transformation matrix M then warps I_R to restore vertical alignment:

$$I_R^{corrected} = WarpAffine(I_R, M), M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\Delta y_{final} \end{bmatrix} \quad (2)$$

2.3 WAFT-Stereo Depth & 3D Point Cloud Reconstruction

To eliminate camera tilt angles and isolate true surface variations, a planar equation is fitted to points within a surface ROI using RANSAC. To stabilize the reference plane across frames, plane parameters are estimated for every frame and grouped into consecutive windows of 50 frames. The median of the plane parameters within each 50-frame window is used to define a representative reference plane:

$$Z(u,v) = \frac{f_x \cdot B}{d(u,v) + \epsilon} \tag{3}$$

$$X = \frac{(u-c_x)Z}{f_x}, Y = \frac{(v-c_y)Z}{f_y}, Z = Z(u,v) \tag{4}$$

where B is the stereo baseline, $d(u,v)$ is the disparity at pixel (u,v), ϵ is a small positive constant to prevent division by zero, and f_x, f_y, c_x, c_y are the camera intrinsic parameters obtained through calibration.

2.4 RANSAC Reference Plane Alignment

A reference plane is fitted to the reconstructed surface points within the ROI using RANSAC. The plane is represented as

$$aX + Y + bZ + d = 0 \tag{5}$$

where aa, bb, and dd are the plane parameters estimated by RANSAC.

$$\tilde{a} = median(a_1, \dots, a_{50}), \tilde{b} = median(b_1, \dots, b_{50}), \tilde{d} = median(d_1, \dots, d_{50}) \tag{6}$$

Plane parameters are estimated for each frame and grouped into consecutive windows of 50 frames. The median values of the plane parameters within each window are used to define a representative reference plane, which is then fixed for the corresponding 50-frame interval.

$$R = I + [w]_\times + [w]_\times^2 \left(\frac{1-n\cdot v}{\|w\|^2}\right), where\ w = n \times v \tag{7}$$

Rotated points $P_{rotated} = R \cdot P$ are offset-corrected to yield the true sea surface elevation $\eta\,(X,Z)$:

$$\eta(X,Z) = -(Y_{rotated} - y_{offset}) \tag{8}$$

2.5 Time-Series Processing & Wave Estimation

Points within a spatial target region are averaged over time to yield the estimated elevation series. Concurrently, raw clearance readings from a custom physical wave gauge are inverted and preprocessed using a rolling median–standard deviation filter to remove splash outliers.

$$\mu_k = median\{\eta_{GT,raw}(j) : j \in W_k\} \tag{9}$$

$$\sigma_k = std\{\eta_{GT,raw}(j) : j \in W_k\} \tag{10}$$

$$\eta_{GT}(k) = \begin{cases} \mu_k, & |\eta_{GT,raw}(k) - \mu_k| > n_k\sigma_k \\ \eta_{GT,raw}(k), & otherwise \end{cases} \tag{11}$$

where μ_k and σ_k are the rolling median and standard deviation calculated within the centered window W_k, respectively. In this study, a window size of $W = 7$samples and a threshold of $n_\sigma = 2.5$ were used. With the sensor sampling rate of 2 Hz, the 7-sample window corresponds to 3.5 s of sampled data.

3. RESULTS AND DISCUSSION

3.1 Dataset, Ground Truth, and Evaluation Setup

The experimental dataset was collected at Haeundae Beach, Busan, on 2026-02-12 using a stereo camera system with a 0.6 m baseline. Images were captured at 2 fps with a resolution of 2560 × 1440 pixels, and the target water-surface region was approximately 5 m from the cameras. For Ground Truth (GT) validation, an ultrasonic sensor was mounted on a fixed pole 1.0 m above the ground and used to measure the distance to a vertically moved styrofoam ball. The measurements were converted to relative water-surface elevations and synchronized with the stereo frames using timestamps. The sensor operated at 2 Hz with an official measurement accuracy of $\pm 0.3mm - \pm 0.5mm$. A total of 192 frames were evaluated after excluding frames containing a person in the measurement area (15:27:30–15:27:31.5, 15:28:16.5–15:28:18.5, and 15:28:21.5–15:28:25.0). Median-based outlier filtering was applied to the GT signal, which ranged from 0.0031 to 0.7992 m. Three pipeline configurations were compared: (1) RAFT-Stereo [9] + Fixed Plane, (2) WAFT-Stereo + Dynamic Plane with frame-wise plane fitting, and (3) the Proposed System, which applies inter-frame reference-plane stabilization to WAFT-Stereo. These configurations were used to evaluate the effects of the stereo disparity method and plane-fitting strategy on wave-height estimation.

