Pertanika J. Sci. & Technol. 33 (S4): 99 - 114 (2025)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Evaluation of Multimodal Dataset for Continuous Air-writing Multidigit Number Recognition Using Wearable Sensor

Yusuke Takei[1] and Eiichi Inohira[2]*

[1]*Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan*
[2]*Faculty of Engineering, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan*

ABSTRACT

In recent years, there has been a significant increase in the research and development of technologies and products, such as wearable devices. Gesture input has emerged as a promising new input method well-suited to these technologies. However, developing a reliable method for recognizing continuous gestures that lack contextual relationships poses a significant challenge. This study introduces a novel approach for recognizing continuous multidigit numbers to address this issue. This method utilizes a deep learning model equipped with sample-level dense labeling and leverages a multimodal dataset comprising inertial measurement unit data from wearable sensors and camera data. The outcomes of our recognition experiment reveal that using a multimodal dataset to produce accurate training data enhances recognition accuracy by 13% compared to approaches that do not use a multimodal dataset. Additionally, using our two proposed methods, the recognition of continuous digit gestures comprising 5, 8, and 10 digits achieved a correct recognition rate exceeding 90%. These results underscore the efficacy of our proposed method in recognizing continuous air-writing character gestures.

Keywords: Air writing, gesture recognition, multimodal dataset, wearable sensor

ARTICLE INFO

Article history:
Received: 01 April 2024
Accepted: 10 April 2025
Published: 10 June 2025

DOI: https://doi.org/10.47836/pjst.33.S4.06

E-mail address:
takei.yusuke468@outlook.jp (Yusuke Takei)
inohira.eiichi402@mail.kyutech.jp (Eiichi Inohira)
*Corresponding author

INTRODUCTION

The advent of wearable technology has revolutionized how we interact with digital devices, paving the way for innovative methods of communication. Zhang et al. (2018) have applied a gesture input technique to off-the-shelf smartwatches. Schäfer et al. (2022) have proposed a gesture-based UI for controlling continuous

e-ISSN: 2231-8526

locomotion in Virtual Reality applications. Kim et al. (2023) have presented a gesture-based control system for an industrial manipulator with a robotic hand. Among these advancements, the development of air-writing gesture recognition stands out as a significant leap, offering a seamless and intuitive input mode for a wide range of applications. We have focused on wearable devices because they have no problem with occlusions with cameras and poor illumination. Numerous studies have explored the recognition of air-writing gestures using a variety of wearable devices. Tripathi et al. (2022) introduced a high-precision technique for recognizing English air-writing gestures, employing wristband-type inertial measurement unit (IMU) sensors. This method transforms sensor data into an image form for enhanced recognition accuracy. Zhang et al. (2022) have developed a wearable system attached to a finger and a CNN (convolutional neural network)-based classification model for digits and letters. Kim et al. (2014) tackled the challenge of limited hardware resources in air-writing gesture scenarios. They developed a computationally efficient and cost-effective approach using a stepwise lower-bound dynamic time warping algorithm. Wu et al. (2009) have used a Wiimote, which is the controller of the Nintendo Wii equipped with a 3-axis accelerometer and implements a gesture recognition system, including the air-writing gesture of Lu et al. (2014), and proposed a gesture recognition system using a mobile phone for digits. Dash et al. (2017) have proposed a gesture recognition system using a Myo armband and a model with CNN and GRU (gated recurrent unit). This system can recognize a single-digit number. Yin et al. (2019) focused on improving recognition accuracy in user-independent scenarios by converting sensor data into gesture contours. However, these studies have predominantly focused on recognizing gestures as isolated events. A significant issue in continuous gesture recognition is the need for pauses between gestures, significantly reducing input speed and degrading user experience. Studies have also investigated the recognition of gestures performed in a continuous sequence without pauses.

Lin et al. (2018) devised a technique for recognizing air writing gestures for text input on small-screen smartwatches, where conventional touch input is impractical. In their study, gestures were not performed continuously; short pauses were introduced between gestures to aid word recognition. Additionally, they improved tolerance for ambiguity in air-writing gestures by offering suggestions for four words based on input prefixes, utilizing a language database, and implementing automatic word completion. Amma et al. (2014) proposed applying methods used in gesture recognition to speech recognition. In this approach, the entire dataset of consecutively performed gestures is processed as a single input, rather than recognizing each gesture individually for character recognition. This method employs a hidden Markov model (HMM) decoding and language modeling to enable the recognition of words and sentences by interpreting a continuous stream of gesture data as input. Zhang et al. (2021) introduced an innovative end-to-end word-level recognition methodology.

This approach eschews the segmentation of each character and instead employs a deep learning framework, utilizing a connectionist temporal classification (CTC) decoder for streaming character output. They integrated an additional decoder with an attention layer within a CNN architecture to further refine character accuracy. This configuration ensures that a more precise word recognition result is achieved upon completion of character processing. Additionally, they incorporated a language model to enhance overall accuracy. Chen et al. (2021) developed ViFin, a novel approach for small-motion finger handwriting gesture recognition. This technique captures vibrations generated during air-drawn letters using a smartwatch with a finger and uses CTC and spell checking for continuous letter recognition. These studies prioritize words and sentences as their recognition targets and utilize the contextual dependency of letters and words. They employ advanced language models such as transformers and propose recognition methods that utilize dictionary-based interpolation. However, these methods may exhibit limited efficacy for newly coined words or for numbers and symbols that lack contextual dependencies.

