

Article

Sequence-Information Recognition Method Based on Integrated mDTW

Boliang Sun  and Chao Chen *

Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410073, China; sunboliang1986@126.com

* Correspondence: chenc1997@nudt.edu.cn

Abstract: In the fields of machine learning and artificial intelligence, the processing of time-series data has been a continuous concern and a significant algorithm for intelligent applications. Traditional deep-learning-based methods seem to have reached performance ceilings in certain specific areas, such as online character recognition. This paper proposes an algorithmic framework to break this deadlock by classifying time-series data by evaluating the similarities among handwriting samples using multidimensional Dynamic Time Warping (mDTW) distances. A simplified hierarchical clustering algorithm is employed as a classifier for character recognition. Moreover, this work achieves joint modeling with current mainstream temporal models, enabling the mDTW model to integrate modeling results from methods like RNN or Transformer, therefore further enhancing the accuracy of related algorithms. A series of experiments were conducted on a public database, and the results indicate that our method overcomes the bottleneck of current deep-learning-based methods in the field of online handwriting character recognition. More importantly, compared to deep-learning-based methods, the proposed method has a simpler structure and higher interpretability. Experimental results demonstrate that our proposed method outperforms existing state-of-the-art models in handwriting character recognition, achieving a top-1 accuracy of 98.5% and a top-3 accuracy of 99.3%, thus confirming its effectiveness in overcoming the limitations of traditional deep-learning models in temporal sequence processing.

Keywords: computational intelligence; multidimensional dynamic time warping (mDTW); online character recognition; deep-learning model integration



Citation: Sun, B.; Chen, C.

Sequence-Information Recognition Method Based on Integrated mDTW.

Appl. Sci. **2024**, *14*, 8716. <https://doi.org/10.3390/app14198716>

Academic Editors: Ugo Vaccaro and Luis Javier García Villalba

Received: 15 May 2024

Revised: 20 September 2024

Accepted: 23 September 2024

Published: 27 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Measuring the similarities of different samples is fundamental to many fields of computational intelligence. Many algorithms take distances to value similarities of samples for classification, detection, and other pattern-recognition tasks. In most of the research, samples are represented by a set of vectors through encoding processes. Then, the Minkowski distances of those vectors could be computed as similarities. Minkowski distances are defined as $dist(x, y) = \sum_{k=1}^d (x_k - y_k)^p$, which is one of the most popular metrics to value samples distances. When $p = 1$, Minkowski distance is named Manhattan distance; when $p = 2$ it is called Euclidean distance. Another popular measure is Mahalanobis distance, which could be viewed as distorted Euclidean distance. Overall, these studies highlight the need for most of the current research.

In the past decade, significant progress has been made in sequence-information processing. Sequence-information processing refers to the analysis and manipulation of data sequences, such as text, speech, or biological sequences. This field involves techniques like natural language processing (NLP) and sequence alignment, which enable machines to understand, generate, and transform sequential data efficiently. However, there is still room for improvement in related methods, especially in terms of interpretability, model integration, and other aspects. Existing methods do not perform well in explaining the

complex temporal structure inherent in the data [1,2]. As a remedy, Dynamic Time Warping (DTW) [3] has emerged as a compelling alternative [4–6]. DTW's most notable advantage lies in its resilience against signal warping, encompassing shifts and scaling along the time axis or even the Doppler effect. Consequently, DTW has evolved into one of the most favored metrics for tasks involving pattern matching. For example, when dealing with signals from different sampling frequencies, the conventional point-wise Euclidean distance may yield different results, especially when one signal is merely a compressed version of another. Without using mDTW for sequence point registration, it is not possible to determine which two points are directly used to calculate the Euclidean distance. Therefore, for two time series, the use of DTW is necessary. In this case, DTW performs well by expertly capturing the scale variation and the minimum distance between the generated signals. What sets DTW apart is not only its ability to produce distance values but also its ability to reveal the exact alignment of two sequences. DTW is unique in that it not only computes distance values but also aligns two sequences precisely, which is particularly critical in intelligent computing applications. Moreover, DTW is not just a similarity measure. As a robust feature-extraction tool, it enhances its versatility and value in the context of intelligent computing.

The processing of online character data is a typical and formidable challenge in the field of time-series processing. Accurate and efficient recognition of handwritten characters is of great significance, especially in an environment where smart devices are constantly evolving. Over recent decades, the domain of sequence-information processing, including non-rigid sparse point matching, Handwritten Chinese Character Recognition (HCCR), and Human Action Recognition (HAR), has become a research hotspot and has made remarkable achievements and progress. Furthermore, the intricacies embedded in the Chinese written language, characterized by a vast number of character classes, introduce an additional layer of complexity to the recognition task. The sheer volume of characters, each with its unique features, poses a formidable challenge for researchers and practitioners alike. The intricate structure inherent in most Chinese characters further amplifies the difficulty of the HCCR task, contributing to its status as a persistent and demanding challenge within the research community.

The proposed model presented in this paper offers high interpretability, integrating traditional sequence modeling methods with deep-learning techniques to achieve superior performance. Its transparent framework ensures a clear understanding of each component and parameter, exemplified by the straightforward calculation process of mDTW. Furthermore, the model's seamless integration with deep learning enables powerful feature extraction while maintaining interpretability. With low computational complexity and minimal parameterization, the model remains efficient and reliable, even for large-scale datasets. Overall, this approach combines interpretability, deep-learning integration, and computational efficiency, making it well-suited for a wide range of intelligent computing tasks.

Our Contribution

Development of the mDTW Algorithm: We developed a multidimensional Dynamic Time Warping (mDTW) algorithm specifically designed for handling sequence information, significantly enhancing accuracy in tasks such as online handwriting recognition.

Integration with Deep-Learning Models: We proposed the integration of mDTW with modern deep-learning models such as Recurrent Neural Networks (RNNs) and Transformers, which improved sequence-data classification across multiple domains, including text recognition, action classification, and point matching.

Hierarchical Clustering Model: We introduced a simplified hierarchical clustering model that groups characters and reduces the computational time for DTW distance calculations, making it feasible to efficiently process large handwriting datasets.

2. Related Works

2.1. Related Works of DTW

DTW (Dynamic Time Warping) [3] distance is a method employed to calculate the similarity between two time series. It was initially proposed and utilized in the field of speech recognition and has subsequently found wide application in various domains, including time-series classification, pattern recognition, anomaly detection, clustering analysis, and more. DTW excels at finding the best match between two sequences by elastically stretching or compressing the time series, enabling effective comparison and matching even when the sequences exhibit offsets or rate changes on the time axis. This is achieved by identifying the best alignment (i.e., the optimal path) between the two sequences, which minimizes the sum of the pairwise distances along this path, therefore accurately reflecting the similarity between the two sequences. This alignment allows for nonlinear stretching of the sequences on the time axis, making DTW particularly suitable for handling time series of different lengths or with varying rates.

In time-series classification tasks, DTW serves as a valuable similarity measure, aiding in the determination of what category a sequence belongs to. By calculating the DTW distance between the sequence to be classified and sequences of known categories, the category with the smallest distance can be confidently selected as the classification result. Furthermore, DTW can be utilized to identify specific patterns in time series or detect anomalies. By comparing a time series with a known pattern or a sequence within the normal range and calculating the DTW distance between them, it becomes possible to determine whether the time-series matches the expected pattern or deviates from the normal range.

In time-series clustering tasks, DTW distance proves to be an effective distance measure in clustering algorithms. By calculating the DTW distance between different sequences, similar time series can be clustered together, enabling the discovery of sequence groups with similar morphologies but possible offsets on the time axis.

DTW stands out as a potent and versatile tool within the realm of temporal data mining, finding applications across various domains due to its unique capabilities. One illustrative domain where DTW demonstrates its prowess is in acoustic data analysis, where it exhibits invariance to Doppler effects. This particular attribute makes DTW invaluable for tasks involving the precise alignment and comparison of acoustic signals, regardless of potential frequency shifts induced by motion or other factors.

Expanding the scope of DTW's applications, it plays a crucial role in the analysis of biological signals such as Electrocardiogram (ECG) [7] or Electroencephalogram (EEG) [8,9]. Researchers leverage DTW to effectively characterize and discern patterns within these signals, aiding in the identification and understanding of potential diseases. The flexibility of DTW in handling diverse types of temporal data makes it an indispensable tool for extracting meaningful insights from complex biological signals.

