Deep Convolutional Neural Network-Based Recognition of Air-Writing

Dhongde.V.S
nvvvvd@gmail.com

Saykar Punam Mahadev
saykarpunam1999@gmail.com

Vishwabharati Academy's College of Engineering

Abstract

Air-writing, the act of writing characters or words in the air using hand gestures, presents a novel and intuitive approach to human-computer interaction. As touchless interfaces gain popularity, the ability to accurately recognize air-written text has become increasingly important. This paper proposes a robust and efficient air-writing recognition system using Deep Convolutional Neural Networks (CNNs). The proposed model extracts spatial and temporal features from video sequences of hand movements and classifies them into corresponding characters or words. The system is evaluated using a custom dataset collected from multiple users under various conditions. Results demonstrate that the CNN-based model achieves high accuracy and generalizes well across different writing styles and environments, making it suitable for real-time applications in augmented reality, virtual keyboards, and assistive technologies.

Keywords: *Air-Writing Recognition, Deep Learning, Convolutional Neural Networks (CNN), Gesture Recognition, Human-Computer Interaction (HCI)*

---

1. Introduction

Air-writing recognition aims to interpret characters written in mid-air using hand or finger gestures, providing an alternative to conventional input methods such as keyboards or touchscreens. This technique has promising applications in ubiquitous computing, smart environments, and accessibility systems, especially for users with limited mobility or in situations where physical contact with devices is not feasible. Unlike handwriting recognition on surfaces, air-writing introduces additional challenges, including motion blur, variability in gesture speed, and the absence of tactile feedback, all of which complicate accurate character recognition.

Traditional approaches to gesture and handwriting recognition rely heavily on handcrafted features and rule-based systems, which often struggle with generalization and robustness. Recently, deep learning techniques, particularly Convolutional Neural Networks (CNNs), have shown remarkable success in various pattern recognition tasks, including image and video analysis. CNNs can automatically learn hierarchical feature representations from raw input data, significantly outperforming traditional feature extraction methods.

This paper explores the use of Deep CNNs for air-writing recognition by designing a model that takes sequential frames of hand movements as input and classifies them into predefined character or word categories. We also discuss the construction of a suitable dataset, the preprocessing steps involved, and the performance evaluation of the system under different conditions.

2. Related Work

Air-writing recognition has attracted significant research interest in recent years. Early methods primarily used depth sensors, accelerometers, or gyroscopes to capture motion trajectories and relied on techniques such as Hidden Markov Models (HMMs) or Dynamic Time Warping (DTW) for sequence classification. While these approaches were effective to some extent, they suffered from limitations in scalability and accuracy due to their reliance on handcrafted features and sensor-specific constraints.

With the rise of computer vision and deep learning, camera-based methods have become more popular. CNNs have been employed in recognizing static hand gestures and written characters on surfaces, but their application in dynamic air-writing remains relatively underexplored. Some recent studies have begun integrating CNNs with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to capture spatiotemporal dependencies in gesture sequences.

Our work builds upon these advances by focusing solely on the power of deep CNNs for feature extraction and classification, demonstrating that a well-designed CNN architecture alone can effectively handle the temporal complexity of air-writing sequences when combined with appropriate preprocessing techniques.

3. Methodology

The proposed air-writing recognition system consists of three main components: data acquisition, preprocessing, and the CNN-based classification model.

3.1 Data Acquisition

We collected video sequences of users writing alphanumeric characters in the air using their index finger. The dataset includes variations in writing speed, size, and orientation to simulate real-world usage. Each video was recorded using a standard RGB camera at 30 frames per second under diverse lighting and background conditions.

3.2 Preprocessing

The video frames were first converted to grayscale to reduce computational complexity. We applied background subtraction and hand segmentation using a skin color model to isolate the hand region. Optical flow was then used to extract the trajectory of the finger, and each sequence was resized and normalized to ensure uniform input to the neural network.

