Time-Frequency Joint Modeling for Few-Shot 3D Chinese Air-Writing Recognition with IMU

Authors: Shuo Xu, Haoxi Shangguan, Jiayin Wang, Haokai Kang, Yixin Bao, Yifeng Wang, Yixin Bao, Yifeng Wang

Date: 2026-03-09T11:59:50+00:00

Abstract

With the development of natural human-computer interaction technology, the recognition of three-dimensional Chinese character writing based on inertial sensors has become a highly potential contactless text input method. However, the huge category scale and subtle structural differences of Chinese character systems, coupled with the difficulty of obtaining large-scale high-quality data in practical applications, pose a unique and severe challenge in this field. The existing sequence classification methods are often difficult to effectively balance the local stroke details and the overall global structure when dealing with such complex motion trajectories with high categories and scarce samples, which is prone to feature recognition confusion and model overfitting. In order to break through this technical bottleneck, this study proposes a novel action recognition framework based on joint time and frequency modeling. Firstly, we design a data augmentation strategy that conforms to the physical laws of real motion, which greatly enriches the diversity of samples while preserving the unique dynamic characteristics of writing. Then, the model constructs a parallel feature extraction structure to capture the temporal dynamic evolution law and the global frequency rhythm pattern of the trajectory sequence respectively, and realizes the dynamic fusion of time-frequency high-dimensional features through the attention mechanism. A large number of experiments show that the dual-path method fundamentally enhances the fine-grained identification ability of similar glyphes, and significantly improves the accuracy and stability of multi-class classification. It is worth mentioning that, thanks to the effective optimization of the feature space by introducing the frequency domain dimension, the proposed model achieves excellent recognition performance while naturally maintaining a very low parameter overhead, which provides an efficient and reliable algorithm foundation for resource-constrained wearable devices.

Full Text

Preamble

Time-Frequency Joint Modeling for Few-Shot 3D Chinese Air-Writing Recognition with IMU

Shuo Xu, Haoxi Shangguan, Jiayin Wang, Haokai Kang, Yixin Bao* and Yifeng Wang College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China; School of Science, Harbin Institute of Technology, Harbin 150080, China; Department of Fundamental Education, Harbin Institute of Technology, Shenzhen 518055, China; Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. *These authors contributed equally to this work and share first authorship.

Abstract

With the development of natural human-computer interaction technology, the recognition of three-dimensional Chinese character writing based on inertial sensors has become a highly promising non-contact text input method. However, the massive category scale and subtle structural differences of Chinese character systems, coupled with the difficulty of obtaining large-scale high-quality data in practical applications, pose a unique and severe challenge. Existing sequence classification methods often struggle to balance local stroke details with global structure when dealing with complex motion trajectories characterized by high category counts and scarce samples, leading to feature confusion and model overfitting. To overcome this technical bottleneck, this study proposes a novel action recognition framework based on joint time and frequency modeling. First, we design a data augmentation strategy that adheres to the physical laws of real motion, enriching sample diversity while preserving unique dynamic writing characteristics. Second, the model constructs a parallel feature extraction structure to capture temporal dynamic evolution and global frequency rhythm patterns, respectively, achieving dynamic fusion of high-dimensional time-frequency features through an attention mechanism. Extensive experiments demonstrate that this dual-path method fundamentally enhances the fine-grained identification of similar glyphs and significantly improves multi-class classification accuracy and stability. Notably, by optimizing the feature space through the frequency domain, the proposed model achieves excellent recognition performance with very low parameter overhead, providing an efficient algorithm foundation for resource-constrained wearable devices.

Keywords

Air-writing recognition; Inertial sensor; Time-frequency fusion; Data augmentation; Attention mechanism

1. Introduction

With the rapid development of pervasive computing and Internet of Things (IoT) technology, human-computer interaction is evolving from traditional graphical user interfaces (GUI) to natural user interfaces (NUI). Traditional modalities like keyboards and mice require physical contact and visual attention, limiting interaction freedom. As augmented reality (AR), virtual reality (VR), and wearable computing become more popular, computing environments are extending into physical space. This shift requires portable, robust, and ubiquitous input methods. Digital reconstruction and intelligent recognition of handwriting are critical in this context. In 3D free-space handwriting, users write motion trajectories in the air without physical support, enabling immersive interaction. Handwriting is a dynamic process involving motion planning, muscle control, and temporal logic. Converting these natural motions into computer instructions is essential for next-generation immersive systems.

As an ideographic system, Chinese characters reflect unique cultural connotations and individual movement characteristics through stroke sequences and rhythms. Using Inertial Measurement Units (IMUs) to capture and recognize 3D writing can enable efficient character input and contribute to cultural preservation by transforming traditional writing into quantifiable motion data. This technology helps preserve Chinese character culture in digital forms, enhancing its vitality in the global ecosystem.

Currently, handwriting recognition follows two main routes: visual and inertial. Vision-based methods face physical constraints such as lighting requirements, limited camera fields of view, and privacy concerns. In contrast, Micro-Electro-Mechanical System (MEMS) inertial sensors offer a low-cost, low-power, and privacy-preserving alternative. However, applying this to Chinese handwriting is challenging due to the massive character set (over 3,000 commonly used characters) and the high variability of 3D motion, including differences in speed, strength, and posture. Training high-precision classifiers for thousands of classes typically requires massive datasets, which are difficult to collect for 3D air-writing. This leads to data sparsity and overfitting.

