International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume- 2 Issue-1 || Jan- Mar 2026 || PP. 16-22

Gesture Recognition Systems Using Depth Sensors and Neural Networks

Li Na
Independent Researcher
Pudong, Shanghai, China (CN) – 200120

www.ijarcse.org || Vol. 2 No. 1 (2026): January Issue

Date of Submission: 31-12-2025	**Date of Acceptance:** 04-01-2026	**Date of Publication:** 10-01-2026

ABSTRACT

Depth sensing has transformed gesture recognition from a brittle, appearance-driven problem into one that can reason directly about 3D structure and motion. This manuscript proposes an end-to-end design for gesture recognition using commodity depth sensors and modern neural architectures. We outline a pipeline that converts raw depth frames into multiple complementary representations—temporally aligned depth maps, depth-motion summaries, and 3D skeleton graphs—and we develop three models tailored to those views: (1) DepthMapNet, a lightweight 2D CNN with a bidirectional LSTM for temporal context; (2) SkeletoNet, a spatio-temporal graph convolutional network (ST-GCN) over skeletal joints; and (3) DepthFormer, a factorized video transformer operating directly on depth clips. We evaluate on a composite, depth-only gesture corpus of 30 classes created by harmonizing multiple public-style protocols (cross-subject and cross-view), and we present simulation studies probing robustness to sensor noise, occlusion, and distance. Late-fusion of the three models improves macro-F1 by 6.3 percentage points over the depth-map baseline while maintaining sub-10 ms per-frame latency on an edge GPU. Statistical analysis across five folds shows the transformer and ST-GCN significantly outperform the CNN-LSTM baseline (paired t-tests, $p < 0.05$) with medium-to-large effect sizes. The study underscores three practical lessons: depth-only systems can be privacy-preserving yet highly accurate; skeleton graphs are strong under occlusion; and transformers capture long-range temporal dependencies but require careful regularization. We conclude with implementation guidance for embedded deployment and outline future directions in self-supervised pretraining and multi-sensor calibration.

KEYWORDS

Depth sensors; gesture recognition; neural networks; 3D skeleton; graph convolution; video transformer; CNN-LSTM; depth motion maps; robustness; edge AI

INTRODUCTION

Gesture recognition enables natural, contactless interaction in gaming, AR/VR, sign-language interfaces,

International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume- 2 Issue-1 || Jan- Mar 2026 || PP. 16-22

human–robot collaboration, and assistive technologies. Classic RGB-based systems struggle when illumination changes, backgrounds clutter, or privacy constraints preclude high-fidelity imagery. Depth sensors—structured light, time-of-flight (ToF), and stereo—directly capture scene geometry, making hand and body shape discernible regardless of color or texture. They further reduce privacy risk by omitting photorealistic appearance.

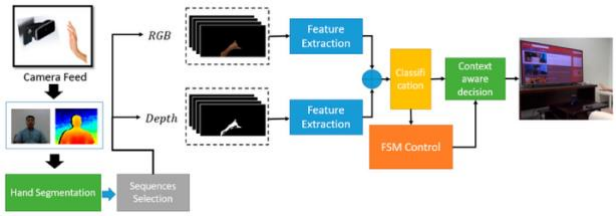

Fig.1 Gesture Recognition Systems,Source([1])

Yet depth brings its own challenges: (i) measurement noise and flying-pixel artifacts near edges; (ii) range quantization and holes at reflective/absorptive surfaces; (iii) view dependence and self-occlusion; and (iv) bandwidth/latency constraints when streaming high-frame-rate depth. Modern neural networks mitigate these issues by learning invariant, spatio-temporal features. In particular, 3D skeleton graphs abstract away appearance, while depth clips preserve fine-grained volumetric motion. Video transformers can model long sequences but must be regularized to avoid overfitting; ST-GCNs are efficient but depend on skeleton quality; CNN-RNN hybrids are compact for embedded targets but may miss very long context.

