A Depth Video Dataset and Neural Framework for Real-Time Handwriting Character Generation and Guidance

Saran S[1*], *Sathiesh* Kumar V[2] and *Jayabalan* J[3]

[1]*Student, Department of Electronics Engineering*, *Madras Institute of Technology Campus, Anna University*
Chennai, Tamil Nadu, India.
[2]*Assistant Professor(Sl. Gr.), Department of Electronics Engineering*, *Madras Institute of Technology Campus, Anna University*
Chennai, Tamil Nadu, India.
[3]*Research Scholar, Department of Electronics Engineering, Madras Institute of Technology Campus, Anna University*
Chennai, Tamil Nadu, India.
*Corresponding author:*indpsk2002@gmail.com*

Abstract. The acquisition of handwriting requires movement in space and time to be monitored. This research introduces a depth-sensing system which registers and recognizes the handwritten signs through the analysis of the 3D movement. By using the Intel RealSense D405 camera, we computed a new dataset to record fine fixed hand and pen trajectories for each of the 26 English alphabetic characters. Each recording was converted to sequences of standardized point-cloud data and annotated with information on action label: pen_down, pen_lift, hand_position. An LSTM network has been used to train this temporal 3-D sequences with an accuracy of 82%, precision of 0.78, a recall of 0.82 and an F1 score of 0.79. By spatiotemporally transcending 2D trajectories, this methodology captures spatial as well as kinematic cues, which features important subtle dynamics of handwriting, which can be useful for real time feedback, educational applications, and assistive technologies. Potential research will extend the data target and explore transformer-based models with the aim of making it better than the current performance words.

1. Introduction

Handwriting requires the skilled use of visuomotor complexity where ocular control and fine motor execution is differentially and temporally paired with aim. Children develop movements based on iterative movements with central nervous system commands that are sent to the peripheral action and are compared with the performed stroke and the discrepancy is assessed. Individuals with motor impairments, especially those that are a result of neuropathological conditions such as cerebral palsy, experience wrist instability, misplacement of finger and pen-jerking, forming distorted lettering, erratic strokes and writing speed.

Traditional methods of assessment process rely predominantly on real-time observation or video still recordings. Although these methods got gross kinematic patterns, they do not take into account the magnitude of the forces and could not account for variation in wrist positioning during the writing process. Digital tablet-based instrumentation can register the pressure and pen trajectory but is still limited to a 2D representation and thus an incomplete representation of the 3D dynamics inherent to handwriting. As a result, the insights derived are limited hindering comprehensive analysis of gestures, and the timely intervention from automation is not possible.

Recent development of the depth sensing technology allows for the reconstruction of the hand--pen interaction on a three--dimensional space, with good precision. But nonetheless, popular handwriting datasets are mostly preserving two-dimensional data obtained from tablet

interfaces and thereby missing a lot of depth data. Moreover, temporal annotations to mark pen-down, lift off or in between stroke transitions are only with rare exceptions. Contemporary handwriting recognition systems are further based on the assumption of simplified plane line patterns, which makes performance on complex, temporal distributed spatial sequences difficult.

The current research fills in these gaps through a collection of a rich set of 3D depth video data of hand-drawn capital English letters with a neural architecture that is temporal resolvable for motion pattern interpretation. Recordings, which had been made with a camera (Intel RealSense D405), were transformed into sequential point clouds using timestamps that marked important phases in the writing. Subsequent preprocessing included normalization, noise suppression and voxelisation for separation into volumetric grids, which were then divided into equally sized parts that could be used for frameworks in sequence learning.

The novelties offered by an end-to-end pipeline for depth sensing, point clouds processing, structured labeling and analyzing temporal sequences for handwriting actions recognition are the main contributions of this work. Our LSTM based implementation shows that depth cues could be used to reliably discriminate motion signatures about different characters to support applications such as handwriting evaluation, motor rehabilitation and enabling real-time feedback provision.