Moreover, in the context of time-series classification problems, DTW serves as a feature extractor when combined with predefined patterns or features, as highlighted in the work by Kate et al. [10]. This approach allows DTW to contribute significantly to the classification of time-series data, where the alignment of sequences is essential for accurate pattern recognition. The use of Hamming distance [1,11] in conjunction with DTW alignment in this setting is referred to as Edit distance, a well-studied variant that enhances our understanding of the relationships between different temporal sequences [12].

In essence, the versatility of DTW spans diverse applications, from mitigating Doppler effects in acoustic data to aiding in the diagnosis of diseases through the analysis of biological signals. Its role as a feature extractor in time-series classification further underscores its adaptability and utility in extracting meaningful information from temporal datasets. The continued exploration and application of DTW in various domains highlight its enduring significance in the field of temporal data mining and analysis.

The proposed framework presented in this paper is combined with two parts: a multi-dimension DTW (mDTW) [13] algorithm and a hierarchical cluster model. The core

idea of our work is very simple: compute similarities among online handwriting samples and use them to classify characters. However, a major problem with the experimental method is that the training time cost is too long. In particular, when we are dealing with a handwriting database that has N_c characters and each character written by N_w writers, the DTW process should be repeated $(N_c \times N_w)^2 / 2$ times to establish a cluster model. Since a Chinese handwriting database often has thousands of characters and hundreds of writers, computing the DTW so many times would take almost a decade on most computing devices, which is not acceptable for researchers. Therefore, we put forward the above two methods to decrease time costs.

2.2. Related Works of Sequence Data

Time-series classification and regression are important techniques for analyzing and modeling data that changes over time. These methods have wide-ranging applications, including Human Activity Recognition, Handwritten Chinese Character Recognition, Earth Observation, Medical Diagnosis, and more. In this study, we will focus on the application of time-series classification and regression in Human Activity Recognition and Handwritten Chinese Character Recognition and provide an overview of the latest developments and challenges in these fields.

Human activity recognition (HAR) is the identification or monitoring of human activity through the analysis of data collected by sensors or other instruments [14]. The recent growth of wearable technologies and the Internet of Things has resulted not only in the collection of large volumes of activity data [15] but also in the easy deployment of applications utilizing these data to improve the safety and quality of human life. HAR is, therefore, an important field of research with applications including healthcare, fitness monitoring, smart homes [16], and assisted living [17].

Handwritten Chinese Character Recognition (HCCR) is a field of study focused on the automatic recognition of handwritten Chinese characters. This technology plays a crucial role in various applications, such as document digitization, optical character recognition (OCR) systems for forms, checks, and IDs, and handwriting input devices. Since the 1980s, HCCR has been a significant research area in pattern recognition due to the complexity and wide variety of Chinese characters. For instance, in Liu et al.'s paper [18], discriminative feature-extraction methods and discriminative learning quadratic discriminant function classifiers were employed, achieving outstanding recognition rates on challenging online and offline HCCR datasets CASIA-OLHWDB and CASIA-HWDB. The best online recognition rates for single characters were 95.28% (DB1.0, 4037 characters), 94.85% (DB1.1, 3926 characters), and 95.31% (ICDAR 2013 Competition DB, 3755 characters). The best offline recognition rates were 94.20% (DB1.0), 92.08% (DB1.1), and 92.72% (ICDAR 2013 Competition DB). More recently, Ren et al. [19] utilized an RNN, combining Bidirectional Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), which processes raw sequential data directly without converting online handwriting trajectories into image-like representations, achieving an end-to-end recognition rate of 98.15% on the ICDAR-2013 dataset.

Compared to other related works, it has been challenging to achieve better performance through deep-learning training. Collectively, there seems to be some evidence to indicate that existing Deep-learning-based HCCR methods have reached the performance bottleneck. Zhang et al. [20] achieved groundbreaking results in the field of HCCR by integrating traditional direction-decomposed feature maps with deep convolutional neural networks; however, it is important to note that these achievements, while groundbreaking, are now seven years old. Additionally, Li et al. [21] introduced a novel approach in the HCCR domain leveraging a Siamese neural network to predict the similarity between handwritten Chinese characters and template images, showing promising generalization to new classes not seen in training, but the method may struggle with limited data scenarios and the complexity of optimizing Siamese networks. However, by integrating mDTW and

Transformer, a significant performance breakthrough was successfully achieved with this framework, demonstrating its potential in other specific sequence-data-related domains.

3. Methodology

As mentioned earlier, the mDTW method presented in this paper serves as a preliminary processing step for classifying data and selecting the top-N results (the top N most likely classifications) to form a rough classification probability distribution. Subsequently, this rough distribution undergoes further refinement through techniques such as cGRU to achieve a more accurate classification. As shown in Figure 1.

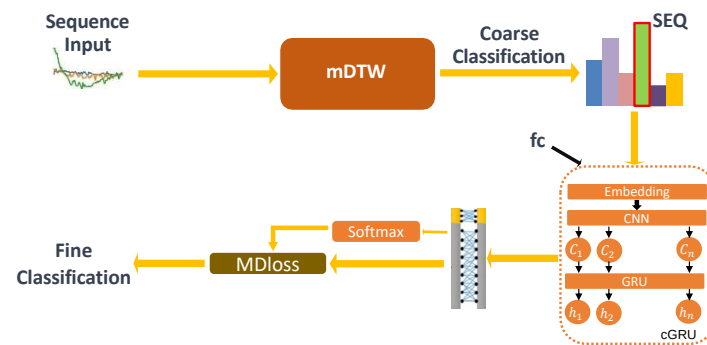


Figure 1. Schematic diagram of the core framework process.

3.1. Multidimensional DTW

Similar to the conventional one-dimensional Dynamic Time Warping (DTW) approach, the method presented here encompasses two primary processes: the forward process and the backward process. In the fundamental stage, the algorithm computes the point-to-point distance matrix, a critical step in establishing the foundational structure for subsequent analyses. This matrix serves as the basis for understanding the dissimilarity between corresponding points in the sequences. Following this, the warping process comes into play, where the optimized point-matching relationships are computed.

In the forward process, the algorithm systematically evaluates the distances between each point in the sequences, creating a comprehensive matrix that encapsulates the pairwise dissimilarities. This matrix lays the groundwork for subsequent alignment, forming the foundation upon which the algorithm can identify the optimal alignment path. The backward process is an equally crucial component, focusing on the derivation of the optimal warping path based on the accumulated point-to-point distances. By iteratively backtracking through the matrix, the algorithm refines the matching relationship between points, ultimately arriving at an alignment that minimizes the overall dissimilarity. Together, these dual processes, the forward and backward stages, synergistically contribute to the effectiveness of the proposed method in capturing the nuanced relationships and achieving optimal alignment between two sequences using the Dynamic Time Warping technique.

Given a couple of sequences $\{SEQ_1 \in \mathcal{R}^{n \times d}, SEQ_2 \in \mathcal{R}^{m \times d}\}$, mDTW first computes a Dm matrix by Equation (1)

$$Dm_{ij} = dsum(\alpha, Seq_{1i}, Seq_{2j}) \quad (1)$$

The variables in formulas are as follows: α is a weight vector of different values in Seq , and $dsum$ is a user-defined formula. Data format of online Chinese handwriting sample is configured as $seq \in \mathcal{R}^{n \times 7}$, each temporal slice $seq_i = \{x_i, y_i, \Delta x_i, \Delta y_i, s_i\}$. Where n is the number of points in a handwriting sequence, $x_i, y_i \in \mathcal{R}$ are the location of a point, $\Delta x_i, \Delta y_i \in \mathcal{R}$ are pen tip displacement of the same point and $s_i \in \{0, 1\}$ is pen tip state. Here $s_i = \{1, 0, 0\}$ indicates the pen tip will continue the writing process, $s_i = \{0, 1, 0\}$ indicates

stroke ends and the pen tip will be raised and moved to another location, $s_i = \{0, 0, 1\}$ indicates the end of the character. Then $dsum$ is defined as Equations (2)–(5).