In human activity recognition, segmentation of time-series data is important to identify the exact boundaries of an activity. Yao et al. (2018) have proposed dense labelling, which is a method for predicting the classification label of each sample (i.e., timestep). They pointed out that the conventional Sliding Window technique had problems with the best window length, sampling stride (window overlapping), and windows' labeling strategy. The windows lapping causes a multi-class window problem. Dense labelling avoids this problem because it has no windows and generates the label sequence of the same length as the input data. Zhang et al. (2019) have proposed a U-Net-based human activity recognition method.

We introduce a deep learning model with sample-level dense labeling, focusing on individual character recognition and accurately identifying numbers independent of preceding or subsequent characters. In this approach, the meticulous creation of precise sample-level labels is paramount for effectively training deep learning models with sample-level dense labeling. However, generating accurate labels from solely IMU data during gesture performance presents a challenge, owing to the inherent difficulty in discerning the exact duration of the gesture using only the IMU data without any cue data. In the previous work, the arm was intentionally immobilized after each gesture, exclusively during training data collection. This strategy facilitated the precise determination of gesture duration solely from IMU data. However, this study diverges from previous methodologies, such as the Sliding Window technique, since it directly inputs the entire dataset into the deep learning model for recognition. Consequently, sensor data obtained during the transitional phase between gestures could potentially influence the training of the deep learning model. To counteract this and improve the accuracy of the deep learning models, we advocate for creating precise labels using a multimodal dataset, consisting of IMU data from wearable sensors and camera data. After training the deep learning model, only the IMU data is

used for continuous air-writing multi-digit recognition. Then, our proposed approach is free from occlusions with cameras and poor illumination.

METHODS

Continuous Air-writing Multi-digit Recognition System

Our approach employed a deep learning model for continuously recognizing air-written multidigit numbers using wearable sensors, as depicted in Figure 1. The model used, U-Net (Ronneberger et al., 2015), is renowned for semantic segmentation and has demonstrated high accuracy in time-series data recognition. The configuration of the deep-learning mode allowed it to output two elements for each sample: a recognition label and a sine wave, as illustrated in Figure 2. The sine wave served to ascertain the number and duration of performed gestures.

The rationale for producing both sine wave and recognition labels is elucidated through specific examples. Consider the scenario where a user airwrites the two-digit number 55 without any temporal gap between the fives in the tens and units places. If the sensor's sampling frequency is not exceptionally high, the data representing the transitional period

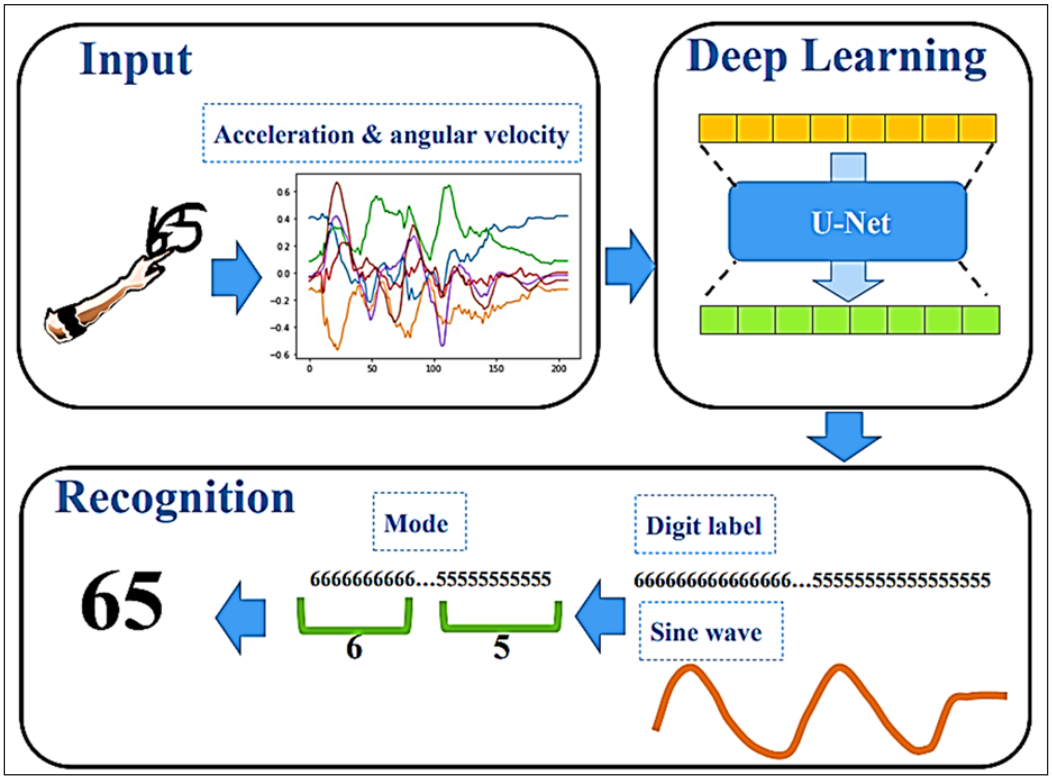

Figure 1. Overview of the continuous air-writing multidigit recognition system utilized in this study