2. Literature Review

This part places the current research into the wider academic debate surrounding handwriting recognition, rehabilitative approaches and three-dimensional motion capture. It outlines important theoretical frameworks, reviews the antecedent investigations, and identifies areas of lacunae to be addressed in the current study.

2.1 Handwriting Learning, Motor Control, and Rehabilitation:

Federmeier and Federmeier [1] considered the interaction between cognition and motor activity in the writing process and the importance of feedback to the fingers in addition to visual-motor coordination. Their results suggest need for instruments which are able to reproduce fine locomotor patterns in letter construction.

Contreras- Vidal and Grossman [2] focused on neural dynamics involved in handwriting revealing us that real time error corrections helps recovery of orthographic skills after neurological injury. These results highlight the importance of having rigorous and empirical-based interventions in rehabilitation settings.

Smits,-Engelsman and Hill [3] evaluated handwriting problems of children with developmental coordination disorders, whom have deficits of proprioception and visuomotor integration that impair script production. Consequently, technologies that can be used to monitor the hand kinematics are justified for a detailed analysis.

Chen et al. [4] created an intelligent system that gives personalised feedback to improve handwriting and the learning machine was used to track the process. Their approach is a powerful way adaptive algorithms can complement therapeutic interventions.

Taborri et al. [5] reviewed digital rehabilitation systems which noted that instrument-centric systems with sophisticated sensing devices enhance fine motor performance. This change has resulted in a growing need for camera based computational techniques for making handwriting easy and tracking it.

2.2 Depth Sensing and 3D Hand Motion Capture:

Pisharady and Saerbeck [6] studied gesture recognition based on video streams and their RGB camera is confronted with the problem of occlusion and spatial ambiguity compared to depth sensing devices, for which a better performance can be claimed. Their observations suggest that RGB -D sensors capture the dynamics of the hand in a more stable way in writing tasks.

Liang, Yuan and Thalmann [7] discussed three-dimensional hand-poses reconstruction using depth information, showing that structured-light and time-of-flight cameras could be used to capture accurate articulations. Their results confirm the efficacy of depth technology in accurate motion capture in handwriting.

2.3 Deep Learning for Point-Cloud Representation:

Qi et. al. [8] introduced the PointNet, a neural architecture that operated on unordered point - cloud data for 3 - D shape recognition and part segmentation. This approach led to the choice of the spatial encoders for the three-dimensional handwriting signals.

Building upon this foundation, Qi et al. [9] expanded PointNet to PointNet++ by using hierarchical layers which capture fine and coarse level geometries. As a result the model performs better at noticing slight deviations in the movement trajectories between the hands.

Uy and Lee [10] had investigated deep learning methods for processing representation of point cloud data where they evaluated the progress in three-dimensional spatial recognition under the condition of dynamic. By combining structural clues with time cues, they ended up in better performances for temporally different scenes, which is consistent with the state of the art in handwriting analysis which highlights structure of the scene in addition to motion patterns.

Fan, Wang, and Sun [11] have used deep learning models to track spatio temporal trajectories to achieve the detection of 3d motion. Their methodology makes use of the ability of recurring neural networks to analyze a sequenced data, which in turn makes the analysis of the hand-drawn shapes as segmented blocks possible.

2.4 Temporal Modeling for Handwriting and Gesture Recognition:

Graves and Schmidhuber [12] came up with the idea of multidimensional recurrent neural networks for offline handwriting recognition, and they have shown that those models are good at obtaining temporal dependencies on stroke level. Their results are in favor of the application of LSTM architectures in the analysis of sequences of handwritten symbols.

Yang, Li, Zhang [13] designed a transformer-based approach for concurrent processing of gesture and handwriting, where the attention layers are used to follow the temporal pattern. While preliminary, this approach indicates some possible extensions to more general instances.

2.5 Handwriting Datasets, Annotation Frameworks, and Toolchains:

Liwicki and Bunke [14] published the IAM-OnDB corpus that is one of the first handwriting corpora with temporally annotated pen trajectories, but it is 2-dimensional planar motion without depth information. This limitation highlights the need for depth-enabled datasets.

