

Original Research Article

A lightweight fall-detection device integrating ToF laser ranging and visual sensing

Liangli Wang, Yulin Song, Jinfei Nong, Kai Guo, Xuan Pan*Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China*

Abstract: To address the instability of vision-only fall detection under low-light conditions and occlusion, as well as the constraints of edge computing, we designed a lightweight device that integrates multi-zone time-of-flight (ToF) laser ranging with low-resolution visual sensing and established a reproducible verification workflow based on digital human simulation. The ToF branch extracts distance, vertical position, descent velocity, effective ranging ratio, and near-ground region ratio, while the visual branch extracts the human body bounding box, centroid, tilt angle, motion amplitude, detection confidence, and brightness. Features from the two branches are gated and fused according to real-time quality scores, after which a depthwise separable temporal convolution identifies the continuous upright-descent-near-ground progression. The simulator generated 9,600 sequences across 32 virtual body profiles, including 3,840 fall sequences and 5,760 daily-activity sequences, under low light, backlighting, occlusion, invalid ToF echoes, and structured unimodal distortions. In composite stress tests using splits that were disjoint by virtual body profile, the proposed method achieved accuracy, precision, recall, and F1 scores of 96.89%, 96.89%, 95.28%, and 96.08%, respectively, with 1.22 false alarms per hour and an average alarm latency of 88.7 ms. The model contains 14,532 parameters and 1.42 M MACs; under the equivalent-cost model for a quad-core ARM processor, the INT8 end-to-end latency was 12.8 ms. The results indicate that quality gating can mitigate the performance degradation caused by unimodal disturbances, although these findings still require validation using physical prototypes and real-world scenarios.

Keywords: ToF laser ranging; visual sensing; fall detection; multimodal fusion; edge intelligence; simulation verification

1. Introduction

Home-based elderly care, wards, and rehabilitation institutions require continuous non-contact fall monitoring. Wearable inertial devices can directly measure motion, but long-term adherence to wearing and charging them is often poor; cameras can capture complete movement sequences, but are susceptible to illumination, occlusion, background changes, and privacy concerns. Artificial intelligence and the Internet of Things are driving the development of fall monitoring toward continuous, localized monitoring at the edge^[1]. Existing vision-based methods use human bounding boxes, contours, skeletons, and spatiotemporal networks, but cross-scene generalization, false-alarm control, and limited computing resources on end devices remain barriers to deployment^[2-3]. Multimodal fusion can exploit complementary sensor information. Federated multimodal learning has demonstrated that joint modeling of visual and inertial information can improve recognition performance^[4]; studies on skeleton kinematics and audio-visual fusion have shown that fall detection should focus on the descent process and rely on one modality to compensate when the other is degraded^[5-6]. However, simple feature concatenation or fixed weighting cannot determine which sensor is more reliable in the current frame. Low light and furniture occlusion weaken visual features, while reflective materials, the edges of the field of view, and echo loss can cause ToF holes. If low-quality modalities are still equally weighted in decision-making, errors may be amplified. Research on mobile platforms and edge systems shows that fall detection must consider frame rate, model size, power consumption, and alarm-delivery mechanisms simultaneously^[7-8]; continuous video analysis is

more suitable than single-frame classification for distinguishing between sitting down quickly, lying down normally, and falling^[9]. Low-resolution infrared arrays can also support low-cost detection using features such as distance, speed, area, and variance. In addition, time-synchronization errors and inaccurate extrinsic calibration between heterogeneous sensors directly affect fusion performance. To evaluate the algorithm architecture and parameter sensitivity before prototype fabrication, this study develops a parameterized digital human and indoor-scene simulator to generate a synchronized ToF matrix, low-resolution grayscale features, and quality scores.

2. Lightweight fall-detection device and method based on ToF–vision fusion

2.1. Overall device architecture and multisource sensing design

The proposed device consists of a multi-zone ToF sensor, a low-resolution grayscale camera, an edge computing unit, an audible and visual alarm, and a wireless communication module. Raw images are processed locally, and only anonymized person regions or geometric features are retained. **Figure 1** shows the simulation verification process: A synchronized ToF matrix and grayscale observation are generated from interactions between the digital human model and the indoor scene. After calibration, quality estimation, gating fusion, and temporal judgment, detection results are produced, and edge resource requirements are estimated using an operator cost model. This process can independently vary illumination, occlusion, echo gaps, and unimodal distortion.