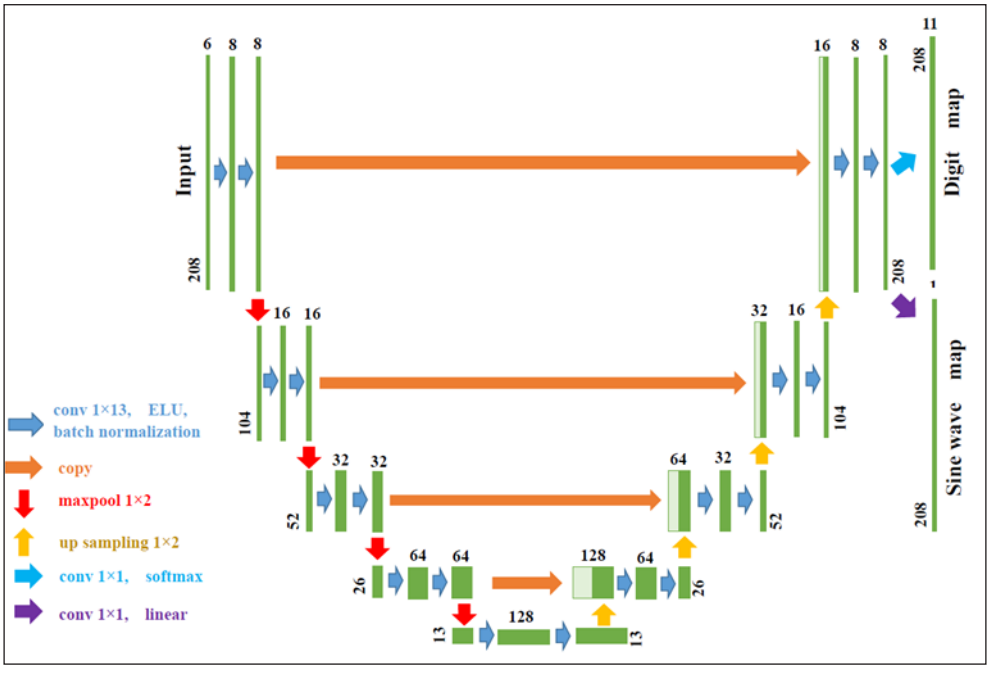

Figure 2. U-Net architecture was employed in this study

might be minimal, potentially as scarce as a single sample. In such instances, the output recognition labels from the deep learning model might only indicate a series of fives, represented as [5,5,5,5,5,...]. This output alone makes it challenging to determine whether the air-written number is a single five or a larger digit like 55 or 555.

To enhance the performance of the deep learning model, it was trained to produce one sine wave for each character. This configuration enables the recognition of any number of digits in a continuous sequence without restricting the count of consecutive digit inputs. The model generates two outputs: recognition labels and sine waves, both of which are per-sample outputs matching the temporal length of the input data. The recognition label is an 11-dimensional output, encompassing digits 0 to 9 and an additional value, 10, to signify non-gesture periods. The sine wave is trained to output values ranging from -1 to 1 for each sample. Therefore, obtaining precise data for each sample is crucial.

Experimental Setup

A multimodal dataset with accurate labels was constructed by integrating IMU and camera data. The process commenced with collecting video data using a webcam on a personal computer during the gesture performance, in conjunction with gathering IMU data from a wearable sensor on the arm. This process is depicted in Figure 3. We used a webcam on a laptop PC MSI GF63-11UC. The obtained video data has a resolution of 720 × 480

and a frame rate of 30 Hz. A participant wore a Myo armband on the forearm near the elbow and performed air-writing gestures in front of the webcam. We collected the IMU data using the Myo armband. The IMU data have three-dimensional acceleration sensor data, gyroscope sensor data, and a sampling rate of 50 Hz. A participant performed the designated air-writing gesture after a sound cue.

We recorded the IMU and camera data simultaneously for the predefined duration tailored to the number of digits, as shown in Table 1. The duration was so short that a pause during air-writing gestures was not allowed. The data for each gesture was recorded in separate files. As illustrated in Figure 4, Step 1 involved the utilization of Mediapipe, an open-source machine learning library. Mediapipe facilitates the implementation of Google's image recognition technology to detect faces, poses, and fingers in video recordings of air-written text gestures. In this study, the air-writing gestures were executed using only the index finger, allowing for determining the index fingertip's position coordinates from the identified landmarks. We obtained the index fingertip's position coordinates from each raw video data image via Mediapipe. The time series of these coordinates was obtained to trace the index finger's trajectory. In Step 2, the duration of each gesture in the video was ascertained from this trajectory, leading to