Xie, Zhu and Chen [15] investigated RGB-D gesture datasets and found that there is a lack of annotations specifically related to writing such as stroke boundaries or touch timestamps. Consequently, the creation of a handwriting corpus which is carefully studied in depth and carefully labeled temporally becomes an imperative.

2.6 Summary of Literature Gaps:

To date, the salience of sensory cues, spatial cues and temporal dynamics in handwriting evaluation have been illustrated in literature. However, most available datasets contain no depth information which impedes a proper dissection of particular stages of strokes. Furthermore, there are few sophisticated models that combine three dimensional formulations of point clouds with mechanisms of motion tracks for the interpretation of hand written characters. In response, the current work presents a novel clip-based data set that contains depth views and an LSTM-based architecture processing spatial shape descriptors and temporal flow in parallel.

3. Methodology

This section outlines the processes used for data acquisition, point cloud processing, systematic annotation, construction of spatiotemporal sequences and the neural architecture developed for the detection of three-dimensional hand motions. The methodology focuses on consistent depth capture, labelling structure, careful preprocessing and the reproducibility of training methodology.

3.1 Research Question:

The study examines whether 3D hand movements captured over time can be converted into data streams that can be interpreted by neural networks, and thereby making the real-time recognition of capital letters as they are written possible. Instead of using static images, the method makes use of changing spatial points in successive frames and associates inter-frame displacement with different writing dynamics.

3.2 Block Diagram:

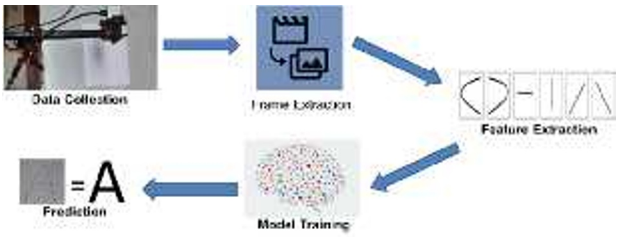

Fig. 1. Block Diagram

Figure 1 shows the block diagram representing the successive stages of the handwriting action - detection system: data acquisition, frame extraction, feature extraction, model training and result classification. First, a RealSense depth camera captures the hand-movements when writing (attached to a dedicated rig to record the hand movement in 3D). Each recording session is saved as .bag file in use of synchronised colour and depth video streams. These files are then processed into point cloud snapshots (.pcd files), one per temporal slice, to give accurate geometric and distance information from a variety of different locations.

From these point clouds are extracted elementary form of shapes (slants, uprights, flats and sweeps) as a symbolic substitute for letters, reducing noise while maintaining temporal and spatial fidelity.

The extracted shape trajectories are then fed to LSTM network that captures the way that the pen moves from one momentum to the next. Because the model can have a chronological dependence, the sequential unfolding of the character through the character's strokes can be detected. In the classification part, the trained network guesses the letter that was the best match on an incoming stroke trajectory. From the first capture of a 3D video to the temporal pattern tracking, the system is able to achieve accurate recognition of the handwritten symbols, making use of both the motion and depth cues, which makes it suitable for real-time feedback applications.

3.3 Data Acquisition and Dataset Construction:

3.3.1 Hardware Setup:

Handwriting motions were recorded by an Intel RealSense D405 RGB-B preserved in 1280 * 720 pixels and 30 fps. The sensor was placed directly above the writing surface and the following operational guidelines were followed for the entire recording sessions:

- Keep your hand between 15 and 25 centimeters from the camera.
- Lighting: Uniform ambient illumination to minimize IR interference.
- Flat surface, not much detail - keeps depth glitches minimal.

These conditions ensured stable 3-D point precision over each recording with regular quality checks being made in order to maintain reliability.

3.3.2 Recording Protocol:

Participants traced all 26 letters of the English alphabet as capital letters, which were either in the air or on a flat surface with a pen shaped tool. Each of the sessions was stored as a.bag file of synchronised colour and depth data in time. For one written letter there was one contiguous stretch of depth images, with variable length for each character.