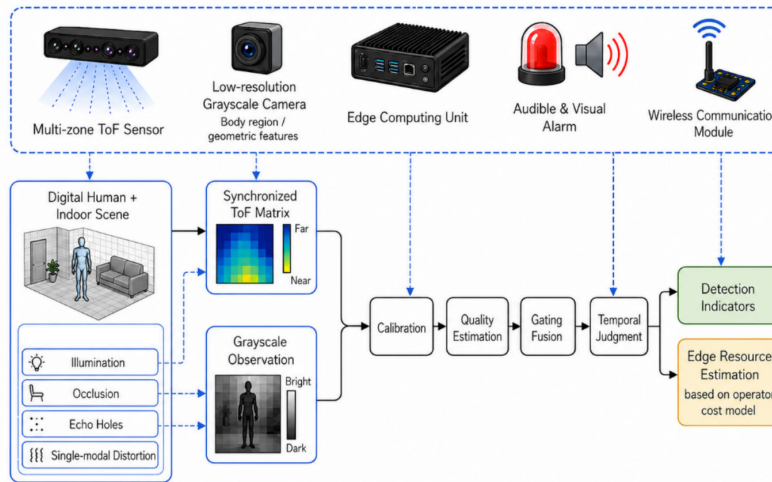


Figure 1. Simulation-based validation workflow for the proposed ToF–vision fall detection device.

Using the edge clock as the reference, the visual frame with the smallest time difference is matched to each time-of-flight (ToF) timestamp, with the synchronization error defined as:

$$\Delta t_i = |t_i^D - t_{j(i)}^V|. \quad (1)$$

In the formula, t_i^D is the ToF timestamp of the i -th frame, $t_{j(i)}^V$ is the timestamp of the matched visual frame, $j(i)$ is the matching index, and Δt_i is the synchronization error. If Δt_i exceeds the threshold τ , the system switches to inference using the more reliable modality.

The mapping from ToF spatial points to the image plane is as follows:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}|\mathbf{T}] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2)$$

In the formula, (X, Y, Z) are the ToF coordinates, (u, v) are the pixel coordinates, s is the homogeneous scaling factor, $\mathbf{K}$ is the camera intrinsic matrix, and $\mathbf{R}$ and $\mathbf{T}$ are the rotation matrix and translation vector, respectively. Calibration requires simultaneous checks of time synchronization, field-of-view overlap, and installation stability.

2.2. Dual-branch feature extraction and confidence-gated fusion

The raw ToF matrix is processed by removing out-of-range points, applying median filtering, and

subtracting the background:

$$\widetilde{D}_t(u,v)=M_t(u,v)[D_t(u,v)-B(u,v)]. \quad (3)$$

In the formula, $D_t(u,v)$ and $B(u,v)$ are the distances measured at time t and in the empty scene, respectively, at grid location (u,v) , $M_t(u,v) \in \{0,1\}$ is the validity mask, and $\widetilde{D}_t(u,v)$ is the correction residual. The ToF feature of the human body region is represented as follows:

$$\mathbf{f}_t^D=[d_{\min}, d_{\text{mean}}, \sigma_d, v_d, r_{\text{valid}}, h_d, r_{\text{ground}}]^T. \quad (4)$$

Here, $d_{\min}$, d_{mean} , and σ_d are the minimum distance, average distance, and standard deviation, v_d is the rate of change in distance or height, r_{valid} is the proportion of valid points, h_d is the normalized vertical position, r_{ground} is the proportion of grid points near the ground, and the superscript T denotes transposition.

The ToF confidence score is defined as follows:

$$q_t^D=r_t^{\text{valid}}\exp(-\beta_D\sigma_t^D). \quad (5)$$

In the formula, q_t^D is the ToF confidence score, r_t^{valid} is the proportion of valid points, σ_t^D is the normalized distance dispersion, and $\beta_D>0$ is the penalty coefficient. The visual confidence score is defined as:

$$q_t^V=p_t^{\text{det}}(1-o_t)g(l_t). \quad (6)$$

In the formula, q_t^V is the visual confidence score, p_t^{det} is the person-detection probability, o_t is the occlusion ratio, l_t is the normalized brightness, and $g(\cdot)$ is the brightness quality function.