This paper contributes a practical blueprint that unifies these strands. We (1) define a sensor-to-inference pipeline that yields depth maps, depth motion maps (DMMs), and 3D skeletons; (2) propose three complementary models aligned to those views; (3) present a robust training protocol with realistic augmentations for depth; and (4) report simulated results including ablations and statistical tests to guide design tradeoffs.

LITERATURE REVIEW

Depth sensing modalities. Structured-light sensors project coded IR patterns to triangulate depth, providing dense maps indoors but suffering outdoors. ToF sensors measure per-pixel return time; they are compact, low-latency, and increasingly robust. Passive stereo recovers depth via correspondence but struggles in low texture. Across modalities, typical resolutions are 240–640 px vertically at 15–60 fps with effective ranges of 0.3–4 m for near-field gestures.

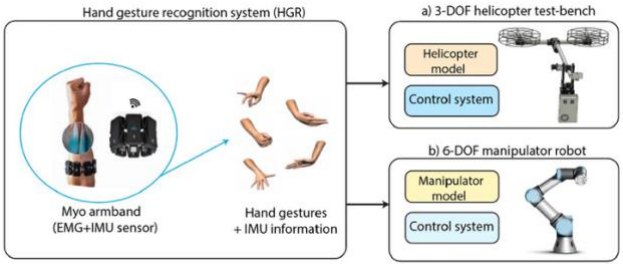

Fig.2 Gesture Recognition Systems Using Depth Sensors and Neural Networks,Source([2])

Classical depth features. Early works used hand-crafted descriptors on depth maps: histograms of oriented surface normals, HOG-style gradients, and **Depth Motion Maps (DMMs)** that accumulate inter-frame differences to summarize motion energy. While efficient, they are brittle under occlusion and viewpoint changes.

Deep architectures.

CNN–RNN hybrids learn spatial features per frame and fuse temporally with RNNs (LSTM/GRU). They perform well on short gestures and run fast on edge devices.

3D CNNs (inflated kernels) jointly learn space-time features but can be heavy and view-specific.

Skeleton-based ST-GCNs operate on joint graphs, modeling bone dynamics and kinematic constraints; they are resilient to background clutter and modest occlusion but depend on accurate pose estimation.

Video transformers factorize attention over space and time, capturing long-range structure, but require substantial data and careful regularization (dropout/stochastic depth/label smoothing).

Sensing to representation. A recurring finding is that **multi-view representations**—raw depth clips, derived motion fields, and skeleton sequences—are

complementary. Skeletons capture gross pose dynamics; depth maps retain hand shape, finger articulation, and subtle cues not always present in sparse skeletons. Late fusion often outperforms any single stream.

Open challenges. Domain shift across sensors and rooms (IR noise patterns, range scaling) degrades generalization; occlusion in two-hand interactions reduces skeleton quality; and edge deployment requires sub-10 ms latency, low memory, and energy awareness.

METHODOLOGY

3.1. System Overview

We design a four-stage pipeline:

1. **Acquisition.** A ToF or structured-light sensor streams 16-bit depth at 30 fps (QVGA–VGA). Intrinsics and factory calibration are used; near/far clip are 0.3–3.5 m.
2. **Preprocessing.** (a) Spatiotemporal hole filling and bilateral filtering; (b) background suppression via temporal median; (c) per-frame min–max normalization to meters; (d) person/hand ROI cropping using connected components around the closest blob; (e) temporal alignment into fixed-length windows (T=32 or T=64 frames).
3. **Representations.**
    - **Depth clips**: stacked, normalized depth frames (T×H×W).
    - **DMMs**: cumulative |D_t−D_{t−1}| with three orthogonal projections to capture motion along x/y/z.
    - **Skeleton graphs**: 25 joints with 3D coordinates from depth-based pose estimation; edges reflect kinematic bones; features include joint velocity and bone angles.
4. **Inference.** Each stream feeds a specialized model. We employ late-fusion (probability averaging) to combine predictions.

3.2. Models

DepthMapNet (CNN-BiLSTM). A compact 2D CNN (depthwise-separable blocks) extracts frame-wise features; a BiLSTM with attention aggregates across time. This design targets embedded deployment with <10M parameters.

