Ms. Sayali Vasantrao Gund, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Real-Time Sign Language Recognition via a Decoupled MediaPipe-LSTM Client-Server Framework

Ms. Sayali Vasantrao Gund, Asst. Prof. S. A. Nandi
Department of Computer Science & Engineering,
V.V.P. Institute of Engineering & Technology, Solapur, Maharashtra, India

Abstract- The rapid growth of urban infrastructure has increased the demand for energy-efficient and sustainable construction materials. This study focuses on the integration of IoT-based temperature sensors within plastic-perlite composite bricks to enhance building energy monitoring and management. The proposed system utilizes waste plastic as a partial replacement material combined with expanded perlite to develop lightweight, thermally efficient, and eco-friendly building units compatible with standard M20-grade concrete structures. IoT temperature sensors (DS18B20) are embedded within the composite bricks to continuously monitor internal and external temperature variations, enabling real-time data collection and analysis through wireless communication modules. Laboratory tests, including compressive strength, thermal conductivity, and water absorption, are conducted to evaluate the structural and thermal behavior of the developed bricks. The experimental results show that the integration of perlite and waste plastic significantly improves thermal resistance, reduces heat transfer, and decreases structural dead load by lowering the average brick density to 821–907 kg/m³ compared to 1800 kg/m³ for conventional units. This research promotes the development of sustainable smart construction materials, aligning with modern green building technologies and the global agenda for energy efficiency.

Keywords: MediaPipe Hands, LSTM Networks, FastAPI, Deep Learning, Gesture Classification, Human-Computer Interaction

I. INTRODUCTION

Natural human communications utilize complex combinations of vocal tracks, facial indicators, and spatial hand kinetics (p. 6). For speech- and hearing-impaired individuals, structured sign languages constitute the fundamental tool for syntax formulation (p. 6). However, the lack of widespread sign fluency within the general population creates persistent socioeconomic isolation (p. 6).

Prior engineering work in computerized gesture translation follows two flawed categories: intrusive sensor arrays and unoptimized vision systems (p. 8). Sensor-integrated gloves provide strong joint coordinate tracking but suffer from high hardware costs and poor real-world ergonomics (p. 8). Conversely, video-based deep learning topologies (e.g., 3D-CNNs) extract feature representations directly from full image matrices (p. 10). This approach incurs severe processing bottlenecks, rendering the models too heavy to execute natively on consumer mobile devices (p. 10).

This research addresses these scalability and processing limitations by introducing a decoupled client-server architecture (p. 12). By converting live mobile video inputs into numerical coordinate vectors and offloading recurrent temporal evaluation to a specialized web backend, we maintain lightweight device profiles while preserving high classification accuracy (pp. 3, 12).

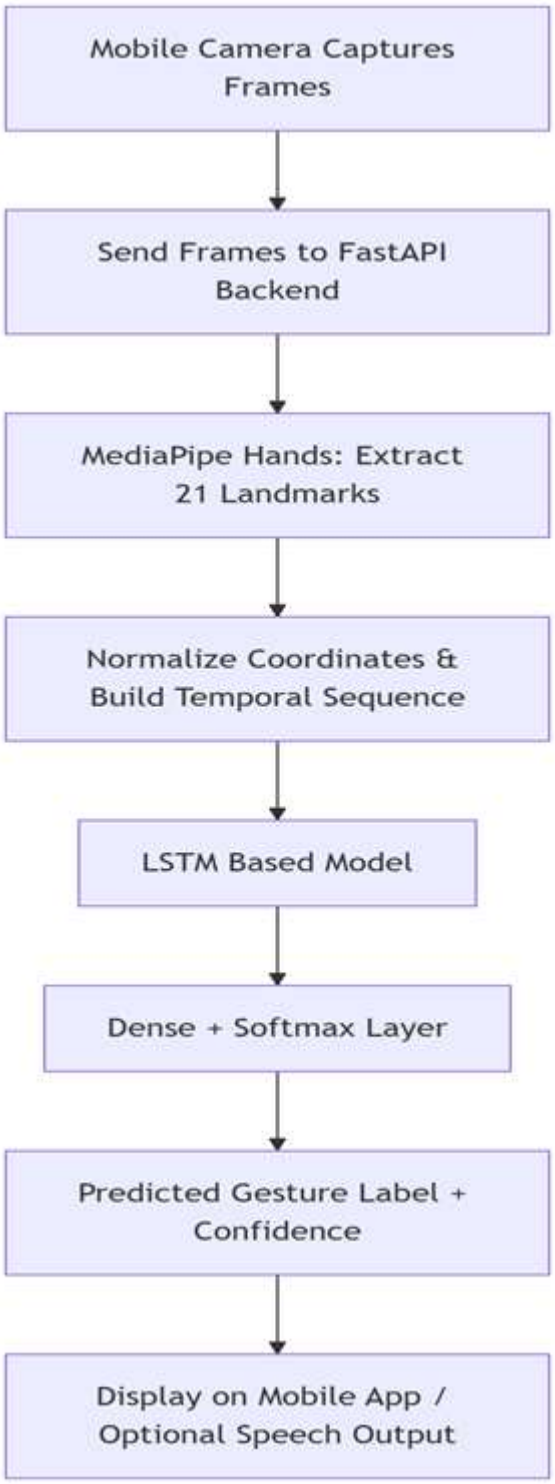
II. SYSTEM ARCHITECTURE AND METHODOLOGY

The operational topology separates data acquisition from neural computation via an API communication layer