Table 1
Measurement time of multidigit numbers

Digits	**2**	**5**	**8**	**10**
Measurement time [s]	4	8	13	15

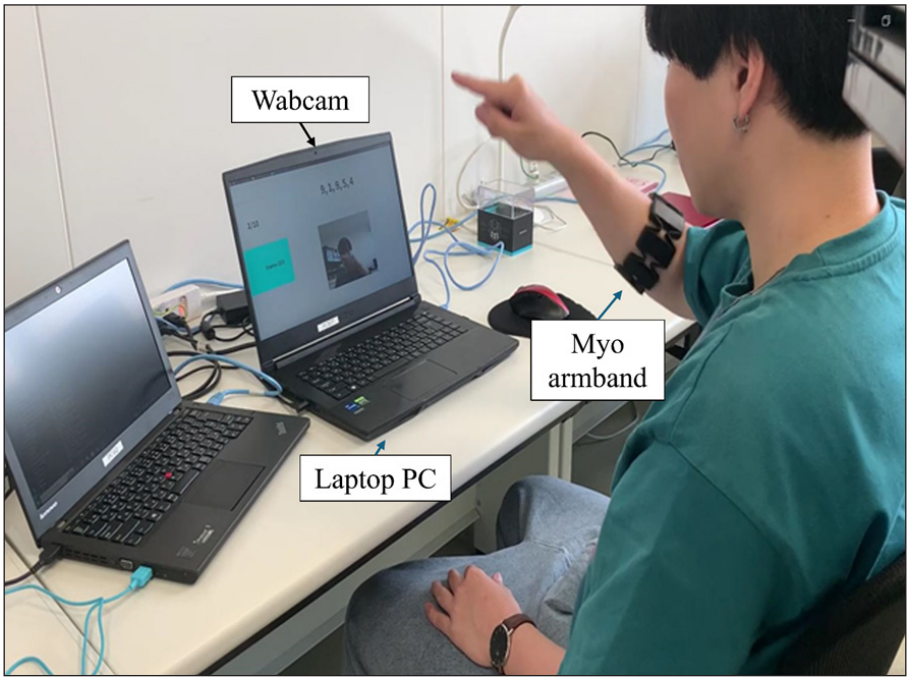

Figure 3. Experimental setup for data collection

the identification of the start and end times for each gesture. In Step 3, the corresponding start and end times in the sensor data for each gesture captured in the video were identified, enabling the creation of precise sample-level dense labels.

Datasets

To assess the multimodal dataset's effectiveness, an "Only-IMU dataset" was constructed, where only IMU data was collected using the wearable sensor. Labeling this dataset involved manual identification of corresponding waveform segments in the IMU data for each gesture. As shown in Step 3 of Figure 4, it is difficult to identify the start and end time of each continuous gesture because the motions before writing the first digit and between the first and second digits were recorded together, and the boundaries between the motions were unclear on the IMU data without camera data. Therefore, the Only-IMU dataset has an inaccuracy in the start and end times of the sample-level dense labels. In the Only-IMU dataset, the air-writing gestures of single-digit and continuous two-digit numbers were recorded. We collected data on the air-writing gesture of the 100 two-digit numbers ranging from 00 to 99 in a randomized order. This resulted in creating an Only-IMU dataset from 8 participants (8 male, 8 right-handed, 22 to 23 years old) and a multimodal dataset from 2 participants (2 male, 2 right-handed, 22 to 24 years old). These datasets were then utilized to train a user-dependent deep neural network model. The model's accuracy was evaluated using a 4-fold cross-validation approach. Each participant executed 100 trials comprising two types of air-writing gestures, forming 150 training and 50 test datasets per participant.

Moreover, two methodologies were employed to validate the effectiveness of recognizing continuous digit sequences in air-writing text gestures: one using a deep learning model with sample-level dense labeling, and another improving model accuracy by generating precise labels with a multimodal dataset combining IMU data and camera data. Recognition experiments involved collecting 5-digit, 8-digit, and 10-digit number gestures from two participants. The experiment included 50 trials for 5-digit numbers, 45 for 8-digit numbers, and 40 for 10-digit numbers, with 20 trials from each set designated as test data. As shown in Table 2, the multi-digit numbers were chosen as the appearance frequencies of each digit were equal. Consequently, 75 and 60 trials were allocated for training and testing, respectively.

Additionally, we used another multimodal dataset from 4 participants (2 males/1 female, 4 right-handed, 22 to 24 years old, 3 days) to evaluate the robustness of the recognition for another participant. Two participants of this dataset were in common with the previous multimodal dataset. Their data collection was conducted after several weeks.