3.2 Quantitative Evaluation

Table 1. Quantitative comparison of water-surface elevation estimation accuracy against ground-truth sensor measurements (N = 192 samples)

	Wave Height MAE (m)	Wave Height RMSE (m)	R^2
RAFT-Stereo + Fixed Plane	0.0366	0.0456	0.3574
WAFT-Stereo + Dynamic Plane	0.0460	0.0591	-0.0805
Proposed	0.0352	0.0440	0.4017

Table 1 summarizes the water-surface elevation estimation accuracy across different stereo backbones and plane-fitting strategies. Here, "Wave Height" refers to the estimation error of the real-time water-surface elevation time series ($\eta(t)$) at ($X = 1.5m, Z = 5.0m$) , rather than significant wave height (H_s) or other wave-height statistics. Performance is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2), representing average error, sensitivity to larger deviations, and overall agreement with the ground-truth measurements, respectively. WAFT-Stereo + Dynamic Plane shows the lowest performance ($R^2 = -0.0805$) with MAE = 0.0460 m and RMSE = 0.0591 m, indicating temporal instability caused by frame-wise plane fitting. RAFT-Stereo + Fixed Plane improves the results ($R^2 = 0.3574$), with MAE = 0.0366 m, but remains affected by disparity estimation uncertainty, which propagates into the 3D reconstruction. In contrast, the Proposed Framework applies WAFT-Stereo with a stabilized reference plane across frames, suppressing per-frame tilt noise and achieving the best overall accuracy (MAE = 0.0352 m, RMSE = 0.0440 m, $R^2 = 0.4017$). The estimated water-surface elevation and frame-wise residuals are compared with the ground-truth measurements to visualize temporal tracking performance. Although the R^2 value is 0.4017, pointwise comparison is sensitive to small surface fluctuations and slight temporal phase differences. Nevertheless, the absolute errors remain at a few centimeters. The comparison with RAFT-Stereo + Fixed Plane isolates the effect of the stereo backbone under the same fixed-plane strategy, while the comparison with WAFT-Stereo + Dynamic Plane evaluates the effect of the reference-plane strategy using the same backbone. The Proposed Framework achieves the best overall performance in both comparisons, demonstrating the effectiveness of the WAFT-Stereo backbone and stabilized reference-plane strategy.

Figure 2. compares the estimated and ground-truth water-surface elevation time series obtained using the Proposed system. The estimated signal generally follows the temporal variation of the ground-truth measurements, although differences in amplitude and slight temporal phase shifts are observed in some intervals. These local differences contribute to the R2R^2 value of 0.4017, while the absolute errors remain at a few centimeters, with an MAE of 0.0352 m and an RMSE of 0.0440 m.

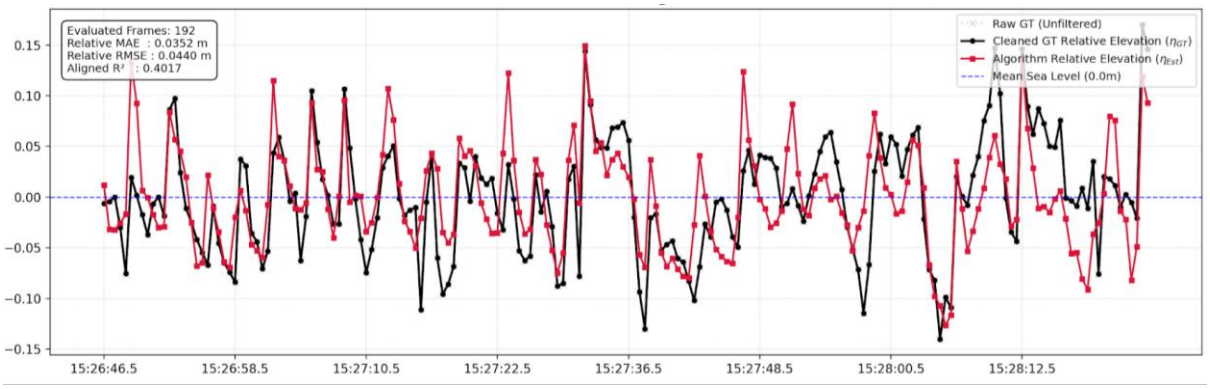


Figure 2. Comparison of the estimated and ground-truth water-surface elevation time series using the Proposed System.

3.3 Qualitative Evaluation

Figure 3 compares the disparity maps obtained using RAFT-Stereo and WAFT-Stereo. The RAFT-Stereo result in Figure 3(b) shows local disparity variations even on sea-surface regions at similar distances, with noticeable noise caused by light reflection. In contrast, Figure 3(c) shows that WAFT-Stereo provides more uniform disparity values and better surface continuity while reducing reflection-induced noise. This indicates that the monocular depth features used in WAFT-Stereo contribute to more stable disparity estimation for marine scenes[6]. Consequently, this structural stability in disparity estimation directly serves as a reliable foundation for subsequent 3D reference plane fitting and surface elevation mapping.