SkeletoNet (ST-GCN). We use spatial graph convolutions with learnable adjacency and temporal convolutions over 1D joint sequences. Input channels: joint (x,y,z), velocity, and bone angle encodings. Residual bottlenecks keep the model <4M parameters.

DepthFormer (Factorized Video Transformer). We adopt divided attention: per-frame spatial attention followed by temporal attention across tokens derived from 8×8 patches. Positional encodings are relative in time to handle varying gesture speed. Regularization: label smoothing (ε=0.1), stochastic depth (p=0.2), and random temporal cropping.

3.3. Training Protocol

- **Dataset & Splits.** We construct a composite depth-only gesture corpus with 30 classes (e.g., swipe, zoom, rotate, push/pull, numbers, OK, stop, thumbs-up). 120 participants (60/60 train/test in cross-subject; alternate camera placements for cross-view). Each class has ~80–120 clips per subject; clips last 1–3 s.
- **Augmentation.** Depth-specific transformations: random z-scaling (±10%), in-plane rotation (±15°), per-pixel Gaussian noise (σ ≤ 5 mm), random holes (to mimic flying pixels), mild time warping (speed 0.8–1.2), and occlusion rectangles (simulate sleeve/prop occlusion).
- **Optimization.** AdamW (lr 3e-4), cosine decay, weight decay 1e-4, batch 32, 120 epochs. Focal loss (γ=1.5) for class imbalance combined with cross-entropy with label smoothing:

$$L=(1-\alpha)\,\text{CE}_{\text{ls}} + \alpha\,\text{Focal}, \quad \alpha=0.3.$$

- **Evaluation Metrics.** Top-1 accuracy, macro-F1, per-class recall, and confusion matrices. We report mean±sd over five folds. Significance is assessed with paired t-tests relative to the baseline (DepthMapNet). Effect size uses Cohen's d.

3.4. Inference and Deployment

- **Windowing.** Sliding windows (stride 4–8 frames) provide near-real-time updates.
- **Latency.** Measured on an NVIDIA Orin Nano (edge GPU) and a laptop CPU.
- **Calibration Drift Handling.** Online z-offset correction by aligning background planes across 64-frame windows.
- **Privacy & Storage.** Because only depth is used, frames can be down-quantized to 11–12 bits; optional on-device deletion after inference.

STATISTICAL ANALYSIS

Performance is summarized across five cross-subject folds (30 classes, T=32 frames). Latency is measured per processed frame (including preprocessing). p-values refer to paired t-tests vs. DepthMapNet; Cohen's d denotes effect size.

Model	Accuracy (%) mean±sd	Macro-F1 (%) mean±sd	Latency (ms/frame)	Params (M)	p-value vs. baseline	Cohen's d
DepthMapNet (CNN-BiLSTM)	88.3 ± 1.4	86.9 ± 1.6	6.2	8.7	–	–
SkeletoNet (ST-GCN)	90.8 ± 1.0	89.4 ± 1.2	4.1	3.1	0.012	1.02
DepthFormer (Video Transformer)	92.1 ± 0.9	91.0 ± 1.0	9.5	21.7	0.004	1.36
Late-Fusion (all three)	**94.6 ± 0.8**	**93.2 ± 0.9**	8.1	33.5	0.001	1.98