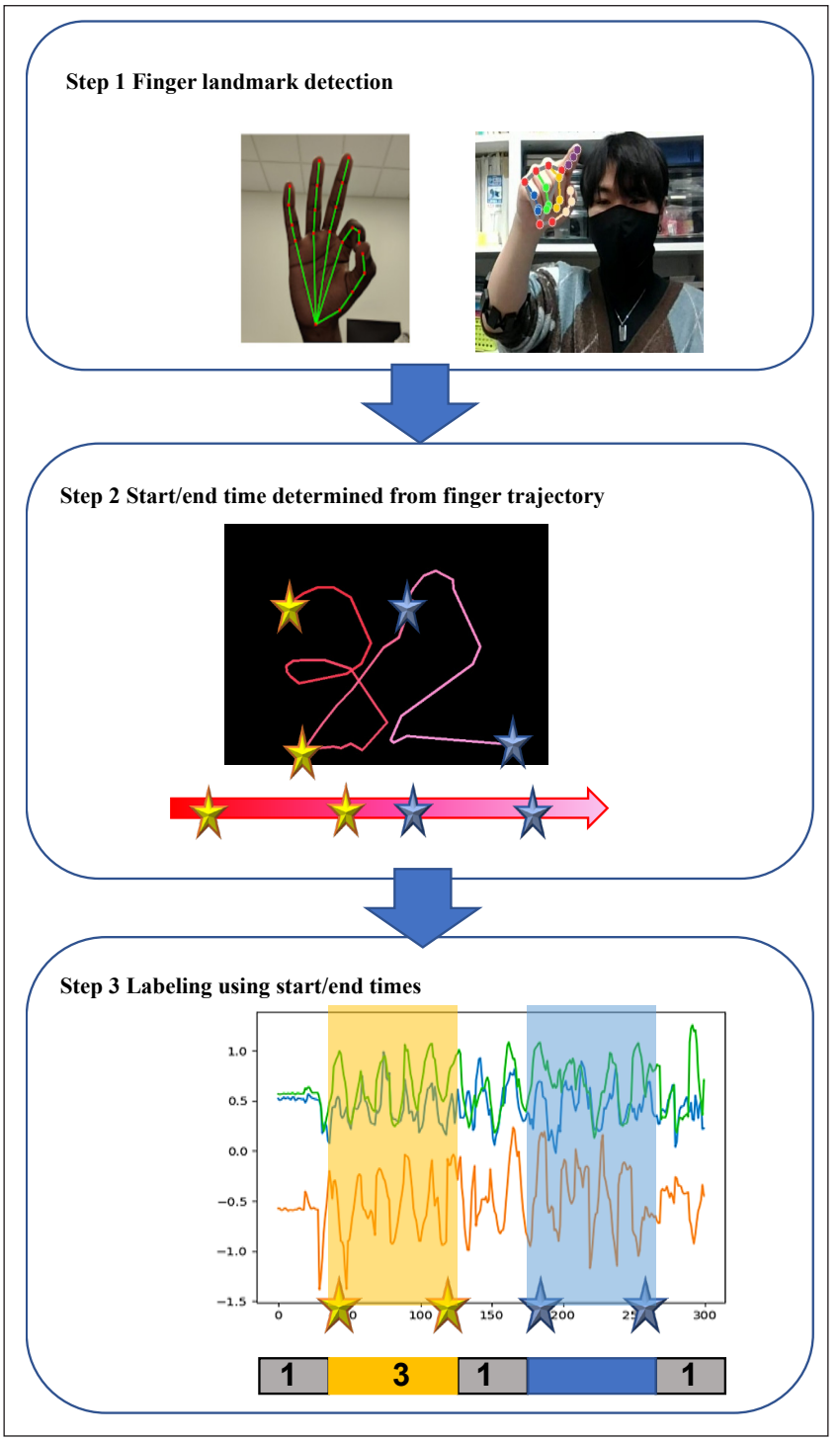

Figure 4. Process of time series data labeling using camera data

Table 2
The multidigit numbers for training and testing

Digits	Training	Test
2	None	23, 41, 80, 96, 98, 25, 12, 19, 10, 20, 30, 45, 54, 77, 66, 33, 69, 88, 75, 47
5	56053, 03783, 52234, 82673, 42638, 02683, 90472, 96916, 23584, 86313, 83775, 81943, 06085, 12721, 90291, 90072, 05710, 81922, 65407, 44156, 86898, 51056, 69472, 94177, 97407,	08391, 63777, 61493, 63836, 95357, 09376, 55061, 82477, 80858, 98761, 54872, 82105, 09631, 90382, 15535, 70161, 04240, 42122, 42699, 44924,
8	54050611, 79161824, 78205757, 09033936, 98378568, 38254293, 07109104, 93891258, 85801886, 60666899, 66532268, 52842347, 56707414, 86926224, 06341128, 83491202, 78851040, 44715470, 69990556, 00766734, 09229570, 65429370, 25254451, 31413553, 41391777, 91337311,	84714903, 63716832, 52518516, 96660113, 88159682, 57819026, 37346495, 06574521, 08142977, 00993294, 93328847, 78332134, 22403008, 82974858, 22234867, 64796720, 33716979, 54605015, 15579940,
10	2546030623, 8999541525, 2491348039, 0172482836, 2922403988, 9396843658, 0207013197, 8645101973, 9222887954, 0810865143, 5073511680, 5382770921, 7682468955, 6850667241, 7319567533, 0942800351, 9460773561, 4967727780, 2455417469, 1713464316,	2052286307, 5219346410, 6013820328, 7372505753, 2345757747, 8536225349, 5099736910, 3710321047, 4241288622, 2671683185, 9014071205, 1746011720, 0901145029, 0131256438, 8665665118, 4064895399, 5564398438, 7768833445, 8897798799, 4699696894,

RESULTS AND DISCUSSION

Utilizing 10 datasets—eight from the Only-IMU group and two from the multimodal group—the training and evaluation process of the user-dependent deep neural network model was repeated 15 times, employing 4-fold cross-validation. This repetition was crucial to ascertain the true recognition accuracy, considering the initial parameter values of the deep neural network model and the potential variations in final parameters due to the training process. Consequently, accuracy was determined by averaging the correct answer percentages from the test dataset across 15 iterations.

The accuracy evaluation in this study is not based on the percentage of correctly identified individual digits within each number. Instead, it is determined by the accuracy of the entire number; a two-digit number is considered correctly recognized only if both digits precisely match the correct answer. Conversely, any discrepancy in even one digit results in the entire number being classified as incorrect.

Recognition of Two-digit Numbers

As depicted in Figure 5, the average percentages of correct responses using the Only-IMU dataset from the eight participants were as follows: 80.6%, 86.1%, 84.1%, 75.9%, 84.1%, 76.6%, 80.9%, and 75.7%. In contrast, when utilizing the multimodal dataset from two participants, the average percentages of correct responses were significantly higher at 94.3% and 92.9%. This represents an approximate 13% improvement in the average percentage of correct answers for the multimodal datasets compared to the Only-IMU datasets.