3.4 Conversion to Point Cloud:

Every frame from the .bag file file was converted to .pcd format using Intel RealSense SDK. Depth pixels are represented in the three-dimensional space by linking the value of each pixel with the coordinates in space:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = d(u,v) \cdot K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (1)$$

where K is the intrinsic camera calibration matrix.

3.5 Annotation Procedure:

Annotations were made with Supervisely platform.

3.5.1 Annotation Policy:

Each shot was labeled based on predefined criteria:

- start_frame: a frame after which signs of hand movements such as initiation of the writing process can be detected.
- end_frame: a frame when the hand stops moving during the writing process.
- pen_down: a moment when the height of the pen from the surface of the writing medium is equal to or

below 1.5 millimeters for at least three consecutive frames.

- pen_lift: a moment when the pen lifts from the surface of the writing medium and the movement exceeds 1.5 millimeters
- hand bounding box: a 3D space that fits around the hand doing the writing.
- keystroke regions: hand-labeled moments showing sharp turns or shifts in path.

All labels were saved as the in json format with the link to the corresponding frame numbers.

3.5.2 Inter-Annotator Consistency:

Two trained annotators independently labelled 20% of the dataset. Inter-annotator reliability was computed using Cohen's Kappa, yielding:

$$\kappa=0.86$$

indicating strong agreement and validating the consistency of annotation policies.

3.6 Data Preprocessing and Feature Extraction:

3.6.1 Noise Filtering:

Depth noise was reduced using:

(a) Bilateral Spatial Filtering:

$$P_f(i) = \frac{1}{W_i} \sum_{j \in \mathcal{N}(i)} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_s^2} - \frac{\|P(i) - P(j)\|^2}{2\sigma_r^2} \right) P(j) \quad (2)$$

(b) Temporal Median Filtering:

$$P_t' = \mathrm{median}(P_{t-k}, \ldots, P_t, \ldots, P_{t+k}) \quad (3)$$

These filters reduce spatial jitter and abrupt depth fluctuations.

3.6.2 Normalization:

Each point cloud was normalized to fit within a unit cube for model stability:

$$P_n = \frac{P - \mu_P}{\sigma_P} \quad (4)$$

where μ_P and σ_P are the mean and standard deviation of the coordinates across all points.

3.6.3 Voxelization:

Each normalized cloud was voxelized into a 32×32×32 occupancy grid:

$$V(x,y,z) = \begin{cases} 1 & \text{if a point exists in voxel} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The grid was then flattened into a vector for LSTM ingestion.

3.7 Temporal Sequence Construction:

Handwriting recordings differ in length; therefore, fixed-length sequences were constructed using the following procedure:

3.7.1 Maximum Sequence Length N:

A target length of N = 64 frames was selected based on the median writing duration across participants.

3.7.2 Handling Short Sequences:

Sequences with T < N were zero-padded and masked:

$$P_t = 0 \quad \text{for } t > T \quad (6)$$

A masking layer in the LSTM ensured padded frames were ignored.

3.7.3 Handling Long Sequences:

For sequences with T > N, frames were uniformly downsampled:

$$t_i = \left\lfloor \frac{i \cdot T}{N} \right\rfloor, \quad i = 1, \ldots, N \quad (7)$$

This preserves temporal proportionality while standardizing sequence length.

3.8 Neural Network Architecture:

An LSTM-based spatio-temporal model was selected due to its ability to capture sequential dependencies in handwriting motion.

3.8.1 Architecture Overview:

A Long Short-Term Memory (LSTM) network was selected to model temporal dependencies between successive 3D frames. The architecture consists of:

Table 1. Model Design

Layer	Description
Input Layer	3D feature vector (voxelized point cloud)
LSTM Layer 1	128 hidden units
LSTM Layer 2	64 hidden units
Dense Layer	Fully connected, ReLU activation
Output Layer	Softmax classifier over 26 alphabet classes

Table 1 shows the operation of the new LSTM model in the detection of 3D handwritten letters. Rather than working with images, the model uses 3D data that has been generated using point clouds, discretised into voxels. These voxelised vectors are recursively fed through two layers of LSTM networks (first consisting of 128, second, of 64 hidden neurons) that learn feature changes of motion over time. Following the recurret stages a fully connected layer with a ReLU activation function is used to compute unique signal patterns and a Softmax layer is employed to obtain the discrete class label ranging from A - Z based on the entire sequence. This layered architecture, in turn, allows the model to maintain fine details of shape that coexist with dynamic details of time in handwriting.