$$dsum = \alpha_1 \cdot ld + \alpha_2 \cdot dd + \alpha_3 \cdot sd \quad (2)$$

$$ld = \sqrt{(x_{1i} - x_{2j})^2 + (y_{1i} - y_{2j})^2} \quad (3)$$

$$dd = \sqrt{(\Delta x_{1i} - \Delta x_{2j})^2 + (\Delta y_{1i} - \Delta y_{2j})^2} \quad (4)$$

$$sd = \beta_1 |s_{i1} - s_{j1}| + \beta_2 |s_{i2} - s_{j2}| + \beta_3 |s_{i3} - s_{j3}| \quad (5)$$

ld, dd, sd are location distance, displacement distance, and state distance, respectively. ld represents the Euclidean distance between two points from different sequences, which directly reflects the spatial distance relationship between the relevant points. It is the most important constraint for our mDTW point matching. dd represents the Euclidean distance between the difference vectors of two points and their previous sampling time steps, which can be understood as the similarity in direction. Its calculation formula is $\Delta X = X_t - X_{t-1}$. For example, in a plane, if the difference vectors of two sequence points both indicate that the movement direction of the curve/trajectory at this time is to the left, then they should naturally obtain a certain degree of matching similarity. A higher value of α_i means we place more attention on the similarity of the corresponding factor to match the points. Since the information on location is contained in displacement data, much of the deep-learning-based handwriting research only relies on displacement. However, the cluster model in this paper does not have a parameter optimizing process to learn the information of location from displacement, so we added ld in Equation (3) to enhance the matching performance. β is the weight of writing states. Considering one character has almost 100 points and 5–10 strokes means that the rate of end stroke is about 10% and character end is 1%, β is set as $[1, 10, 100]$ to balance the infect of different writing states.

The dimension of a computed matrix Dm is $n \times m$, and each element represents $dsum$ distance of two points. The second step of mDTW is computing a matrix G for final point matching through an iterative process. This process is computed point-by-point by Equation (6)

$$G(i, j) = \min \begin{cases} G(i-1, j) + Dm(i, j) \\ G(i-1, j-1) + \lambda \cdot Dm(i, j) \\ G(i, j-1) + Dm(i, j) \end{cases} \quad (6)$$

λ is the weight of temporal similarity, which is configured as a constant 2 in traditional DTW algorithms. In this paper, we made it flexible to fit the multi-dim point matching task. G is computed from bottom to the top, from left to right. As shown in Figure 2, DTW has following steps in summary:

- Step 1: compute Dm Equations (2)–(5)
- Step 2: initialize G as a zero matrix
- Step 3: compute the first row (G_1) of G by Equation (6) from G_{11} to G_{1n} (green values in Figure 2)
- Step 4: repeat Step 3 m times until all elements in G are updated
- Step 5: start from G_{mn} , jump to a temporal location via $\min\{G_{i,j-1}, G_{i-1,j-1}, G_{i-1,j}\}$.
- Step 6: Repeat Step5 until terminal G_{11} (red arrows in Figure 2)

Completing the above steps will return the matrix G , where G_{mn} could be recognized as the final distance of two sequences $DTWdist(seq_i, seq_j)$. Meanwhile, the computing process has a very high time cost in most of the corresponding tasks especially when dealing with complex sequence problems. Therefore, the algorithm is accelerated by parallel computing for practical application. The method of acceleration uses GPU to do high-dimension matrix multiplications. In the case of ensuring that the GPU memory limit is not exceeded, we could set a batch of sequences to do DTW computation parallel. This method further reduces the calculation time cost of the system.

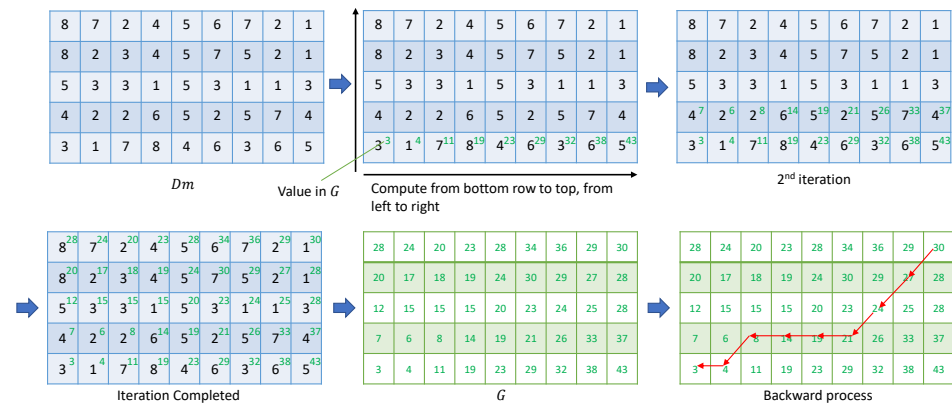


Figure 2. Process of dynamic warping, including forward and backward steps.

3.2. Sequence Clustering

The clustering process outlined in this paper can be interpreted as a form of supervised modeling. Specifically, the method introduced here does not involve the computation of distances between distinct characters; rather, it focuses solely on samples within the same character classes. As elucidated earlier, the computation of complete inner Dynamic Time Warping (DTW) distances for an extensive Chinese handwriting database demands a considerable amount of time, rendering it impractical for experimental purposes. Consequently, we propose the adoption of a simplified hierarchical clustering model designed to compute distances exclusively within the confines of the same character classes. Configuring the model in this manner not only addresses a specific challenge but also confers a distinct advantage, particularly in circumventing the potential computational overhead associated with unnecessary DTW calculations. This streamlined approach serves as a pivotal strategy in mitigating the time cost incurred during the training of the model. The careful restriction of computations to inner class distances within our proposed hierarchical clustering model is a deliberate choice that significantly optimizes the efficiency of the training process.

By narrowing the scope of calculation to inner class distances, we strategically eliminate the need for superfluous DTW calculations, leading to a more efficient and resource-friendly model training procedure. This deliberate simplification ensures that the model focuses its efforts on the most relevant aspects, contributing to a streamlined training phase. As a result, the proposed hierarchical clustering model not only expedites the training process but also alleviates the computational burden associated with analyzing the entire database. In essence, the methodology we propose is designed to strategically minimize the intricacies typically involved in DTW computations. The model, by concentrating exclusively on inner class distances, enhances both the expediency and effectiveness of the clustering process. This approach signifies a purposeful optimization, prioritizing efficiency without compromising the model's ability to discern meaningful patterns within the data. By adopting this strategic simplification, the proposed model emerges as a more efficient and scalable solution, making it well-suited for applications where computational resources are a critical consideration. In summary, the deliberate narrowing of focus within our hierarchical clustering model is a strategic choice that pays dividends in terms of both time efficiency and computational effectiveness, ultimately contributing to an enhanced clustering performance.

Here, we introduce an exclusive concept of this paper: useless DTW calculation. The definition of this is the computation of DTW distance between two Chinese characters with obvious differences in stroke structures. For instance, when we distinguish Chinese characters from dissimilar handwriting samples, the cluster centers could hardly effect testing results since the results (value of DTW) of these samples are too big to be recognized as the same class. Therefore, the computation of those DTW distances is useless. On the contrary, choices of cluster centers are very important for similar Chinese characters; appropriate cluster centers could remarkably increase recognizing accuracy, as shown in

Figure 3. Digits are DTW values between the testing center and samples. For characters with high dissimilarities (gray samples), the value of DTW is much higher than similar ones (blue samples). The value destabilization caused by center changing could not affect the final results for gray samples but have a significant impact on similar samples.

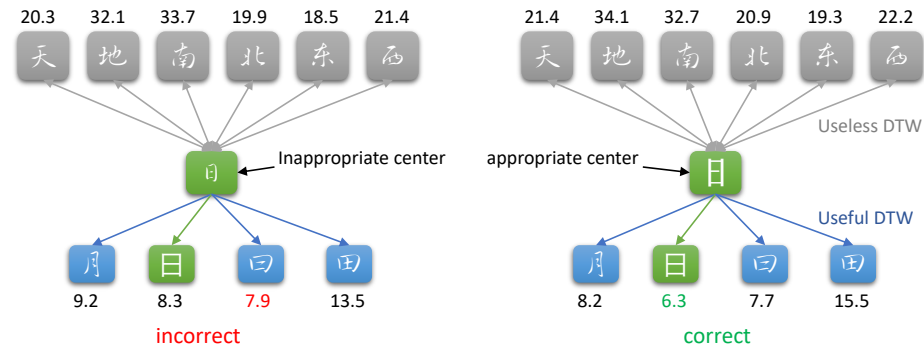


Figure 3. The definition of useless DTWs.