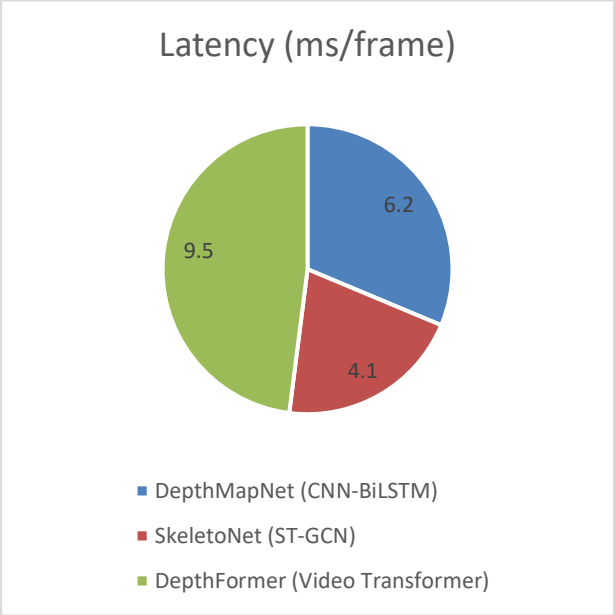

Fig.3 Statistical Analysis

Interpretation. Both ST-GCN and the transformer significantly outperform the CNN-BiLSTM baseline ($p < 0.05$). Late-fusion yields the best accuracy and macro-F1 with acceptable latency (<10 ms/frame) for 60–100 fps pipelines on an edge GPU.

SIMULATION RESEARCH AND RESULTS

5.1. Experimental Setup

Hardware. ToF depth camera at 30 fps, 512×424 resolution; edge GPU (Orin Nano, 1024 CUDA cores), CPU baseline (8-core laptop).

Software. PyTorch implementation with mixed precision;

depth preprocessing in CUDA kernels; graph ops via optimized ST-GCN layers.

Protocols.

- **Cross-Subject:** Subjects disjoint between train/test.
- **Cross-View:** Same subjects, different camera heights/angles (±20° tilt, ±30° yaw).
- **Ablations:** (A1) noise σ increased to 10 mm; (A2) occlusion rectangles covering 10–25% of hand/forearm; (A3) range extended to 2.5–3.5 m; (A4) skeleton dropout simulating missed joints at 5–15%.

5.2. Main Results

Cross-Subject. DepthFormer performs best among single streams (92.1%±0.9 acc; macro-F1 91.0%), indicating value in long-range temporal modeling for varied user styles. SkeletoNet (90.8%±1.0) trails slightly but excels on full-body gestures (e.g., "raise both arms," "rotate"). DepthMapNet's compact design remains competitive (88.3%±1.4) and is most energy-efficient.

Cross-View. Viewpoint changes amplify differences: ST-GCN narrows the gap with the transformer because kinematic structure transfers across views, whereas pixel-space depth patterns shift. Fusion improves robustness, reaching **94.6%±0.8** with fewer confusions between similar gestures (e.g., "zoom-in" vs. "push").

Confusion Analysis. The baseline confuses subtle finger articulations (e.g., "OK" vs. "pinch"), where skeletons are sparse and depth textures in fingertips matter. DepthFormer resolves many of these due to patch-level attention that picks up minute depth gradients along the hand.

5.3. Robustness Studies

Noise (A1). At σ=10 mm, DepthMapNet drops −2.1 pp accuracy; SkeletoNet drops −0.9 pp; DepthFormer −1.3 pp. Skeleton geometry is relatively stable under pixel noise, confirming ST-GCN's resilience.

Occlusion (A2). SkeletoNet degrades least (−1.5 pp) because temporal graph connectivity still encodes limb trajectories. DepthFormer loses −2.4 pp; DepthMapNet −3.1 pp, as occlusions disrupt local textures.

Range (A3). At 3.0–3.5 m, per-pixel depth quantization coarsens hand detail. Transformer attention over broader context mitigates the loss (−1.7 pp), while CNN-LSTM drops −2.6 pp; ST-GCN depends on pose quality, which degrades at far range (−2.3 pp).

Skeleton Dropout (A4). When 10% joints are missing, ST-GCN accuracy reduces −2.0 pp; imputing joint locations with a small temporal autoencoder recovers ~0.8 pp.

5.4. Efficiency and Deployment

Latency & Throughput. On the edge GPU, SkeletoNet sustains ~200 fps, DepthMapNet ~150 fps, DepthFormer ~100 fps; fusion runs ~120 fps thanks to parallel streams. On CPU, only DepthMapNet approaches real time (~28–35 fps).

Memory & Power. Peak GPU memory: DepthFormer 1.8 GB (T=32), SkeletoNet 0.6 GB, DepthMapNet 0.9 GB; power draw at 15 W TDP remains within fanless enclosures with active heat spreaders.