The gating module uses only the quality scores as input:

$$\alpha_t=\text{sigmoid}(\mathbf{w}_q^T[q_t^D, q_t^V]+b_q). \quad (7)$$

In the formula, $\alpha_t \in [0,1]$ is the weight assigned to the visual branch, $\mathbf{w}_q$ is the two-dimensional weight vector, b_q is the bias, and $\text{sigmoid}(\cdot)$ is the sigmoid function. The fused feature is then:

$$\mathbf{f}_t^F=\alpha_t\mathbf{f}_t^V+(1-\alpha_t)\mathbf{f}_t^D. \quad (8)$$

In the formula, $\mathbf{f}_t^F$ represents the fused feature. When low light or occlusion reduces q_t^V , the weights shift to ToF; when ToF holes increase, the weight shifts toward vision.

3. Experimental design and results

3.1. Simulation platform, data generation, and evaluation methods

Using a fixed random seed of 20260730, 32 virtual human subjects were created to generate a total of 9,600 motion sequences (each 40 steps long), covering various types of falls and daily activities under conditions of normal lighting, low light, backlighting, and occlusion. The data were partitioned by subject into a training set (6,000 sequences), a validation set (1,800 sequences), and a test set (1,800 sequences), ensuring no subject overlap between sets; the test set comprised 720 fall sequences and 1,080 daily activity sequences. Additionally, stress tests involving unimodal corruption (visual or ToF) were conducted. The model was trained using AdamW with a window size of 40, a batch size of 128, and a maximum of 30 epochs; performance was evaluated based on accuracy, precision, recall, F1-score, specificity, false alarms per hour, and alarm latency.

3.2. Overall performance and comparative results

Table 1 presents the results of the composite stress test. The Time-of-Flight (ToF) thresholding method achieved an F1 score of only 71.94% under echo distortion; the ToF classifier and visual classifier achieved 83.67% and 84.95%, respectively, indicating that a single modality is insufficient to cover both types of distortion. Feature concatenation improved the F1 score to 95.38%, while fixed fusion achieved 94.03%, but neither method used quality scores to down-weight unreliable modalities.

Table 1. Event-level results on the composite stress test.

Method	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	FA/h	Delay (ms)
ToF threshold	77.94	73.24	70.69	71.94	10.33	176.4
ToF classifier	87.94	91.30	77.22	83.67	2.94	142.7
Vision classifier	88.39	88.19	81.94	84.95	4.39	134.9
Vision-TCN	87.72	89.29	78.75	83.69	3.78	118.6
Feature concat.	96.33	96.19	94.58	95.38	1.50	103.8
Fixed fusion	95.33	96.22	91.94	94.03	1.44	96.5
Proposed	96.89	96.89	95.28	96.08	1.22	88.7

The proposed method achieved accuracy, precision, recall, F1 score, and specificity of 96.89%, 96.89%, 95.28%, 96.08%, and 97.96%, respectively, with 1.22 false alarms per hour and an average alarm latency of 88.7 ms. Compared to fixed fusion, the F1 score improved by 2.05 percentage points, and the false alarm rate decreased by 15.4%; compared to feature concatenation, the F1 score improved by 0.70 percentage points, and the latency decreased by 15.1 ms. Of the 720 fall scenarios, 34 were missed, mainly due to the simultaneous occurrence of slow descent and structured ToF distortion; 22 false positives were concentrated in rapid lying-down, sitting, and deep-kneeling actions. Quality gating cannot eliminate all conflicts, but it maintained a good balance between recall and false alarms when distortions alternated between modalities. Alarm latency also includes continuous near-ground confirmation.

3.3. Ablation, robustness, and edge-deployment results

Table 2 summarizes the ablation and cost results. Removing the ToF or visual branch reduced the F1 score to 83.69% and 83.67%, respectively; removing the temporal module resulted in an F1 score of 88.29%, with false positives increasing to 5.83 per hour. In the difficult subset consisting of rapid sitting, bending, squatting, kneeling, sitting on the ground, and normal lying down, the false positive rate was 2.72% for the complete method, 12.96% for the non-temporal model, and 3.21% for the fixed fusion method, indicating that temporal constraints and quality gating mitigated errors caused by action similarity and modality-specific distortion, respectively.

Table 2. Ablation, robustness, and edge-cost results.