Therefore, the proposed method ignores computing useless DTWs on dissimilar characters and only considers similar samples instead. In order to further reduce the training time cost, the clustering process is set to compute DTWs between samples with the same characters. As discussed above, a database with $N_c \times N_w$ samples urges repeating DTW $(N_c \times N_w)^2 / 2$ times at the origin. The simplified method needs to repeat the DTW process $N_c \times N_w^2 / 2$ times only. For a handwriting database with thousands of characters ($N_c > 1000$), this step could reduce more than 99.9% of total origin training time cost.

Unlike the cluster methods based on Minkowski distances, a major problem within the sequence cluster domain is that it is very hard to generate a virtual sequence as a “center” in the corresponding space. Therefore, the proposed method should take some of the real samples as cluster centers by a hierarchical model. Then, the concrete steps for clustering a $N_c \times N_w$ database of this work are arranged as follows:

- Step 1 Separate the whole database into N_c groups by their characters. Thus, each group has N_w handwriting samples with the same character.
- Step 2 Choose a cluster number $n > 0 \in \mathcal{R}$ and then divide each group to cl_n clusters by minimizing Equation (7)

$$\arg \min [\sum_{i=1}^n \sum_{k=1}^{cl_i} DTWdist(Cen_i, Seq_k)] \quad (7)$$

where $Center_i$ is the center of i_{th} cluster, cl_i is the number of sequence samples in i_{th} cluster, Seq_k is the k_{th} sequence sample of i_{th} cluster. The significance of this step is finding a set of centers that could minimize the sum of DTW distances between training samples and the nearest appropriate centers. In other words, this step finds n real samples as the representation of a whole sample group.

- step3 Repeat step2 N_c times to obtain all characters clustered. After that, we obtain a center set with $N_c \times n$ sequence samples.

Step 2 is the core content of the proposed cluster framework. To realize this, the following calculation steps were adopted, as listed in Algorithm 1. Where D is the storage of DTW distance for every two samples in the group, Equation (8) shows the computation of this process.

$$D_{ij} = DTWdist(seq_i, seq_j) \quad (8)$$

where seq_i represents the i th time-series vector. It should be noted that for temporal information, its vector may be discontinuous. For example, in handwriting, a word may be composed of multiple strokes. Therefore, when performing mDTW calculation on two time series of text, we need to consider the matching relationships between different strokes.

Noticing that D here is a symmetric matrix, and all the diagonal elements are 0. Then a temporal matrix D' is generated from D by Equation (9).

$$D_i' = D_i \times \frac{iter}{iter + 1} + D_j \quad (9)$$

After the iteration process in line2–line6 of Algorithm 1, $C \in \mathcal{R}^{n \times N_w}$ is calculated. Each line of C shows the sequence index of a cluster. Then, we calculate the centers of every cluster. For instance, the non-zero elements in i_{th} row of C are $c_{i1}, \dots, c_{im}$. Finally, find the center of this cluster by Equation (10).

$$Cen(k) = \arg \min(\sum_{l=1}^m DTWdist(Seq_k, Seq_{c_{il}})) \quad (10)$$

Algorithm 1: Hierarchical cluster for a Chinese handwriting sample group

Input: A group of training sequences $SEQ\{seq_1, \dots, seq_{N_c}\}$

0: Initialize a cluster index matrix $C \in \mathcal{R}^{N_c \times N_c}$, set $C_{i1} = i$ and others as 0

1: Compute a DTW distance matrix $D \in \mathcal{R}^{N_c \times N_c}$ via Equation (8).

2: for iter in $N_c - n$

3: find the location $i, j = \operatorname{argmin}(D), (i < j)$

4: compute a temporal matrix D' by Equation (9) and replace D by D'

5: merge the j_{th} cluster to the i_{th}

6: end for

7: for iter in n

8: find the sequence sample in $iter$ row of C by Equation (10) and mark it as $seq_{ct_{iter}}$

9: end for

Output: n sequences as centers $Cen\{seq_{ct_1}, \dots, seq_{ct_n}\}$

3.3. Joint Modeling with Sequence Processing Networks

In the framework proposed herein, sequence data are initially processed using multivariate Dynamic Time Warping (mDTW), followed by clustering of the processed data to achieve a coarse categorization. Subsequently, the results of this coarse categorization are utilized as conditional inputs along with the sequence data (SEQ) for processing by the Conditional Gated Recurrent Unit (cGRU). The role of the cGRU is to enable the model to consider additional conditional or contextual information, therefore enhancing adaptability and performance for specific tasks. In this model, the additional condition provided is the coarse categorization result obtained from the initial clustering, while the sequence data itself offers contextual information. The processed data are then fed into a fully connected neural network, where fine categorization is performed using the SoftMax function, resulting in the final categorization outcomes after balancing the loss. As shown in Figure 4, the formulas for cGRU are as follows:

$$d_t = \tanh(W_d d_t + b_d) \quad (11)$$

$$s_t = \tanh(W_s s_t + b_s) \quad (12)$$

$$r_t = \sigma(W_r h_{t-1} + U_r d_t + V_r s_t + M_r fc + b_r) \quad (13)$$

$$z_t = \sigma(W_z h_{t-1} + U_z d_t + V_z s_t + M_z fc + b_z) \quad (14)$$

$$\tilde{h}_t = \tanh(W(r_t \odot h_{t-1}) + U d_t + V s_t + M fc + b) \quad (15)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (16)$$

$$o_t = \tanh(W_o h_t + U_o d_t + V_o s_t + M_o fc + b_o) \quad (17)$$

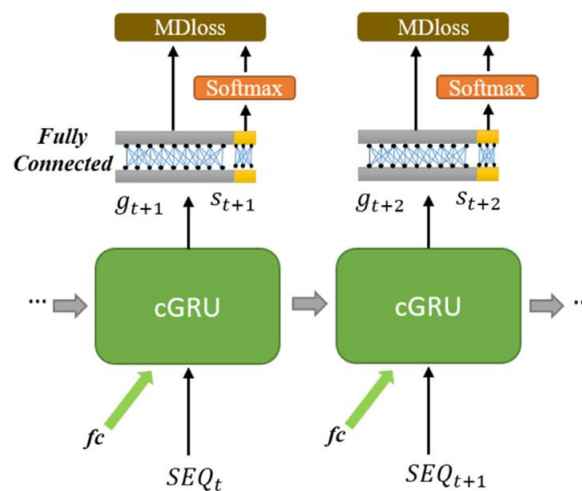


Figure 4. Detailed presentation of the core parts of the framework.

The variables utilized in Equations (11)–(17) are defined as follows: W , U , V , and M represent weight matrices, which are crucial parameters that govern the influence of various inputs on the neurons within a neural network layer. b denotes the biases, which are added to the weighted sum of inputs to adjust the output of the activation function. h_t signifies the hidden state at time t , encapsulating the internal representation of the network based on the inputs received up until that point. o_t represents the output vector at time t , which is generated by the network after processing the input and hidden state. The input vector d_t corresponds to the t -th state of the writing motion path signature, denoted as $\{d_{1,i}, d_{2,i}\}$. Here, i varies from 1 to n , indicating that the signature is comprised of a sequence of n points. s_t represents the pen's up–down state at time t , indicating whether the pen is touching the writing surface or not. The variable fc stands for the conditional input, which provides supplementary information to the network to influence its output.

The output of the conditional Gated Recurrent Unit (cGRU), o_t , is subsequently fed into two fully connected layers. These layers further process the output, ultimately yielding the system's output, which is a $M \times 5 + 3$ vector denoted as g_{t+1}, s_{t+1} . Here, g_{t+1} is a vector of size $M \times 5$, which serves as the input data for a Gaussian Mixture Model (GMM). The GMM is employed to predict pen movements, considering the probability distribution of potential movements. s_{t+1} is a 3×1 vector that is used for predicting the pen tip state, indicating its orientation or other pertinent properties. To enhance the performance of our networks, we utilize a Mixture Density Loss function (MDLoss). This loss function is specifically tailored to handle the output of mixture models, such as the GMM used in this context. By minimizing the MDLoss, we aim to refine the accuracy of the predicted pen movements and pen tip states, ultimately bolstering the overall performance of the system in generating realistic and precise handwriting simulations.