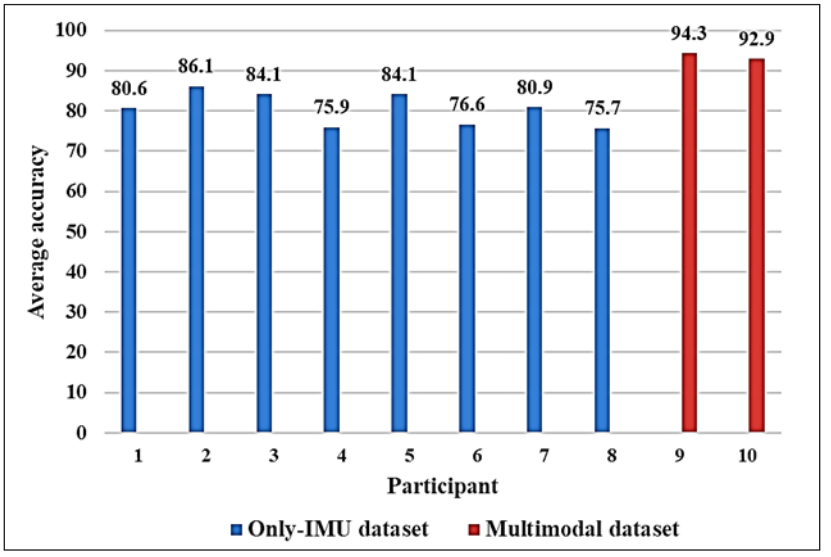

Figure 5. Comparative analysis of average accuracy: Only-IMU dataset vs. multimodal dataset

Additionally, the study examined confusion matrices of character recognition. The matrices, presented in Figures 6 and 7, indicated a notable reduction in misidentifications among numbers with similar gestures, such as 0, 6, 2, and 3. Figure 6 illustrates the results obtained using the Only-IMU dataset. The air-writing gesture for 0 was misrecognized as 6, a total of 63 times in the Only-IMU dataset. In contrast, as shown in Figure 7, the same misrecognition occurred only once when using the multimodal dataset. The misrecognition of 0, 2, and 3 also decreased similarly due to the multimodal dataset. This improvement is attributed to the more precise label data creation in the multimodal dataset. In the Only-IMU dataset, correct data was derived by manually identifying similar waveform segments in the IMU data for each gesture. However, this method proved less accurate, particularly for gestures with shared characteristics. In cases like 0, 6, 2, and 3, the IMU data waveforms were predominantly similar, with only minor differences at the beginning or end of the waveforms. These subtle differences, often represented by only a few samples, were occasionally overlooked in the labeling process, leading to the same waveform being

assigned different numerical labels. Consequently, the deep learning model was trained to map identical waveforms to diverse labels, potentially confusing the training phase. These findings underscore the effectiveness of the method that enhances the accuracy of deep-learning models by generating precise labels through a multimodal dataset, which integrates IMU data from wearable sensors and camera data.

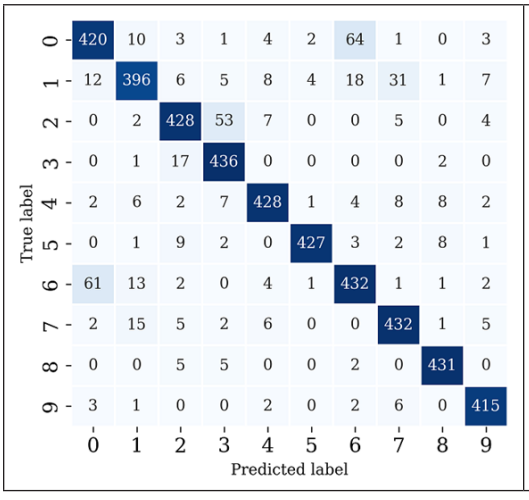

Figure 6. Confusion matrix from an experiment using the Only-IMU dataset

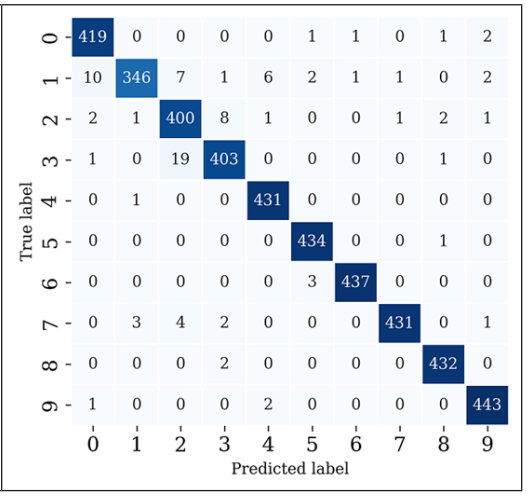

Figure 7. Confusion matrix from an experiment using a multimodal dataset

Recognition of Multidigit Numbers

In a study focused on continuous air-writing of multidigit number gestures, 5-digit, 8-digit, and 10-digit sequences were collected from two participants. This experiment replicated the methodology of the previous study, including 15 training trials of the deep neural network model. The criterion for correct recognition remained consistent; a response was deemed accurate only if all digits of a number precisely matched the correct answer. Table 3 presents the average accuracy rates: 99% and 96.3% for 5-digit numbers, 99.6% and 97.6% for 8-digit numbers, and 99.3% and 93.3% for 10-digit numbers. These results suggest that the methods employed in this study were effective for recognizing air-writing gestures in continuous digit sequences.