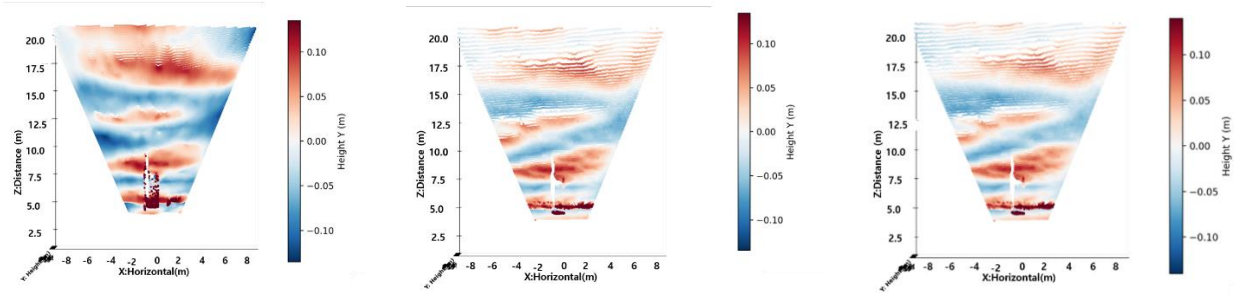


(a)Left RGB Input Image (b)RAFT-Stereo Disparity (c)WAFT-Stereo Disparity

Figure 3. Comparison of Input Image and Stereo Disparity Estimation Results.

To evaluate spatial reconstruction performance, Figure 4 presents top-view 3D visualizations of estimated wave heights, where surface elevations are represented via pseudo-color mapping. As shown in Figure 4(a), the RAFT-Stereo + Fixed Plane baseline exhibits noisy 3D surface reconstructions with irregular spatial distributions. These artifacts originate from noise and uncertainty in the estimated disparity itself, which propagate through the 3D reconstruction process and result in noisy and unstable spatial structures. Although a fixed reference plane is applied, the disparity-level uncertainty remains and directly affects the quality of the reconstructed 3D surface. In contrast, WAFT-Stereo + Dynamic Plane (Figure 4(b)) produces a visually cleaner 3D surface reconstruction due to its more reliable disparity estimation; however, its per-frame dynamic plane fitting slightly underestimates the reference plane height, resulting in systematically overestimated wave elevations. The Proposed System (Figure 4(c)) successfully achieves both clean 3D structural reconstruction and robust reference plane

alignment. By stabilizing the reference plane across consecutive frames, our approach accurately represents the absolute wave height distribution while maintaining well-defined elevation contours, thereby providing a more consistent representation of the actual sea-surface dynamics.



(a) RAFT-Stereo + Fixed Plane (b) WAFT-Stereo + Dynamic Plane (c) Proposed

Figure 4. Three-dimensional visualization of the estimated wave height map.

4. CONCLUSIONS

In this paper, we proposed a stereo vision-based system for water-surface elevation estimation. By integrating Y-offset alignment, WAFT-Stereo disparity estimation, inter-frame reference-plane stabilization, and temporal outlier filtering, the proposed framework achieved an MAE of 0.0352 m, an RMSE of 0.0440 m, and an R2R^2 of 0.4017 in the field experiment at Haeundae Beach. The results indicate that reference-plane stabilization improves the temporal consistency of water-surface elevation estimation. A limitation of this study is that the stereo backbone and plane strategy are not fully decoupled in the current comparison design, and evaluation was conducted using a limited field dataset and target region. Future work will evaluate the method under a wider range of wave and environmental conditions.

Acknowledgement

This research was financially supported by the Korea Ministry of Climate, Energy and Environment as Waste to Energy-Recycling Human Resource Development Project; the ANCHOR program through the ANCHOR Center, Gyeongsangnam-do, funded by the Ministry of Education(MOE) and the Gyeongsangnam-do Provincial Government, Republic of Korea.(2026-ANCHOR-16-002); and by the Technology Innovation Program (RS-2026-25546292, Development of experimental and measurement technologies to estimate ice loads and derive design parameters for vessels navigating Arctic shipping routes) funded by the Ministry of Trade, Industry and Resources(MOTIR, Korea)

References

[1] Huang, W., Liu, X., & Gill, E. W. (2017). Ocean wind and wave measurements using X-band marine radar: A comprehensive review. *Remote Sensing,* 9(12), 1261.

[2] Jebari, O., Hong, S. H., Hunsucker, T., Kim, D. K., & Jin, C. (2025). Wave-field reconstruction using stereo cameras on a floating platform in a synthetic environment. *Ocean Engineering,* 327, 120958.

[3] Li, D., Xiao, L., Wei, H., Li, J., & Liu, M. (2022). Spatial-temporal measurement of waves in laboratory based on binocular stereo vision and image processing. *Coastal Engineering*, 177, 104200.

[4] Wang, F., & Zhu, Q. (2025). Artificial texture-free measurement: A graph cuts-based stereo vision for 3D wave reconstruction in laboratory. Journal of Marine Science and Engineering, 13(9), 1699.

[5] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

[6] Wang, Y., & Deng, J. (2026). WAFT-Stereo: Warping-alone field transforms for stereo matching. *European Conference on Computer Vision (ECCV), 2026.*

[7] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. 2011 *International Conference on Computer Vision (ICCV)*, 2564–2571.

[8] Lipson, L., Teed, Z., & Deng, J. (2021). RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. 2021 *International Conference on 3D Vision (3DV)*, 218–227.