To address these challenges, this study proposes a Time-Frequency Cross-attention GRU (TF-CGRU) framework. We address the problem at both the data and model levels. At the data level, we design an augmentation strategy that respects physical constraints to expand sample distribution. At the model level, we construct a unified time-frequency feature modeling structure to capture both dynamic evolution and spectral patterns. The main contributions are:

1. A motion-consistent data augmentation framework for IMU-based 3D air-writing under few-shot conditions.
2. A unified time–frequency modeling architecture that integrates temporal representations with frequency-domain features via a structured fusion mechanism.

3. A reproducible IMU-based 3D Chinese character recognition framework for large-category, small-sample scenarios.

2. Related Work

IMU signal processing focuses on modeling multivariate continuous time series data. In air-writing, research has evolved from manual statistical features to trajectory recovery and end-to-end deep learning.

2.1. Statistical modeling method

Early methods relied on hand-crafted features and probabilistic models like Hidden Markov Models (HMM) or Support Vector Machines (SVM). While efficient, these methods struggle with the complex structural differences of large-scale Chinese character sets.

2.2. Classification method based on trajectory reconstruction

These methods attempt to recover 2D or 3D trajectories through pose estimation and integration, treating recognition as an image classification task. However, performance is highly sensitive to sensor drift and cumulative errors during long stroke sequences.

2.3. End-to-end approach based on Deep Learning

Modern approaches model raw IMU sequences directly. Convolutional Neural Networks (CNNs) extract local patterns, while Recurrent Neural Networks (RNNs) like GRUs and LSTMs capture temporal dependencies. Attention mechanisms and Transformers have further improved global modeling. However, challenges remain in balancing local details and global structures for high-category 3D Chinese character recognition.

3. Methodology

3.1. Overall Framework

The TF-CGRU framework (Figure [Figure 1: see original paper]) processes raw IMU signals through data augmentation before splitting them into Time Domain and Frequency Domain paths. GRU encoders extract features from each path, which are then fused via a cross-attention mechanism. The time domain path captures stroke order dependencies, while the frequency domain path uses Fast Fourier Transform (FFT) to extract rhythmic patterns. This dual-path design is complementary: the time domain handles sequence dependence, while the frequency domain highlights global patterns to distinguish similar characters.

[Figure 1: see original paper]

3.2. Data Augmentation

To mitigate small sample sizes, we apply eight physical transformations: Gaussian noise, global amplitude scaling, axis-independent scaling, random rotation, bias addition, channel attenuation, and Gaussian smoothing.

- **Gaussian Noise & Bias:** Simulates sensor jitter and zero drift. $S'_{i,j} = S_{i,j} + b_j$, where $b_j \sim \mathcal{N}(0, 0.03)$.
- **Scaling:** Simulates variations in writing strength and posture. $S' = k \cdot S$ (global) or $S'_{i,j} = k_j \cdot S_{i,j}$ (independent).
- **Random Rotation:** Simulates hand orientation changes using rotation matrices.
- **Smoothing:** Reduces high-frequency noise using a Gaussian kernel to focus on stroke structure.

3.3. Feature Enhancement

The feature enhancement module (Figure [Figure 2: see original paper]) uses parallel GRU encoders. The frequency domain path is transformed via FFT:

$$F_k = T^{-1} \sum_{n=0}^{T-1} s_n \exp\left(-i\frac{2\pi kn}{T}\right)$$

The GRU encoder captures temporal patterns using update (z_t) and reset (r_t) gates:

$$\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\
\tilde{h}_t &= \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
\end{aligned}$$

Cross-attention fusion integrates the paths:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \quad O_{\text{time}} = AV$$

The final features are concatenated and passed to a softmax classifier:

$$y = \text{softmax}(W_c \text{Feat} + b_c)$$

[Figure 2: see original paper]

4. Experiment

4.1. Dataset and experimental setup

The dataset contains 496 Chinese character categories, each with 10 samples written by different individuals. Each sample consists of 6-axis IMU data (triaxial acceleration and angular velocity). We use 8 samples per category for training and 2 for testing, averaging results over five runs.

4.2. Data augmentation validation

Figure [Figure 3: see original paper] illustrates that the augmentation strategy introduces reasonable perturbations while maintaining stroke morphology, effectively simulating real-world writing fluctuations.

[Figure 3: see original paper]

4.3. Frequency domain validation

Spectral analysis (Figure [Figure 4: see original paper]) shows that different characters exhibit distinct energy distributions. Low-complexity characters concentrate energy in low frequencies, while complex characters show multi-peak distributions in middle and high frequencies, particularly in angular velocity.

[Figure 4: see original paper]

4.4. Comparative Studies

As shown in Table , the proposed TF-CGRU model achieves an accuracy of 0.8810 and an MCC of 0.8809, outperforming mainstream models like DeepConvLSTM and Multi-scale Transformers.

Importantly, Table shows that TF-CGRU maintains a small parameter footprint (2.39 MB), making it more efficient than larger Transformer-based models.

4.5. Ablation Studies

Ablation results (Table) confirm that reducing augmentation strength, removing the convolutional layers, or using only the time domain significantly degrades performance.

5. Discussion

The TF-CGRU framework effectively handles 496 categories with limited samples. Future work will address larger class sizes (3,000+ characters) using hierarchical classification or metric learning. We also plan to explore non-stationary modeling (e.g., wavelet transforms) and end-to-end segmentation for continuous writing.

6. Conclusions

This study presents a robust framework for 3D Chinese air-writing recognition. By fusing time and frequency domain features and employing motion-consistent data augmentation, we achieve high accuracy and stability in large-category, few-shot scenarios. The model’ s lightweight nature makes it ideal for wearable interaction systems.

Note: Figure translations are in progress. See original paper for figures.