Table 3
Average accuracy in multidigit recognition experiments (user-dependent)

Participant number	Accuracy rate for 5-digit numbers (%)	Accuracy rate for 8-digit numbers (%)	Accuracy rate for 10-digit numbers (%)
9	99.0	99.6	99.3
10	96.3	97.6	93.3

Compared to the accuracy rates for one- and two-digit numbers in the multimodal dataset verification experiment, it is noteworthy that the increased number of digits and the consequent rise in recognition complexity did not hinder accuracy. In the multimodal dataset experiment, the training data comprised 150 samples, divided equally between one-digit and two-digit numbers, amounting to 225 data points when considered individually. In contrast, the 5-digit, 8-digit, and 10-digit cases involved 30, 25, and 20 training data samples, culminating in 550 data points. This indicates that the training incorporated double the amount of data, which likely contributed to the improved accuracy rates despite the heightened complexity.

However, a persistent challenge was the tendency to overlook the recognition of the digit one in 5-digit, 8-digit, and 10-digit sequences. This issue is attributed to the variation in the time required to complete the gesture for each digit. Notably, the gesture time for digit one is significantly shorter compared to other digits, which potentially led to its under-recognition in these scenarios.

The deep learning model employed in this research was U-Net, with the kernel size of the convolutional layer maintained constant throughout the model. This fixed configuration might have contributed to the inadequate feature extraction of Gesture 1 during the deep learning process. As a result, enhancements in data preprocessing and the deep learning models are needed to ensure robust recognition of variations in the duration of each gesture.

Additionally, this study utilized user-dependent recognition. However, for the practical implementation of gesture input systems, it is preferable to distinguish between the users providing training data and those performing the actual gesture recognition. Relying on users to collect training data prior to actual use and subsequently training a deep learning model with this data is a time-consuming and labor-intensive process. This approach significantly detracts from the user experience as an input method, making it less feasible for practical applications.

Consequently, there is a pressing requirement for developing user-independent recognition systems. Such systems need to account for the diverse nature of user data, which can vary significantly in aspects like writing style, writing speed, arm position, and arm movement patterns. Therefore, designing a recognition system resilient to these user-specific variations is essential for successfully implementing gesture input technology.

Robustness

To investigate the robustness of multidigit number recognition, we conducted three-fold cross-validation, where the three participants were separated into two groups for training and one for testing. Table 4 shows the results of the cross-validation. In our method and dataset, the robustness is very low. The reason is that our dataset is too small. Air-writing gestures were highly dependent on participants. For instance, in the case of 2-digit numbers,

the accuracy varied significantly from 0.355 to 0.777. Another factor is the writing speed because we ask them to write multidigit numbers very fast. For instance, a participant should perform an air writing gesture of 10-digit numbers in 15 seconds. It is difficult to write digits fast and neatly.

Table 4
Accuracy in user-independent conditions

Test participant	2-digit	5-digit	8-digit	10-digit
P1	0.642	0.315	0.162	0.200
P2	0.777	0.357	0.260	0.277
P3	0.355	0.052	0.013	0.013
Average	0.591	0.241	0.163	0.092

Comparison with Other Studies

We compared our results with other studies using IMU-based approaches for recognizing digits. Singh and Koundal (2024) have achieved a recognition accuracy of 99.50% for the RealSense-Based 3D Trajectory Digit and Character (RTD-RTC) datasets (Image-based), which Alam et al. (2019; 2020) have collected, and a recognition accuracy of 96.30% for the 6DMG dataset (IMU-based), which Chen et al. (2012) have collected. Zhang et al. (2022) have achieved a recognition accuracy of 97.95% for the dataset using an IMU sensor attached to an index finger. Dash et al. (2017) have achieved a recognition accuracy of 96.7% for the user-dependent recognition and 91.7% for the user-independent recognition. Lamaakal et al. (2024) have achieved a testing recognition accuracy of 98.7%. These studies employed the sliding window technique. In the case of recognizing 10-digit numbers, the multi-class windows problem causes a decrease in the recognition accuracy. Here, a denotes a recognition accuracy of a single digit. A recognition accuracy of a 10-digit number is estimated to be a^{10}, which means 10 consecutive successes of recognition. Therefore, our recognition accuracy of 10-digit numbers in the user-dependent evaluation was higher because $0.987^{10} \approx 0.877$.

CONCLUSION

This study proposed a novel method for consecutive digit recognition, employing a deep learning model with sample-level dense labeling, diverging from the traditional Sliding Window technique of mapping a single label to one window. Additionally, the study advocated for enhancing deep-learning model accuracy by generating precise labels through a multimodal dataset that integrates IMU data from wearable sensors with camera data. The results of the recognition experiments conducted to validate the model's effectiveness demonstrated an average accuracy improvement of approximately 13% when using the

deep learning model trained on the multimodal dataset, compared to the model trained solely on the IMU dataset. The confusion matrix analysis revealed a significant reduction in misrecognition among numbers sharing similar gestures, such as 0, 6, 2, and 3, confirming the positive impact of accurate data creation using a multimodal dataset on the training of deep learning models. The participants in the study achieved an average correct response rate of over 90% in the recognition experiments for 5-digit, 8-digit, and 10-digit number gestures.

Notably, even with up to 10 consecutive digits, the correct answer rate was as high as 99.3% for the subject with the highest correct answer rate and 93.3% for the subject with the lowest correct answer rate, exceeding 90%. This indicated the proposed method's effectiveness for recognizing air-writing gestures in continuous digit sequences, employing a deep learning model with sample-level dense labeling and a multimodal dataset. However, future work is necessary to further verify and enhance the robustness of the recognition method in relation to variations in gesture duration and individual user differences. This will involve focusing on data preprocessing and advancing deep learning models to address these challenges.

