Stereo Vision-Based Skeleton Pose Estimation with Proximity Index for Environment

Sukwoo Jung
Contents Convergence Research Center
Korea Electronics Technology Institute
Gyeonggi-do, South Korea
swjung@keti.re.kr

Youn-Sung Lee
Contents Convergence Research Center
Korea Electronics Technology Institute
Gyeonggi-do, South Korea
yslee@keti.re.kr

Jung Wook Wee
Contents Convergence Research Center
Korea Electronics Technology Institute
Gyeonggi-do, South Korea
jwwee@keti.re.kr

***Abstract*— Ensuring worker safety requires accurate monitoring of human movement and detection of potential hazards in workplace environments. This study presents a stereo vision-based skeleton pose estimation framework that incorporates a Proximity Index (PI) to evaluate the distance between human body joints and predefined risk zones. Using the ZED 2i stereo camera, both RGB images and depth maps were acquired to enable reliable three-dimensional pose estimation. The proposed method extracts skeleton data from static images, computes a joint-to-zone distance metric, and classifies safety states based on threshold PI values. Unlike conventional 2D pose estimation approaches, the integration of stereo depth information improves accuracy in proximity evaluation and risk detection. To validate the framework, experiments were conducted in a office environment, simulating scenarios where workers approached restricted or hazardous areas. The results confirm that the method effectively detects abnormal proximity behaviors and supports real-time safety monitoring in workplace settings.**

Keywords— Stereo Vision, Human Pose Estimation, Safety Monitoring, Depth Map, Object Detection

I. INTRODUCTION

Ensuring worker safety in industrial environments requires accurate detection of human pose and proximity to hazardous zones. Recent advances in computer vision have enabled robust extraction of human pose features using RGB and depth data. However, extending these methods to accurately assess three-dimensional proximity remains a challenge.

Pose estimation research has progressed significantly with deep learning. Mask R-CNN, originally designed for instance segmentation and object detection, has been extended to support human keypoint detection, serving as a strong foundation for pose-based safety analysis[1]. For example, Jung *et al.* introduced a method that integrates Mask R-CNN instance segmentation with IMU data to detect moving objects using a single moving camera[2]. Similarly, monocular camera and IMU fusion has been used for real-time sensor pose tracking, mitigating drift and improving stability[3]. Depth-enhanced pose estimation using time-of-flight (ToF) or stereo vision has also gained traction. Methods that fuse ToF with stereo matching achieve more accurate high-resolution depth maps through efficient data integration[4,5].

Despite these advancements, existing frameworks seldom focus on static image-based skeleton pose estimation within industrial safety contexts, particularly leveraging stereo depth to quantify proximity to danger zones.

This study proposes a novel framework, termed Stereo Vision-Based Skeleton Pose Estimation with Proximity Index, which integrates stereo camera-derived depth maps with 2D pose estimation to calculate a Proximity Index (PI). The PI quantitatively evaluates the three-dimensional distance between human body joints and predefined risk areas.

II. PROPOSED METHOD

This section describes the proposed approach for skeleton-based pose estimation and proximity index calculation using stereo vision. The method is designed to detect potentially hazardous situations in indoor environments by combining depth information and human pose keypoints extracted from stereo image pairs.

A. System Overview

The proposed system employs a ZED 2i stereo camera to capture synchronized left and right RGB frames along with a disparity map. The stereo disparity is converted into a depth map, which provides per-pixel distance information in metric units. A 2D human pose estimation model is applied to the left RGB image to extract body keypoints K, where each keypoint ki is associated with a pixel coordinate (u,v).

Using camera intrinsic parameters (fx, fy, cx, cy) and the depth value for each keypoint, the 3D position P_i in the camera coordinate frame is computed via:

$$X_i = \frac{(u_i - c_x)\cdot d_i}{f_x}, Y_i = \frac{(v_i - c_y)\cdot d_i}{f_y}, Z_i = d_i \quad (1)$$

B. Proximity Index Definition

The **Proximity Index** (PI) is introduced to quantify the closeness of a detected person to predefined hazardous zones or objects. Given the 3D coordinates P_i of all keypoints and a set of reference points r_j representing hazard locations, the minimum Euclidean distance is computed:

$$d_{min} = \min_{i,j} \left\| P_i - r_j \right\|_2 \quad (2)$$

The Proximity Index is then normalized into a range [0,1] :

$$PI = \max(0, 1 - \frac{d_{min}}{d_{th}}) \quad (3)$$

where d_{th} is a threshold distance defining the boundary of the safety zone. PI value close to 1 indicates a high-risk proximity, while a value near 0 indicates a safe distance.

C. Processing Pipeline

The processing flow of the proposed method is as follows:

1. **Stereo Image Acquisition**: Capture synchronized stereo RGB frames and disparity map using ZED.

2. **Depth Map Computation**: Convert disparity map to metric depth map.
3. **2D Skeleton Extraction**: Apply pose estimation network to detect keypoints in the left RGB image.
4. **3D Keypoint Reconstruction**: Use camera intrinsics and depth values to obtain 3D coordinates for each keypoint.
5. **Proximity Index Calculation**: Compute the minimum distance to hazard zones and normalize to obtain PI.
6. **Alert Generation**: Trigger warnings if PI exceeds a predefined safety threshold.