Setting	Scenario	F1 (%)	FA/h	Latency (ms)	Size (MB)
Full FP32	Composite stress	96.08	1.22	18.9	0.064
Full INT8*	Composite stress	96.08	1.22	12.8	0.016
w/o ToF	Composite stress	83.69	3.78	10.6	0.043
w/o vision	Composite stress	83.67	2.94	6.9	0.039
Fixed gate	Composite stress	94.03	1.44	17.5	0.064
w/o temporal	Composite stress	88.29	5.83	7.4	0.027
Proposed	Low light	99.65	0.17	12.8	0.016
	40% occlusion	99.65	0.17	12.8	0.016
	40% invalid ToF	99.45	0.33	12.8	0.016

4. Conclusion

This paper proposes a lightweight fall detection device that integrates multi-zone time-of-flight (ToF) and low-resolution vision, and validates it through simulations using digital human models and structured modality-specific distortions. The method synchronizes the distance matrix with the corresponding person region, uses dual-branch features to describe distance motion and visual pose, and employs quality gating to adjust weights based on ranging validity, distance dispersion, brightness, occlusion, and detection probability. Temporal convolution and continuous near-ground conditions are used to distinguish falls from similar daily activities. In composite stress tests, the proposed method achieves an F1 score of 96.08%, a false alarm rate of 1.22 false alarms/h, and an alarm latency of 88.7 ms, outperforming fixed fusion and feature-concatenation methods, while maintaining a lightweight model size. The next step is to complete physical-system calibration, measure device latency and power consumption, and conduct real-world validation under appropriate ethical approval and safety protocols.

Fundings

Project Name: Seamless Navigation – A Revolutionary Smart Real-Time Warning White Cane for Cross-Scenario Adaptation. Project Number: 202610595111X.

References

- [1] Mohan, D., Al-Hamid, D. Z., Chong, P. H. J., Sudheera, K. L. K., Gutierrez, J., Chan, H. C. B., & Li, H. (2024). Artificial intelligence and IoT in elderly fall prevention: A review. *IEEE Sensors Journal*, 24(4), 4181–4198. <https://doi.org/10.1109/JSEN.2023.3344605>.
- [2] Alam, E., Sufian, A., Dutta, P., & Leo, M. (2022). Vision-based human fall detection systems using deep learning: A review. *Computers in Biology and Medicine*, 146, 105626. <https://doi.org/10.1016/>

j.compbimed.2022.105626.

[3] Gaya-Morey, F. X., Manresa-Yee, C., & Buades-Rubio, J. M. (2024). Deep learning for computer vision based activity recognition and fall detection of the elderly: A systematic review. *Applied Intelligence*, 54(19), 8982–9007. <https://doi.org/10.1007/s10489-024-05645-1>.

[4] Qi, P., Chiaro, D., & Piccialli, F. (2023). FL-FD: Federated learning-based fall detection with multimodal data fusion. *Information Fusion*, 99, 101890. <https://doi.org/10.1016/j.inffus.2023.101890>.

[5] Amsaprabhaa, M., Jane, Y. N., & Nehemiah, H. K. (2023). Multimodal spatiotemporal skeletal kinematic gait feature fusion for vision-based fall detection. *Expert Systems with Applications*, 212, 118681. <https://doi.org/10.1016/j.eswa.2022.118681>.

[6] Jiao, S., Li, G., Zhang, G., Zhou, J., & Li, J. (2024). Multimodal fall detection for solitary individuals based on audio-video decision fusion processing. *Heliyon*, 10(8), e29596. <https://doi.org/10.1016/j.heliyon.2024.e29596>.

[7] Lafuente-Arroyo, S., Martín-Martín, P., Iglesias-Iglesias, C., Maldonado-Bascón, S., & Acevedo-Rodríguez, F. J. (2022). RGB camera-based fallen person detection system embedded on a mobile platform. *Expert Systems with Applications*, 197, 116715. <https://doi.org/10.1016/j.eswa.2022.116715>.

[8] Lin, B.-S., Yu, T., Peng, C.-W., Lin, C.-H., Hsu, H.-K., Lee, I.-J., & Zhang, Z. (2022). Fall detection system with artificial intelligence-based edge computing. *IEEE Access*, 10, 4328–4339. <https://doi.org/10.1109/ACCESS.2021.3140164>.

[9] Romeo, L., Marani, R., D'Orazio, T., & Cicirelli, G. (2023). Video based mobility monitoring of elderly people using deep learning models. *IEEE Access*, 11, 2804–2819. <https://doi.org/10.1109/ACCESS.2023.3234421>.