3.7.2 LSTM Formulation:

The LSTM update equations are as follows:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (8)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (9)$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (10)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (11)$$
$$h_t = \tanh(c_t) \quad (12)$$

where x_t is the input at time t, h_t the hidden state, and σ the sigmoid activation.

3.9 Training and Evaluation:

3.9.1 Dataset Split:

The dataset was divided into 70% training, 15% validation, and 15% testing sets, stratified across all 26 characters.

3.9.2 Loss Function:

Categorical cross-entropy was used:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{13}$$

Where y_i and $\hat{y}_i$ are the true and predicted probabilities for class C and $C = 26$.
Model performance was evaluated using accuracy, precision, recall, and F1-score.

3.9.3 Implementation Details:

The system was implemented in Python 3.10 using:

- Framework: PyTorch
- Batch size: 32
- Learning rate: 1×10^{-3}
- Epochs: 100
- Point-cloud operations: Open3D
- Annotation management: Supervisely
- Depth extraction: Intel RealSense SDK

3.10 Replicability:

All preprocessing scripts, along with model setups and data layouts, follow a split design so everything can be copied exactly. To get the same outcomes, people just run that system on fresh RealSense .bag recordings.

4. Results and Discussions

This section examines how the novel approach of the LSTM-based algorithm performs on the synthesised 3D handwritings dataset by taking into account the precision, performance of the group-wise approaches, as well as the effect of the temporal depth cues on the recognition of the gesture.

4.1 Network Architecture Summary:

The last configuration includes another masking part to take care of padded sequence entries, followed by an LSTM layer with 256 units to take care of temporal dependencies. Regularization is used in the form of dropout layers which are scattered throughout the architecture. A next dense layer containing 128 neurons uses ReLU activation function. Rather than ending there, the representations produced by the network are passed on to a softmax classifier (that is, a classifier which assigns the input to one of n classes, in this case 26). In total, there are about 305,433 learnable parameters in the model, which gives it enough capacity to train quickly, but is lightweight enough to be potentially used for real-time applications.

4.2 Classification Performance:

The model gave an accuracy of 82 % on the held out test set suggesting temporal depth information provides useful discriminative information for handwritten character recognition. The main performance measures which can be derived from the classification report are listed below:

Table 2. Classification report

Metric	Value
Accuracy	0.82
Precision	0.78
Recall	0.82
F1-score	0.79

The results presented in the Table - 2 show that the LSTM handwriting model is consistent across the groups of utterances and the values of precision closely match the recall and they provide a reliable identification of the characters without spurious predictions.

4.3 Confusion Matrix Analysis:

The confusion matrix shows a high degree of concentration of the values along the diagonal line which indicates that the predictions are mostly consistent with the ground truth labels. Some of the characters have especially strong diagonal dominance such as C, D, G and P, which are due to the different visual characteristic and the fact that they have curved features making them easier to remember.
Misclassifications mostly occur between characters with a similar three-dimensional motion trajectories. For example, letters I, J, L and T have linear vertical or horizontal stroke patterns, while K or R have crossing lines of movement.
Individual handwriting styles and the magnitude of a stroke cause temporal inconsistencies which contributes to these is another cause of these errors. These observations lead one to consider that increasing temporal resolution by adding more spatial descriptors, for example, by adding curvature, speed, or angular variation, might lead to better class separability.