3.3 CNN Architecture

The CNN architecture comprises multiple convolutional layers with ReLU activations, followed by max-pooling layers to reduce spatial dimensions. Batch normalization is used to stabilize training, and dropout layers help prevent overfitting. The final layers include a global average pooling layer and a softmax classifier to output the probabilities of each character class. The network is trained using cross-entropy loss and optimized with the Adam optimizer.

4. Experimental Setup

The dataset consists of 36 classes (26 letters and 10 digits), with each class containing 100 samples. The data is split into training (70%), validation (15%), and testing (15%) sets. Training is conducted over 50 epochs with a batch size of 32. Data augmentation techniques such as random rotation and scaling are employed to improve generalization. The implementation is done in Python using TensorFlow and Keras libraries.

Evaluation metrics include accuracy, precision, recall, and F1-score. We also analyze confusion matrices to identify commonly misclassified characters and explore their underlying causes.

5. Data Augmentation Techniques

To improve the robustness and generalizability of the CNN model, several data augmentation strategies were applied to the training data. These techniques aimed to simulate real-world variations in writing style, hand motion, and camera positioning. The augmentation methods included:

- **Random Rotation:** Sequences were rotated by ±15 degrees to mimic slight wrist or arm rotations during writing.
- **Scaling:** Input frames were scaled up or down by a factor ranging from 0.8 to 1.2 to account for variations in writing size.
- **Translation:** Horizontal and vertical shifts were introduced to simulate off-centered air-writing.
- **Noise Injection:** Gaussian noise was added to some frames to simulate low-quality or noisy video input.
- **Frame Dropping:** Some frames were randomly removed from sequences to simulate motion blur or frame rate drops.

These augmentations increased the diversity of training samples, reducing overfitting and improving the model's performance on unseen data. Empirically, the use of data augmentation led to a 4–6% improvement in test accuracy, highlighting its critical role in training deep models for air-writing recognition.

6. Comparison with Other Models

To benchmark the performance of our CNN-based approach, we implemented and evaluated several baseline models, including:

- **Support Vector Machine (SVM):** Applied on handcrafted trajectory features extracted from optical flow.
- **Hidden Markov Model (HMM):** Used for sequence modeling based on fingertip coordinates over time.
- **CNN+LSTM Hybrid Model:** Combined CNN feature extraction with LSTM-based temporal modeling.
- **Transformer-Based Model:** Applied temporal self-attention over extracted frame-level features.

Model	Accuracy (%)	Inference Time (ms)
SVM with handcrafted features	78.3	25
HMM	74.9	30
CNN+LSTM	92.5	85
Transformer-Based Model	93.8	95
Proposed CNN Model	**94.2**	**48**

Our CNN-only model slightly outperformed the hybrid and transformer models in accuracy while offering significantly lower inference time, making it better suited for real-time deployment on resource-constrained devices.

7. Applications and Future Work

The proposed air-writing recognition system opens the door to a wide array of real-world applications. Some promising use cases include:

- **Virtual and Augmented Reality (VR/AR):** Enabling intuitive text input in immersive environments without the need for controllers or physical keyboards.
- **Smart Environments:** Facilitating gesture-based commands in homes, classrooms, or workplaces.

- **Assistive Technologies:** Providing communication tools for individuals with mobility impairments or speech disorders.
- **Public Interfaces:** Allowing hygienic, touchless input for public kiosks or ATMs, especially relevant in post-pandemic settings.

Future Work will focus on:

- Incorporating **depth sensing or 3D hand pose estimation** to improve recognition in cluttered scenes.
- Developing **user-independent models** that adapt to different writing styles through transfer learning or domain adaptation.
- Extending the system to recognize **entire words or sentences**, potentially using sequence-to-sequence architectures or attention mechanisms.
- Deploying the model on **embedded devices or mobile platforms** for real-world usability tests.