Unlike CGRU, other neural networks we utilize, such as LSTM or Transformer, do not possess a designated conditional input c . To address these discrepancies, we employ a method where we concatenate the output probability distribution of mDTW with the normalized original sequence information before inputting it into the LSTM or Transformer model, as listed in Equation (18)

$$i_t = \text{concat}(\text{seq}_t, c) \quad (18)$$

Among the many methods of time-series analysis and pattern recognition, dynamic time warping (DTW) and its multivariate version (mDTW) have proven to be effective tools, especially when dealing with time series of unequal lengths and varying speeds. However, traditional mDTW methods may be limited in handling large-scale, high-dimensional, and complex pattern data. In recent years, deep-learning models, such as Long Short-

Term Memory Networks (LSTM) and Transformer, have achieved remarkable success in processing sequence data. These models can automatically extract useful features from raw data, allowing them to handle complex patterns and long-term dependencies. To combine the flexibility of mDTW with the powerful representation capabilities of deep-learning models, we propose a joint modeling method. The core idea of this method is to embed mDTW as a differentiable module into the deep-learning framework so that the parameters of mDTW and the parameters of the deep-learning model can be optimized during the end-to-end training process. In the process of joint modeling, model selection is a crucial step because it will directly affect the model's ability to handle specific tasks. For the problem we face, namely the analysis of time-series data, we need to choose among multiple deep-learning models. Long short-term memory networks (LSTM) have proven to be very effective in processing time-series data and are able to capture long-term dependencies in the data, which is crucial for many time-series tasks. However, the Transformer model's advantages in handling parallelization and self-attention mechanisms make it more efficient when processing large-scale data sets and capable of capturing complex interactions between different parts of the input sequence. After selecting a suitable deep-learning model, the next step is to embed the mDTW algorithm into this model. The key to this step is to ensure that mDTW can be combined with the deep-learning model in a differentiable way so that we can use the backpropagation algorithm to train the entire model. To achieve this, we need to make appropriate modifications to the mDTW algorithm so that the gradients generated during the calculation process can be effectively passed to the deep-learning model. In this way, the deep-learning model can update its parameters based on the gradient information provided by mDTW, therefore achieving more efficient processing of time-series data.

In the model training phase, we need to design a joint loss function that will consider both the distance metric of mDTW and the classification or regression loss of the deep-learning model. In this way, we can ensure that the parameters of the mDTW algorithm and the deep-learning model are optimized simultaneously during the training process. This joint optimization method will enable the entire model to take into account the local shape characteristics of the data and capture the global structural information of the data when processing time-series data.

$$L = - \sum_{i=1}^N y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (19)$$

In the equation, L is the value of the loss function, N is the number of samples, y_i is the actual label of the i th sample, p_i is the predicted probability that the i th sample belongs to the positive class (with label 1), and $\log$ denotes the natural logarithm.

Finally, in terms of model optimization, we will use appropriate optimization algorithms to update the model parameters. The gradient descent method or its variants are commonly used optimization algorithms that can adjust the parameters of the model based on the gradient information of the loss function, therefore gradually reducing the training error. In this process, we need to choose appropriate learning rates and other hyperparameters to ensure the stability and efficiency of the optimization process. Through continuous iteration and optimization, we will eventually obtain a joint model that cannot only utilize the flexibility of mDTW to process time-series data but also leverage the powerful representation capabilities of deep-learning models.

4. Experiments

Three experiments' results are shown in this paper. The first is a system of point matching for temporal sequences; the second is a system of online character recognition; The third one is human activity recognition. These experiments correspond to the three major components in our algorithm: mDTW, hierarchical cluster, and sequence processing performance.

4.1. Non-Rigid Sparse Point Matching

It is important to introduce the data we utilized. CASIA-HWDB, with its high-resolution color images and well-defined binary representation, stands as a valuable resource for offline handwritten character recognition research.

Point matching constitutes a classical and intricate challenge within the expansive domain of image registration research. When juxtaposed with the task of matching feature points in images, the endeavor of aligning points within handwritten characters introduces three distinctive characteristics that significantly shape the complexity of the problem. There is the imposition of a nonlinear spatial transformation when attempting to match points in handwriting characters. This nonlinear transformation induces a notable disturbance in the distribution of the two sets of matching points, introducing intricacies that are not present in linear transformations. This spatial nonlinearity poses a formidable obstacle, demanding sophisticated techniques to navigate and effectively align the points within the given framework. Especially as depicted in Figure 5, the significance of the point-matching process becomes strikingly evident when confronted with the notably distinct writing styles of various authors for the same character. This figure illustrates a scenario where two different individuals have written the same character, yet their unique writing styles result in significantly different representations. The red and blue line segments in the figure represent two online handwritten datasets, which are handwritten text samples that include time information. These datasets capture the intricate details of each author's writing style, encompassing the speed, pressure, and direction of the pen strokes. The green line segments indicate the point alignment relationships calculated using the modified Dynamic Time Warping (mDTW) algorithm. This algorithm allows us to match corresponding points between the two datasets, even when they exhibit different lengths and writing styles. Upon examining the point alignment process, it is evident that our results are highly accurate, as the green line segments closely adhere to the contours of the red and blue line segments, indicating a remarkable degree of similarity between the matched points. This underscores the effectiveness of our approach in addressing the challenges posed by diverse writing styles and emphasizes the crucial role of accurate point matching in online handwritten text recognition.

The nature of handwriting introduces an idiosyncratic challenge: individuals often connect different strokes for the sake of convenience during the writing process. This tendency results in a considerable number of errors in the sets of matching points. The interconnection of strokes complicates the matching task, requiring an intelligent and adaptable approach to correctly align points despite these inherent irregularities.

The feature points of Chinese characters exhibit a geometric dispersion that adds an additional layer of difficulty to the matching process. Unlike more regularly structured shapes, the geometric dispersion of feature points in Chinese characters makes it challenging to rely on local spatial distribution similarity for effective point matching. Consequently, conventional methods that hinge on local spatial characteristics may fall short when confronted with the intricacies of matching points within these dispersed geometric patterns. Point matching in handwriting characters is distinguished by the introduction of nonlinear spatial transformations, the prevalence of stroke connections leading to errors, and the geometric dispersion of feature points in Chinese characters. Addressing these challenges requires the development and application of advanced methodologies that can adeptly handle the intricacies inherent in the matching process within this unique context.

While point-matching is not the central focus of our investigation, its significance lies in its capacity to serve as a reflective indicator of the performance of mDTW. An extensive exploration of point-matching experiments has been conducted to comprehensively assess the prowess of mDTW. The findings reveal that mDTW takes into account various aspects such as handwriting trajectory, direction distribution, moving displacement, and stroke connection in tandem to determine the outcomes of the matching process. A detailed analysis of the experimental results, as depicted in the figures, allows for an intuitive description of the prevailing scenario wherein correctly matched points dominate the

major portions of the matching results. The proposed mDTW algorithm has effectively met our stipulated requirements, demonstrating a reliable point-matching performance essential for subsequent clustering endeavors. The amalgamation of trajectory, distribution, displacement, and stroke considerations in the mDTW framework contributes to its efficacy in delivering accurate and meaningful matching outcomes, therefore substantiating its suitability for the specified clustering objectives. To give a further analysis and evaluation on mDTW, we compare our method with two traditional Chinese handwriting points matching methods [22,23]. These two algorithms are image-based. In other words, they do not consider temporal information while matching. Similar to some other Chinese handwriting research, in addition, we also compared two of the latest point-matching methods, LVMs [24] and PMT [25]. LVMs, or Large Vision Models, leverage semantic features derived from large-scale vision models to enhance geometry-based shape feature learning, making them highly effective for non-rigid point cloud matching. PMT, or Proxy Match Transform, introduces a high-order feature transform layer to efficiently match geometric shapes in 3D point clouds, particularly excelling in assembling geometric shapes by establishing reliable local correspondences between part surfaces. We divide the sequence samples into three groups by their complexities, which are simple (0–10 points), normal (20–50 points) and complex (more than 50) points. The summary statistics for comparison are illustrated in Table 1. APM refers to Asymmetric Point Matching, while CGE stands for Constrained Global Energy function.