4.4 Strengths observed:

The LSTM architecture is good at grabbing temporal patterns in depth images, thus aiding the recognition of stroke dynamics. Even at only providing only 5 examples per character, the model shows good performance which emphasizes the consistency of 3D stroke formation across samples. The parameter count of around 1.2 are projected makes it possible to deploy it on resource limited devices, such as embedded devices or applications with real-time requirements. Using 3D point cloud data provides the model with depth characteristics not available with traditional 2D handwriting data, which can be used to add more discriminatory power.

4.5 Limitations identified:

- ***Dataset scale:*** The insize of the dataset impacts the model's ability to generalize to other styles of writing, hence a decrease in performance for those who have not seen before. The lack of variety of the input leads to a poor adaptability which is very apparent when the model is applied on a larger scale.
- ***Class imbalance and motion similarity:*** Some of the characters do not have unique depth signature and they become visually indistinguishable for the similar strokes.
- ***Spatial noise in point clouds:*** here are small deviations in point clouds caused by the depth sensor that may create static that may interfere with the stability of the 3D pixel characteristics.
- ***Limited feature representation:*** The current approach uses only occupancy grid as input without taking into consideration the shape features, such as bend, motion change or pattern change. Consequently, in this analysis basic spatial occupancy and not finer descriptors of movement and surface details are focused on.

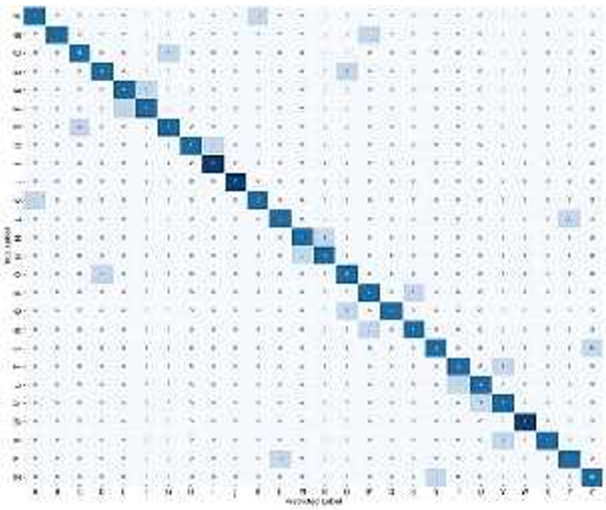

Fig. 2. Confusion Matrix

Figure 2 shows the confusion matrix for the model that is used to determine the handwritten letter gestures. Covering all 26 upper case letters of english language, each row represents true label and each column denotes predict label. Five video samples per class based on depth were used. Right answers fill the lead diagonal; variances show off to the off-diagonal cells. The concentration of correct predictions on the diagonal is related to consistency and would imply that this system makes consistent discrimination between most characters with little confusion. Occasionally, the visually similar letters or the ones having similar 3D motion are misclassified, mainly because of the overlapping of trajectories. The overall classification accuracy of 82% shows that by combining LSTM to learn the movement over time is able to provide a robust handwriting recognition from depth video.

4.6 Implications for Handwriting Recognition and Rehabilitation:

The results indicate that 3D depth signals boost the performance in handwriting detection since they hold information of wrist orientation, height of the hand and distance between pen and paper that are lost in traditional 2D inputs. These benefits have a number of promising applications:

Rehabilitation systems: Monitor the consistency of strokes, the height of each mark, and the timing stability all through automated analysis.

Educational environments: It is possible that providing immediate corrective feedback would enhance the ability of the learners to write by providing feedback regarding both the spatial accuracy and the movement dynamics.

***Human–computer interaction*:** Wireless input of text via hand motions is gathered using 3D sensors and has potential for immersive simulations and assistive technologies.

4.7 Future Works:

In order to improve the model performance over new data sets and to increase the system capabilities, future studies will explore:

- Using additional contextual factors in combination with various handwriting samples to increase the data variability.
- Shining light on motion with additional features: motion descriptors are typically simple, i.e., they only describe the position and the time of events. Motion can be described with more complex characteristics, e.g. velocity, curvature, and steadiness.
- Evaluating contemporary spatio - temporal architectures - like Transformers or Graph Neural Networks - to better represent extended temporal dependencies.
- Increasing integration of the detection framework with real-time feedback mechanisms to improve the praxis training and motor recovery for learning through the instantaneous correction to errors.