III. Experiments

A. Experimental Setup

The proposed method was implemented in C++ on a Windows 11 workstation equipped with an NVIDIA RTX 4080 Ti GPU. A ZED 2i stereo camera was used to acquire synchronized left and right RGB images and disparity maps. The 2D pose extraction module uses a pretrained pose estimator executed on the left RGB image; mapping to 3D is performed using per-keypoint depth values from the ZED depth map and camera intrinsics. The Proximity Index (PI) computation and safety logic are implemented in C++ for real-time evaluation.

Data were collected in an office environment that emulates typical indoor industrial settings (tabletops, chairs, stationary equipment, and defined restricted areas). Ten independent acquisition runs were performed; each run consisted of multiple static frames captured while subjects simulated typical and risky behaviors (approaching restricted zones). Ground truth for "risky proximity" was labeled manually by an expert based on known hazard zone boundaries.

B. Evaluation Metrics

For comparison, a representative baseline method was selected that reflects common practice in pose-based proximity detection:

Baseline: 2D Pose Estimation with 2D Proximity Metric — a framework that uses the same 2D pose estimator but computes proximity using only image-plane distances between 2D keypoints and projected 2D hazard ROI. This baseline omits any depth information and uses heuristic thresholds on pixel distance to classify risky proximity.

Precision is used as the primary metric to measure the correctness of hazardous-proximity detections:

$$Precision = \frac{TP}{TP + FP}$$

For each acquisition run the precision was computed; aggregate statistics (mean and standard deviation) were calculated across the 10 runs.

C. Results

The following fig.1 shows a prediction of the precision values across 10 acquisition runs for both the Proposed method (Stereo + PI) and the Baseline (2D Pose only). The baseline method shows mean precision 0.802, StdDev 0.036, otherwise the proposed method shows mean precision 0.896, StdDev 0.022.

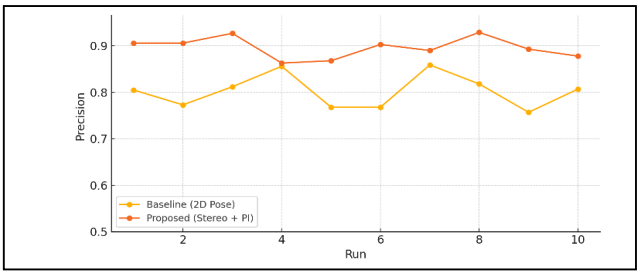

Fig.1 Precision Result of the Experiments

IV. Conclusion

This paper proposed a stereo camera–based hazardous proximity detection method that integrates 2D human pose estimation with depth-enhanced 3D keypoint reconstruction and a Proximity Index (PI) metric. The method leverages per-keypoint depth information from a ZED stereo camera to overcome the scale and distance ambiguities inherent in purely image-plane approaches.

Experiments demonstrated that the proposed approach consistently outperformed a conventional 2D proximity-based baseline, achieving higher precision and lower variance across multiple acquisition runs.

Future work will extend the method to handle dynamic environments with moving cameras, integrate multi-camera fusion for occlusion handling, and evaluate the approach in industrial scenarios with more diverse hazard zone configurations.

Acknowledgment

This research was supported in part by the Technology Innovation Program (No.20023743, Development of Digital Twin-based AMR Operation Service Technology for Industrial Use Cases, 50%) and in part by the Korea Evaluation Institute of Industrial Technology (KEIT) (RS-2024-00453509, Development of Lightweight AR/MR Devices Based on High Optical Characteristics Optical Engine and Establishment of Evaluation System, 50%) both funded by the Ministry of Trade, Industry & Energy (MOTIE), South Korea.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.* (ICCV), pp. 2961–2969, Oct. 2017.

[2] S. Jung, Y. Cho, KT Lee, and M. Chang, "Moving object detection with single moving camera and IMU sensor using mask R-CNN instance image segmentation," *J. Precision Engineering and Manufacturing*, vol. 22, no. 6, pp. 1049–1059, June. 2021, doi: 10.1007/s12541-021-00527-9.

[3] S. Jung, S. Park, and KT Lee, "Pose tracking of moving sensor using monocular camera and IMU sensor," *KSII Trans. On Internet and Information Sysems*, vol. 15, no. 8, pp. 3011–3024, Dec. 2021.

[4] S.Jung, H. Yun, KT Lee, "Accurate pose estimation method using ToF-stereo camera data fusion," *Int. Conf. ICT*, pp. 644–645, Oct. 2023, doi: 10.1109/ICTC58733.2023.10392469.

[5] X. Zhang, Y. Yang, and Z. Wei, "Stereo and ToF data fusion for high-resolution depth mapping," *IET Comput. Vis.*, vol. 13, no. 3, pp. 245–253, May 2019, doi: 10.1049/iet-cvi.2018.5476.