

ABSTRAK

Interaksi manusia-komputer terus bergeser menuju modalitas yang lebih alami dan bebas kontak fisik, dengan pengenalan gestur tangan sebagai salah satu pendekatan yang paling banyak dieksplorasi untuk mewujudkan interaksi tanpa sentuh tersebut. Penelitian ini membandingkan dua arsitektur deep learning, CNN+LSTM dan TimeSformer, untuk mengenali 5 kelas gesture tangan dari *dataset* 20BN-Jester (12.000 sampel, 2.400 sampel per kelas) melalui *grid search* sistematis dengan 48 eksperimen. Evaluasi pada 4.500 data uji independen menunjukkan CNN+LSTM (EXP31) mencapai akurasi 95,76% dan latensi inferensi 14,72 ms, sedangkan TimeSformer (EXP43) mencapai akurasi 95,44% namun mencatat latensi 217,35 ms yang melampaui ambang batas real-Time 100 ms sebesar 2,17 kali lipat. CNN+LSTM ditetapkan sebagai arsitektur optimal karena satu-satunya yang memenuhi seluruh syarat kelayakan secara simultan: akurasi tertinggi, latensi jauh di bawah ambang batas, dan efisiensi komputasi yang memungkinkan deployment pada perangkat keras kelas konsumen. Temuan ini memberikan gambaran empiris mengenai *trade-off* akurasi dan latensi antara pendekatan berbasis rekurensi dan berbasis *attention*, sekaligus menjadi acuan bagi pengembangan sistem interaksi berbasis gestur tangan yang lebih luas.

Kata kunci : Pengenalan gestur tangan, CNN+LSTM, TimeSformer, antarmuka tanpa sentuh, deep learning, inferensi real-time, *dataset* Jester

ABSTRACT

Human-computer interaction is increasingly shifting toward more natural, contact-free modalities, with hand gesture recognition emerging as one of the most widely explored approaches for enabling such touchless interaction. This study compares two deep learning architectures, CNN+LSTM and TimeSformer, for recognizing 5 hand gesture classes from the 20BN-Jester *dataset* (12,000 samples, 2,400 per class) through a systematic *grid search* comprising 48 experiments. Evaluation on 4,500 independent test samples shows that CNN+LSTM (EXP31) achieves 95.76% Accuracy with an inference latency of 14.72 ms, while TimeSformer (EXP43) achieves 95.44% Accuracy but records a latency of 217.35 ms, exceeding the 100 ms real-Time threshold by 2.17 times. CNN+LSTM is established as the optimal architecture, being the only model that simultaneously satisfies all viability criteria: highest Accuracy, latency well below the threshold, and computational efficiency enabling deployment on consumer-grade hardware. These findings provide empirical insight into the accuracy-latency *trade-off* between recurrence-based and *attention*-based approaches, and serve as a reference for the broader development of hand gesture-based interaction systems.

Keywords: Hand gesture recognition, CNN+LSTM, TimeSformer, touchless interface, deep learning, real-time inference, Jester *dataset*