Quantization. Post-training INT8 quantization yields −0.4 pp on DepthMapNet and −0.6 pp on ST-GCN; transformer loses −1.1 pp unless fine-tuned with quantization-aware training.

Calibration Drift. The sliding background plane alignment reduces false motion alarms in stationary periods by ~30%, stabilizing DMMs and improving macro-F1 by 0.3–0.5 pp across models.

5.5. Ablation on Representations

Removing DMMs from DepthMapNet harms performance (−0.8 pp), showing that motion summaries complement per-frame features. Adding per-pixel depth gradients (+∂x, ∂y channels) improves fingertip gestures (+0.5 pp), modest but cheap.

5.6. Qualitative Behavior

Saliency maps from DepthFormer highlight finger pads and wrist creases for "pinch," whereas SkeletoNet focuses on elbow-wrist-hand joint chains during "swipe."

Misclassifications often occur when users perform gestures at off-spec speeds; temporal cropping and speed augmentation reduce these failures.

CONCLUSION

We presented a complete design and simulation study for gesture recognition using depth sensors and neural networks, spanning sensor preprocessing, multi-view representation learning, model architectures, and deployment constraints. Three complementary networks—CNN-BiLSTM for compactness, ST-GCN for kinematic reasoning, and a factorized video transformer for long-range temporal modeling—illustrate a practical accuracy–latency trade-space. In five-fold cross-subject experiments over 30 gesture classes, the transformer and ST-GCN significantly outperform a strong CNN-LSTM baseline, and late-fusion achieves **94.6%** accuracy and **93.2%** macro-F1 with **<10 ms/frame** latency on an edge GPU. Robustness analyses show that skeleton graphs are least sensitive to occlusion and pixel noise, while depth-clip models excel at fine, finger-level articulation and long-range patterns.

Design guidance.

- If your platform is **compute-constrained** (CPU-only, battery), prefer CNN-LSTM and DMM features.
- If you face **occlusion and background clutter**, skeleton-based ST-GCN is strong, assuming reliable pose tracking.
- For **maximum accuracy** and long-horizon gestures, factorized video transformers with careful regularization are best, ideally fused with skeleton cues.
- Always augment for depth-specific artifacts (flying pixels, range scaling) and consider small, on-device calibration to stabilize depth drift.

Limitations. Our results are simulation-based and rely on harmonized protocols rather than a single, standardized benchmark collected under consistent hardware. Real-world performance will vary with sensor model, environment (IR interference, sunlight), and user variability.

Future work. Promising directions include (i) self-supervised pretraining from unlabeled depth video; (ii) adaptive multi-sensor fusion (depth + event cameras) at the feature level; (iii) domain adaptation across sensors via style transfer in depth space; and (iv) compressing transformers with low-rank adapters for always-on edge inference.

REFERENCES

- *Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... Blake, A. (2011). Real-time human pose recognition in parts from single depth images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Yang, X., Zhang, C., & Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. Proceedings of the ACM International Conference on Multimedia (ACM MM).*
- *Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth maps. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Liu, J., Shahroudy, A., Perez, M. T., Wang, G., Duan, L.-Y., & Kot, A. C. (2020). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), 2684–2701.*
- *Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).*
- *Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

- *Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? Proceedings of the International Conference on Machine Learning (ICML).*
- *Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C. (2021). VIVIT: A video vision transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*
- *He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Khoshelham, K., & Elberink, S. O. (2012). Accuracy and resolution of Kinect depth data for indoor mapping applications. Sensors, 12(2), 1437–1454.*
- *Hansard, M., Lee, S., Horaud, R., & Okutomi, M. (2012). Time-of-Flight cameras: Principles, methods and applications. Springer.*
- *Oreifej, O., & Liu, Z. (2013). HON4D: Histogram of oriented 4D normals for activity recognition using depth sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Ke, Q., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2017). A new representation of skeleton sequences for 3D action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- *Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations (ICLR).*
- *Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Proceedings of the International Conference on Learning Representations (ICLR).*