8. Results

The proposed CNN model achieved an overall test accuracy of **94.2%**, with individual character accuracies ranging from 88% to 98%. Characters with similar shapes, such as 'O' and 'Q', or '5' and 'S', showed higher misclassification rates, which aligns with findings in other handwriting recognition tasks. Precision and recall values averaged above 90%, indicating a strong balance between detection capability and reliability.

The model's inference time per sample was under 50 milliseconds on a standard GPU, confirming its potential for real-time applications.

9. Discussion

The results validate the effectiveness of deep CNNs in recognizing air-written characters without the need for recurrent architectures or sensor-based inputs. The use of only visual data ensures that the system is cost-effective and easier to deploy in everyday environments. However, the system still faces challenges with ambiguous characters and performance in low-light or cluttered backgrounds.

Future work could explore integrating temporal models like LSTMs or Transformer-based networks to further enhance sequence understanding. Expanding the dataset to include cursive and connected writing, as well as multi-lingual characters, will also be valuable for broader applicability.

10. Conclusion

This paper presented a Deep Convolutional Neural Network-based system for recognizing air-written characters using visual input. The model demonstrates high accuracy and efficiency, highlighting CNNs' capability to extract and interpret complex spatiotemporal features in gesture-based input. The proposed approach lays the groundwork for intuitive, touchless interaction in various technological domains, contributing to the advancement of natural user interfaces.

References:

1. A. K. Bhaga, G. Sudhamsu, S. Sharma, I. S. Abdulrahman, R. Nittala and U. D. Butkar, "Internet Traffic Dynamics in Wireless Sensor Networks," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1081-1087, doi: 10.1109/ICACITE57410.2023.10182866.

2. Butkar, M. U. D., Mane, D. P. S., Dr Kumar, P. K., Saxena, D. A., & Salunke, D. M. (2023). Modelling and Simulation of symmetric planar manipulator Using Hybrid Integrated Manufacturing. Computer Integrated Manufacturing Systems, 29(1), 464-476.
3. Butkar, U. D., & Gandhewar, N. (2022). Accident detection and alert system (current location) using global positioning system. Journal of Algebraic Statistics, 13(3), 241-245.
4. Vaishali Rajput. (2024). "Quantum Machine Learning Algorithms for Complex Optimization Problems". International Journal of Intelligent Systems and Applications in Engineering, 12(4), 2435 –. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/6663
5. D. S. Thosar, R. D. Thosar, P. B. Dhamdhere, S. B. Ananda, U. D. Butkar and D. S. Dabhade, "Optical Flow Self-Teaching in Multi-Frame with Full-Image Warping via Unsupervised Recurrent All-Pairs Field Transform," 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), Wardha, India, 2024, pp. 1-4, doi: 10.1109/IDICAIEI61867.2024.10842718.
6. P. B. Dhamdhere, D. S. Thosar, S. B. Ananda, R. D. Thosar, D. S. Dabhade and U. D. Butkar, "A Semantic Retrieval System Using Imager Histogram Computation to reduce Trademark infringement based on the conceptual similarity of text," 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), Wardha, India, 2024, pp. 1-6, doi: 10.1109/IDICAIEI61867.2024.10842701.
7. Butkar, M. U. D., & Waghmare, M. J. (2023). Crime Risk Forecasting using Cyber Security and Artificial Intelligent. Computer Integrated Manufacturing Systems, 29(2), 43-57.
8. Ahmed, S., Baig, M. S., & Kim, H. (2018). A Vision-Based Air-Writing Recognition System Using Convolutional Neural Network. In Applied Sciences, 8(11), 2170. https://doi.org/10.3390/app8112170
9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/
10. Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2013.6638947
11. Wang, W., & Popović, J. (2009). Real-Time Hand-Tracking with a Color Glove. In ACM Transactions on Graphics (TOG), 28(3), 1–8. https://doi.org/10.1145/1531326.1531387