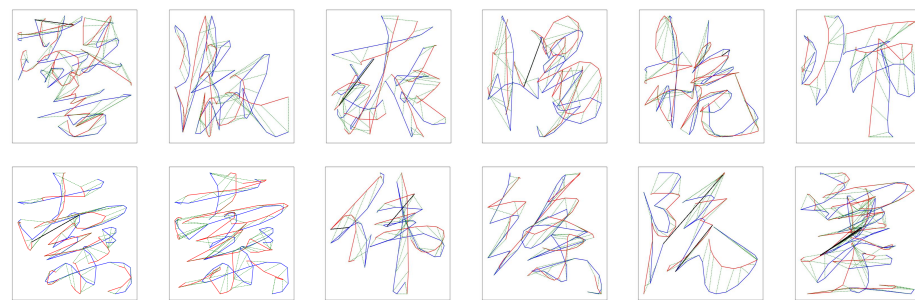


Figure 5. Point-matching results, black line indicates wrong, green indicates right. Red and blue lines are two handwriting characters.

We randomly selected 200 pairs of Chinese handwriting samples from the OLHWDB1.1 dataset. In this dataset, the shortest sample contains 3 points, while the longest one has 216 points. OLHWDB (Online Handwritten Chinese Character Database) is the online handwritten version of HWDB (Handwritten Chinese Character Database), capturing handwritten point sequences with both spatial and temporal information. In contrast, HWDB represents characters in image format and lacks sequential data. The OLHWDB1.1 dataset comprises 3755 distinct Chinese characters, each represented by numerous handwriting samples. For our point registration experiment, we randomly selected 200 of these characters to ensure a diverse and representative subset for analysis. Since the purpose of this experiment is to demonstrate the accuracy of point registration, it is not necessary to select all Chinese characters. The experiment data are labeled and checked manually. Results reported in Table 2 indicate that the proposed mDTW has an advantage against state-of-the-art methods in all complexity conditions. A possible explanation for this is that mDTW takes temporal information as a significant factor. Since online handwriting trajectories may have severe spatial noises in different samples, the writing order for every stroke should be the same. Thus, sequence-based mDTW performs much better than image-based methods in such kinds of tasks.

Table 1. Comparison of point matching results.

Method	Simple	Normal	Complex
APM [22]	34.2%	73.4%	82.7%
CGE [23]	87.1%	91.2%	85.3%
LVMs [24]	90.5%	92.0%	88.7%
PMT [25]	89.8%	93.5%	87.4%
mDTW	93.1%	94.2%	95.7%

4.2. Handwriting Character Recognition Systems

We present the settings of the model hyperparameters to facilitate reproducibility. For the mDTW+Transformer model, we employ a 6-layer Transformer architecture to efficiently process temporal sequence data, leveraging the modified Dynamic Time Warping (mDTW) algorithm to enhance its capability to handle sequences with varying lengths and temporal distortions. Each layer of the Transformer is designed with a hidden state space dimension of 512. This configuration is crucial for balancing computational efficiency with the model's ability to capture complex temporal relationships. The choice of six layers is informed by preliminary experiments that indicate this depth provides a suitable trade-off between model complexity and performance on our targeted tasks without leading to significant overfitting or computational inefficiency. The hidden state dimension of 512 for each layer was selected based on its performance in capturing the intricacies of the data while maintaining a manageable model size and computational demand. These hyperparameters—specifically, the depth of the Transformer and the dimensionality of the hidden states—are essential components of our model's architecture that significantly influence its performance and efficiency in processing time-series data.

The ultimate proposal of this paper is the HCCR experiment. We compare our method with state-of-the-art methods, including Benchmark, ICDAR-2011 Winner: VO-3, UWarwick, DropSample and DropSample-Ensemble-9 [26], NET4-subseq30 and NET123456 [27] and RNN-attention [19]. The illustrated results of those methods are recorded from the corresponding reference. In Table 2, “top-1”, “top-3”, and “top-10” refer to metrics for model prediction accuracy. “Top-1” accuracy means the proportion of times the model's first choice is the correct answer; “top-3” accuracy indicates the chance the correct answer is within the model's top three predictions; similarly, “top-10” accuracy refers to the correct answer being among the model's top ten predictions.

As can be seen from Table 2, the result of mDTW+Transformer architecture is the current best performance among state-of-the-art methods. It should be mentioned that the results we utilize to compare are from all deep-learning-based end-to-end methods. Our method fails in the top-1 test against other state-of-the-art since we used the mDTW-cluster model to do the class task, which means that a test sample only refers to the nearest centers. Nevertheless, the top-3 result shows the achievements our method has made, indicating that the test samples find the correct centers as top-3 in 3755 classes. All these experiment outcomes indicate that our method has a huge potential in sequence processing tasks. We then combined it with the optimized Transformer to form the mDTW+Transformer architecture. As can be seen, this new architecture integrates the strengths of both components, surpassing the accuracy of all other compared SOTA methods.

In recent times, there have been notable strides in the realm of sequence modeling, particularly with the advancements achieved by RNN-attention-based algorithms. These methodologies, leveraging attention mechanisms, have demonstrated significant progress across various domains. A noteworthy example is the RNN:Ensemble [19], which effectively reduced model scale while maintaining competitive performance. However, upon scrutinizing the data presented in Table 2, it becomes apparent that deep-learning-based models appear to have reached a performance plateau. The observed trend indicates that new methodologies have only marginally improved accuracy, with gains typically falling

below the threshold of 0.3%. Despite the collective efforts to enhance these models, the incremental improvements have been relatively modest.

Table 2. Comparison of HCCR methods.

Methods	Top-1	Top-3	Top-10
Benchmark 2013 [18]	95.3%	97.3%	98.8%
VO-3 2013 [28]	95.8%	97.8%	99.0%
UWarwick 2013 [28]	97.4%	98.4%	99.3%
Runner-up 2013 [28]	96.9%	98.2%	99.2%
DropSample 2016 [26]	97.2%	98.4%	99.4%
Ensemble-9 2016 [26]	97.5%	98.5%	99.5%
NET4-subseq30 2017 [27]	97.9%	98.9%	99.5%
NET123456 2017 [27]	98.1%	99.0%	99.5%
RNN:Single 2019 [19]	97.2%	98.4%	99.4%
RNN:Ensemble 2019 [19]	97.6%	98.6%	99.5%
MSCS+ASA+SCL 2022 [29]	96.7%	97.6%	98.9%
SqueezeNext+CCBAM10 2023 [29]	97.4%	98.6%	99.2%
DLHCR 2024 [29]	97.8%	98.7%	99.4%
mDTW-Cluster	68.7%	88.8%	94.6%
mDTW+cGRU	97.4%	99.1%	99.6%
mDTW+Transformer	98.5%	99.3%	99.8%

Breaking away from this performance bottleneck, our proposed method has significantly altered the landscape. Our approach introduces a paradigm shift, outperforming existing models and achieving a notable breakthrough in accuracy. This transformative outcome underscores the potential for further advancements in sequence modeling, challenging the limitations imposed by the perceived performance ceiling in deep-learning-based methodologies. Except for higher HCCR accuracy, the proposed one has an advantage in the real meanings of model parameters when compared to these state-of-the-art methods. For instance, in this paper, the proposed model is a sequence group, where each sequence stands at the center of the character. On the contrary, parameters in deep-learning-based models are weights and biases. It is very hard to figure out the relationships between these parameters and real samples in a black box model.

4.3. Human Activity Recognition Experiments

We present the experimental evaluation of our proposed method, mDTW+Transformer, on three human activity datasets: UCI-HAR, PAMAP2, and OPPORTUNITY. The datasets were chosen due to their popularity and representative nature in the field of human activity recognition. The experiments aim to demonstrate the effectiveness of our approach in comparison to the state-of-the-art (SOTA) methods. For the experimental evaluation, we followed a standard protocol for data preprocessing, which is borrowed from [30]. We compared our proposed method, mDTW+Transformer, with several SOTA methods on each dataset. The results of our experiments are summarized in Tables 1–3 for point matching and HCCR, respectively. In Table 3, the hyphens indicate that the given algorithm was not tested on the specific dataset.

From the tables, it is evident that our proposed method, mDTW+Transformer, outperforms all the compared SOTA methods on all datasets in terms of accuracy. The results demonstrate the effectiveness of our approach in recognizing human activities from various datasets. The superior performance of our method can be attributed to several factors: (i) the combination of mDTW and Transformer allows for a robust alignment between different activity instances, (ii) the Transformer architecture captures long-range dependencies effectively, and (iii) the use of appropriate evaluation metrics ensures a fair comparison with SOTA methods. Furthermore, our method is generic and can be applied to various

human activity recognition tasks without any modifications, making it a versatile approach for practical applications.