ACKNOWLEDGEMENT

The authors acknowledge Mr. Kazuma Kodama for the data collection of air-writing gestures from eight participants.

REFERENCES

Alam, M. S., Kwon, K. C., & Kim, N. (2019). Trajectory-based air-writing character recognition using convolutional neural network. In *2019 4th International Conference on Control, Robotics and Cybernetics (CRC)* (pp. 86-90). IEEE. https://doi.org/10.1109/CRC.2019.00026

Alam, M. S., Kwon, K. C., Alam, M. A., Abbass, M. Y., Imtiaz, S. M., & Kim, N. (2020). Trajectory-based air-writing recognition using deep neural network and depth sensor. *Sensors*, *20*(2), Article 376. https://doi.org/10.3390/s20020376

Amma, C., Georgi, M., & Schultz, T. (2014). Airwriting: A wearable handwriting recognition system. *Personal and Ubiquitous Computing*, *18*, 191-203. https://doi.org/10.1007/s00779-013-0637-3

Dash, A., Sahu, A., Shringi, R., Gamboa, J., Afzal, M. Z., Malik, M. I., Dengel, A., & Ahmed, S. (2017). Airscript-creating documents in air. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 908-913). IEEE. https://doi.org/10.1109/ICDAR.2017.153

Kim, D. W., Lee, J., Lim, H., Seo, J., & Kang, B. Y. (2014). Efficient dynamic time warping for 3D handwriting recognition using gyroscope equipped smartphones. *Expert Systems with Applications, 41*(11), 5180-5189. https://doi.org/10.1016/j.eswa.2014.03.011

Kim, E., Shin, J., Kwon, Y., & Park, B. (2023). EMG-based dynamic hand gesture recognition using edge AI for human–robot interaction. *Electronics*, *12*(7), Article 1541. https://doi.org/10.3390/electronics12071541

Lamaakal, I., Ouahbi, I., El Makkaoui, K., Maleh, Y., Pławiak, P., & Alblehai, F. (2024). A TinyDL model for gesture-based air handwriting arabic numbers and simple arabic letters recognition. *IEEE Access, 12,* 76589-7660. https://doi.org/10.1109/ACCESS.2024.3406631

Lin, X., Chen, Y., Chang, X. W., Liu, X., & Wang, X. (2018). Show: Smart handwriting on watches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(4), Article 151. https://doi.org/10.1145/3161412

Lu, Z., Chen, X., Li, Q., Zhang, X., & Zhou, P. (2014). A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. *IEEE Transactions on Human-Machine Systems*, *44*(2), 293-299. https://doi.org/10.1109/THMS.2014.2302794

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells & A. F. Frangi (Eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI* (pp. 234-241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Schäfer, A., Reis, G., Stricker, D. (2022). Controlling continuous locomotion in virtual reality with bare hands using hand gestures. In G. Zachmann, M. A. Raya, P. Bourdot, M. Marchal, J. Stefanucci & X. Yang (Eds.) *International Conference on Virtual Reality and Mixed Reality* (pp. 191-205). Springer International Publishing. https://doi.org/10.1007/978-3-031-16234-3_11

Singh, A. K., & Koundal, D. (2024). A temporal convolutional network for modeling raw 3D sequences and air-writing recognition. *Decision Analytics Journal*, *10*, Article 100373. https://doi.org/10.1016/j.dajour.2023.100373

Tripathi, A., Mondal, A. K., Kumar, L., & Prathosh, A. P. (2022). ImAiR: Airwriting recognition framework using image representation of IMU signals. *IEEE Sensors Letters, 6*(10), Article 7003704. https://doi.org/10.1109/LSENS.2022.3206307

Wu, J., Pan, G., Zhang, D., Qi, G., & Li, S. (2009). Gesture recognition with a 3-d accelerometer. In D. Zhang, M. Portmann, A. H. Tan & J. Indulska (Eds.) *Ubiquitous Intelligence and Computing: 6th International Conference, UIC* (pp. 25-38). Springer Berlin Heidelberg.

Yao, R., Lin, G., Shi, Q., & Ranasinghe, D. C. (2018). Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. *Pattern Recognition*, *78*, 252-266. https://doi.org/10.1016/j.patcog.2017.12.024

Yin, Y., Xie, L., Gu, T., Lu, Y., & Lu, S. (2019). AirContour: Building contour-based model for in-air writing gesture recognition. *ACM Transactions on Sensor Networks (TOSN)*, 15(4), 1-25. https://doi.org/10.1145/3343855

Zhang, H., Chen, L., Zhang, Y., Hu, R., He, C., Tan, Y., & Zhang, J. (2022). A wearable real-time character recognition system based on edge computing-enabled deep learning for air-writing. *Journal of Sensors, 2022*(1), Article 8507706. https://doi.org/10.1155/2022/8507706

Zhang, Q., Wang, D., Zhao, R., Yu, Y., & Jing, J. (2021). Write, attend and spell: Streaming end-to-end free-style handwriting recognition using smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(3), Article 138. https://doi.org/10.1145/3478100

Zhang, Y., Gu, T., Luo, C., Kostakos, V., & Seneviratne, A. (2018). FinDroidHR: Smartwatch gesture input with optical heart rate monitor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(1), Article 56. https://doi.org/10.1145/3191788

Zhang, Y., Zhang, Y., Zhang, Z., Bao, J., & Song, Y. (2018). Human activity recognition based on time series analysis using U-Net. *arXiv preprint arXiv:1809.08113*. https://doi.org/10.48550/arXiv.1809.08113