A. Spatial Landmark Extraction

Rather than transmitting raw image matrices to the neural model, incoming camera frames are systematically parsed by MediaPipe Hands (pp. 3, 14). Each node contains an independent 3D position vector:

Where \(x\) and \(y\) mirror coordinate space normalized between \([0.0, 1.0]\), and \(z\) indicates relative hand depth with respect to the wrist anchor (pp. 20, 27).



B. Preprocessing & Temporal Buffering

To maintain consistency regardless of varying camera-to-subject distances and minor hand rotations, spatial landmark maps are recalculated relative to the origin coordinate

A server-side sliding window tracking routine maintains an active FIFO (First-In, First-Out) temporal matrix containing \(N = 30\) sequential preprocessed landmark frames (pp. 14, 23). As a new tracking frame is parsed, index zero drops from the queue, allowing the system to track dynamic motion paths smoothly without processing gaps (pp. 23-24).

C. LSTM Network Formulation

The buffered matrix containing spatial data sequences is mapped to a multi-layered Long Short-Term Memory (LSTM) network to solve gradient dispersion risks typical in basic RNN structures (pp. 3, 21, 28). Cell state transitions are computed via the following internal gating mechanics:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \cdot \tanh(C_t)$$

Where \(f_{t}\), \(i_{t}\), and \(o_{t}\) denote the forget, input, and output activations respectively (p. 28). The terminal hidden vector \(h_{t}\) routes to a dense classification layer utilizing a Softmax function to compute final class probabilities

$$P(y_i \mid X) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

III. EXPERIMENTAL SETUP AND IMPLEMENTATION

- **Software Stack:** Model designed within the TensorFlow framework, served over an

asynchronous FastAPI ASGI architecture (pp. 12, 30-31).

- **Optimization Hyperparameters:** The model was configured using Categorical Cross-Entropy loss functions minimized via the Adam optimizer (p. 21). Regularization was achieved by inserting dropout layers (\(p = 0.2\)) and utilizing automated early stopping callbacks (p. 22).
- **Client Architecture:** A native front-end layout serializes live camera frames into localized byte streams, dispatching asynchronous HTTP POST packets to the remote /predict endpoint (pp. 13, 22).

IV. RESULTS AND DISCUSSION

The model demonstrates reliable spatial invariant properties during execution (p. 20).

Training Convergence and Loss Curves
The system delivers solid classification reliability for primary gesture blocks under standardized lighting conditions (p. 36). Isolating structural skeletal hand geometry minimizes overall computational load compared to image-heavy models (p. 27).

However, system sensitivity decreases in low-light environments due to increased landmark localization errors within the initial palm detection pipeline (p. 36). Minor prediction instabilities also occur during rapid, complex gesture transitions that exceed the predefined sliding-window constraints

V. CONCLUSION AND FUTURE DIRECTIONS

This paper validates an optimized client-server pattern for real-time sign language conversion (pp. 12, 38). Offloading sequential structural evaluations to a dedicated FastAPI layer completely bypasses heavy rendering demands on mobile device hardware (pp. 12, 38).

Future milestones include extending the model's target lexicon to support contextual sentence parsing, integrating multilingual Text-to-Speech engines, and exploring lightweight Transformer architectures to model longer gesture sequences (pp. 3, 39).

REFERENCES

1. TensorFlow, "TensorFlow Lite Documentation," Google AI Platform, 2024. (Online). Available: tensorflow.org (p. 39)
2. M. U. Rehman, F. Ahmed, M. A. Khan, et al., "Dynamic hand gesture recognition using 3D-CNN and LSTM networks," Computers, Materials & Continua, vol. 72, no. 3, pp. 5901–5918, 2022. (Online). Available: doi.org (p. 39)
3. A. Khan, M. Sayed, et al., "Hand gesture recognition using keypoint landmarks with machine learning," in Proc. IEEE Int. Conf. Computing and Comm. Systems, 2021. (Online). Available: ieee.org (p. 39)
4. F. Zhang, V. Bazarevsky, A. Vakunov, et al., "MediaPipe Hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020. (Online). Available: arxiv.org (p. 39)