Table 3. Comparison of HAR methods.

Methods	UCI-HAR	WISDM	PAMAP2	UniMib	OPPOR
DCNN 2015 [31]	95.1%	-	-	-	-
tFFT+Convnet 2016 [32]	95.7%	-	-	-	-
b-LSTM-S 2016 [33]	-	-	-	-	92.7%
DeepConvLSTM 2016 [34]	-	-	-	-	93.0%
CondConv 2018 [35]	-	98.7%	94.0%	77.31%	81.1%
CIRF 2018 [36]	-	-	-	-	89.1%
LSTM+CT+CSA 2018 [37]	-	-	89.96%	-	-
AttnSense 2019 [38]	-	-	89.3%	-	-
The Layer-wise CNN 2020 [39]	96.9%	98.8%	92.9%	78.0%	81.0%
Lego CNN 2021 [40]	91.4%	97.5%	93.5%	74.4%	88.1%
Real+Virtual 2023 [41]	-	-	69.9%	-	-
TinyTNAS 2024 [42]	96.7%	96.5%	93.7%	77.9%	93.2%
mDTW+Transformer	88.7%	98.5%	94.2%	78.5%	93.5%

Our experimental results with the mDTW+Transformer method outperform the state-of-the-art (SOTA) due to a key characteristic of our proposed approach: the coarse-to-fine classification method applied to sequences. The results are listed in Table 3. This method involves a two-step classification process that enhances both the accuracy and efficiency of classification.

In the first step, coarse-grained classification, sequences are initially grouped into broad categories based on their overall features and patterns. This initial stage focuses on capturing the overall characteristics of the sequences, reducing the complexity of subsequent processing. In the second step, fine-grained classification, sequences within each broad category are further classified with greater precision. This stage emphasizes the subtle features and patterns within sequences, aiming to improve the accuracy of classification. By employing this coarse-to-fine classification strategy, our mDTW+CGRU method comprehensively considers the features of sequences and effectively captures complex patterns. Additionally, the CGRU architecture, with its powerful expression capabilities, effectively learns long-term dependencies within sequences, further enhancing the performance of classification. Therefore, we attribute the superior experimental results of our mDTW+CGRU method compared to SOTA to the coarse-to-fine classification approach. This strategy not only improves classification accuracy but also better adapts to the complexity of various sequence data.

4.4. Ablation Study

The target of Experiment 2 was to investigate the effect of model hyperparameters on different complexity of characters. We manually selected 100 handwriting sample couples, including 33 simple, 34 normal, and 33 complex. We investigated 7 groups of hyperparameters, where we only change α , β is configured as a constant vector $\{1, 10, 100\}$. Table 4 compares the data we obtained from experiments. It could be concluded that the value of α_2 plays a significant role in our algorithm. On the contrary, the other two α variables are not key factors. Meanwhile, appropriate gaining of α_1 and α_3 could increase matching accuracy. Considering that the normal complexity is a major part of all characters, we set α as $\{0.1, 0.8, 0.1\}$ for clustering. The challenge of data significance arises from the inherent properties of diverse variables. One particularly well-established feature in the realm of handwriting analysis research is the pen's moving displacement, denoted as Δx . This variable has gained widespread recognition for its efficacy in capturing important variations in writing directions. Researchers have extensively employed it to unveil nuanced aspects of handwriting intricacies.

Table 4. Results of the effect of different model hyperparameters α on different complexity of characters.

α	Simple	Normal	Complex
{0, 1, 0}	89.5%	92.2%	89.3%
{0.1, 0.8, 0.1}	93.1%	94.2%	95.7%
{0.45, 0.45, 0.1}	66.3%	42.2%	32.7%
{0.8, 0.1, 0.1}	50.3%	62.2%	48.6%
{0.1, 0.45, 0.45}	43.6%	38.1%	30.2%
{0.45, 0.1, 0.45}	56.7%	18.8%	21.9%
{0.1, 0.1, 0.8}	68.0%	18.2%	15.3%

However, it is essential to delve into the relationship between pen-moving displacement and pen location. The latter can be conceptualized as the integral of displacement, offering a holistic perspective on the trajectory of the writing instrument. Yet, a notable drawback surfaces: the directional variation information tends to be subdued within the integral representation. Consequently, while pen location provides an overall picture of the writing path, it may lack the granularity required to discern subtle changes in direction. In stark contrast, the utilization of pen-moving displacement emerges as a superior choice in many instances. By focusing on the instantaneous changes in position, this approach excels at capturing the finer nuances of writing directions. The importance of these variations becomes particularly evident when seeking a comprehensive understanding of the intricate dynamics involved in handwriting.

We conducted ablation experiments to validate the effectiveness of our method, experimenting on both the HCCR dataset and the Human Action Recognition dataset. In these experiments, we selected three sequence modeling models: cGRU, LSTM, and Transformer. We first conducted experiments using these three models individually, and then we integrated them into the mDTW framework. By comparing and observing the impact of the mDTW framework on the accuracy of the experimental results, it can be seen from Table 5 that the experimental results with mDTW achieved an accuracy improvement of about 1% on most datasets. This demonstrates the effectiveness of our proposed method.

Table 5. Ablation of mDTW Joint modeling methods.

Methods	HCCR	UCI-HAR	WISDM	PAMAP2	UniMib	OPPOR
cGRU	95.6%	84.1%	94.8%	92.9%	75.3%	88.1%
LSTM	95.1%	86.3%	94.5%	93.7%	75.8%	89.6%
Transformer	96.6%	86.2%	95.5%	93.2%	76.4%	90.4%
mDTW+cGRU	97.4%	87.6%	97.1%	94.1%	77.1%	92.1%
mDTW+LSTM	97.2%	88.1%	97.9%	94.1%	77.7%	92.7%
mDTW+Transformer	98.5%	88.7%	98.5%	94.2%	78.5%	93.5%

In summary, the preference for pen-moving displacement over pen location stems from its ability to accentuate crucial directional changes, therefore outperforming the integral representation. This nuanced perspective enables a more detailed and insightful analysis of the complexities inherent in the act of writing.

5. Conclusions

The purpose of the current study was to develop a sequence algorithm for temporal data modeling applications. To achieve this, we proposed a simplified hierarchical cluster model and designed a mDTW method to value the similarities between sequences. By taking HCCR as a specific research area, the proposed method has achieved a big improvement in recognizing accuracy (Top-3 and Top-10). This work contributes great confidence that reforming, improving, and appropriately using some simple methods could defeat deep learning in large-scale data-sequence processing tasks. We can outperform deep-learning approaches in the realm of intelligent computing for large-scale data-sequence processing

tasks. Moreover, we envision significant potential for the expansion of our method into other research domains focused on sequence analysis within the purview of intelligent computation. Additionally, the methodology presented in this article seamlessly lends itself to joint modeling with various deep neural network models, further enhancing its adaptability and effectiveness in intelligent computing applications. Deep models include main sequence-oriented methods, including RNN, Tf, and other models. Experimental results show that our method can improve the modeling performance of the model to a certain extent when combined with a deep model. It has achieved good performance in human action recognition and handwriting recognition tasks, and some results exceed SOTA.

Through joint modeling, we can effectively combine the flexibility of mDTW with the powerful representation capabilities of deep-learning models. This method has significant advantages when dealing with complex time-series analysis tasks, especially when the shape and global structure of the time series need to be considered simultaneously. Future work will include validating the effectiveness of this joint modeling approach in a wider range of application scenarios and exploring further optimization strategies to improve model performance and efficiency.