5. Conclusion

This research built a full depth-driven method for identifying 3D handwriting actions, combining data collection, organized labeling, point cloud cleaning, along with time-sequence analysis. Thanks to the Intel RealSense D405 camera, precise hand and pen paths were recorded then turned into aligned point clouds tagged with major writing stages like pen_down, pen_lift, start_frame, end_frame. A uniform processing chain - made up of noise removal, scaling, voxel grouping, plus sequence standardization - ensured stable spatial and temporal expression of movement.

The LSTM model identified temporal patterns in 3D voxel inputs, achieving 82% accuracy; precision, recall, and F1-score averaged 0.78, 0.82, and 0.79 respectively. Results suggest that depth-sensitive features improve gesture detection compared to conventional 2D approaches - particularly during monitoring of hand elevation, motion changes, or wrist orientation.

Even with limited data and personal writing quirks, results show that depth sensors along with loop-based networks can pick up tiny hand motions. Next steps include gathering more examples, adding different kinds of handwriting, followed by testing newer setups like transformer models or space-time graphs. This setup gives instant feedback on writing, supports recovery through guided tasks, while boosting digital tools that respond in 3D.

References

1. M. L. Federmeier, and K. D. Federmeier, "The neuroscience of handwriting: motor learning and cognitive control in written language," Frontiers in Psychology, vol. 14, pp. 1–15, 2023.
2. L. Contreras-Vidal and M. Grossman, "Neural control of handwriting: lessons for rehabilitation and brain–computer interface applications," Neuroscience & Biobehavioral Reviews, vol. 146, pp. 104–120, 2023.
3. H. Smits-Engelsman, and E. Hill, "Handwriting difficulties in children with developmental coordination disorder: A systematic review," Research in Developmental Disabilities, vol. 95, pp. 103530–103542, 2019.
4. J. F. Chen, S. Y. Lin, and W. C. Hsu, "Artificial intelligence-based occupational therapy in handwriting rehabilitation," Journal of Hand Therapy, vol. 35, no. 4, pp. 589–599, 2022.
5. K. T. Taborri, E. Palermo, and S. Rossi, "Digital rehabilitation systems for handwriting and fine motor control: a review of AI-based solutions," Sensors, vol. 21, no. 19, pp. 6481–6498, 2021.
6. P. K. Pisharady and M. Saerbeck, "Recent methods and databases in vision-based hand gesture recognition: a review," Computer Vision and Image Understanding, vol. 155, pp. 152–165, 2017.
7. H. Liang, J. Yuan, and D. Thalmann, "3D hand pose estimation using depth data: a review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 682–701, 2022.
8. KC. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1939–1951, 2019.
9. C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 12, pp. 4338–4354, 2021.
10. M. A. Uy and G. H. Lee, "Point-cloud deep learning for 3D shape understanding: A survey," Computers & Graphics, vol. 104, pp. 72–88, 2022.
11. H. Fan, X. Wang, and J. Sun, "Deep spatio-temporal networks for 3D human action recognition," IEEE Transactions on Image Processing, vol. 32, pp. 1904–1917, 2023.
12. A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," Neural Networks, vol. 21, no. 9, pp. 1323–1333, 2008.
13. Z. Yang, Q. Li, and J. Zhang, "Transformer-based spatio-temporal modeling for dynamic gesture and handwriting recognition," Pattern Recognition Letters, vol. 165, pp. 165–174, 2023.
14. Liwicki and H. Bunke, "IAM-OnDB: An on-line English sentence database acquired from handwritten text on a whiteboard," Pattern Recognition Letters, vol. 31, no. 19, pp. 2305–2312, 2010.
15. S. Xie, L. Zhu, and C. Chen, "RGB-D datasets for gesture and hand action recognition: A comprehensive review," Information Fusion, vol. 91, pp. 252–275, 2023.