Author Contributions: Conceptualization, B.S. and C.C.; methodology, B.S.; software, B.S.; validation, C.C.; formal analysis, B.S.; investigation, B.S.; resources, C.C.; data curation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, C.C.; visualization, B.S.; supervision, C.C.; project administration, C.C.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is sponsored by the General Program of the National Natural Science Foundation of China (No. 62376279).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://nlpr.ia.ac.cn/databases/handwriting/Online_database.html.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cai, X.; Xu, T.; Yi, J.; Huang, J.; Rajasekaran, S. DTWNet: A Dynamic Time Warping Network. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11640–11650.
2. Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [\[CrossRef\]](#)
3. Qu, Y.; Yang, M.; Zhang, J.; Xie, W.; Qiang, B.; Chen, J. An outline of multi-sensor fusion methods for mobile agents indoor navigation. *Sensors* **2021**, *21*, 1605. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Luo, Z.; Qi, R.; Li, Q.; Zheng, J.; Shao, S. ABODE-Net: An Attention-based Deep Learning Model for Non-intrusive Building Occupancy Detection Using Smart Meter Data. In Proceedings of the International Conference on Smart Computing and Communication, New York, NY, USA, 18–20 November 2022; pp. 152–164.
5. Song, C.; Lu, M.; Wang, Y.; Lu, W. A dynamic time warping loss-based closed-loop CNN for seismic impedance inversion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5925313. [\[CrossRef\]](#)
6. Middlehurst, M.; Schäfer, P.; Bagnall, A. Bake off redux: A review and experimental evaluation of recent time series classification algorithms. *Data Min. Knowl. Discov.* **2024**, *38*, 1958–2031. [\[CrossRef\]](#)
7. Shen, J.; Bao, S.D.; Yang, L.C.; Li, Y. The PLR-DTW method for ECG based biometric identification. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August–3 September 2011; pp. 5248–5251.
8. Schirrneister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggenberger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [\[CrossRef\]](#)
9. Lerogeron, H.; Picot-Clément, R.; Rakotomamonjy, A.; Heutte, L. Approximating dynamic time warping with a convolutional neural network on EEG data. *Pattern Recognit. Lett.* **2023**, *171*, 162–169. [\[CrossRef\]](#)
10. Kate, R.J. Using dynamic time warping distances as features for improved time series classification. *Data Min. Knowl. Discov.* **2016**, *30*, 283–312. [\[CrossRef\]](#)

11. Zhang, H.; Dong, Y.; Li, J.; Xu, D. An efficient method for time series similarity search using binary code representation and hamming distance. *Intell. Data Anal.* **2021**, *25*, 439–461. [\[CrossRef\]](#)
12. Gold, O.; Sharir, M. Dynamic Time Warping and Geometric Edit Distance: Breaking the Quadratic Barrier. *ACM Trans. Algorithms* **2016**, *14*, 50. [\[CrossRef\]](#)
13. Ibrahim, M.Z.; Mulvaney, D. Geometry based lip reading system using multi dimension dynamic time warping. In Proceedings of the 2012 Visual Communications and Image Processing, San Diego, CA, USA, 27–30 November 2012; pp. 1–6.
14. Gupta, N.; Gupta, S.K.; Pathak, R.K.; Jain, V.; Rashidi, P.; Suri, J.S. Human activity recognition in artificial intelligence framework: A narrative review. *Artif. Intell. Rev.* **2022**, *55*, 4755–4808. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Ramanujam, E.; Perumal, T.; Padmavathi, S. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors J.* **2021**, *21*, 13029–13040. [\[CrossRef\]](#)
16. Lockhart, J.W.; Pulickal, T.; Weiss, G.M. Applications of mobile activity recognition. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 1054–1058.
17. Vaizman, Y.; Ellis, K.; Lanckriet, G. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Comput.* **2017**, *16*, 62–74. [\[CrossRef\]](#)
18. Liu, C.L.; Yin, F.; Wang, D.H.; Wang, Q.F. Online and offline handwritten Chinese character recognition: Benchmarking on new databases. *Pattern Recognit.* **2013**, *46*, 155–162. [\[CrossRef\]](#)
19. Ren, H.; Wang, W.; Liu, C. Recognizing online handwritten Chinese characters using RNNs with new computing architectures. *Pattern Recognit.* **2019**, *93*, 179–192. [\[CrossRef\]](#)
20. Zhang, X.Y.; Bengio, Y.; Liu, C.L. Online and offline handwritten Chinese character recognition: A comprehensive study and new benchmark. *Pattern Recognit.* **2017**, *61*, 348–360. [\[CrossRef\]](#)
21. Li, Z.; Xiao, Y.; Wu, Q.; Jin, M.; Lu, H. Deep template matching for offline handwritten Chinese character recognition. *J. Eng.* **2020**, *2020*, 120–124. [\[CrossRef\]](#)
22. Lian, W.; Zhang, L.; Yang, M. An Efficient Globally Optimal Algorithm for Asymmetric Point Matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1281–1293. [\[CrossRef\]](#)
23. Zhao, B.; Yang, M.; Pan, H.; Zhu, Q.; Tao, J. Nonrigid point matching of Chinese characters for robot writing. In Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau, China, 5–8 December 2017; pp. 762–767.
24. Chen, Z.; Jiang, P.; Huang, R. Unsupervised Non-Rigid Point Cloud Matching through Large Vision Models. *arXiv* **2024**, arXiv:2408.08568.
25. Lee, N.; Min, J.; Lee, J.; Kim, S.; Lee, K.; Park, J.; Cho, M. 3D Geometric Shape Assembly via Efficient Point Cloud Matching. *arXiv* **2024**, arXiv:2407.10542.
26. Yang, W.; Jin, L.; Tao, D.; Xie, Z.; Feng, Z. DropSample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition. *Pattern Recognit.* **2016**, *58*, 190–203. [\[CrossRef\]](#)
27. Zhang, X.Y.; Yin, F.; Zhang, Y.M.; Liu, C.L.; Bengio, Y. Drawing and recognizing chinese characters with recurrent neural network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 849–862. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Yin, F.; Wang, Q.F.; Zhang, X.Y.; Liu, C.L. ICDAR 2013 Chinese handwriting recognition competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1464–1470.
29. Kriuk, B.; Kriuk, F. Deep Learning-Driven Approach for Handwritten Chinese Character Classification. *arXiv* **2024**, arXiv:2401.17098.
30. Wang, X.; Zhang, L.; Huang, W.; Wang, S.; Wu, H.; He, J.; Song, A. Deep convolutional networks with tunable speed–accuracy tradeoff for human activity recognition using wearables. *IEEE Trans. Instrum. Meas.* **2021**, *71*, 2503912. [\[CrossRef\]](#)
31. Jiang, W.; Yin, Z. Human activity recognition using wearable sensors by deep convolutional neural networks. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1307–1310.
32. Ronao, C.A.; Cho, S.B. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **2016**, *59*, 235–244. [\[CrossRef\]](#)
33. Hammerla, N.Y.; Halloran, S.; Plötz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv* **2016**, arXiv:1604.08880.
34. Ordóñez, F.J.; Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [\[CrossRef\]](#)
35. Ignatov, A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Appl. Soft Comput.* **2018**, *62*, 915–922. [\[CrossRef\]](#)
36. Hu, C.; Chen, Y.; Hu, L.; Peng, X. A novel random forests based class incremental learning method for activity recognition. *Pattern Recognit.* **2018**, *78*, 277–290. [\[CrossRef\]](#)
37. Zeng, M.; Gao, H.; Yu, T.; Mengshoel, O.J.; Langseth, H.; Lane, I.; Liu, X. Understanding and improving recurrent networks for human activity recognition by continuous attention. In Proceedings of the 2018 ACM International Symposium on Wearable Computers, Singapore, 8–12 October 2018; pp. 56–63.
38. Ma, H.; Li, W.; Zhang, X.; Gao, S.; Lu, S. AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, Macao, China, 10–16 August 2019; pp. 3109–3115. [\[CrossRef\]](#)

39. Teng, Q.; Wang, K.; Zhang, L.; He, J. The layer-wise training convolutional neural networks using local loss for sensor-based human activity recognition. *IEEE Sensors J.* **2020**, *20*, 7265–7274. [[CrossRef](#)]
40. Tang, Y.; Teng, Q.; Zhang, L.; Min, F.; He, J. Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors. *IEEE Sensors J.* **2020**, *21*, 581–592. [[CrossRef](#)]
41. Leng, Z.; Kwon, H.; Plötz, T. Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition. In Proceedings of the 2023 ACM International Symposium on Wearable Computers, Cancun, Mexico, 8–12 October 2023; pp. 39–43.
42. Saha, B.; Samanta, R.; Ghosh, S.K.; Roy, R.B. TinyTNAS: GPU-Free, Time-Bound, Hardware-Aware Neural Architecture Search for TinyML Time Series Classification. *arXiv* **2024**, arXiv:2408.16535.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.